ISSN 2395-8618

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Algorithm Optimization Using Features In SVD & Classification In Eigenspace

Chaman Lal Sabharwal
Computer Science Department
Missouri University of Science and Technology
Missouri, Rolla, 63128 chaman@mst.edu

Abstract Singular Value Decomposition (SVD) is ubiquitous in a range of applications including computer science, economics, engineering, geology, oceanography, psychology, social networking etc. It is an unsupervised modeling technique that creates latent vectors for a subspace that reduces the dimensionality of observed data from n to k (k<<n) dimensions. Latent variables are uncorrelated variation of attribute values that are correlated in the original space. Moreover, SVD can be used to detect/remove noise/outliers, cluster similar entities and make predictions. On the other hand, classification tree is a supervised technique that accomplishes the similar tasks. It models decision trees from training data in order to make intelligent predictions. There is a close connection between SVD and decision trees, but differ in purpose, algorithm design and error analysis techniques. We present a hybrid algorithm bridges the gap between these standalone algorithms and adaptively supersedes their outcomes. For experimental analysis, we use real-world benchmark data, wines, publicly available from UCI machine learning repository. The algorithm is implemented in Matlab, supported by decision trees in Weka software, on MacOS Seirra Version 10.12.3 8GB 160MHZ.

Keywords. PCA, SVD, MDS, Dimensionality Reduction, Classification Tree.

I. INTRODUCTION

SINGULAR Value Decomposition (SVD) is ubiquitous in a range of areas including computer science, image processing (compression, enhancement), engineering, economic and social behavior models, geology, oceanography, psychology, psychophysics, social networks, visualization, and natural language processing [1], [2]. Singular Value Decomposition (SVD) is a study of the underlying structure of objects and their properties for efficient knowledge exploration.

Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are interchangeably referred in the literature. In fact, PCA is for real symmetric matrices whereas SVD is generalization of PCA applicable to any rectangular positive semi-definite matrices. However, PCA and SVD are equivalent for symmetric positive semi-definite matrices. In Section 2, we will elaborate in detail on the difference between the two, PCA and SVD. While SVD is used for unsupervised modeling, the Multi Dimensional Scaling (MDS) and Decision Tree are supervised modeling techniques for making reliable classification predictions. Both of these techniques have complimentary roles in extracting knowledge embedded in data. For example, for recommendation of a product, it may be authenticated with large consumer survey in the face of high dimensional large question answer data. We will present a hybrid algorithm that takes advantage of functionality of SVD to optimize Decision trees, resulting in substantial reduction in computational effort and reduced storage space at insignificant cost in accuracy.

The paper is organized as: Section II is background Section III is on related work, Section IV is algorithm design, prediction and accuracy measurement metrics, Section V implementation, experiments and discussion of results. Section VI is conclusions, Section VII is references and Section VIII includes the linear algebra appendix.

II. BACKGROUND

The real world business data is in the form of tables. Mathematically [see Appendix] speaking, a table is an m×n matrix whose m rows are observation vectors and n columns are properties/attributes of the observation. Principal Components are new latent vectors blending features in original vectors [3]. Because row rank and column rank of a matrix are equal, the rank of an m×n matrix A is k, where $k \leq \min(m, n)$ [4]. One of the advantages of SVD is that it is applicable to positive semi-definite rectangular matrices including symmetric square matrices [5], [6]. Linear algebra is the backbone of PCA development. Some users use it blindly without understanding the algebraic implications. We clarify it with the following examples.

A. What is Eigen-Decomposition?

If A is an n×n real matrix and has eigenvalues and eigenvectors, then (1) the eigenvectors are normalized to unit vectors and (2) the eigenpairs are arranged in the descending order of eigenvalues [7]. Let V be the matrix of eigenvectors of A. If V is invertible, algebraically $VV^{-1} = I$, and geometrically V is a latent orientation of original axes for data visualization. Let D be the matrix of eigenvalues of A arranged in descending order. The goal is to express A in terms of eigenvectors matrix V and eigenvalues matrix D as $A = VDV^{-1}$. For example, (1) the matrix $\begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}$ has two identical eigenvalues: -1, -1; and only one eigenvector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. It does not have eigen-decomposition. (2) The matrix $\begin{bmatrix} 1 & 0 \\ -2 & -1 \end{bmatrix}$ has two distinct eigenvalues: 1, -1, and eigenvectors: $\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

https://doi.org/10.17562/PB-56-1

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

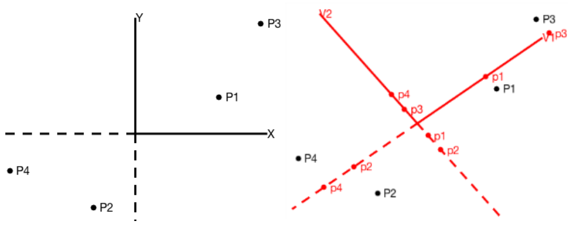
which are linearly independent, but *not orthogonal*. The eigen-decomposition is A = VDV^{-1}

$$\begin{bmatrix}1 & 0\\ -2 & -1\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}} & 0\\ \frac{-1}{\sqrt{2}} & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}\sqrt{2} & 0\\ 1 & 1\end{bmatrix}, \text{ but } V^{-1} \neq V^T.$$

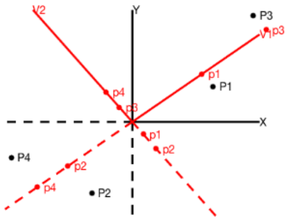
If eigenvectors are not orthogonal, this decomposition is not useful for data mining. If A is symmetric, then V is orthogonal matrix of eigenvalues and the inverse is simply the transpose of V, i.e., $V^{-1}= V^T$.

B. What is PCA?

The basic concept of PCA is to create smaller set new latent variables in place of original large set of variables/attributes used in observations [8]. Principal Component Analysis (PCA) is a generalization of eigen-decomposition to symmetric matrices leading to orthogonal eigenvectors such that A = VDV^T= VDV^{-1}. However, PCA uses the eigenvectors of covariance matrix, A^TA = A^2 [9]. The eigenpairs are arranged in descending order of eigenvalues. The first eigenvector gives the *direction of the largest spread*, and the first eigenvalue is *the largest spread*, see Figure2. It is equivalent to the least squares deviation meaning data points are at shortest distance from the computed eigenvector. PCA determines eigenpairs for A^2 such that the eigenvectors are (1) pairwise orthogonal, (2) normalized to unit vectors and (3) arranged in the descending order of eigenvalues, (4) that the variance along the eigenvectors are maximum in descending order. Let V be the matrix of *eigenvectors of A^2* arranged in the descending order of eigenvalues of A, then algebraically V is orthogonal matrix and is invertible. Let D be the diagonal matrix of eigenvalues of A arranged in descending order provided eigenvalues of A are non-negative. For negative eigenvalues of A, sign has to be taken into consideration. PCA reconstructs A from V and D as A = VDV^T = VDV^{-1}. Figure 1(a) shows four data points along with the standard coordinate axes. Figure 1(b) has eigenvectors and projections of data points on the eigenvectors to show the variance of data along the eigenvectors as new computed axes. It is a rotation of the vase coordinates system. Figure 1(c) displays standard and new coordinate systems, data points and their projections on the eigenvectors.



(a) (b)



(c)

Figure 1. (a) Four data points {P_1, P_2, P_3, P_4}, (b) eigenvectors, projections of data points on the eigenvectors to show data spread along new axes, (c) representative of uv, and xy frames with data points and projections.

We generalize the previous example to symmetric matrix for the PCA. The matrix $\begin{bmatrix}1 & 0\\ 0 & -2\end{bmatrix}$ has eigenvalues λ =1, -2 and the eigenvectors are $\begin{bmatrix}1\\ 0\end{bmatrix}, \begin{bmatrix}0\\ 1\end{bmatrix}$ which are linearly independent, and *orthogonal* eigenvectors. Then the eigen-decomposition A = VDV^{-1} = VDV^T is $\begin{bmatrix}1 & 0\\ 0 & -2\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$.

But PCA uses AA^T and A^TA which are A^2 =$\begin{bmatrix}1 & 0\\ 0 & 4\end{bmatrix}$ (A, A^2 symmetric) for computing the eigenpairs. Except for signs, the eigenvalues of A are square roots of the eigenvalues of A^2 that are 4 and 1, the corresponding eigenvectors are eigenvectors of A^2 are $\begin{bmatrix}0\\ 1\end{bmatrix}$, $\begin{bmatrix}1\\ 0\end{bmatrix}$. Matlab svd does not reconstruct A. In our algorithm, we include the proper signs. Here we used the proper sign for *square root of 4 to -2*, because -2 is eigenvalue of A.

Using eigenvectors of A^2, the eigen-decomposition

A = VDV^{-1} = VDV^T is $\begin{bmatrix}1 & 0\\ 0 & -2\end{bmatrix} = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -2\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$.

C. What is SVD?

Singular Value Decomposition (SVD) is a generalization of PCA to include (1) non-square and (2) positive semi-definite matrices [10], [11]. PCA and SVD are equivalent for symmetric positive semi-definite matrices. By definition, an m×n real matrix A is positive semi-definite, if $\mathbf{v}^T A\mathbf{u} \geq 0$ for all vectors **u** and **v**. The matrices AA^T and A^TA are symmetric and positive semi-definite. For symmetric matrices AA^T and A^TA, the eigenvalues are de facto non-negative and eigenvectors are orthogonal. SVD uses covariance matrices AA^T and A^TA to determine two orthogonal matrices of eigenvectors U, V and a diagonal matrix S for eigenvalues such that the eigenvectors in U, (and V) are (1) pairwise orthogonal, (2) normalized to unit vectors and (3) arranged in the descending order of eigenvalues. Then SVD decomposes A into three factors U, V and S such that A = USV^T. The examples where A is not both symmetric and positive semi-definite are shown in the Table 1 to understand SVD and PCA are not equivalent in general.

Example. To accommodate both PCA and SVD, we generalize the previous example matrix to symmetric, positive semi-definite A= $\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}$ whereas again AA^T=A^TA=$\begin{bmatrix}1 & 0\\ 0 & 4\end{bmatrix}$.

The eigenvalues of A are 2,1; so D= $\begin{bmatrix}2 & 0\\ 0 & 1\end{bmatrix}$ which is same as S. Thus for PCA/SVD of A, the eigenpairs of AA^T, A^TA are U

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

= $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, V= $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, and S = $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$.
And A = USV^T becomes
$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ implying A = USV^T.

Summarizing this discussion, we see that several possibilities exist for an arbitrary matrix A. We show these examples in Table 1.

Consequently, we formulate the sufficiency conditions for the existence and equivalence of PCA and SVD. If a matrix A is symmetric and positive semi-definite, then A has SVD decomposition. Recall that for any matrix A, the covariance matrices AA^T, A^TA are symmetric and positive semi-definite. We can always get around the exceptional cases for A. In Table 1, there are some cases where A is (1) symmetric and (1.1) has PCA SVD decomposition equivalent on PSD (1.2) has PCA SVD decomposition, but not equivalent on not PSD, and (2) not symmetric and (2.1) PCA SVD decomposition equivalent on PSD (2.2) has PCA SVD decomposition not equivalent on not PSD or non-square PSD. Also for non-square matrices, there is SVD, not PCA.

TABLE1. FOUR POSSIBLE CASE OF MATRIX

Data Matrix	Positive Semi-Definite	Not Positive Semi-Definite
Symmetric	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ PCA≡SVD	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ PCA≢SVD
Not Symmetric	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ SVD but no PCA	$\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & 0 \end{bmatrix}$ no SVD, no PCA

D. What is MDS? [10], **[11]**, [12] Multidimensional Scaling (MDS) is a twist of PCA/SVD and resorts to minimization or regression optimization for solution to this problem, whereas PCA/SVD resorts to eigenpair solution to this problem. The new variables are linear combinations of original variables. Mathematically, observations samples are points in higher dimensional space. If $\mathbf{x}^{\mathbf{T}} = [x_1, x_2, ..., x_n]$ is a point in base coordinate system, then $\mathbf{y}^{\mathbf{T}} = [\, y_1, y_2, ..., y_k]$, k<<n, is the same point in new coordinate system where for p =1,…,k $y_p = \sum_i w_{ip}$ x_i the coefficients w_{ip} are also called the weights associated with x_i, namely, $w_{\mathbf{p}}^{\mathbf{T}} = [w_{1p}, w_{2p}, ..., w_{np}]$ is a weight vector. This means y_p is the projection of **x** on $\mathbf{w_{\bullet p}}$. Algebraically, the matrix $[\mathbf{w_{\bullet p}}]$ is a linear transformation of **x** to **y**. Geometrically, it is a matrix of rotation from a standard vector space basis to principal component vector basis. The vector $\mathbf{w_{\bullet p}}$ is determined in such at way that variation of data points along this direction is maximum, then the data projection is $A\mathbf{w_{\bullet p}} = b_p$, $A[\mathbf{w_{\bullet p}}] = [b_p]$. Note that this $\mathbf{w_{\bullet p}}$ is the same as vector $\mathbf{v_p}$ in PCA/SVD as such $V = [\mathbf{v_p}] = [\mathbf{w_{\bullet p}}]$, AV = B expresses m×n matrix A to smaller m×k matrix B. V is the matrix of eigenvectors and amount of spread along each vector is the corresponding eigenvalue.

E. What is a Decision Tree? A decision tree is a tree where the paths from the root to the leaf nodes generate rules for classification of data items [13],[14]. The conventional way to classify multidimensional data is to start with a feature space whose dimensionality is the same as that of the data. The attributes are rearranged by leveraging entropy at each step in the process of tree construction from the root node to proceed to child nodes. Entropy [15Kdnugg2017] is a computational technique for determining the best possible attribute to be used in the decision tree construction. For example, if an attribute has n values x_i, we compute the probability $p(x_i)$, to associate the entropy with **x**:

$$\text{Entropy}(x) = -\sum_{i=1,n} p(x_i) \log_2 p(x_i)$$

Decision is based on entropy of the feature values. At each stage of tree construction, we use the conditional entropy based on attributes that have been already selected.

How is decision tree used? The decision tree can be displayed graphically for easy visualization and intuitive understanding quickly. In decision tree approach prediction, search in the decision table is replaced by decision tree traversal algorithm to determine the classification. All machine learning algorithms have flaws. For a list of such flaws see [15]. It is an interesting article on what is right and what is wrong with the algorithms. So one has to understand the fundamental underpinnings of the algorithm and has to be vigilant in the selection of algorithm before using it. This paper adheres to this conviction.

III. RELATED WORK

A. What is meant by Size Reduction, Noise Reduction, and Clustering Data? The primary purpose of reduction is to optimize storage space and computation time, not necessarily to recreate the original data. The most useful contribution PCA/SVD is that one can ignore eigenvalues or variations below a certain threshold [11]. For this purpose, it is desirable to uncorrelated the correlated. Let A be an m×n real and symmetric data matrix. We may want to reduce the number of variables/properties/attributes of data (row size) or number of observations/objects for size of data (column size) or both. We create two orthonormal matrices U and V where U and V are eigenvector matrices of covariance matrices AA^T and A^TA. Since AA^T and A^TA are square, symmetric, and positive semi-definite, there exist (1) orthonormal matrices U, V, such that $UU^T = 1$ and $VV^T = 1$ and (2) a diagonal matrix S such that is $AA^T = USU^T$, $A^TA = VSV^T$. Consequently, the positive semi-definite rectangular matrix A can be recreated from U, V, S : $A = USV^T$. The covariance matrix A^TA is not necessarily a diagonal, but covariance matrix of rotated data AV is diagonal: $(AV)^TAV = (US)^TUS = SU^TUS = SS = S^2$ which is a diagonal matrix with non-negative values. Similarly $U^TA(U^TA)^T = SV^T(SV^T)^T = SV^TVS^T = SS = S^2$. Recall that $U^TU = I$ implies $UU^T = I$ This confirms that the orthonormal transformation matrices U and V un-correlate the correlated values because the covariance of AV and U^TA are diagonal. It follows from the diagonal matrix that (1) the pairwise covariance is zero, hence variables are uncorrelated in the new, rotated coordinate system. (2) eigenvalues of AA^T and A^TA are never negative. (3) trace(AA^T) is equal to the sum of eigenvalues of AA^T.

Now we have two orthonormal matrices U_{mxm} and V_{nxn}. The matrix A_{mxn} can be reduced by constraining U and V to (1) fewer columns AV_{nxk} by reducing the variables/attributes from n to k, k<<n, (2) fewer rows $U_{mxk}{}^TA$ reducing size of data from m observations/rows to k rows, k<<m. In case, we want to reduce both attributes and size, then we can use kxk instead of mxn A: $U_{nxk}{}^TAV_{mxk}$. The data size can be further reduced if the number of non-zero eigenvaluesis less than k<<min(m,n).

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

These concepts and results are used in error analysis for determining the optimal number of eigenvectors sufficient for dimension reduction. However, (1) it is difficult to extract a small number of features for any learning algorithm, and (2) PCA/SVD cannot determine exactly which of the original uncorrelated variables/attributes can be dropped. Thus it cannot determine which original variables are important [1]. As we calculate eigenvectors, it becomes trivial that $A = USV^T$ implies $S = U^TAV$ that means A is diagonalizable.

When we transform the dataset, error is natural to occur due to projection for reducing dimensions. For example, for Wine data discussed in section 4.1, we find that as a result of SVD, using eigen space of dimension 1,2,3, etc. the error decreases, see Figure 2. It shows that just four dimensional transformed space data accounts for 98% of the data in the original space.

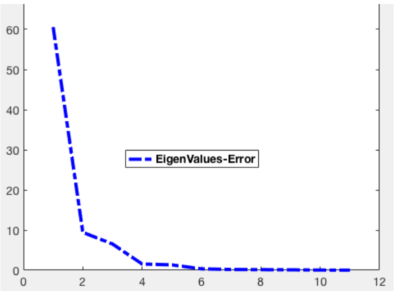


Figure 2. This shows errors due to reduced dimensionality when we use 1,2,3…,11 principal components. Reduction in data space that accounts for increased computational efficiency and increased error.

Most challenging part of PCA/SVD is interpretation of components in terms of original coordinates. Most of the time we reduce only the number of attributes (row length). It is more useful to apply SVD transformation before applying classification tree algorithm. This can be done more efficiently in transformed space. Moreover, we apply classification learning in the transformed space with substantial reduction in computational effort and storage space at insignificant cost in accuracy.

B. Algebraic Foundations and Dimension Reduction

For a positive semi-definite matrix A, the SVD decomposition $A = USV^T$ is full spectrum decomposition. If we use fewer, say k, columns of V, it reduces to k features in the new frame, k << min(m,n), in the m×n data where k indicates the number of columns used, S_k is a k×k diagonal matrix. Then $A_k = U_kS_kV_k^T$, A_k has least error from A.

C.1 Eigen-Decomposition Analysis

For a symmetric matrix A, eigenvalues exist and are real. The eigenvectors generate the vector space of matrix rows. The eigenvectors corresponding to different eigenvalues are orthogonal.

Theorem. Let A be a real symmetric matrix. Show that there exists an orthogonal matrix V and a diagonal matrix D such that $A = VDV^T$ where V is the matrix of eigenvectors of A; and D is the diagonal matrix of eigenvalues of A.

Proof. Let $A\ \mathbf{v_i} = \lambda_i\ \mathbf{v_i}$ for which $(\lambda_i, \mathbf{v_i})$ is a corresponding eigenvalue and eigenvector pair. We arrange the eigenvalues and eigenvectors on descending order of eigenvalues.

Let D be diagonal matrix of eigenvalues i.e., $d_{ii} = \lambda_i$ or $D = [\lambda_i]$, and V be the matrix whose columns are eigenvectors, i.e., $V = [\mathbf{v_i}]$. Since V is orthogonal matrix, $V^{-1} = V^T$. From eigenpair equation

$$A\ \mathbf{v_i} = \lambda_i\ \mathbf{v_i}$$

using matrix notation, it becomes

$$A\ [\mathbf{v_i}] = [\lambda_i\ \mathbf{v_i}] = [\mathbf{v_i}\,\lambda_i] = [\mathbf{v_i}][\lambda_i]$$

Since A is symmetric $V^{-1} = V^T$

$$AV = VD \rightarrow A = VD\,V^{-1} = VDV^T$$

Note that the symmetry constraint on A is sufficient, but not necessary.

Example: The matrix A *has eigen-decomposition*, *even when A* is *not symmetric*.

The matrix $\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix}$ has eigenvalues: 0,1. Eigenvectors: $\begin{bmatrix}1\\-1\end{bmatrix}$, $\begin{bmatrix}0\\1\end{bmatrix}$ which are linearly independent, but *non-orthogonal* eigenvectors. The eigen-decomposition is $\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix} = \begin{bmatrix}0 & \frac{1}{\sqrt{2}}\\1 & -\frac{1}{\sqrt{2}}\end{bmatrix} \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} \begin{bmatrix}1 & 1\\\sqrt{2} & 0\end{bmatrix}$.

This example shows that eigenvectors are not orthogonal, and $V^{-1} \neq V^T$, thus $A = VDV^{-1}$, but $A \neq VDV^T$.

For PCA, the eigenvectors of A^TA and AA^T are used.

The matrix $A^TA = \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$ has eigenvalues:0, 2. Eigenvectors: $\begin{bmatrix}1\\-1\end{bmatrix}$, $\begin{bmatrix}1\\1\end{bmatrix}$ which are linearly independent, *orthogonal.*

The matrix $AA^T = \begin{bmatrix}0 & 0\\0 & 2\end{bmatrix}$ has eigenvalues:0,2. Eigenvectors: $\begin{bmatrix}1\\0\end{bmatrix}$, $\begin{bmatrix}0\\1\end{bmatrix}$ which are linearly independent, *orthogonal* forming $U = \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}$, $V = \frac{\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}}{\sqrt{2}}$ here $U \neq V^T$, $S = \begin{bmatrix}\sqrt{2} & 0\\0 & 0\end{bmatrix}$,

Note 1. The PCA is $\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix} = \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix} \begin{bmatrix}\sqrt{2} & 0\\0 & 0\end{bmatrix} \frac{\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}}{\sqrt{2}}$ which eigen-decomposition, SVD, but different.

Note 2. The matrix $\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix}$ does not have PCA but SVD, $\begin{bmatrix}0 & 0\\-1 & -1\end{bmatrix} = -\begin{bmatrix}0 & 0\\1 & 1\end{bmatrix}$ has no PCA , or SVD

C.2 PCA Analysis

Here we see that for a symmetric matrix, eigen-decomposition is guaranteed. For a square, symmetric matrix A, A^2 is also a square symmetric matrix, eigenvalues of A^2 are real and non-negative. Let $(\alpha, \mathbf{u})$ be eigenpair for A^2, i.e., $A^2\ \mathbf{u} = \alpha\ \mathbf{u}$. then $\alpha \geq 0$, because

$$\alpha = \alpha\ (\mathbf{u}, \mathbf{u}) = (\alpha\ \mathbf{u}, \mathbf{u}) = (A^2\,\mathbf{u}, \mathbf{u})$$

$$= (A\,\mathbf{u}, \mathbf{A^Tu}) = (A\,\mathbf{u}, \mathbf{Au}) \geq 0 \text{ from symmetry of A.}$$

Theorem. If A is a symmetric square matrix and **u** is a unit eigenvector of A^2 with eigenvalue α, then

(1) A**u** is also an eigenvector of A^2 with the same

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

eigenvalue α,

(2) $A\mathbf{u} = \sqrt{\alpha}\ \mathbf{v}$ and $A\mathbf{v} = \sqrt{\alpha}\ \mathbf{u}$ where **u, v** are unit eigenvectors of A^2.

Proof: let $A^2\mathbf{u} = \alpha\ \mathbf{u}$, then $AA^2\mathbf{u} = \alpha\ A\mathbf{u}$
then

$$AA^2\mathbf{u} = \alpha\ A\mathbf{u}$$
$$A^2(A\mathbf{u}) = \alpha\ A\mathbf{u}$$

Since $\mathbf{u} \neq 0$, and $A\mathbf{u} \neq 0$, $A\mathbf{u}$ is an eigenvector of A^2 for the same eigenvalue.
Let us denote the unit eigenvector by **v**
The $A\mathbf{u} = \beta\ \mathbf{v}$ for some beta.
We show that $\beta = \sqrt{\alpha}$ □□

$$\alpha\ □ = \alpha\ (\mathbf{u}, \mathbf{u}) = (\alpha\ \mathbf{u}, \mathbf{u}) = (A^2\ \mathbf{u}, \mathbf{u}) = (A\ \mathbf{u}, A^T\mathbf{u}) = (A\ \mathbf{u}, A\mathbf{u}) = (\beta\ \mathbf{v}, \beta\ \mathbf{v}) = \beta^2\ (\mathbf{v},\ \mathbf{v}) = \beta^2$$

therefore $\alpha\ □ = \beta^2$
Since $\alpha\ □ = \beta^2$

therefore $A\mathbf{u} = \pm\sqrt{\alpha}\ \mathbf{v}$

Let us use + symbol for simplicity. Also it can be verified that $A\mathbf{v} = \sqrt{\alpha}\ \mathbf{u}$. For example,
$A\mathbf{u} = \sqrt{\alpha}\ \mathbf{v}$ implies $A^2\mathbf{u} = \sqrt{\alpha}\ A\mathbf{v}$ or $\alpha\ \mathbf{u} = \sqrt{\alpha}\ A\mathbf{v}$ or $A\mathbf{v} = \sqrt{\alpha}\ \mathbf{u}$

Example This example shows that each eigenvector of A^2 is an eigenvector of A.

let $A = \begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}$, $A^2 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$,

eigenvalues of A^2 are 1, 1 and eigenvectors $\mathbf{u} = \begin{bmatrix}1\\0\end{bmatrix}$, $\mathbf{v} = \begin{bmatrix}0\\1\end{bmatrix}$,

eigenvalues of A are 1, -1 and eigenvectors $\mathbf{u} = \begin{bmatrix}1\\0\end{bmatrix}$, $\mathbf{v} = \begin{bmatrix}0\\1\end{bmatrix}$,

$$A^2\ \mathbf{u} = \mathbf{u}\ ,\ A\ \mathbf{u} = \mathbf{u}.$$
$$A^2\ \mathbf{v} = \mathbf{v}\ ,\ A\ \mathbf{v} = -\ \mathbf{v}.$$

$A = USU^T$

Trivially, $\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

Theorem. Let A be symmetric matrix. Show that there exists an orthogonal matrix U and a diagonal matrix S such that $A = USU^T$ where U is matrix of eigenvectors and S is the matrix of square roots of eigenvalues of A^2.

Proof. In PCA, we arrange the eigenvalues and eigenvectors according to descending order of eigenvalues. Let $S = [\lambda_i]$ be diagonal matrix of square roots of eigenvalues and $U = [\mathbf{u_i}]$ be matrix of eigenvectors of A^2. Since U is orthogonal matrix, $U^{-1} = U^T$ and

$$A\ \mathbf{u_i} = \lambda_i\ \mathbf{u_i}$$

which implies that

$$A\ [\mathbf{u_i}] = [\lambda_i\ \mathbf{u_i}] = [\mathbf{u_i}\,\lambda_i]$$

becomes

$AU = US \rightarrow A = US\ U^{-1}$ or $A = USU^T$

C.3 SVD Analysis

If A is not positive semi-definite, we may not have SVD decomposition. For example, let A be a rectangular matrix. Then AA^T and A^TA are symmetric and positive semi-definite. The eigenvalues of AA^T and A^TA exist, are non-negative and identical. Let U be matrix of eigenvectors of AA^T, and V be matrix of eigenvector of A^TA, also $A^TA\ \mathbf{v} = \lambda□\mathbf{v}$ implies $AA^T\ (A\ \mathbf{v}) = \lambda□(A\ \mathbf{v})$ or $A\mathbf{v}$ is eigenvector of AA^T. Since **u** and **v** are unit vectors, the relation between A^Tu and **v** is $A^Tu = \mu\ \mathbf{v}$ for some scalar μ; and the relation between $A\mathbf{v}$ and **u** is $A\mathbf{v} = \mu\ \mathbf{u}$ where $\mu = \sqrt{\lambda}$. For example, if $A^Tu = \mu\ \mathbf{v}$, then

$$\lambda = \lambda\ (\mathbf{u}, \mathbf{u}) = (\lambda\ \mathbf{u}, \mathbf{u}) = (AA^T\ \mathbf{u}, \mathbf{u})$$
$$= (A^T\ \mathbf{u}, A^T\mathbf{u}) = (\mu\ \mathbf{v}, \mu\ \mathbf{v})$$
$$= \mu^2\ (\mathbf{v},\ \mathbf{v}) = \mu^2$$

Thus if we know U or V, the other is readily available.
We have now

$A\mathbf{v_i} = \mu_{\mathbf{i}}\mathbf{u_i}$ for each non-zero eigenvalue $\mu_{\mathbf{i}}$,

zero value contribute nothing to the matrix of U, S, V.

$$A[\mathbf{v_i}] = [\mu_{\mathbf{i}}\mathbf{u_i}\ ]$$
$$A[\mathbf{v_i}] = [\mathbf{u_i}\,\mu_{\mathbf{i}}] = [\mathbf{u_i}\ ][\ \mu_{\mathbf{i}}]$$
$$AV = US$$

This is called the *polar* decomposition. An interesting outcome of this equation is that projection of right singular vectors are scores of left singular vectors. Eigenvectors are called factors and eigenvalues are called loads. Since U and V are orthogonal, the equation AV = US leads to spectral decomposition of A:

$$A = USV^T$$

Example. The matrix A is *not positive semi-definite*, but A is symmetric. The PCA and SVD exist and are not equal to eigen-decomposition.

The matrix $A = \begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}$ has eigenvalues 1 and -1 and eigenvectors as $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$. It shows that the eigenvectors for $AA^T = A^TA = A^2$ are same, but the eigenvalues are 1,1. This results in

$U = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} = V = V^T$, $D = \begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}$ and, $S = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ and not $A \neq USV^T$ though we have $A = UDU^{-1} = UDU^T$.

If A is positive definite, square or rectangular, we can have $A = USV^T$, where V is not necessarily U.

IV. THE ALGORITHM AND THE METRICS FOR ACCURACY ANALYSIS

A. The Algorithm

Routinely the algorithms for classification and dimensionality reduction are used standalone to accomplish the intended tasks. As a result, computational efficiency and storage efficiency suffer. We create a hybrid algorithm to take advantage of the computational and storage efficiency first and then apply the classification tree algorithm to data in the compressed domain. As a result we get substantial efficiency, on the average, we gain 18% computational efficiency at the cost of .08% accuracy reduction. Further optimization depends on the size of data and data properties. More the number of attributes, better the performance of the algorithm. Figure 3 describes this strategy.

```
Input: m×n data matrix A
Output: classification tree for $A_{mxm}$ from reduced data $B_{mxk}$
The enhanced SVD algorithm is as follows
    Ignore the missing values from consideration.
    Create covariance matrices $AA^T$ and $A^TA$ normalized by
        the dimensions
    Calculate eigenvalues and eigenvectors (normalized to
        unity) of $AA^T$ and $A^TA$
    Calculate eigenvalues of A to track the negative
```

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

```
        eigenvalues
    Calculate the square roots of values of $AA^T$ and $A^TA$
    Update the signs according to the sign for eigenvalues of
        A
    Rank the updated eigenvalues in descending order and
        form a diagonal matrix S
    Rank the eigenvectors for $AA^T$ using the order of
        eigenvalues and form matrix U
    (Once U is computed V is readily available)
    Rank the eigenvectors for $A^TA$ using the order of
        eigenvalues and form matrix V
    Determine k<<n for the number of eigenpairs acceptable
        for data reduction.
    Transform m×n A to reduced m×k data set $B = AV_k$
The Decision tree algorithm in general is
    Convert the classification numeric attribute to intelligent
        interval categorical attributes,
    Apply Entropy based Greedy attribute selection
    Use entropy to determine the attribute selection at each
        node of the tree construction using
      max entropy gain
    Create classification tree nodes
    The leaf nodes are classes and internal nodes are tree
        paths to generate rules.
Hybrid approach
    Collect data, clean the data items with incomplete values
    Apply the enhanced PCA that may reduce classifier
        dimension based
    Reduce the data set by ignoring contribution of
        unacceptable eigenvalues (near zero eigenvalues)
    Apply Classification Tree algorithm in the reduced
        domain and original base domain
    Analyze and compare the results.
```

Figure3. Algorithm for Hybrid approach

The enhanced PCA/SVD algorithm is implemented in Matlab and decision tree algorithm J48 is used from Weka software. Benchmark data on wines is obtained from UCI ML dataset repository. The following metrics are used to benchmark the accuracy.

B. *The Metrics for Accuracy Analysis*

B.1 *Metrics for Decision Trees*

The Data mining community uses *gold* standard, *precision, recall and F measures*, to determine the predictive accuracy of the classification [15]. For example, *Precision* is related to the accuracy of positive prediction, predicting negative as positive that means false positive, a false alarm, whereas, *Recall is* related to the accuracy of prediction on positive data, predicting positive to negative that meaning false negation, a missed case.

$$Precision = P = \frac{TP}{TP+FP}$$

$$Recall = R = \frac{TP}{TP+FN}$$

One can err on one side or the other: if one decreases the number of false negatives to ensure that it is less likely to miss an actual value, or one can reduce the number of false positives at the cost of misses. One can "fine tune" the detection algorithms. One way to tune is the *F-measure,* which is the *weighted Harmonic mean* of Precision and Recall. For example, the F-measure [15], is defined by Harmonic mean of P and R,

$$F = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} \quad or \quad F = \frac{PR}{(1-\alpha)\,P + \alpha\,R}$$

where α ∊ [0, 1]. The weight α can be fine-tuned by decreasing the missed at the cost of increasing the false alarms. The *default* balanced F-measure is with equally weighted precision and recall, which means making α = 1/2 making

$$F = \frac{2PR}{P+R}$$

B.2 Metrics for Dimensionality Reduction SVD

Based on k<<n, k highest eigenvalues out of n eigenvalues are selected to analyze dimensionality reduction. Recall A is the original matrix; V is the matrix whose columns are eigenvectors of A^TA. There are three equivalent ways to measure errors; experimental simulations verify this and the theorems confirm it. The value of k is determined as follows: First by dropping near zero eigenvalues, the error is the based on the dropped eigenvalues λ_p, p=k+1…n. Second since corresponding eigenvector are dropped, the projection in eigenspace becomes AV and projection space constrained to k dimensions becomes newAV. Thirdly, the projection translated in the base space becomes newA.

Error is measured in three ways:

$$\frac{\Sigma p{=}k{+}1{,}n\ \lambda p}{\Sigma p{=}1{,}n\ \lambda p}$$ relative error in the eigenvalues

$$\frac{|AV-newAV|}{|AV|}$$ relative error in projection space.

$$\frac{|A-newA|}{|A|}$$ relative error in the original space

For logical data, the data is coerced to numerical integer data, thus quantization is performed before error checking.

V. IMPLEMENTATION AND EXPERIMENTATION

The SVD algorithm, implemented in Matlab, is enhanced by prioritizing the directions of eigenvectors produced by the Matlab svd function. In addition to entropy heuristic used in the Weka tree construction algorithm J48, we updated it to include svd as part of the tree construction. The metrics in section 3 are used to benchmark the overall accuracy obtained in the confusion matrix.

Benchmark data on Wine dataset is obtained from UCI [16] where it has been extensively studied. Besides the classic site, there are numerous sites where the same data sets are available. For our purposes data is ready to use for analysis. Many times it may still not be practical to use directly due to the data size and computational bottlenecks. Data customization may need to be performed, e.g. data reduction in size and removal of attributes irrelevant to classification. If data gets transformed, it may be mapped in the original space for analysis or the analysis may be performed in the transformed domain.

We experimented it both ways and found that either

ISSN 2395-8618

way the algorithm performs better that brute force decision tree algorithm. We show results of experiments in Table 2 and Table 3. Our purpose is to show the improvement in reduced computations of the hybrid algorithm on classification without loss of accuracy.

V. WINES DATASET

A predictive model on Wine dataset is useful to provide guidance to vineyards regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters [16]]. Two datasets are available, of which one dataset is on red wine and has 1599 different varieties. All wines are produced in a particular area of Portugal. Dataset has 12 attributes: (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality). Quality is based on sensory data and the rest are function of chemical properties of the wines. All 11 chemical properties of wines are real variables, whereas Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Since Weka J48 uses non-numeric categorical attribute for classification, we replaced the numbers 1 to 10 by number strings One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Ten.

Before we use our algorithm, we analyze by using standalone Weka algorithm J48, classification model is constructed and confusion matrix for accuracy calculated for baseline comparison. We used SVD to determine the eigenvalues and found that the ratio of two smallest eigenvalues to largest eigenvalue is .0004 (all 11 eigenvalues are 60.573958, 9.433888, 6.540730, 1.550106, 1.338038, 0.346191, 0.188387, 0.148977, 0.101789, 0.040841, 0.024550). We decided to drop the smallest two eigenvalues out of 11 eigenvalues and apply the classification to the reduced dataset. The result show the there is 18% reduction in data with no loss of accuracy, in fact accuracy improved slightly. This may account for noise reduction. Computation is also speeded up by applying the algorithm in the compressed space with retaining slightly better accuracy because conversion to base space is eliminated. Thus we reduced the data by 18% and applied the decision tree algorithm in two ways: in original space and in compressed space. The results are shown as confusion matrix in Table 2 and Table 3

The latent variables do not tell much about the original variables. Prediction of Quality ranking from the chemical properties of the wines can be inferred from the latent variables directly.

TABLE 2. CONFUSION MATRIX FOR DECISION TREE GENERATED BY J48 ALGORITHM

Precision	Recall	F-Measure	Class
0.2	0.111	0.143	Eight
0.492	0.472	0.482	Seven
0.609	0.635	0.622	Six
0.71	0.72	0.715	Five
0.054	0.038	0.044	Four
0.167	0.1	0.125	Three
0.612	**0.622**	**0.616**	**weighted average**

TABLE 3. DATA REDUCTION WITH SVD AND J48 ALGORITHM IN RAW, REDUCED, AND TRANSFORMED DOMAIN.

	PRECISION	RECALL	F-MEASURE
Decision Tree J48 Raw data	0.612	0.622	0.617
Decision Tree in Original domain after 18% data reduction	0.615	0.625	0.620
Decision Tree in reduced domain after 18% data reduction	0.617	0.626	0.621

VI. CONCLUSION

We have given a hybrid algorithm that takes advantage of reducing data by dropping near zero eigenvalues and applying classification algorithm on latent variables. We applied the hybrid algorithm to a well-known benchmark dataset of a collection of Wines. For comparison the classification algorithm is applied in both the full base space and reduced eigenspace. We have shown that hybrid algorithm outperforms the usual standalone algorithm applied to data reduction and classification on their own for the intended tasks. Though accuracy result shown by confusion matrices are comparable, but the reduction in computational effort and storage space is significant. The application developers working in this area will find it useful in real time computations.

VII. REFERENCES

[1] Kirk Baker, Singular Value Decomposition Tutorial https://www.ling.ohiostate. edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf , January 2013.

[2] Jonthan Shlens A Tutorial on Principal Component Analysis, arXiv:1404.1100 [cs.LG], pp. 1-15,2014]

[3] Karen Bandeen-Roche Nov 28, 2007, An Introduction to Latent variable Models, http://www.biostat.jhsph.edu/~kbroche/Aging/Intro to Latent VariableModels.pdf

[4] Jim Hefferon, Linear Algebra, Free Book, http://joshua.smcvt.edu/linearalgebra, 2014.

[5] Abdi, Hervé, Beaton, Derek, Principal Component and Correspondence Analyses Using R, Springer, ISBN 978-3-319-09256-0, Digitally watermarked, DRM-free, 2017

[6] Suresh kumar Gorakala, Principal Component Analysis in R, Percept Psychology (2016)78:3-20

[7] Caroline J Anderson, Psychology Lecture Notes: Principal Component Analysis 2017

[8]Sebastian Raschka Principal Component Analysis in 3 Simple Steps LSA-Least Squares Approximation http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html, 2015.

[9] Jure Leskovec, Anand Rajaraman, Jeffrey D Ullman, Datamining of Massive Datasets, 2014

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

[10] Patrick J.F. Groenen, Michel van de Velden, Multidimensional Scaling, Econometric Institute EI 2004-I5, Erasmus University Rotterdam, Netherlands, 2015.

[11] An SVD-Entropy and Bilinearity based product ranking algorithm using heterogeneous data, Journal of Visual Languages & Computing, https://doi.org/10.1016/j.jvlc.2017.06.001, 2017.

[11] Christopher K. I. Williams 1998 - On a Connection between Kernel PCA and Metric, Multidimensional Scaling, Machine Learning, Volume 46, Issue 1–3, pp 11–19, January 2002.

[12] Rita Osadchy , Unsupervised learning 2011 DIMENSIONALITY REDUCTION: PCA, MDS http://www.cs.haifa.ac.il/~rita/uml_course/lectures/PCA_MDS.pdf

[13] Matthew Mayo, All machine learning models have flaws, http://www.kdnuggets.com/2015/03/all-machine-learning-models-have-flaws.html

[14] Matthew Mayo, Simplifying Decision Tree Interpretability with Python & Scikit-learn, 2017 http://www.kdnuggets.com/tag/decision-trees.

[15] Wikipedia, the free encyclopedia https://en.wikipedia.org/wiki/Precision_and_Recall

[16] Center for Machine Learning and Intelligent Systems, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data, https://onlinecourses.science.psu.edu/stat857/node/223, 2016.

VIII. APPENDIX A

The linear algebra definitions used here can be found in any textbook [4]. Briefly we describe in frequent terms definitions. We adopt the convention that the vectors are *column vectors* by default. The linear algebra concepts of *vector*, *transpose* of a vector, *scalar* product of a vector, Euclidean *norm or length* of a vector, *unit* vector, *sum* of two vectors, *dot product* of two vectors, *orthogonal* vectors, *Gram-Schmidt* orthogonalization, *matrix*, *square* matrix, *identity* matrix, *diagonal* matrix, *transpose* of matrix, *symmetric* matrix, *sum* of matrices, *product* of matrices, *inverse* of a matrix, *orthogonal* matrix, *norm* of a matrix, *orthogonal* matrix, *rotation* matrix, *rank* of a matrix, *determinant* of a matrix, *vector space*, and *basis* of a vector space, are standard terms in linear algebra. Additional terms that we use are an *eigenvector*, and an *eigenvalue*. By conventions of linear algebra, all vectors are column vectors unless categorically and specifically stated. For details on linear algebra, reader may consult references

For convenience in reading this paper, the notation necessary for readability is: matrices are described in *uppercase*, vectors are in *lowercase bold*, and the elements in matrices and elements in vectors are *lowercase italic*. For example,

Definition. If **u** and **v** are both *row* vectors or both *column* vectors, and **v** is a *unit* vector, then the scalar *projection* of **u** on **v** is given by **u•v** which is expressed as a matrix product
if **u** and **v** are column vectors, then **u•v** = $\mathbf{u}^T$ **v** =

$$[u_1 \quad \cdots \quad u_n]\begin{bmatrix} v_1 \\ \cdots \\ v_n \end{bmatrix}$$

Definition. The vectors u and v are *orthogonal* if **u•v = $\mathbf{u}^T\mathbf{v}$ = 0.**

Definition. A set of vectors is *orthonormal* if each vector is a unit vector, and any two different vectors are mutually orthogonal.

A matrix is a 2D array consisting of rows and columns. For a matrix A, we use the shorthand notation for matrix A = $[a_{ij}]$, where a_{ij} is the ij-th element of A. If A is a matrix, $\mathbf{a}_{i\bullet}$ is a row vector representing the i-th row of matrix A, and $\mathbf{a}_{\bullet j}$ is a column vector representing the j-th column of A**.** Thus ith row is $\mathbf{a}_{i\bullet}$ = [a_{i1}, a_{i2},..., a_{in}]. Similarly jth column is $\mathbf{a}_{\bullet j}$ = $\begin{bmatrix} a_{1j} \\ a_{2j} \\ \cdots \\ a_{mj} \end{bmatrix}$.

Definition. The matrix A= $[a_{ij}]$ is said to be *symmetric* if $a_{ij} = a_{ji}$ for all i,j. That is, the elements across the main diagonal are identical, that is, A is equal to its transpose, A = A^T.

Definition. The norm of a matrix A is defined as the square root of the sum of squares of all elements in A. If A = $[a_{ij}]$ then $|A| = \sqrt{(\sum_{i,j} a_{ij}^2)}$

Definition. The trace of a matrix A is defined as the sum of entries on its main diagonal. If A = $[a_{ij}]$ then trace(A) = $\sum_i a_{ii}$

Definition. The matrix A is *orthogonal*, if the rows (and columns) are pairwise orthogonal, and $AA^T = A^TA = I$, the identity matrix.

There are three statistical terms associated with a matrix A: trace , rank, determinant,

Proposition. For a square matrix A,

$$\mathrm{trace}(AA^T) = \mathrm{trace}(A^TA) = \mathrm{trace}(A^2) = |A|^2$$
$$\mathrm{trace}(AA^T) = \textstyle\sum_i \boldsymbol{a}_{i\bullet}\, a_{i\bullet}{}^T = \sum_i \boldsymbol{a}_{i\bullet} \bullet a_{i\bullet}$$
$$\mathrm{trace}(A^TA) = \textstyle\sum_i \boldsymbol{a}_{\bullet i}{}^T\, a_{\bullet I} = \sum_i \boldsymbol{a}_{\bullet i} \bullet a_{\bullet i}$$
$$= \textstyle\sum_i \sum_k a_{ik}\, a_{ik}, \quad = \sum_i \sum_k a_{ki}\, a_{ki},$$
$$\text{In either case} \quad = \textstyle\sum_i \sum_k a_{ik}{}^2 = |A|^2$$

It is the sum of squares of all the entries in A, |A| = $\sqrt{}$trace(AA^T) = $\sqrt{}$trace(A^TA) = $\sqrt{}$trace(A^2)

Definition. The *rank* of a matrix A is the number of linearly independent rows (and columns) in a matrix. It is equivalent to number of non-zero eigenvectors.

Definition. The *determinant* of a matrix A is denoted by det(A) and is defined as product of the eigenvalues of A.

Definition. For a matrix A, if there exists a non-zero vector **u** and a real number λ□such that A**u** = λ **u**, then λ is called an *eigenvalue* and **u** is called the corresponding *eigenvector*. The term singular value and singular vecror are also used.

If λ is an eigenvalue, it is computed by solving the *characteristic* equation det(A- λ I)=0.

Note. Any non-zero multiple of a eigenvector is again an

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

eigenvector. To make them unique, they are normalized to unit vectors, see Figure A1 (a). But if **u** is unit eigenvector, then –**u** is also a unit vector. In the literature, they use the convention of making the first non-zero component positive in the eigenvector, see Figure A1(b). Since eigenvectors are ordered, we propose to make the k-th element of k-th vector to be positive, see FigureA1(c) that makes the vectors look like a right handed system.

Theorem. If the matrix is *symmetric*, then for different eigenvalues, the eigenvectors are orthogonal.

Definition. The matrix A is diagonalizable if there is an invertible matrix V such that A = V D V^{-1} where D is diagonal matrix of eigenvalues. If there are n distinct eigenvalues for n×n matrix, then it is diagonalizable.

Theorem. If the matrix is *symmetric or diagonalizable*, then there are as many eigenvectors as eigenvalues.

Proposition The eigenvalues of a real symmetric matrix are real.

Let A **u** = $\square\ \lambda\ \mathbf{u}$, **u** is a unit vector.

$$\square\square\ \lambda\ \square = \lambda(\mathbf{u},\mathbf{u})\ \square = (\mathbf{u},\lambda\mathbf{u}) = (\mathbf{u},A\mathbf{u})$$
$$= \square\ \overline{(A\mathbf{u},\mathbf{u})} = \overline{(\lambda\mathbf{u},\mathbf{u})} = \square\ \overline{\lambda(\mathbf{u},\mathbf{u})}$$
$$= \square\ \bar{\lambda}$$

Therefore $\lambda = \bar{\lambda}$, that is λ is equal to its complex conjugate, thus eigenvalues are real.

Definition. A matrix A is positive semi definite if $\mathbf{v}^T A\mathbf{u} \geq 0$ for all vectors **u,** and **v**.

Proposition. For a positive semi definite matrix A, eigenvalues must be non-negative.

Proof. If **u** is an eigenvector of A, then A**u** = λ **u**, that is, $\lambda = \mathbf{u}^T A\mathbf{u} \geq 0$, thus eigenvalues are non-negative .

Proposition. If A is a positive semi-definite matrix, then

- *Eigenvalues of* AA^T *and* A^TA *are non-negative*

Let λ be eigenvalue of A^TA and AA^T.

$\lambda = \lambda\,(\mathbf{v},\mathbf{v}) = (\lambda\mathbf{v},\mathbf{v}) = (A^TA\ \mathbf{v}, \mathbf{v}) = (A\mathbf{v}, A\mathbf{v}) \geq 0$ and is the variance of A**v**

- *Eigenvalues of* AA^T *and* A^TA *are identical.*

Let λ be eigenvalue of A^TA.

If $A^TA\ \mathbf{v} = \lambda\ \mathbf{v}$, then $AA^TA\ \mathbf{v} = \lambda\ A\mathbf{v}$

Since $\mathbf{v} \neq 0$, $A\mathbf{v} \neq 0$, and $AA^T(A\mathbf{v}) = \lambda\ (A\mathbf{v})$

Therefore A**v** is an eigenvector of AA^T, hence λ is an eigenvalue of AA^T.

Similarly, if λ be eigenvalue of AA^T, then λ be eigenvalue of A^TA

- *Eigenvectors for different eigenvalues of* AA^T *and* A^TA *are orthogonal.*

Since AA^T and A^TA are symmetric, the eigenvectors are orthogonal.

Also, let **u**, **v** be eigenvectors and α, β be corresponding eigenvectors of AA^T and A^TA respectively. Then

$$\alpha(\mathbf{u},\mathbf{v}) = (\alpha\ \mathbf{u},\mathbf{v}) = (AA^T\mathbf{u},\mathbf{v})$$
$$= (\mathbf{u}, AA^T\mathbf{v}) = (\mathbf{u},\beta\mathbf{v}) = \beta\ (\mathbf{u},\mathbf{v})$$
$$\alpha(\mathbf{u},\mathbf{v}) = \beta\ (\mathbf{u},\mathbf{v})$$

if $\alpha \neq \beta$, (**u**, **v**) must be zero.

- If $AA^T\ \mathbf{u} = \lambda\ \mathbf{u}$ and $A^T\ \mathbf{u} = \mu\ \mathbf{v}$, **v** is unit vector then $\mu = \sqrt{\lambda}$

$$\lambda\ \square = \lambda(\mathbf{u},\mathbf{u})$$
$$= (\lambda\ \square\mathbf{u}, \mathbf{u}) = (AA^T\mathbf{u},\mathbf{u})$$
$$= (A^T\mathbf{u}, A^T\mathbf{u}) = (\mu\mathbf{v},\mu\mathbf{v})$$
$$= \mu^2\ (\mathbf{v},\mathbf{v}) = \mu^2$$

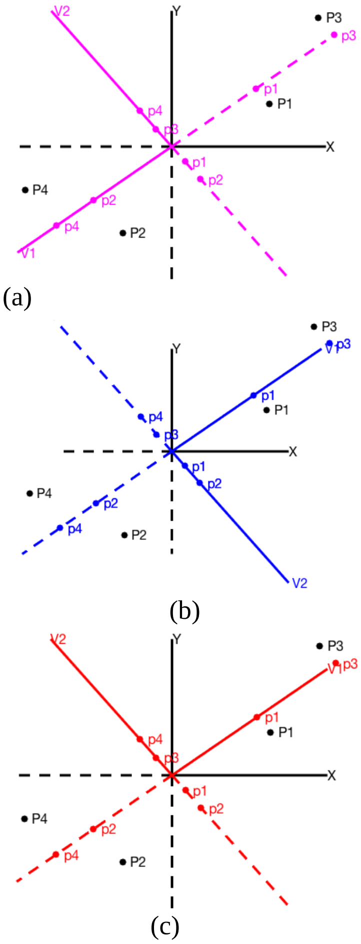


Figure A1. Matrix A representing four points P_1, P_2, P_3, P_4, axes are in black. V1 and V2 vectors resulted by using Matlab functions (a) Magenta: eig() function is used, we get arbitrary signs of vectors (b) Blue: svd() function is used and first component coerced to be positive for all eigenvectors, (c) Red: we organized it to look more natural right handed system, adaptively making the sequential component positive.

Proposition. For a symmetric matrix A, trace(A) = trace(UDU^T) = trace(D) = sum of eigenvalues of A, that is, trace(A) = $\sum_i \lambda_i$

Proof.

$$\text{trace}(A) = \text{trace}(UDU^T) =$$
$$= \textstyle\sum_i \boldsymbol{u_{i\bullet}}\, D\boldsymbol{u_{i\bullet}}^T$$
$$= \textstyle\sum_i\ [\ u_{ik}\ \lambda_k]\ \boldsymbol{u_{i\bullet}}^T\ \textit{where}\ [\ u_{ik}\ \lambda_k]$$

is a row vector with index k

$$= \textstyle\sum_i\sum_k\ u_{ik}\,\lambda_k\,u_{ik}$$
$$= \textstyle\sum_k\ \lambda_k \sum_i u_{ik}\,u_{ik}$$

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

$= \sum_k \lambda_k \boldsymbol{u}_{\bullet k} \bullet \boldsymbol{u}_{\bullet k} \quad \mathbf{u}_{\bullet k}$ is a unit column vector.

$= \sum_k \lambda_k$

Thus it shows that trace(A) is the sum of eigenvalues of A.

Chaman L. Sabharwal was born at Ludhiana, Punjab, India, in 1937. He received his B.A.(Hons) in 1959, M.A.(Math) in 1961 from Panjab University, Chandigarh, India. He received his M.S.(Math) in 1966 and Ph.D.(Math) from the University of Illinois, Urbana, Champaign, Illinois, USA, in 1967.

He is professor of Computer Science at Missouri University of Science and Technology (1986-). He was assistant professor (1967-971), associate professor (1971-9175), full professor (1975-1982) at Saint Louis University. He was Senior Programmer Analyst(1982), Specialist(1983), Senior Specialist(1984) Lead Engineer(1985) at Boeing Corporation. He published several technical reports and journal articles on CAD/CAM. He was consultant at Boeing (1986-1990). He was National Science Foundation fellow (1979) at Boeing and NSF Image Databases Panelist (1996).

Dr. Sabharwal has been member of American Mathematical Society, Mathematical Society of America, IEEE Computer Society, ACM, and ISCA. He has been on editorial board of International Journal of Zhejiang University Science (JZUS), Editorial Board CAD(Computer Aided Design), Progress In Computer Graphics Series, Modeling and Simulation, Instrument Society of America. He has been a reviewer for numerous books, journals and conferences. He was awarded service awards by NSF Young Scholars George Engelmann Institute, and ACM Symposium on Applied Computing for Multimedia and Visualization track. He is on program committees of several conferences including IEEEicSoftComm, MICAI, MIKE, MDS etc. including general Chair of MIKE2017.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Comparison of Classification Algorithms for File Type Detection A Digital Forensics Perspective

Konstantinos Karampidis, Ergina Kavallieratou, and Giorgos Papadourakis

***Abstract*— Digital forensics is a relatively new field in Computer Science and it focuses on the acquisition, preservation and analysis of digital evidence, in a way that that these evidences are suitable for presentation in a court of law. Forensic investigators follow a standard set of procedures. One major and difficult problem is the correct identification of file types. Criminals often hide evidence in a digital device, by changing the file type. It is very common, a child predator to try to hide image files with immoral content in order to fool police authorities. In this paper we examine a methodology for file type identification, which uses computational intelligence techniques for feature selection and classification. This methodology was applied to the three most common image file types (jpg, png and gif). In order to ascertain the method's accuracy, different machine learning classifiers were utilized. A three stage process involving feature extraction (Byte Frequency Distribution), feature selection (genetic algorithm) and classification (decision tree, support vector machine, neural network, logistic regression and k-nearest neighbor) was examined. Experiments were conducted having files altered in a digital forensics perspective and the results are presented. The examined methodology showed -in most cases- very high and exceptional accuracy in file type identification.**

***Index Terms*— artificial neural network, computer crime, digital forensics, file type detection, genetic algorithms, machine learning algorithms**

I. Introduction

DIGITAL forensics concerns the recovery and investigation of possible evidence found in digital devices. A digital forensics examination typically consists of four major phases; data collection, examination, analysis and report. Data examination is very critical because in this phase the evaluation of the collected data will be made. Therefore misjudgment of possible evidence may lead a court of law to wrong conclusions about criminal's activities. For example, it is very common, a child predator to try to hide image files with immoral content in order to fool police authorities. Typically they change file's extension, file's signature (magic bytes) or even both. On the other hand, forensic examiners use specialized forensic software to identify those forged files but in some cases even the best forensic software cannot identify correctly a file type. File type detection methods can be classified into three categories: extension, magic and content based methods. A lot of scientific methods have been proposed [1] but although some of them show more advantages than weaknesses, none is comprehensive or reliable enough to fulfil all the requirements. Due to the fact that extension-based methods are easy enough to be spoofed – just by a simple renaming of the file- and since the magic bytes in files cannot precisely determine true file type (because there is no predefined standard for the developers), the most significant methods are those who focus on the content of files. To achieve this, the content of each file is examined thoroughly and statistical modeling techniques are utilized to detect the correct file type. These methods are the most promising ones and proved to show the best results. M. McDaniel et al. [2], [3] suggested three algorithms for content-based file type detection. The correctness varied from 23 % to 96 % depending upon the algorithm used. W.J. Li et al. [4] worked on these algorithms and proposed to compute a set of centroid models and use clustering in order to find a minimal set of centroids with good performance while the use of more pattern data was considered necessary. Their methodology had a result of 82 % to 89.5% accuracy (one or multi centroid respectively) and 93.8 % accuracy when they examined a larger number of files in the dataset. Supervised learning techniques were used by J. Dunham et al.[5]. More specifically they used neural networks for classification and reached 91.3 % accuracy. M.C. Amirani et al. [6] used the Principal Component Analysis and unsupervised neural networks for the automatic feature extraction. They also used a neural network for classification, succeeding an accuracy of 98.33 % which was the best so far. D. Cao et al. [7] used Gram Frequency Distribution and vector space model with results of 90.34 % accuracy. I. Ahmed et al. [8] proposed two very interesting methods. Primary they used the cosine distance as a similarity metric when comparing the file content. Subsequent they decomposed the identification procedure into two steps. They used 2000 files of 10 file types as a dataset and achieved an accuracy of 90.19 %. I. Ahmed et

This work was submitted for review on 8/8/2016.

Konstantinos Karampidis is with Department of Information & Communication Systems Engineering, University of the Aegean, 83200 Karlovasi Samos, Greece (e-mail: karampidis@ outlook.com).

Ergina Kavallieratou is with Department of Information & Communication Systems Engineering, University of the Aegean, 83200 Karlovasi Samos, Greece (e-mail: kavallieratou@aegean.gr).

Giorgos Papadourakis is with Department of Informatics Engineering, Technological Educational Institute of Heraklion, 71500 Heraklion Crete, Greece (e-mail: papadour@cs.teicrete.gr).

https://doi.org/10.17562/PB-56-2

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

al. [9] also proposed two new techniques to reduce the classification time. The first method involved a feature selection technique and the K-nearest neighbor (KNN) as a classifier. The second method was the content sampling technique, which used a small portion of a file to obtain its byte-frequency distribution. M.C. Amirani et al.[10] then proposed an improved version of their first method by using a Support Vector Machine classifier and finally succeeded in raising the accuracy of the method to 99.16 %. J. D. Evensen et al. [11] used an n-gram analysis with naïve Bayes classifier to a large dataset of 60000 files (6 file types) with very good results achieving 99.51 % topmost. Finally, we proposed a new method [12] which included a three stage process involving feature extraction (Byte Frequency Distribution), feature selection (genetic algorithm) and classification (neural network). This method was tested to a large dataset in a digital forensics perspective and it showed extremely high accuracy (99.61%). All above papers refer to identification of whole files. Although our method, achieved extremely high accuracy even in a digital forensics perspective, we considered wise to explore whether the utilization of another classification method achieves better results. More specifically, we will reproduce the first two stages of the method and examine five different classification algorithms - decision tree, support vector machine, neural network, logistic regression and k-nearest neighbor- on the same dataset. Eventually an evaluation will be made about which classifier shows the best results for a forensic file type detection. A similar work [13] has been published but the authors examined fragments of files. The number of files they examined were too small (150 files of each type at topmost) in order to estimate accurately the correct file type and they used a different method for feature extraction (PCA). Finally the file types chosen for this study included PDF documents, JPG images, ASCII text files (TXT), Microsoft Word documents (DOC), HTML pages and executable files (EXE) which is far different from our point of interest. The rest of the paper is organized as follows: In Section 2 the proposed methodology is described, in Section 3 the different classifiers and the experimental parameters which utilized are briefly described and in Section 4 the experimental results are presented followed by conclusions.

II. Methodology of the Examined Method

Due to their importance to Digital Forensics, the identification of the most common image file types (jpg,png,gif) will be examined. The scientific method we examine in this paper uses computational intelligence techniques in order to identify the file type and to reveal the correct type if the file is altered. It involves feature extraction, feature selection and classification. Primarily all files from the dataset are loaded and the features are extracted. Byte Frequency Distribution (BFD) is used as feature extraction method. The number of incidences of each byte value in an input file is counted and an array with elements from 0 to 255 is created. Then each element of the array is normalized by dividing with the maximum occurrence. The final result is a file containing 256 features for each instance. Subsequently, feature selection is performed using a genetic algorithm. Genetic algorithms can provide candidate solutions. Each candidate solution (chromosome) is represented by a binary feature vector of dimension 256, where zero (0) indicates that the respective feature is not selected, and one (1) indicates that the feature is selected. The score of each candidate solution is evaluated by a fitness function. As a fitness function the Correlation based Feature Selection (CFS) [14] algorithm is utilized. This algorithm evaluates the candidate solutions from the genetic algorithm and choses those which include features highly associated to the file type category and low correlated with each other, by calculating each candidate's solution merit. Let S be a candidate solution consisting of k features. The merit of each candidate solution is calculated as shown in (1).

$$MeritS_{\mathrm{k}}=\frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \qquad (1)$$

,where:

$\overline{r_{cf}}$ is the average value of all feature-classification correlations and

$\overline{r_{ff}}$ is the average value of all feature-feature correlations.

CFS stops when five consecutive fully expanded candidate solutions show no improvement. The utilization of the genetic algorithm as a search method and CFS as an evaluator leads to the reduction of the 256 extracted features to 44. Finally a one hidden layer neural network using the backpropagation algorithm is used for classification. Caltech 101 [15] from Caltech University is utilized as dataset. This dataset contains 9144 images in jpeg format from 101 categories. From these jpeg images 5519 of them are utilized. One third of these 5519 files are converted to png format and a similar number to gif format. The dataset is divided into a training set (70%) and a test set (30%). Furthermore, 1840 pdf files were added. The created dataset is uniformly distributed and its exact numbers are indicated in Table I.

TABLE I
THE DATASET

Dataset			
Total files		**Training**	**Testing**
jpeg	1840	1288	552
png	1840	1288	552
gif	1839	1287	552
pdf	1840	1288	552
Total	**7359**	**5151**	**2208**

In this paper the first two phases of the experiment will be reproduced (feature extraction and feature selection) and the performance of five different machine learning algorithms - including the neural network originally proposed in [12]- will be evaluated.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

III. Learning Algorithms and parameters setup

Waikato Environment for Knowledge Analysis (Weka) [16], an open source machine learning software developed at the University of Waikato, New Zealand was used for all the conducted experiments. An attribute selected classifier was used in Weka. Furthermore, a genetic algorithm was chosen as a search method. The population size was 256, the number of generations 100, crossover was set to 0.8 and mutation probability to 0.033. CFS was the fitness function, roulette wheel selection was used to probabilistically select individuals and the single-point crossover operator was selected. The use of CFS as a filter selection evaluator and the genetic algorithm as a search strategy resulted to the selection of 44 features i.e. 82.81 % reduction. The classifiers examined in this paper were: decision trees, support vector machines, Neural Network, Logistic Regression, k-Nearest Neighbor. In order to estimate the accuracy of each classification model during the training phase, a stratified 10 fold cross validation was performed.

A. Decision Trees

The algorithm selected for decision tree building was C4.5, developed by R. Quinlan [17]. More specifically an open source implementation of the C4.5 algorithm in Weka known as J48 was utilized. The algorithm has a top down approach. It is a recursive divide and conquer algorithm. The training data are classified instances, while each one of these instances consists of features along with the class the specific instance belongs. One feature is selected as root node and the algorithm creates a branch for each possible feature value. That splits the instances into subsets, one for each branch that extends from the root node. The splitting criterion the algorithm uses is the normalized information gain. The feature with the highest normalized information gain is chosen to make the decision. Then the procedure is repeated recursively for each branch, selecting a feature at each node and only instances that reach that node are used to make the selection. This machine learning algorithm can be fine-tuned by setting up a lot of parameters. The parameters which were optimized in the experiment are shown on Table II.

TABLE II
PARAMETERS OF THE J48 LEARNING ALGORITHM

Parameter	Default Value	Chosen Value
Minimum number of instances per leaf	2	1
Use of unpruned trees	False	False
Confidence factor used for pruning	0.25	0.25
Consider subtree raising operation when pruning	True	True
Use of binary splits on nominal attributes	False	False

B. Support Vector Machines

A Support Vector Machine (SVM) is a machine learning method based on statistic learning theory. SVM try to find the maximum margin hyperplane that separates two classes. An adaptation of the LIBSVM [18] implementation was used in the following. Four types of kernel function linear, polynomial, radial basis function, and sigmoid are provided by LIBSVM. A Support Vector Classification (C-SVC) was used with Radial Basis Function (RBF) kernel. After various conducted experiments, it was found that the optimal value of gamma (G) parameter of the RBF kernel was 2.

C. Artificial Neural Network

A multilayer neural network using the backpropagation algorithm was implemented as a classifier in Weka. The neural network consisted of one hidden layer with 3 nodes. The number of inputs was the 44 selected features and the number of outputs the four possible categories namely jpeg, png, gif and pdf. The learning rate was set to 0.3 and in order to avoid local minimum and to accelerate the learning process, the momentum parameter was set to 0.2. The training time (epochs) after experimentation was set to 500.

D. Logistic Regression

The idea of Logistic Regression (LR) is to make linear regression produce probabilities. Instead of predicting classes, it predicts class probabilities. These class probabilities are estimated directly using the logit transform. In Weka the Logistic algorithm was utilized with the default parameter setup as shown on Table III.

TABLE III
THE DEFAULT PARAMETERS OF THE ALGORITHM

Parameters in Weka	
Maximum Iterations (MaxIts)	-1
Ridge Value (ridge)	1.0E-8

E. k-Nearest Neighbor

The k-Nearest Neighbor (k-NN) is a simple algorithm used for classification. The purpose of the k-NN algorithm is to use a training set - in which each one of instances is already classified- , in order to predict the classification of a new unknown instance in a test set. It is a lazy algorithm as it does not use the instances in training set to do any generalization. When a new instance is presented from a given test set, the algorithm searches the entire training set for the k most similar instances (the neighbors). To determine which of the k instances in the training set are most similar to a new input, a distance measure is used. The distance measure utilized in this implementation was the Euclidean distance. The output then can be calculated as the class with the highest frequency from the k-most similar instances. Each instance votes for their class and the class with the most votes is taken as the prediction. In order to find the optimum number of k, different implementations were done in Weka and it was found that the optimal value of k is 10.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

IV. Experimental Results

Primarily, in order to evaluate the performance of every classifier used, 2208 unseen instances of unaltered files (equally distributed to the four categories as already shown on Table I) were presented to each classification model. The detailed accuracy by class, along with other performance metrics such as true positive rate (TP Rate), false positive rate (FP Rate), precision and recall for every one of the five classifiers are presented on tables IV-XII. Moreover, the resulted confusion matrix for each one of the learning algorithms examined, is presented.

TABLE IV
DETAILED ACCURACY BY CLASS USING DECISION TREE (J48)

Class	TP Rate	FP Rate	Precision	Recall
jpg	0.969	0.040	0.889	0.969
pdf	0.862	0.013	0.956	0.862
png	0.960	0.015	0.955	0.960
gif	0.991	0.004	0.989	0.991

TABLE V
CONFUSION MATRIX – DECISION TREE (J48)

Actual file type	Classified as			
	jpg	pdf	png	gif
jpg	535	9	4	4
pdf	57	476	17	2
png	10	12	530	0
gif	0	1	4	547

TABLE VI
DETAILED ACCURACY BY CLASS USING SVM

Class	TP Rate	FP Rate	Precision	Recall
jpg	1	0.014	0.960	1
pdf	0.953	0.001	0.996	0.953
png	0.986	0.009	0.973	0.986
gif	0.989	0	1	0.989

TABLE VII
CONFUSION MATRIX – SVM

Actual file type	Classified as			
	jpg	pdf	png	gif
jpg	552	0	0	0
pdf	17	526	9	0
png	6	2	544	0
gif	0	0	6	546

TABLE VIII
DETAILED ACCURACY BY CLASS USING NEURAL NETWORK

Class	TP Rate	FP Rate	Precision	Recall
jpg	1	0.002	0.995	1
pdf	0.987	0.002	0.993	0.987
png	0.993	0.005	0.984	0.993
gif	0.986	0.002	0.995	0.986

TABLE IX
CONFUSION MATRIX – NEURAL NETWORK

Actual file type	Classified as			
	jpg	pdf	png	gif
jpg	552	0	0	0
pdf	3	545	2	2
png	0	3	548	1
gif	0	1	7	544

TABLE X
DETAILED ACCURACY BY CLASS USING LR

Class	TP Rate	FP Rate	Precision	Recall
jpg	0.996	0.011	0.968	0.996
pdf	0.975	0.008	0.975	0.975
png	0.955	0.005	0.985	0.955
gif	0.987	0.005	0.986	0.987

TABLE XI
CONFUSION MATRIX – LOGISTIC REGRESSION

Actual file type	Classified as			
	jpg	pdf	png	gif
jpg	550	1	0	1
pdf	9	538	3	2
png	8	12	527	5
gif	1	1	5	545

TABLE XII
DETAILED ACCURACY BY CLASS USING k-NN

Class	TP Rate	FP Rate	Precision	Recall
jpg	0.993	0.018	0.950	0.993
pdf	0.942	0.007	0.979	0.942
png	0.975	0.006	0.982	0.975
gif	0.996	0.001	0.996	0.996

TABLE XIII
CONFUSION MATRIX – k-NN

Actual file type	Classified as			
	jpg	pdf	png	gif
jpg	548	4	0	0
pdf	23	520	8	1
png	6	7	538	1
gif	0	0	2	550

A. *The Digital Forensics Perspective*

In digital forensics it is very common someone to try to alter evidence, like by renaming image files to documents, in order to fool authorities. In order to examine if the proposed method identifies the correct file type when the file was altered, one third of the testing pdf files (168) was replaced with unseen image files whose extension and signature (magic bytes) was changed to pdf. The first test set contained 168 altered pdf files. These 168 altered files were actually jpeg images whose extension and signature was changed to pdf. Likewise, the 168 pdf files of the second dataset were actually png altered images and in the third data set the 168 pdf files

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

were altered gif images. Therefore, three new test sets were created. Subsequently, unseen instances from all categories were presented to the models for evaluation. The resulted confusion matrix for every learning algorithm for each one of the three testing sets is shown on tables XIV-XVIII.

TABLE XIV
CONFUSION MATRIX – DECISION TREE (J48)

Forged file's Actual Type	Classified as			
	jpg	pdf	png	gif
168 jpg	167	1	0	0
168 png	8	3	157	0
168 gif	0	0	0	168

TABLE XV
CONFUSION MATRIX – SVM

Forged file's Actual Type	Classified as			
	jpg	pdf	png	gif
168 jpg	168	0	0	0
168 png	3	8	155	2
168 gif	0	1	0	167

TABLE XVI
CONFUSION MATRIX – NEURAL NETWORK (NN)

Forged file's Actual Type	Classified as			
	jpg	pdf	png	gif
168 jpg	168	0	0	0
168 png	0	2	166	0
168 gif	0	0	0	168

TABLE XVII
CONFUSION MATRIX – LOGISTIC REGRESSION (LR)

Forged file's Actual Type	Classified as			
	jpg	pdf	png	gif
168 jpg	168	0	0	0
168 png	7	6	150	5
168 gif	0	0	0	168

TABLE XVIII
CONFUSION MATRIX – K-NEAREST NEIGHBOR (KNN)

Forged file's Actual Type	Classified as			
	jpg	pdf	png	gif
168 jpg	167	1	0	0
168 png	8	3	157	0
168 gif	0	0	0	168

The combined confusion matrix for every classifier utilized in the experiments is shown on table XIX. The greyed color cells indicate the maximum accuracy achieved.

TABLE XIX
COMBINED CONFUSION MATRIX FOR THE FIVE CLASSIFIERS

Forged File types	**Prediction Accuracy (%)**				
	J48	**SVM**	**NN**	**LR**	**kNN**
jpg	99.40	100.00	100.00	100.00	99.40
png	93.45	92.26	98.81	89.28	93.45
gif	100.00	99.40	100.00	100.00	100.00

It is obvious from table XIX that even a very simple neural network achieved excellent results and identified extremely well almost all the forged files. The other classifiers achieved very high accuracy as well but we have to consider that in digital forensics the misclassification of even one file could be crucial in a court of law and could lead to the issue of an incorrect decision by the court members.

V. CONCLUSIONS

In this paper we examined a methodology for file type identification, which uses computational intelligence techniques for feature selection and classification. More specifically, this methodology was applied to the three most common image file types (jpg, png and gif) due to their significance to digital forensics. In order to ascertain the method's accuracy, different machine learning classifiers were utilized. A three stage process involving feature extraction (Byte Frequency Distribution), feature selection (genetic algorithm) and classification (decision tree, support vector machine, neural network, logistic regression and k-nearest neighbor) was examined. Experiments were conducted having files altered in a digital forensics perspective –by changing both their extension and signature- and the results were presented. The examined methodology showed -in most cases- very high and exceptional accuracy in file type identification, even if someone intentionally changes file's extension and signature. It was found that the classifier with the best results was the artificial neural network. In the future we plan to deploy the model, in fragments of files and examine its behavior. During our research we had strong evidence that the proposed method would work well too, although slight modifications and changes have to be made. Furthermore the correct identification of more file types should be another extension to our research and we plan to examine whether the proposed model depends on file compression.

REFERENCES

[1] K. Karampidis, G. Papadourakis, and I. Deligiannis, "File Type Identification -A Literature Review," in *9th International Conference on New Horizons in Industry Business and Education, NHIBE 2015*, 2015, p. 141.

[2] M. McDaniel, "Automatic File Type Detection Algorithm," James Madison University, 2001.

[3] M. McDaniel and M. H. Heydari, "Content based file type detection algorithms," *36th Annu. Hawaii Int. Conf. Syst. Sci. 2003. Proc.*, 2003.

[4] W. J. Li, K. Wang, S. J. Stolfo, and B. Herzog, "Fileprints: Identifying file types by n-gram analysis," *Proc. from 6th Annu. IEEE Syst. Man Cybern. Inf. Assur. Work. SMC 2005*, vol. 2005, no. June, pp. 64–71, 2005.

[5] J. Dunham, M. Sun, and J. Tseng, "Classifying file type of stream ciphers in depth using neural networks," in *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*, 2005.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

[6] M. C. Amirani, M. Toorani, and a. a B. Shirazi, “A new approach to content-based file type detection,” in *IEEE Symposium on Computers and Communications*, 2008, no. July 2008, pp. 1103–1108.

[7] D. Cao, J. Luo, M. Yin, and H. Yang, “Feature selection based file type identification algorithm,” in *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2010, vol. 3, pp. 58–62.

[8] I. Ahmed, K. Lhee, H. Shin, and M. Hong, “Content-based File-type Identification Using Cosine Similarity and a Divide-and-Conquer Approach,” *IETE Tech. Rev.*, vol. 27, no. 6, p. 465, Nov. 2010.

[9] I. Ahmed, K. Lhee, H. Shin, and M. Hong, “Fast content-based file-type identification,” in *7th Annual IFIP WG 11.9 International Conference on Digital Forensics*, 2011, pp. 65–75.

[10] M. C. Amirani, M. Toorani, and S. Mihandoost, “Feature-based Type Identification of File Fragments,” *Secur. Commun. Networks*, vol. 6, no. 1, pp. 115–128, Jan. 2013.

[11] J. D. Evensen, S. Lindahl, and M. Goodwin, “File-type Detection Using Naïve Bayes and n-gram Analysis,” *Norwegian Information Security Conference, NISK*, vol. 7, no. 1. Fredrikstad, 2014.

[12] K. Karampidis and G. Papadourakis, “File Type Identification for Digital Forensics,” Springer International Publishing, 2016, pp. 266–274.

[13] N. S. Alamri and W. H. Allen, “A comparative study of file-type identification techniques,” in *SoutheastCon 2015*, 2015, pp. 1–5.

[14] M. Hall, “Correlation-based feature selection for machine learning,” The University of Waicato, 1999.

[15] L. Fei-Fei, R. Fergus, and P. Perona, “Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories,” p. 178.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, Nov. 2009.

[17] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993,” *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994.

[18] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Sistema de Información para el Diagnóstico de Adenocarcinoma Gástrico

J. L. Alcaide, M. Patiño, A. Balankin, J. Patiño, M. A. Martínez, T. A. Ramírez

Resumen—**Actualmente el cáncer es considerado como el mayor problema de salud a nivel mundial. El cáncer de estómago o adenocarcinoma gástrico es uno de los canceres causantes de más de medio millón de muertes por año, ocupando el segundo lugar de mortandad en el mundo. El procesamiento de imágenes y su análisis mediante diferentes métodos ha permitido un diagnóstico clínico en etapa temprana.**

El presente trabajo de investigación tiene como objetivo desarrollar un sistema de información, basado en la caracterización y análisis estadístico de las imágenes utilizadas en el diagnóstico del adenocarcinoma gástrico, que permita determinar una respuesta a variables determinísticas y no determinísticas, así mismo, que coadyuve en la detección temprana para el tratamiento oportuno de éste padecimiento.

Para el logro de los objetivos de la presente investigación se utiliza una metodología con un enfoque sistémico y sistemático, se utilizan técnicas de la ingeniería del software para el desarrollo y optimización de sistemas de información, se diseñan los algoritmos de forma que permitan dar robustez al sistema de información. Por último, se analizan los resultados de la aplicación del sistema para la optimización del mismo.

Palabras Clave—**adenocarcinoma gástrico, diagnóstico clínico, procesamiento de imágenes.**

Abstract—**The cancer is now considered as the biggest health problem worldwide. The stomach cancer is one of the causes of more than half a million deaths per year, and is considered as the second place of mortality. The Image processing and analysis, by different methods, has allowed a clinical diagnosis at an early stage.**

This research aims to develop an information system based on the characterization and statistical analysis of the images used in the diagnosis of stomach cancer, should determine a response to deterministic and non-deterministic variables, likewise, It assist with the early detection for timely treatment of this condition.

To achieve the objectives of this investigation a methodology with a systemic and systematic approach is used; techniques of software engineering for the development and optimization of information systems are used; the algorithms are designed to allow robustness to information system. Finally, the results of applying the system for optimization thereof are analyzed.

Key words—**stomach cancer, clinical diagnosis, image processing.**

I. Introducción

En la actualidad muchas de las actividades son realizadas y soportadas por Sistemas de Información (SI) que se apoyan en las Tecnologías de Información y Comunicación (TIC´s), sin embargo, es importante que estas tecnologías, y en específico la computadora, contengan los programas necesarios que coadyuven a llevar a cabo la actividad misma. Así mismo, los programas están basados en algoritmos que permiten la aplicación de pasos lógicos, secuenciales y metódicamente aplicados para dar solución a un problema en cuestión.

El adenocarcinoma gástrico o cáncer de estómago es considerado como una enfermedad neoplásica de gran frecuencia en el mundo, arrojando cifras de más de medio millón de muertes por año, lo que representa más del 8% [1]. Actualmente se cuentan con varios métodos de diagnóstico o pruebas complementarias encaminadas a diagnosticar el cáncer [2-5]. Mientras se avanza en el conocimiento del cáncer se desarrollan nuevas herramientas y se perfeccionan las existentes [6-9]. El correcto diagnóstico de localización y extensión de la enfermedad permite al médico elegir el tratamiento adecuado: cirugía, quimioterapia y/o radioterapia [3].

Para poder caracterizar las imágenes utilizadas en el diagnóstico clínico del adenocarcinoma gástrico se utiliza una metodología con un enfoque sistémico y sistemático, con métodos y técnicas apegados a la Inteligencia Artificial (IA) y la Mecánica Estadística (ME), que permitan conocer los parámetros estadísticos que las gobiernan [10-13].

El presente trabajo de investigación muestra un Sistema de Información, el cual fue desarrollado con el objetivo de coadyuvar en el diagnóstico clínico del adenocarcinoma gástrico o cáncer de estómago. Para el desarrollo del SI y el logro del objetivo, primeramente, se realizó un análisis de la situación actual de los procesos llevados a cabo por los médicos para el diagnóstico del cáncer de estomago, y del estado del arte que se tiene para el análisis de imágenes de dicho padecimiento. Se diseñaron los algoritmos y aplicaciones basadas en métodos y técnicas de la Ingeniería del software, la inteligencia artificial y la mecánica estadística, y se analizaron los resultados de la aplicación del sistema de información para la optimización del mismo.

J. L. Alcaide, M. Patiño, A. Balankin, J. Patiño, M. A. Martinez, T. A. Ramirez, Instituto Politécnico, Unidad Profesional "Adolfo López Mateos"-Zacatenco, Gustavo A. Madero, México D.F. (e-mail: alkaideipn@yahoo.com.mx, mpatino2002@ipn.mx)

https://doi.org/10.17562/PB-56-3

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

II. Objetivo y Metas

Desarrollar un SI con un enfoque sistémico y sistemático, que permita analizar las imágenes utilizadas en el diagnóstico clínico del adenocarcinoma gástrico. La metodología y técnicas utilizadas para el desarrollo del SI están basadas en el modelo de ciclo de vida de desarrollo de los sistemas de información, aplicando técnicas y herramientas de la ingeniería del software, sistemas expertos, filtros y algoritmos que permitan la caracterización de imágenes obtenidas de endoscopias.

Para el desarrollo del sistema, se llevó a cabo la investigación de la situación actual para contar con un conocimiento más claro y amplio de los sistemas de información para el manejo de imágenes. Se diseñaron los elementos del SI y se aplicaron y usaron filtros y algoritmos basados en técnicas de la ingeniería del software e inteligencia artificial para la construcción del mismo.

III. Métodos y Materiales

III.1 Métodos

A. Método de Shannon [14]

Shannon define la entropía como una medida de incertidumbre de la información contenida en un sistema. La entropía de una variable aleatoria está definida en términos de una distribución de probabilidad, la cual puede mostrar una buena medida de incertidumbre.

Se consideran los píxeles de una imagen convertida a 256 niveles de gris (*1*) y se separan en dos niveles principales de gris, el primer plano o *foreground* (*2*) y un fondo de base o *background* (*3*). La variable g denotará esos valores de niveles de gris. Para imágenes de 8 bits g = 0...255

$$I = \{\text{conjunto de pixeles de la imagen de entrada}\} \tag{1}$$

$$F = \{g \in I \,/\, g = 1{:}T\} \tag{2}$$

$$B = \{g \in I \,/\, g = T + 1{:}G\} \tag{3}$$

En el contexto de procesamiento de imágenes, el foreground es el conjunto de pixeles con luminancia menor a un cierto valor T, mientras que el background es el conjunto de pixeles con luminancias por encima de este valor de umbral T.

La función imhist() de Matlab, por ejemplo, calcula el nivel de gris para cada pixel, las frecuencias absolutas para cada píxel g. Calculamos las probabilidades estimadas de cada pixel g haciendo el cociente entre n_g y N, (*4*), siendo n_g el número de veces que se repite el pixel g en la imagen y N la cantidad total de pixeles.

$$(g) = \frac{n_g}{N} \tag{4}$$

$$\sum_{g=1}^{G} \ (g) = 1 \tag{5}$$

$$N = \sum n_g \tag{6}$$

Las probabilidades del foreground y background están expresadas como se indica en las ecuaciones (7) y (8), respectivamente:

$$p_f\,(g),\ 0 \leq g \leq T \tag{7}$$

$$p_b\,(g),\ T + 1 \leq g \leq G \tag{8}$$

Definimos la probabilidad acumulada como lo expresa la ecuación (9)

$$P(g) = \sum_{g=1}^{G} \ (g) \tag{9}$$

Esta función de probabilidad puede ser considerada como una suma o unión de dos funciones de probabilidad, una para zonas claras (foreground) y otra para zonas oscuras (background). Ecuaciones (10) y (11).

$$P_f(T) = P_f = \sum_{g=0}^{T} \ (g) \tag{10}$$

$$P_b(T) = P_b = \sum_{g=T+1}^{G} \ (g) \tag{11}$$

La entropía de Shannon, paramétricamente dependiente del valor umbral T, está definida, para el foreground y background, como: (12) y (13).

$$H_f(T) = -\sum_{g=0}^{T} {}_f(g).\,log\ p_f(g) \tag{12}$$

$$H_b(T) = -\sum_{g=T+1}^{G} {}_b(g).\,log\ p_b(g) \tag{13}$$

La suma de estas dos expresiones puede ser denotada como *H(T)* definida mediante (14).

$$H(T) = H_f(T) + H_b(T) \tag{14}$$

Usando las ecuaciones (12) y (13), se reemplaza obteniendo lo enunciado por la ecuación (15).

$$\begin{aligned} H(T) = {} & \left(-\sum_{g=0}^{T} {}_f(g).\log p_f(g)\right) + \\ & \left(-\sum_{g=T+1}^{G} {}_b(g).\,log\ p_b(g)\right) \end{aligned} \tag{15}$$

Que también puede expresarse como (16).

$$H(T) = -\sum_{g=0}^{G} \ (g).\log(\ \ (g)) \tag{16}$$

El umbral óptimo será entonces aquel que maximice esta entropía global (17).

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

$$T^* = Max\{H(T)\} \tag{17}$$

En imágenes a color pertenecientes al espacio RVA se pueden representar por hipermatrices de m x n x 3, cada una de las tres capas de las hipermatrices contiene los valores de luminancia correspondiente al Rojo, Verde y Azul. Al separar estas capas, presentan sus luminancias en escalas de grises, por lo tanto se pueden aplicar todas las ecuaciones anteriormente descritas a cada capa de cada color.

B. Convolución [15]

La convolución de dos funciones f(x) y g(x) se define mediante la integral:

$$(f * g)(x) = h(x) = \int_{-\infty}^{\infty} f(z)g(x-z)dz \tag{18}$$

La gran importancia de esta operación radica en el hecho de que la Transformada de Fourier de un producto de convolución de dos funciones es igual al producto de las Transformadas de Fourier de dichas funciones, es decir:

$$T(f(x) * g(x)) = F(u)G(u) \tag{19}$$

Este resultado denominado Teorema de Convolución implica que se puede calcular un producto de convolución de dos funciones multiplicando sus correspondientes Transformadas de Fourier y al resultado aplicarle la Transformada de Fourier inversa. En el caso de señales discretas, las longitudes que pudieran tener las sucesiones de puntos de cada una de las funciones son posibles causas de errores en el cálculo final de la convolución, es por ello que ambas funciones han de definirse en una misma cantidad de puntos por cada eje.

Para lograr esto se debe considerar que la función *f*(x) ha sido muestreada sobre un conjunto de puntos de longitud A y la función g(x) lo ha sido sobre un conjunto de longitud B, entonces ambas funciones se rellenarán con ceros hasta que cada una de ellas quede definida en *M=A+B-1* valores. La formula de rellenar con ceros los valores que faltan no es la única manera que existe de fijar dichos valores, aunque si es la más comúnmente usada. Una vez que ambas funciones tienen el mismo rango de definición, la convolución se puede calcular por:

$$f(x) * g(x) = h(x) = \sum_{m=0}^{M-1} f(m)g(x-m) \tag{20}$$

Para *x=0,1,.....M-1*

C. Correlación [15]

La correlación de dos funciones es una operación de características similares a la convolución. La expresión matemática para esta operación es:

$$f(x) \circ g(x) = h(x) = \int_{-\infty}^{\infty} f(z)g(x+z)dz \tag{21}$$

Bajo las mismas condiciones que se establecieron en la convolución en el caso discreto, la expresión de la correlación de funciones discretas reales es:

$$f(x) \circ g(x) = \sum_{m=0}^{M-1} f(m)g(x+m) \tag{22}$$

para *x=0, 1, ..., M-1.*

De forma paralela a como se vio que existía un teorema de convolución, ahora se puede enunciar un Teorema de Correlación. El teorema establece que la Transformada de Fourier de la correlación entre dos funciones es igual al producto de la Transformada de Fourier conjugada de una de ellas por la otra. Es decir:

$$T[f(x) \circ g(x)] = F^*(u)G(u) \tag{23}$$

La demostración en el caso continuo es la siguiente:

$$\begin{aligned} T[f(x) \circ g(x)] &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(z)g(x+z)\,dz \right] e^{-iux} dx \\ &= \int_{-\infty}^{\infty} f(z) \left[\int_{-\infty}^{\infty} g(x+z)\, e^{-iux} dx \right] dz = \\ &= G(u) \int_{-\infty}^{\infty} f(z)\, e^{iuz} dz = G(u) \left[\int_{-\infty}^{\infty} f(z)\, e^{-iuz} dz \right]^* \\ &= G(u)F^*(u) \end{aligned} \tag{24}$$

La demostración en el caso discreto es análoga al anterior.

Igual que la convolución, la correlación es una operación básica del procesamiento de imágenes digitales.

III.2 Materiales

Para el desarrollo del SI se utilizó un equipo de cómputo con software básico y especializado, así mismo, se programaron y emplearon los algoritmos y aplicaciones de los métodos utilizados.

IV. Desarrollo

El SI propuesto considera cuatro etapas: 1) La primera que contiene las vivencias del experto, las cuales son depositadas en una Base de Datos de Conocimiento, misma que contiene cadenas de caracteres con formas, contenidos y propiedades específicas: discretos, representativos y arbitrarios [12], se considera entonces esta base como la parte principal que alimenta al sistema experto en donde se tiene la experiencia del especialista y del cual se obtienen resultados aproximados a la realidad [13]. 2) Los SI que son un conjunto de módulos

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

relacionados que interactúan entre sí con un fin común. Cada uno de estos cuenta con 5 módulos: datos de entrada, datos de salida, transformación, mecanismos de control y objetivos [16]. 3) Inteligencia Artificial (IA) que es la encargada de unir la ciencia y la ingeniería que nos ayudarán a comprender desde una perspectiva informática el comportamiento de la vida diaria para poder mostrarla en un sistema inteligente que da como resultado su uso [10]. 4) Un repositorio electrónico de datos (base de datos) en donde se tendrán todos los registros archivados de manera ordenada y computarizada de los pacientes, se pueden agregar, modificar y consultar dichos datos con la ayuda del mismo sistema [17, 18].

La Figura 1 muestra el proceso a bloques de como el sistema coadyuva con el diagnóstico, en la figura se muestran las actividades llevadas a cabo, como son:

1. Cargar un archivo de imagen valido (en formato y extensión).
2. Hacer una conversión a tipo negativo de la imagen: Donde se resalta la parte lesionada (aplicación de filtros).
3. Extrae el valor de color de cada pixel de la lesión resaltada: se hace un barrido de cada pixel y se compara con los patrones (condiciones guardadas en la Base de Datos de Conocimiento), para guardar el valor de los colores Rojo, Verde y Azul, de la parte lesionada en una base de datos.
4. Se dibuja la lesión, es decir, se resalta cada pixel en color rojo de la parte lesionada.
5. Arroja un diagnóstico: muestra en la pantalla el diagnóstico de la imagen procesada y lo almacena en la base de datos (el diagnóstico está asociado a un paciente).

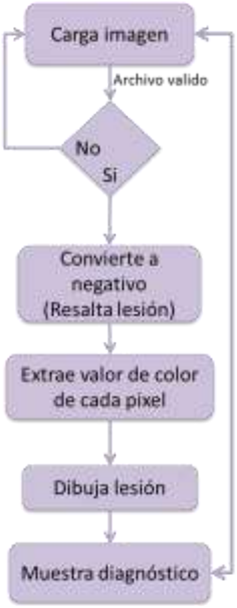

Carga imagen
Archivo valido
No
Si
Convierte a negativo (Resalta lesión)
Extrae valor de color de cada pixel
Dibuja lesión
Muestra diagnóstico

Figura 1. Proceso de detección y diagnóstico.
Fuente: Elaboración propia

Cabe resaltar que el sistema está preparado para proporcionar, junto con el diagnóstico, recomendaciones que proporciona el especialista para el tipo de lesión resultante.

La Figura 2 muestra una imagen que es proporcionada por el usuario del sistema y es la que se carga al mismo, mientras que la Figura 3 muestra la imagen de la Figura 2 después de aplicarle el filtro de negativo.

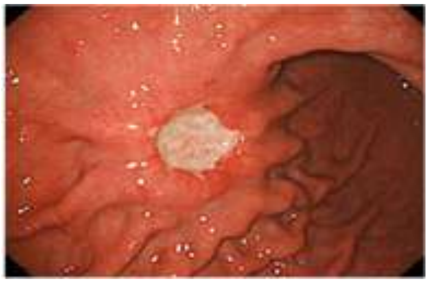


Figura 2. Imagen original de una lesión intestinal
Fuente: Fotografía obtenida de una endoscopia

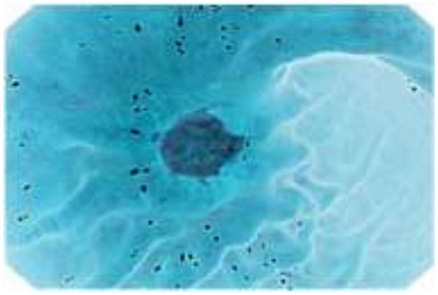


Figura 3. Imagen después de la aplicación del filtro negativo.
Fuente: Elaboración propia.

La imagen que se analiza es la mostrada en la Figura 3, la imagen tratada con el filtro negativo, ya que en ella, y de acuerdo a los expertos en lesiones, las manchas de color oscuro en la imagen filtrada indican una lesión. Sin embargo, para poder determinar con una mayor precisión la lesión, primeramente se le da a la imagen un fondo negro y se filtran los colores Rojo, Verde y Azul, para cada pixel, dando como resultado el dibujo de la lesión, el cual se aprecia en la Figura 4.

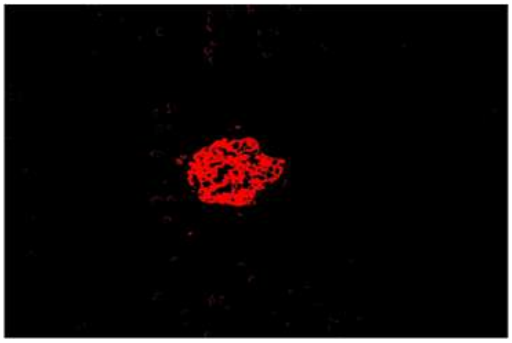


Figura 4. Representación de la lesión.
Fuente: Elaboración propia.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Para obtener el nivel de saturación de los colores que componen a la imagen procesada, se extraen los colores primarios de cada pixel y se toma el valor mínimo y máximo de los valores de cada uno de ellos, es decir, los mínimos y máximos Rojo, Verde y Azul. Con lo anterior se tiene el valor para saber en dónde se encuentra la lesión, y además es lo que interpreta el sistema para poder dibujarla y así determinar un diagnóstico. Se consideran los máximos y mínimos porque es en donde se nota la variación en color.

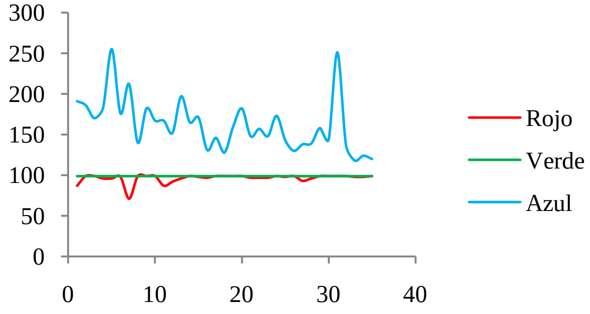


Figura 5. Valores Máximos: Rojo, Verde, Azul.
Fuente: Elaboración propia

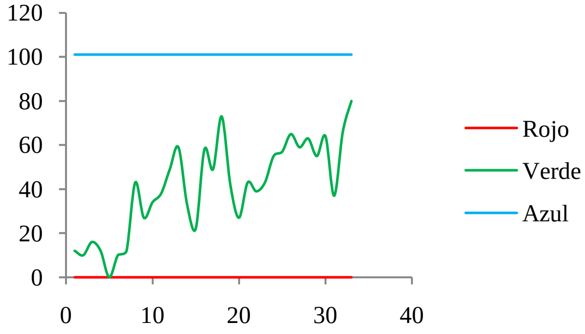


Figura 6. Valores Mínimos: Rojo, Verde, Azul.
Fuente: Elaboración propia

Se observa en la Figura 5 que el valor que tiene una variación significativa es el de color azul, mientras que en la Figura 6 se observa que el valor que tiene variación significativa es el de color verde.

Para poder interpretar el valor de los colores máximos, que es el porcentaje de saturación, se tienen que estandarizar los datos extraídos como valores máximos primarios (R, V, A), lo cual se obtiene haciendo la división del valor del color entre la sumatoria de los valores máximos R, V, A. Es decir:

$$V_{MAX}= \frac{VColor}{\sum_{i=1}^{i=3}(R,V,A)} \qquad (25)$$

Donde:

VColor = Valor de color a obtener porcentaje de saturación
R = Valor de pixel máximo rojo
V = valor de pixel máximo verde
A = valor de pixel máximo azul

Para poder determinar un diagnóstico automático se consideran los criterios utilizados por el especialista, en donde para determinar el grado de lesión se toman en cuenta los datos en donde el porcentaje de color verde varia significativamente (valores mínimos de verde), y se obtiene la media de los niveles de saturación de color verde para determinar el grado de lesión: Grave, Mediana ó Leve.

V. Resultados

Aplicación del Sistema de Información, Filtros y Algoritmos

Los resultados obtenidos, aplicando los filtros y algoritmos desarrollados al conjunto de imágenes de endoscopias, se muestran en la tabla 1, donde se observa que el porcentaje de saturación de los valores mínimos del rojo y azul no varía; los máximos del verde y rojo, tienen poca variación; el valor mínimos verde y máximo azul son los que tienen variaciones significativas.

Analizando el conjunto de imágenes de endoscopias con el sistema, se determina que el más confiable, *en el conjunto específico analizado*, es el verde, considerándolo como referencia, para las imágenes analizadas y filtros aplicados, de esta forma podemos identificar una lesión, y se puede considerar un diagnóstico clínico muy aproximado.

Tabla 1 Valores Aplicando filtros para los colores Rojo, Verde y Azul.

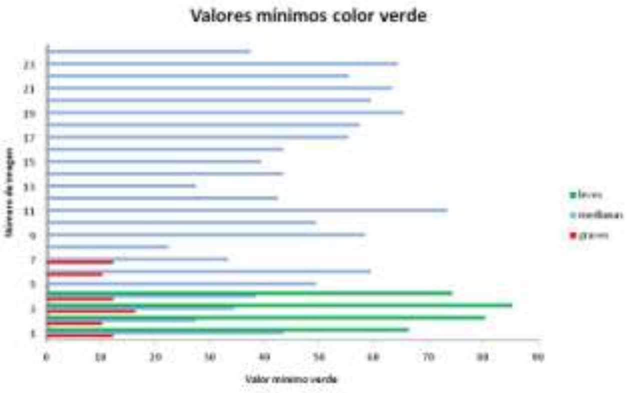
MAXIMOS			***MINIMOS***		
ROJO	*VERDE*	*AZUL*	*ROJO*	*VERDE*	*AZUL*
Graves					
87	99	191	0	12	101
99	99	186	0	10	101
99	99	170	0	16	101
96	99	183	0	12	101
96	99	255	0	0	101
98	99	176	0	10	101
71	99	212	0	12	101
Medianas					
99	99	140	0	43	101
99	99	182	0	27	101
99	99	167	0	34	101
87	99	167	0	38	101
92	99	152	0	49	101
96	99	197	0	59	101
99	99	165	0	33	101
98	99	171	0	22	101
97	99	131	0	58	101
99	99	146	0	49	101
99	99	128	0	73	101
99	99	160	0	42	101

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

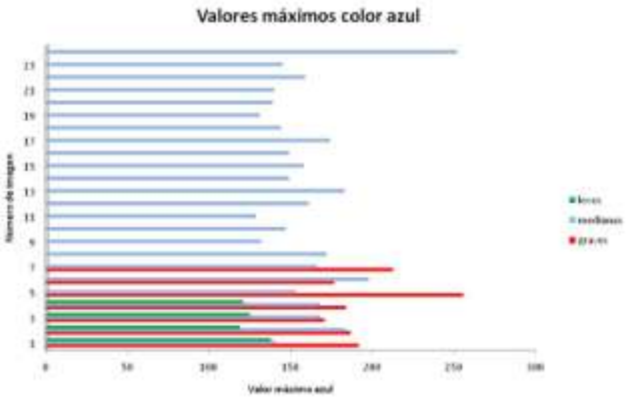
ISSN 2395-8618

99	99	182	0	27	101
97	99	148	0	43	101
97	99	157	0	39	101
97	99	148	0	43	101
99	99	173	0	55	101
98	99	143	0	57	101
99	99	130	0	65	101
93	99	138	0	59	101
96	99	139	0	63	101
99	99	158	0	55	101
99	99	144	0	64	101
99	99	251	0	37	101
Leves					
99	99	137	0	78	101
98	99	118	0	80	101
98	99	124	0	85	101
99	99	120	0	74	101

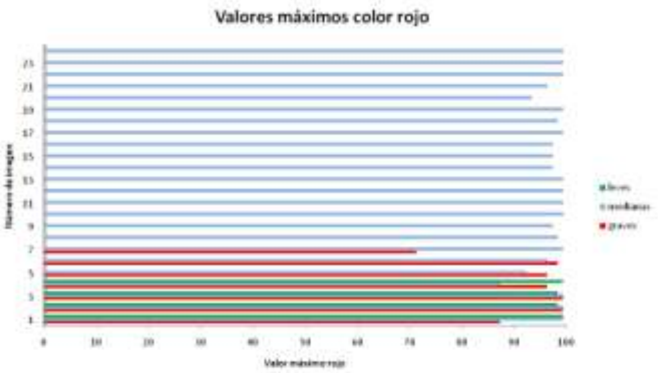
Aplicando los algoritmos de los filtros RVA, se tiene como resultado la variación entre los colores. En las Figuras 7 (a), (b) y (c) se pueden apreciar las variaciones en los colores que van de acuerdo al filtro aplicado y tipo de lesión determinada: leve, mediana y grave.



(a)



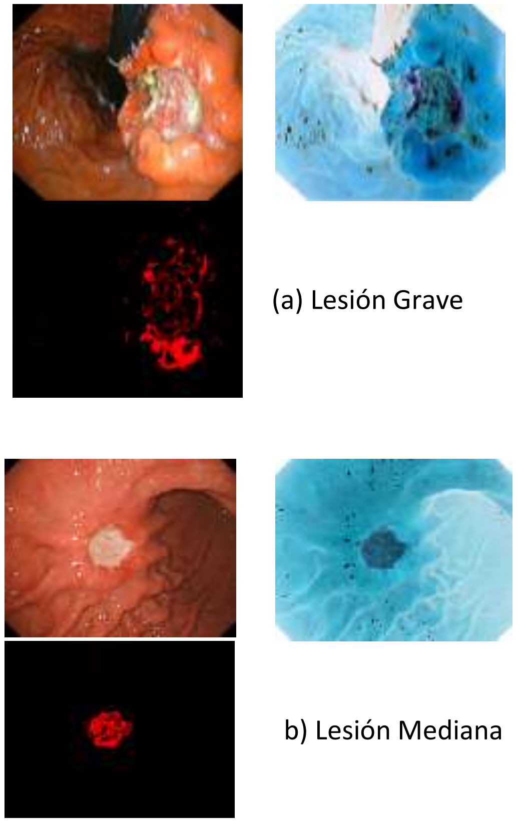
(b)



(c)

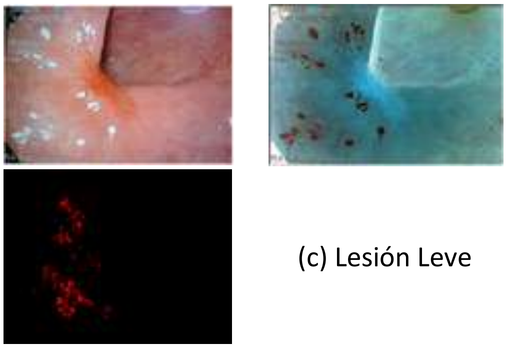
Figura 7. Gráficas de Valores de Saturación en los colores mínimo Verde, máximo Azul y máximo Rojo para un conjunto de imágenes.
Fuente: Elaboración propia

La Figura 8 (a), (b) y (c) muestra resultados obtenidos por el sistema para los casos de una lesión leve, una mediana y una grave. En los resultados se muestra, para cada caso, la imagen de una endoscopia, enseguida a la derecha la imagen filtrada y por último, en la parte de abajo, la imagen resultante para poder coadyuvar con el diagnóstico de la lesión, así mismo, dependiendo del grado de lesión, el sistema hace referencia a una base de datos que busca la mejor alternativa de tratamiento y/o especificaciones del posible tratamiento de la lesión.



IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618



(c) Lesión Leve

Figura 8. Resultados obtenidos con el sistema de información para una lesión grave, mediana y leve.
Fuente: Elaboración propia

VI. Conclusiones

De acuerdo a los filtros utilizados, al conjunto de imágenes utilizadas y a la interpretación del especialista con respecto a los resultados, podemos considerar en ellas que: El porcentaje de saturación en un diagnóstico de una lesión grave, el valor mínimo verde se ubica entre el rango de 0 a 16, mientras que en el máximo azul varia de 170 a 255; para una lesión mediana se tienen los valores mínimos de verde de 22 a 73 y los máximos en azul de 128 a 251; por último, para las lesiones leves los rangos mínimos verdes son 74 a 85 y en el caso de los máximos azules están entre 118 y 137.

Así, para este conjunto de imágenes, y utilizando los filtros propuestos, podemos considerar que entre menor sea el valor mínimo del color verde la lesión es más grave o fuerte, mientras que, si es más alto el valor del verde, la lesión se considera leve o incluso se podría decir que no hay lesión.

Impacto de la investigación

A través del desarrollo de este trabajo se logró que expertos de diferentes áreas integraran y cohesionaran su conocimiento y visiones, bajo un enfoque sistémico, global e integral, que posibilitó otra forma de análisis, interpretación y solución al estudio de sistemas, como el sistema de información propuesto, cuyo objetivo es el de coadyuvar en el diagnóstico clínico del adenocarcinoma gástrico o cáncer de estómago.

Agradecimientos

Los autores agradecen al Instituto Politécnico Nacional (IPN), México, por su apoyo para el desarrollo de este trabajo.

VII. Bibliografía

[1] Barboza Eduardo, *"Cáncer de estómago"*. Revista médica herediana. Vol. 7 No. 2. 1996.

[2] Macarulla, *"Comprender el cáncer"*. Amat editorial. Barcelona. 42p. 2009.

[3] Instituto Nacional de Cancerología. *"El cáncer. Aspectos básicos sobre su biología, clínica, prevención, diagnóstico y tratamiento"*. Ministerio de la protección social. Instituto Nacional de Cancerología E.S.E. 2004.

[4] Moreira y López, *"Endoscopia digestiva alta"*. Revista Española de Enfermedades Digestivas. Madrid. Vol. 100. No. 7. 2008.

[5] American Cancer Society. *"Cáncer de estómago. Guía detallada"*. American Cancer Society. 2014.

[6] Ding-Yun Liu, Tao Gan, Ni-Ni Rao, Yao-Wen Xing, Jie Zheng, Sang Li, Cheng-Si Luo, Zhong-Jun Zhou, Yong-Li Wan. *“Identification of lesion images from gastrointestinal endoscope based on feature extraction of combinational methods with and without learning process”*. Medical Image Analysis, Vol. 32, pp. 281-294, 2016.

[7] Dimitris K. Iakovidis, Dimitris E. Maroulis, Stavros A. Karkanis. *“An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy”*. Computers in Biology and Medicine Vol. 36 pp. 1084–1103, 2006.

[8] Fons van der Sommen, Svitlana Zinger, Erik J. Schoon and Peter H. N. de With. *“Computer-Aided Detection of Early Cancer in the Esophagus using HD Endoscopy Images”*. Proc. SPIE Vol. 8670 86700V-1, 2013.

[9] Liu, D.Y. , Gan, T. , Rao, N.N. , Xu, G.G. , Zeng, B. , Li, H.L. *"Automatic detection of early gastrointestinal cancer lesions based on optimal feature extraction from gastroscopic images"*. Journal of Medical Imaging and Health Informatics, Volume 5, Number 2, pp. 296–302, 2015.

[10] Pino, *"Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva"*. Universidad de Oviedo, Servicio de publicaciones. 1P. 2001.

[11] Van Gigch J. P. *"Teoría general de sistemas"*. Trillas. México. 65p. 2006.

[12] Vilarroya, *"Palabra de robot: Inteligencia artificial y comunicación"*. Publicaciones de la Universidad de Valencia. Valencia. 36p. 2006.

[13] Ascolano, *"Inteligencia artificial: modelos, técnicas y áreas de aplicación"*. Editorial Paraninfo. 4p. 2003.

[14] Eduardo Dvorkin, Marcela Goldschmit, Mario Storti (Eds.), *"Mecánica Computacional"* Vol. XXIX, págs. 6177-6193, Buenos Aires, Argentina, 15-18 Noviembre 2010.

[15] Gustavo M. Murmis, *“Tratamiento de Imágenes Digitales”,* Facultad de Ingeniería – Universidad de Buenos Aires, 2013.

[16] Fernandez Vicenc, *"Desarrollo de sistemas de información: una metodología basada en el modelado"*. Barcelona. 11p. 2006.

[17] Date C. J. *"Introducción a los Sistemas de Bases de Datos"*, Séptima Edición, Prentice Hall, 2001.

[18] Thomas M. Connolly & Carolyn E. Begg: *“Sistemas de Bases de Datos”,* 4a edicion, Addison-Wesley, 2005.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Artificial Intelligence Models to Estimate Biomass of Tropical Forest Trees

Razer Anthom Nizer Rojas Montaño, Carlos Roberto Sanquetta, Jaime Wojciechowski, Eduardo Mattar, Ana Paula Dalla Corte, and Eduardo Todt

***Abstract*—Artificial Intelligence Models (AI) were tested for aboveground dry biomass estimation of 4,004 trees collected in forests throughout the Tropics, and compared to a classic Allometric model of literature. The data come from various countries, in the Neotropics, Africa, Southeast Asia, and Oceania. Statistical analysis of the data showed that they do not have normal distribution and homocedasticity, which violates the regression assumptions. Examination of bias, precision and accuracy of the Allometric model and the AI models revealed that KNN (K Nearest Neighbors), ANN (Artificial Neural Network) and SVM (Support Vector Machines) have strong estimation power of the biomass of tropical trees, comparable or greater than the linear regression (Allometric model), which is considered the state of the art. It was concluded that AI models can be considered an interesting alternative to the regression technique, especially when the data do not show normality and homoscedasticity, which is the case of biomass of tropical forest trees. In particular, SVM showed better accuracy for data considered.** ***Index Terms*—machine learning, allometry, carbon, data mining, neural networks, support vector machines**

I. Introduction

Forests are considered important global carbon reservoirs, storing about $296Gt$ of carbon. Carbon concentrations are found in tropical forest of South America and Central Africa, stocking about $120tC.ha^{-1}$, while the world average is $75tC.ha^{-1}$. However, tropical forests have been the main victims of deforestation and degradation [1]. This has led to increase of accumulated emissions of Greenhouse Gases (GHG) by activities using land and forests, of $490 \pm 180GtCO_2$ in 1970 to $680 \pm 300GtCO_2$ 2010 [2].

The largest fraction of carbon stored in the forests of the world is in your living biomass, with $250GtC$ [1], and still there is great uncertainty about these stocks, mainly due to insecurity of estimation on large scale. A complicating factor is that any model to be applied on a larger scale must be based on direct measures, which are complex, costly and destructive [3].

It is essential to develop precise and accurate models of large-scale carbon stocks, but this is not a simple task. There are several variables that interfere in calculations, such as composition and structure of the vegetation, specific information such as density or specific mass of tissue, the carbon content in tissue, the method for calculation of areas and reliability of forest inventory, among others. One of the most important of these factors is the modeling methodology used to estimate the biomass or individual carbon from dendrometric variables, such as the diameter and height of trees.

Allometry is one of the best known indirect methods of estimation of biomass and individual carbon, which is usually materialized via simple input regression models - only with a diameter at breast height (dbh) as an independent variable - double entry, with dbh and height as independent variables - or triple entry, including the density or specific mass of the specie. Another robust methodological alternative, but less flexible, is the application of so-called growth factors. These modeling approaches are widely used and widely found in the literature [4].

Despite the current and widespread use of allometric models to estimate the biomass of trees, literature alerts that regression should respect some basic assumptions, and you should not use it indiscriminately without these assumptions. These assumptions are four: 1. variables must be normally distributed; 2. should have a linear relationship between the dependent and independent variables; 3. variables must be measured reliably; and 4. the variables must have homogeneous variances [5]. Although these assumptions are crucial to give validity to the estimates, they are rarely investigated and/or reported in quantification studies of forest biomass, which represents a risk to the estimation process.

The use of Artificial Intelligence (AI) techniques are a completely different approach to allometry via regression for individual biomass estimation. These techniques offer flexibility, simplicity and versatility, having the potential to estimate forest biomass in a way comparable to allometric classic models usually employed. Perhaps the most important feature in this context is the fact that the AI techniques in principle not require to attend the regression assumptions [6].

AI techniques have been applied in different scientific fields and sectors of human activities. In the last decade several studies have been published on the use in forest science ([7], [8], [9], [10], [11], [12]). One of the standard techniques of AI (machine learning often used in data mining) was recently explored [6], demonstrating its potential in quantification of

Manuscript received on February 12, 2017, accepted for publication on May 7, 2017, published on July 30, 2018.

Felix Freitag and Leandro Navarro are with the Department of Computer Architecture (DAC) at the Universitat Politècnica de Catalunya (UPC), Jordi Girona 1-3, 08034 Barcelona, Spain (e-mail: {felix, leandro}@ac.upc.edu).

The authors are with DINF – Federal University of Paraná (UFPR), Brazil (email: {razer,todt}@inf.ufpr.br), and BIOFIX – Federal University of Paraná (UFPR), Brazil (email: {sanquetta,jaimewo,eduardomattar,anacorte}@ufpr.br).

https://doi.org/10.17562/PB-56-4

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

individual tree biomass. Although data mining is a promising technique, some issues became to solve: Are there other AI techniques that can also be applied, or that are better than allometry? Do these techniques respond positively in situations where the data have high dispersion, as in case of biomass species of tropical trees?

This study aims to analyze data of more than 4,000 trees, collected in various regions of the tropics, including the Neotropics, Africa and Southeast Asia, with respect to the assumptions of regression models, and analyze AI techniques K-nearest neighbor, artifical neural networks and support vector machines, to estimate the variable of interest compared with a tipical allometric model.

II. Material and Methods

The data used in this study were provided by Jerome Chave, Director of Research at French Scientific Research French Centre - CNRS, France. These data correspond to 4,004 observations of diameter at breast height in cm (dbh), total height in m (ht), total aboveground dry biomass in kg (b), wood density (basic specific mass, ρ). These data were collected in the following countries: Australia, South Africa, Brazil, Cambodia, Cameroon, Central African Republic, Colombia, Cost Rica, French Guiana, Gabon, Ghana, Guadeloupe, India, Indonesia, Mexico, Madagascar, Malaysia, Mozambique, New Guinea, Peru, Puerto Rico, Tanzania, Venezuela and Zambia.

Data were randomly separated into two parts, 70% for adjustment or training and 30% to validate the estimates.

The basic descriptive statistics for these data were calculated: arithmetic average, standard deviation, coefficient of variation, and maximum and minimum of the variables b, dbh, ht and ρ. It was also investigated the linear correlation between these variables, using Pearson's correlation coefficient. It was held the normality test (Lilliefors and Shapiro-Wilk) of these variables.

Four biomass estimation models (b) were evaluated depending on the variables dbh, ht and ρ, having as a witness the allometric model proposed by [13] and computed by Equation 1.

A. Allometric Model

The allometric model used for comparison was proposed by [13], based on Schumacher-Hall's model [14], and described by Equation 1.

$$\lg(\hat{b}) = \beta_0 + \beta_1 \lg(dbh) + \beta_2 \lg(ht) + \beta_3 \lg(\rho) \tag{1}$$

where:

- lg : decimal log of previous variables;
- β_0, β_1, β_2, β_3: model coeficients to be adjusted.

B. KNN (K-Nearest Neighbor)

KNN is a non parametric method used in data mining called *instance based learning*, which employs the values of closest neighbors to be estimated. This method applied to biomass estimation is described in more details in [6]. In this work dbh, ht and ρ were applied as proximity variables, to have a direct comparison of its estimated power with allometric model. We used the Euclidean distance (Equation 2), which showed the best results in experiments compared with Squared, Manhattan and Chebychev, and inverse distance weighting [15], [6]. Such choices depend on the performance of adjustment, that is, several simulations were made using the amount of neighbors, types of distance and types of weighting, and we choose the best performance. This procedure is indicated in the study of Bradzil [16], where estimates are made from one to five closer neighbors.

$$d(p,q) = \sqrt{(dbh_p - dbh_q)^2 + (ht_p - ht_q)^2 + (\rho_p - \rho_q)^2} \tag{2}$$

where:

- p e q : trees variables;
- dbh_p, dbh_q: dbh of tree p and q;
- ht_p, ht_q: total height of tree p and q;
- ρ_p, ρ_q: specific mass of tree p and q.

The method uses a technique known as Cross-Validation where each instance is compared to other from sample, being selected the instance with shorter distance. The biomass estimated for that instance is the biomass of the instance with lowest distance from it. The method allows the use of n nearest neighbors of the tree in question, and the value of the estimated biomass is a balance between the biomass of trees with smaller distances among vectors of weighting by the inverse of the distance ($1/d$). This study employed three nearest neighbors (Equation 3) and weighted by the inverse of distance:

$$\hat{b}_p = \frac{b_1 w_1 + b_2 w_2 + b_3 w_3}{w_1 + w_2 + w_3} \tag{3}$$

- $\hat{b}_p$: estimated biomass of tree p;
- w_n: $\frac{1}{d(p,q_n)}$ closest trees weighted, from tree p to tree q_n;
- $d(p, q_n)$: distance from tree p to one of three closest trees q_n;
- b_1, b_2, b_3: real biomass of three closest trees, mensured by distance $d(p, q_n)$.

C. Artificial Neural Networks

ANN is a machine learning technique used for various purposes, also recently used to address forestry problems [11]. It is a comutational system composed by simple processing units, highly connected. These units, or neurons, compute mathematical functions and their results are processed together in the network layers. The connections simulate biological synapses and have associated weights to inputs. These weights

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

are adjusted as the whole set is trained, that is, learning acquired knowledge [17].

There are several ANN configurations, the main ones are Multilayer Perceptron (MLP), based on the Radial Basis Function (RBF) and based on Vector Quantization (LVQ) [18]. Here we used the MLP networks, commonly used in studies of forest area [19].

A neuron receives values and returns a result. The input values are weighted, combined (added) and submitted to a mathematical function f_a. Thus, if the vector $x = [x_1, x_2, ...x_m]^t$ is the input of a neuron and the vector $w = [w_1, w_2, ..., w_m]^t$ is the weights applied to each input, the result of the neuron $f'(x)$ is:

$$\begin{aligned} u &= \textstyle\sum_{i=3}^{m} x_i w_i \\ f'(x) &= f_a(u) \end{aligned} \tag{4}$$

The f_a function is called *activation function* and it can be of various types, the most used are: linear, threshold and sigmoid. In this case we used the sigmoid, the most used according to [18]. Neurons are arranged in one or more layers, and one neuron receives as input the outputs of the previous layer's neurons, and its output is put in the next layer. The input layer receives data to be processed, and the layer that gives the result is called output layer. The other layers are known as hidden.

To solve nonlinearly separable problems we should use one or more hidden layers [20]. In this work we employed Multilayer Perceptron (MLP) with sigmoidal function in their hidden layers. In the case of regression problems, the output can not be discretized, and a decimal value is returned. For the ANN training we used back-propagation algorithm [21]. It consists of two parts: *forward* and *backward*. In phase forward the object is presented to the network, neurons calculate their values to the specific weights and the activation function produces its output value. This is done until the output neurons have their calculated values. The computed result is compared with the expected result and this difference is the error on the network. The error value is then used in step backward to adjust the weights of neurons.

There are several parameters to be configured to find an ANN that gives acceptable estimates and comparable to allometric model. Since the amount of layers, how many neurons in each layer, learning rate (multiplicative value for weights adjustment in the learning process), number of epochs (number of times network is presented to the input data), among others. In this work, the data were entered into multiple networks containing a hidden layer of neurons ranging from 5 to 100. The learning rate was tested between 0.1 and 0.9, with steps of 0.2. The momentum varied between 0.001 and 1. Using a training set with 30% of population. These data are used to verify the mistake of training, in which the increase of error can stop the process without all the expected number of times to run.

D. Support Vector Machines

Support Vector Machines (SVM) is a machine learning technique used in many situations for pattern recognition, obtaining results superior to those achieved by other learning techniques in various situations, such as categorization of texts in image analysis and bioinformatics [22]. The technique is grounded by statistical learning theory, developed by [23], [24]. SVMs can be applied to problems of classification and regression, with potential use in various forestry issues.

Like other methods, several parameters must be set to obtain a SVM model comparable to other tested models. The main parameters are the cost (C), which gives balance between accuracy and complexity of the model, and the kernel function used to design values for a larger, where data have more probability to be linearly separable [25]. The type of kernel function used here was *RBFKernel*, which has the *gamma* parameter, that controls the shape of the peaks when the data is passed to another dimension. Small values indicate pointed peaks, that is, small bias and high variance, which may cause overfitting when the model learns only the entered values and lose the ability to generalize, giving poor results for new entries. Large values result in soft forms, with high bias and low variance and can harm the learning process.

They were tested more than 150 combinations of C and $gamma$, to find the combination with best correlation and residual values. The C parameter was varied from 1,000 to 10,000, initially with steps of 1,000. When a promising region was identified, the steps were reduced to 500, 100, 50 and 10. The $gamma$ value was varied from 0.01 to 0.09 with steps of 0.02, and varied from 0.1 up to 0.9, with steps of 0.2, in each test performed.

E. Quality Assessment of Estimates

The performances of the estimates obtained with the four techniques were evaluated according to three numerical indicators on the average [26], [27], that is, bias, precision and accuracy.

Bias is given by:

$$\bar{e} = \frac{\sum_{i=1}^{n}(\hat{b}_i - b_i)}{n} \tag{5}$$

$$\bar{e}\% = \frac{\bar{e}}{\bar{b}}100 \tag{6}$$

Precision is given by:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(\hat{b}_i - \bar{e} - b_i)^2}{n-1}} \tag{7}$$

$$s_e\% = \frac{s_e}{\bar{b}}100 \tag{8}$$

Accuracy is given by:

$$m_b = \sqrt{\frac{\sum_{i=1}^{n}(\hat{b}_i - b_i)^2}{n-1}} \tag{9}$$

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

$$m_b\% = \frac{m_b}{\bar{b}}100 \tag{10}$$

where:

- $\hat{b}_i$: estimated biomass (kg);
- b_i: real biomass (kg);
- $\bar{b}$: average of real biomass (kg);
- n: number of observations.

In addition, it was calculated R^2_{adj} and S_{yx} for the allometric model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(b_i - \hat{b}_i)^2}{\sum_{i=1}^{n}(b_i - \bar{b}_i)^2} \tag{11}$$

$$R^2_{adj} = 1 - \frac{n-1}{n-k}(1 - R^2) \tag{12}$$

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(b_i - \hat{b}_i)^2}{n-k}} \tag{13}$$

where:

- k: number of coefficients of the model.

The closer to zero (in module) is $\bar{e}\%$ the smaller is bias and better is the performance of the model. The smaller is the dispersion expressed by $s_e\%$, the greater is accuracy. The smaller $m_b\%$ the more accurate and closer to the target are the estimates. The interpretation of these indicators can be seen in Figure 1, making an analogy with shots at a target.

Estimates were also appreciated by the linear correlation between the estimated and actual values and the absolute residual distribution graphs (real value - estimated value).

III. Results and Discussion

A. Basic statistics of data

Examining the four variables analyzed (b, dbh, ht and ρ), it appears that none of them presents normality by testing Lilliefors and Shapiro-Wilk, as in their original form as the log-transformed. Therefore, thus it is shown that the first regression assumption [5] is not attended.

The data shows wide dispersion of the variables considered in the modeling (Table I). Correlations between biomass (b) and the independent variables dbh and ht were 0.79 and 0.61, respectively, which were significant to 95% of probability. On the other hand, the correlation ρ was -0.05, which are regarded as null. Therefore, changes in biomass may be significantly explained by their respective variations in diameter and height, but the specific mass alone does not explain the variations in biomass. Considering logarithmic transformation of variables, Pearson's correlation coefficients of biomass were 0.96, 0.86 and 0.07, respectively, which indicates the degree of association of variables increases with such transformation.

The relation of the dependent variable (b) presented curvilinear behavior, with greater dispersion in large trees, that is, large trees showed significant variation in their biomass. Something similar occurs with the height variable, that is, taller trees show greater variation in biomass. There is no direct relationship between density and biomass. When the logarithmic transformations of variables are considered, dbh and ht show a linear relationship with biomass, not occurring with ρ. Thus, it appears that, given the logarithmic model presented in [13], two variables follow the regression presupposition of variable linearity assumption [5], and one not.

The adjustment of allometric model of 2,802 data by the method of ordinary least squares resulted in the following equation:

$$\lg(\hat{b}) = -1.21356 + 2.01484\lg(dbh) + 0.888954\lg(ht) + 0.83138\lg(\rho) \tag{14}$$

With results R^2_{adj} = 0.9730 and S_{yx} = 0.1518.

Given this equation, it was possible to verify the normality and homoscedasticity of its residual. The results of the tests Lilliefors and Shapiro-Wilk stated lack of normality at 95% probability and graphical analysis of residues (Figure 3a) showed absence of heteroscedasticity along the estimate line. However, when the residuals of logarithmic variable are converted into biomass values it is noticed that there is variation of dispersion over adjust line, indicating that they do not behave homogeneously (Figure 3b). Therefore, it is evident that the assumptions of normality and homoscedasticity of residuals are not attended when adjusting allometric model.

B. Avaliation of AI Models

The results of the three AI models, alternatives to the allometric model proposed by [13] showed that KNN, ANN and SVM provide estimates with about the same degree of bias, accuracy and precision (Table II). In training with 70% of the data it was found bias below 10% for all models, which indicates that the estimates do not exhibit pronounced trends over or underestimation. Low bias was also observed in validations, which is a positive aspect, considering that the application of models to independent data to those used in training is consistent and trends free.

Given the high variability of the observed data, considering they refer to occurring trees throughout the Tropics and of different species and sizes, all models showed low precision and accurary indexes, although they have not shown bias, high correlation between the observed values and estimated was detected by the Pearson coefficient. The performance of the models was similar in training base, but SVM model excelled in accuracy (that joins bias and accuracy) which showed the smaller percentage values for both, the set of train data and test data. In testing base, ANN presented better correlation, althogh precision and accuracy are not the best.

Observing the graphic distribution of residuals, it is understood that estimates behave more evenly along the estimation line in comparison with test. It also appears that there is a tendency of heterocedasticity of residual along

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

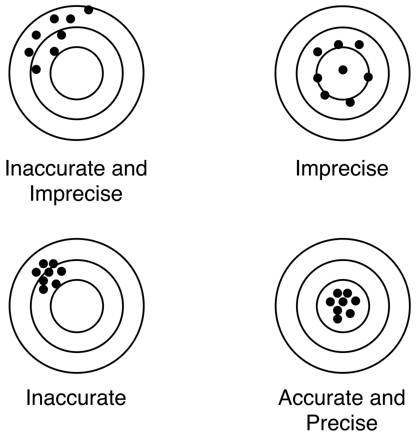


Fig. 1. Precision and Accuracy, where the goal is the center of the target

TABLE I
MEASURES

Variable	Average	Standard deviation	CV%	Minimum	Maximum
Biomass (b in Kg)	1,134.14	3,917.97	345.46	1.23	76,063.52
Diameter at breast height (dbh, in cm)	23.99	24.09	100.41	5.00	212.00
Total height (ht, in m)	16.04	10.77	67.17	1.30	70.70
Specific mass (ρ in $g.cm^{-3}$)	0.63	0.16	25.94	0.09	1.20

TABLE II
RESULTS

Data	Model	Bias%	Precision%	Accuracy%	r
Train (n = 2802)	Allometric	-2.22	100.09	100.10	0.9570
	KNN	1.30	91.13	91.16	0.9479
	ANN	7.75	91.71	92.04	0.9566
	SVM	-3.12	86.43	86.49	0.9619
Test (n = 1202)	Allometric	4.29	84.87	129.71	0.9580
	KNN	6.60	116.69	178.36	0,9092
	ANN	-0.04	133.01	133.01	0.9626
	SVM	-9.86	152.31	152.62	0.9426

the x-axis, increasing the variability for big trees. In the distributions corresponding to the validation it is more evident, with large dispersion of residual, in addition to detect trend of underestimation in large trees.

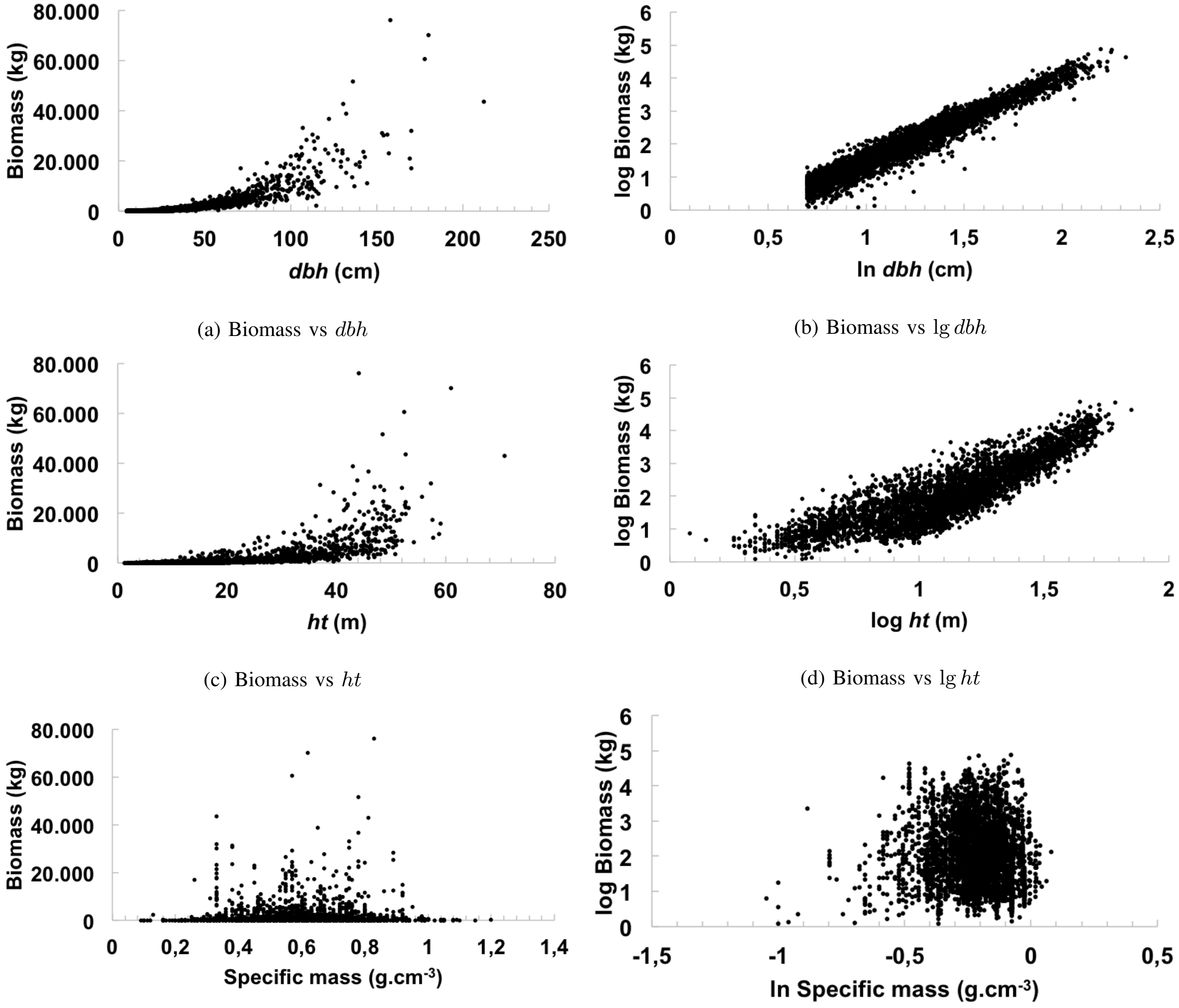
Native forests, particularly tropical, have high structural and dimensional variability. The major cause of this variability, especially in their biomass, is the occurrence of large trees. There was a direct linear relationship between the density of trees with dbh above 70 cm and its biomass in tropical forests [28]. The authors evaluated the importance of large trees in the stock biomass of tropical forests and rated intrinsic and extrinsic aspects related to changes that occur in different regions of the tropics. In this study it is evident that the major cause of loss of accuracy and bias is the increased variability in biomass of large trees. Calculating the biomass of trees above 70 cm dbh, it is found that, although few in number compared to smaller trees (about 6%), these individuals represent about 67% of the biomass.

Models to estimate the biomass of trees should seek to reduce uncertainties. These uncertainties are due to the nature of the data and aspects inherent to modeling. In this article we didn't emphasize the intrinsic or extrinsic variations in biomass of trees, but the modeling itself. However, it is important that these variations are clarified, since the accuracy of the models is highly dependent on the behavior of the data, especially the dispersion of biomass in large trees, a fact also reported in this study.

Regarding the models evaluated here, the main concern is that in many cases the conditions or assumptions of regression might not be attended. Data with large variations, such as those analyzed here, can cheat the requirements for application of regression models, as reported in the related literature. The diametric distribution of tropical forests is decreasing, which implies on non-normality of biomass data. The great variability in the biomass production of large dimensions trees implies heterocedasticity. Due to the nature of forest biomass data, these assumptions may lead to uncertainties in the inferences and the estimated potential of regression models

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618



(a) Biomass vs dbh

(b) Biomass vs $\lg dbh$

(c) Biomass vs ht

(d) Biomass vs $\lg ht$

(e) Biomass vs ρ

(f) Biomass vs $\lg \rho$

Fig. 2. Relation among biomass and dbh, height and specific mass of 4,004 individual trees at tropical forests.

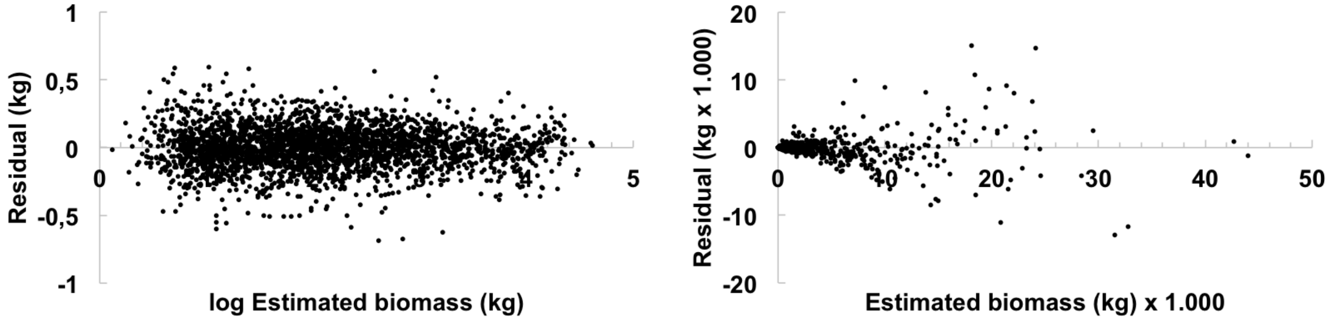
[6]. Therefore, other modeling approaches need to be sought.

In this study we tested some AI methods, K-nearest neighbor, artificial neural networks and support vector machines, as alternative to allometric ones. They are mathematical and computational methods using different principles of statistical regression. The KNN technique showed comparable accuracy to the allometric model. This technique is used in other areas and its application to estimate carbon in trees was recently introduced [9]. The estimation procedure of this technique is based on the average of the known values of the closest neighbors of a point to estimate, and in this case we analyzed three neighbors, but other options could be considered. Comparative modeling studies to quantify biomass in restoration plantings in the Atlantic indicated that this technique can generate accurate estimates and the setting of the number of neighbors affects the results [6]. The authors concluded that 3 to 5 neighboring enable better performance technique and as the number of neighbors increases no loss of accuracy. The amount of data also influences the performance of this technique in predicting biomass, requiring a large mass of data for the technique to work properly [15].

ANN also provided biomass estimates with the same degree of accuracy and precision that models based on regression. Although the technique already known and reasonably used in forestry, its application in quantification of forest biomass is more restricted to applications of geotechnology, such as remote sensing. Estimates of biomass stocks in a fragment of natural forest with satellite images IKONOS employment,

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618



(a) Residual vs lg b

(b) Residual vs b

Fig. 3. Residual of estimation of log decimal variable of biomass with allometric model (Equation 14) and related residual of variable biomass in function of estimates of 2,802 individual trees of tropical forests.

were held recently and with excellent performance [29]. Similar research conducted in tropical forests in Indonesia have also indicated that the estimated biomass ANN applied to Landsat 5 TM satellite images resulted in appropriate and strongly correlated with estimates forest inventory data [30].

SVM is a technique that is poorly explored in forestry, particularly in quantifying forest biomass. Studies on quantification of biomass Juniperus pinchotii in the United States with images derived from its top and SVM as classifier, produced promising results [31]. A review of the different machine learning techniques for applications in estimating biomass pointed to the potential of this technique for this purpose, highlighting its flexibility and ability to properly process large amounts of data is presented in [32].

AI models need to be more widely known and tested in forestry applications. In general, techniques such as KNN, SVM and ANN have the potential development and application in any field, such as forest inventory, forest planning, harvesting and forestry supply systems, etc. These techniques present auspicious prospects in the quantification of biomass and carbon in trees and forests, either as individual estimation strategy or per unit area, with forest inventories or remote sensing data. This potential needs to be further explored, especially when large amounts of data need to be analyzed and interpreted and there are restrictions to the application of conventional techniques such as regression.

IV. Conclusions

In this work, several machine learning algorithms were applied to estimate biomass of tropical forest trees. We trained KNN, SVM and ANN with more than 4,000 trees and results were compared to regression models.

Also, AI methods are presented as a suitable alternative to the state of art allometric techniques, since results obtained here are compared to allometric ones.

The main considerations are:

- Artificial Intelligence Models have strong estimation power for biomass of tropical trees, comparable or superior to regression (allometric model), which is considered state of the art;
- Analysis of a large amount of biomass data tropical forest trees shows that the assumptions underlying the use of regression models are frequently violated and often simply neglected;
- AI models constitute an attractive alternative to the regression technique, especially when the data do not show normality and homoscedasticity, which is the case of biomass of tropical forest trees.

V. Acknowledgements

The authors would like to thank Dr. Jerome Chave for the kind disposition of data used in this research.

References

[1] FAO – Food and Agriculture Organization, “Global forest resources assessments 2015,” FAO, Roma, Itália, Tech. Rep., 2015.

[2] UNFCCC – United Nations Framework Convention on Climate Change, “Fourth assessment report: Summary for policymakers,” Cambridge University, Geneva, Switzerland, Tech. Rep., 2007.

[3] L. F. Watzlawick, F. F. Kirchner, and C. R. Sanquetta, “Estimativa de biomassa e carbono em floresta com araucária utilizando imagens do satélite ikonos ii,” *Ciência Florestal*, vol. 19, pp. 169–181, 2009.

[4] C. R. Sanquetta, A. P. Corte, and F. da Silva, “Biomass expansion factor and root-to-shoot ratio for pinus in brazil,” *Carbon Balance and Management*, vol. 6, no. 1, pp. 1–8, 2011. [Online]. Available: http://dx.doi.org/10.1186/1750-0680-6-6

[5] J. W. Osborne and E. Waters, “Four assumptions of multiple regression that researchers should always test,” *Practical Assessment Research and Evaluation*, vol. 8, pp. 1–8, 2002.

[6] C. R. Sanquetta, J. Wojciechowski, A. P. Dalla Corte, A. Behling, S. Péllico Netto, A. L. Rodrigues, and M. N. I. Sanquetta, “Comparison of data mining and allometric model in estimation of tree biomass,” *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–9, 2015.

[7] M. J. Diamantopoulou, “Predicting fir trees stem diameters using artificial neural network models,” *The Southern African Forestry Journal*, vol. 205, no. 1, pp. 39–44, 2005. [Online]. Available: http://dx.doi.org/10.2989/10295920509505236

[8] H. G. Leite, M. L. M. da Silva, D. H. B. Binoti, L. Fardin, and F. H. Takizawa, “Estimation of inside-bark diameter and heartwood diameter for tectona grandis linn. trees using artificial neural networks,” *European Journal of Forest Research*, vol. 130, no. 2, pp. 263–269, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10342-010-0427-7

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

[9] C. R. Sanquetta, J. Wojciechowski, A. P. D. Corte, A. L. Rodrigues, and G. C. B. Maas, "On the use of data mining for estimating carbon storage in the trees," *Carbon Balance Manag*, vol. 8, pp. 6–6, Jun 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693975/

[10] R. V. O. Castro, C. P. B. Soares, F. B. Martins, and H. G. Leite, "Crescimento e produção de plantios comerciais de eucalipto estimados por duas categorias de modelos," *Pesquisa Agropecuária Brasileira*, vol. 48, pp. 287 – 295, 03 2013.

[11] A. Imada, "A literature review: Forest management with neural network and artificial intelligence," in *Neural Networks and Artificial Intelligence*, ser. Communications in Computer and Information Science, V. Golovko and A. Imada, Eds. Springer International Publishing, 2014, vol. 440, pp. 9–21.

[12] F. A. A. M. N. Soares, E. L. Flores, C. D. Cabacinha, G. A. Carrijo, and A. C. P. Veiga, "Recursive diameter prediction for calculating merchantable volume of eucalyptus clones using multilayer perceptron," *Neural Computing and Applications*, vol. 22, no. 7, pp. 1407–1418, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00521-012-0823-7

[13] J. Chave, C. Andalo, S. Brown, M. A. Cairns, J. Q. Chambers, D. Eamus, H. Fölster, F. Fromard, N. Higuchi, T. Kira, J.-P. Lescure, B. W. Nelson, H. Ogawa, H. Puig, B. Riéra, and T. Yamakura, "Tree allometry and improved estimation of carbon stocks and balance in tropical forests," *Oecologia*, vol. 145, no. 1, pp. 87–99, 2005. [Online]. Available: http://dx.doi.org/10.1007/s00442-005-0100-x

[14] F. Schumacher and D. Hall, "Logarithmic expression of timber-tree volume," *Journal of Agricultural Research*, vol. 47, no. 9, pp. 719–734, 1933.

[15] F. Mognon, A. P. D. Corte, C. R. Sanquetta, T. G. Barreto, and J. Wojciechowski, "Estimativas de biomassa para plantas de bambu do gênero Guadua," *Revista Ceres*, vol. 61, pp. 900 – 906, 12 2014.

[16] P. B. Brazdil, C. Soares, and J. P. da Costa, "Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results," *Machine Learning*, vol. 50, no. 3, pp. 251–277, 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1021713901879

[17] A. Braga, A. C. Carvalho, and T. B. Ludermir, *Redes Neurais Artificiais: Teoria e aplicações*. LTC Editora, 2007, vol. 2. [Online]. Available: http://www.worldcat.org/isbn/9788521615644

[18] S. Haykin, *Redes neurais: princípios e prática*. Porto Alegre, RS: Bookman, 2001.

[19] M. L. M. da Silva Binoti, "Emprego de redes neurais artificiais em mensuração florestal e manejo florestal," Ph.D. dissertation, Universidade Federal de Viçosa, 2012.

[20] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 303–314, 1989.

[21] D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986.

[22] A. C. Lorena and A. C. de Carvalho, "Uma introdução às support vector machines," *Revista de Informática Teórica e Aplicada*, vol. 14, no. 2, pp. 43–67, 2007.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] V. N. Vapnik, *Statistical learning theory*. John Wiley and Sons, 1998.

[25] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, vol. EC-14, pp. 326–334, 1965.

[26] H. E. Burkhart and M. Tomé, *Modeling Forest Trees and Stands*. Springer Netherlands, 2012.

[27] H. Pretzsch, *Forest Dynamics, Growth and Yield*. Springer-Verlag Berlin Heidelberg, 2009.

[28] J. W. F. Slik, G. Paoli, K. McGuire, I. Amaral, J. Barroso, M. Bastian, L. Blanc, F. Bongers, P. Boundja, C. Clark, M. Collins, G. Dauby, Y. Ding, J.-L. Doucet, E. Eler, L. Ferreira, O. Forshed, G. Fredriksson, J.-F. Gillet, D. Harris, M. Leal, Y. Laumonier, Y. Malhi, A. Mansor, E. Martin, K. Miyamoto, A. Araujo-Murakami, H. Nagamasu, R. Nilus, E. Nurtjahya, A. Oliveira, O. Onrizal, A. Parada-Gutierrez, A. Permana, L. Poorter, J. Poulsen, H. Ramirez-Angulo, J. Reitsma, F. Rovero, A. Rozak, D. Sheil, J. Silva-Espejo, M. Silveira, W. Spironelo, H. ter Steege, T. Stevart, G. E. Navarro-Aguilar, T. Sunderland, E. Suzuki, J. Tang, I. Theilade, G. van der Heijden, J. van Valkenburg, T. Van Do, E. Vilanova, V. Vos, S. Wich, H. Wöll, T. Yoneda, R. Zang, M.-G. Zhang, and N. Zweifel, "Large trees drive forest aboveground biomass variation in moist lowland forests across the tropics," *Global Ecology and Biogeography*, vol. 22, no. 12, pp. 1261–1271, 2013. [Online]. Available: http://dx.doi.org/10.1111/geb.12092

[29] A. S. Ferraz, V. P. Soares, C. P. B. Soares, C. A. A. S. Ribeiro, D. H. B. Binoti, and H. G. Leite, "Estimativa do estoque de biomassa em um fragmento florestal usando imagens orbitais," *Floresta e Ambiente*, vol. 21, pp. 286 – 296, 09 2014.

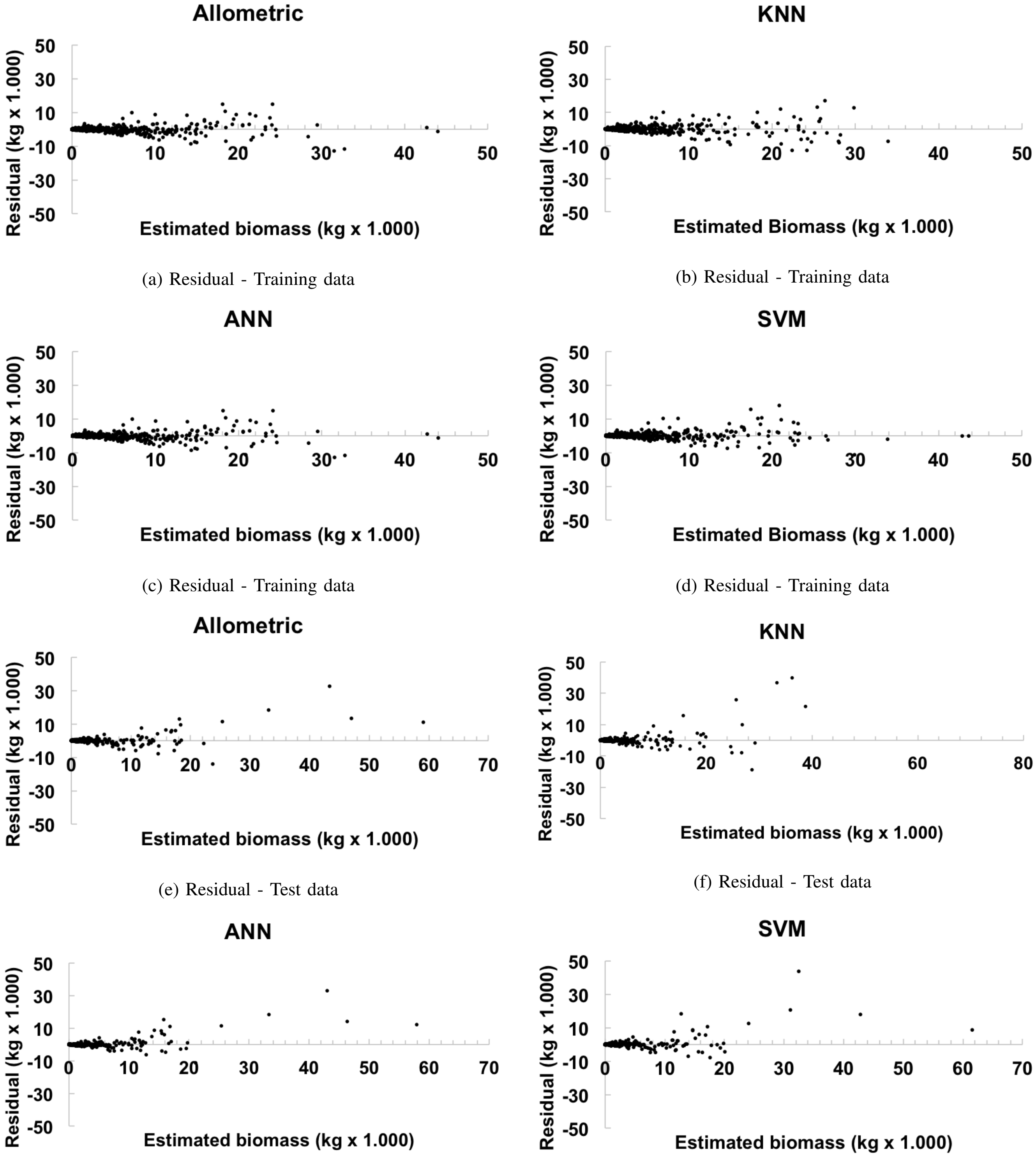
[30] G. M. Foody, M. E. Cutler, J. McMorrow, D. Pelz, H. Tangki, D. S. Boyd, and I. Douglas, "Mapping the biomass of bornean tropical rain forest from remotely sensed data," *Global Ecology and Biogeography*, vol. 10, no. 4, pp. 379–387, 2001. [Online]. Available: http://www.jstor.org/stable/2665383

[31] M. Mirik, S. Chaudhuri, B. Surber, S. Ale, and R. J. Ansley, "Evaluating biomass of juniper trees (juniperus pinchotii) from imagery-derived canopy area using the support vector machine classifier," *Advances in Remote Sensing*, vol. 2, pp. 181–192, 2013.

[32] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann, and C. Notarnicola, "Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data," *Remote Sensing*, vol. 7, pp. 16 398–16 421, Dec. 2015.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618



(a) Residual - Training data

(b) Residual - Training data

(c) Residual - Training data

(d) Residual - Training data

(e) Residual - Test data

(f) Residual - Test data

(g) Residual - Test data

(h) Residual - Test data

Fig. 4. Graphical distribution of residual of four models to estimate dry biomass above ground, *dbh*, total height and specific mass of 4004 individual trees at tropical forests.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

A Complex Network Approach to Identify Potential Financial Scandals: The Colombian Market Case

Jaime Humberto Nino-Pena, Cesar Garcia-Diaz, and German Jairo Hernandez

Abstract—**Financial data is abundant, diverse and generated in large volumes at any time worldwide. Finding fast and reliable ways of analyzing it is key for market actors (regulators, market makers, brokers and investors). In this work, we intend to use graph theory as a vehicle to analyze relationships among owners of publicly traded companies in order to extract latent dynamics that are very difficult to discover otherwise. As a case study we took the bankruptcy of the largest brokerage firm in Colombia in 2012. Network theory uncovered latent dynamics in the relationships among key owners and companies during the period of analysis (2009 - 2012) that could be used in the future as an early warning tool for market participants.**

Index Terms—**computational finance, graph theory, knowledge discovery, time series mining, financial data mining**

I. Introduction

Financial data is being generated almost instantly across the world. It comes in two basic forms: fundamental and market data. On one hand, market data reflects the market dynamics when agents trade any financial asset. On the other hand, fundamental data refers to news, press releases, financial statements and other economic, social or political events that may affect prices of financial assets. Given the fact that prices respond to actions generated by market agents, time becomes a key element since the price formation process happens as agents interact over time. As a result, financial data constitutes an attractive source for researchers, since it exhibits complex systems' properties[13] which are difficult to analyze and understand[14][16][17]. Finding ways to summarize financial data is fundamental to support market agents' decision making process.

a) : Graph theory has proven to be very useful for analyzing complex dynamics in a wide variety of fields[1][4], including financial markets' behavior [2][19][20][21][22], ownership analysis of publicly traded companies[3][9] and systematic risk by contagion models[5]. It facilitates the analysis of inter-dependencies analysis among graph components, and synthesizes huge amounts of data such as those observed in financial markets.

b) : This paper extends ownership structure analysis by specifically analyzing behaviors in owners' sub-networks. Our interest is focused in the preceding time lapse of a bankruptcy by aiming at finding possible patterns in the ownership structure. As a case study we take the bankruptcy of INTERBOLSA, the largest security brokerage firm in Colombia. The paper continues with the following sections: a background of the financial case of analysis, a brief description of key graph theory definitions, the experiment, discussion and conclusions.

II. Financial Context

Market actors use a wide variety of informational sources to negotiate financial assets such as fundamental data (e.g., macroeconomic news, industry analysis, and companies' financial statements), technical information (e.g., statistical analysis over past data) and insights obtained from application of modern techniques such as machine learning) [16][18]. Data processing for any of these categories is a key component for the market agent decision making process. In this paper, we focus our attention on analyzing fundamental market data, particularly ownership dynamics of the top 20 owners of major companies listed in the Colombian Stock Market. Our motivation lies in the 2012 bankruptcy case of the biggest brokerage firm in Colombia (INTERBOLSA). Given the fact that this firm managed a significant part of the money flow in this market, its bankruptcy caused a very negative impact in market confidence. Moreover, the collapse was caused by its own top management team, which in collusion with another important investors in the Colombian market (private funds and individuals), tried to gain control on other public traded company (FABRICATO) via repurchasing agreements (REPOS)[1] using their own money as well as money from INTERBOLSA's clients. Under this scenario we want to apply network based principles to detect patterns from data, not easy observable by other means.

III. Graph theory definitions used

- Bipartite Graph: A graph in which links relate two independent set of nodes (U, V). That is, an element in U could only be linked to an element in V[1].

Manuscript received on January 16, 2017, accepted for publication on May 3, 2017, published on July 1, 2017.

The authors are with the Universidad Nacional de Colombia (email: jhninop@unal.edu.co, gjhernandezp@unal.edu.co), and the Universidad de los Andes (email: ce.garcia392@uniandes.edu.co).

[1] A REPO is an acronym for Repurchase Agreement. In Colombia they are used to facilitate leverage operations on stocks.

https://doi.org/10.17562/PB-56-5

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

- Eigenvector centrality: It is a recursive measure that determines how central is a node within a network. The node's centrality is based on the number of, and the quality of, its connections[12].
- Community detection (modularity): It is a measure that determine modules or communities depending on how nodes are interconnected. It allows to analyze the overall community structure within a graph[1][7].

IV. Experiment setup

Network theory has been used in the extant literature to analyze the dynamics between owners and public traded companies[3][9]. We want to apply specific network concepts such as community detection, bipartite graphs and eigenvector centrality, in order to extract patterns of ownership behavior of firms and individuals with the largest involvement in INTERBOLSA collapse[2] (i.e., INTERBOLSA, VALORES INCORPORADOS, INVERTATICAS, ALESANDRO CORRIDORI, HELADOS MODERNOS DE COLOMBIA, GITECO SAS, MANANTIAL SVP and RENTAFOLIO BURSATIL).

A. Data acquisition

Colombian public listed companies must report information to the market regulator (Superfinanciera de Colombia)[3]. This information is publicly available and it includes information about the Top 20 owners. As a result, data was collected quarterly basis from 2009-01-01 to 2012-09-30 for the major companies of the Colombian Market, including INTERBOLSA.

B. Data pre-processing

Data pre-processing was carried out using Talend Data Integrator and Python. Original data was downloaded in Excel format and includes company name, quarter reported, owner id, owner name, number and class of shares owned, and the percentage of ownership. Data only includes the top 20 of the owners. Given the original formatting of numbers, it was necessary to verify companies' id's, number format (decimal character and thousands separator). Because Gephi was used to build the graphs, the pre-processing output were CSV files containing columns required by Gephi[4].

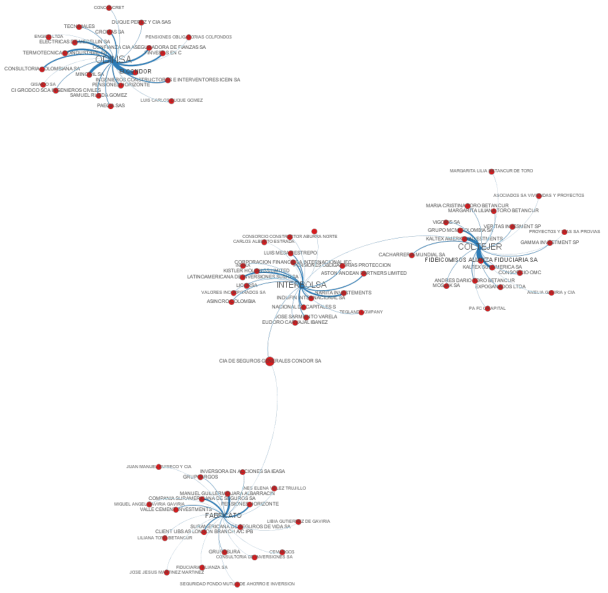

CONCONCRET
DUQUE PEREZ Y CIA SAS
TECNIVIALES
CROMAS SA
ENGIKO LTDA
ELECTRICAS DE MEDELLIN SA
PENSIONES OBLIGATORIAS COLFONDOS
CONFIANZA CIA ASEGURADORA DE FIANZAS SA
TERMOTECNICA COINDUSTRIAL SA
INVESTIS EN C
ODINSA
CONSULTORIA COLOMBIANA SA
EL CONDOR
MINCIVIL SA
INGENIEROS CONSTRUCTORES E INTERVENTORES ICEIN SA
GISAICO SA
PENSIONES HORIZONTE
CI GRODCO SCA INGENIEROS CIVILES
SAMUEL RUEDA GOMEZ
LUIS CARLOS DUQUE GOMEZ
PAECIA SAS
MARGARITA LILIA BETANCUR DE TORO
ASOCIADOS SA VIVIENDAS Y PROYECTOS
MARIA CRISTINA TORO BETANCUR
MARGARITA LILIANA TORO BETANCUR
VIGORIS SA
VERITAS INVESMENT SP
GRUPO MCM COLOMBIA SA
PROYECTOS Y VIAS SA PROVIAS
KALTEX AMERICA INVESTMENTS
GAMMA INVESTMENT SP
COLTEJER
CACHARRERIA MUNDIAL SA
FIDEICOMISOS ALIANZA FIDUCIARIA SA
KALTEX SURAMERICA SA
CONSORCIO OMC
ANDRES DARIO TORO BETANCUR
MOSSIK SA
EXPOGANADOS LTDA
AMELIA GAVIRIA y CIA
PA FC C CAPITAL
CONSORCIO CONSTRUCTOR ABURRA NORTE
CARLOS ALBERTO ESTRADA
LUIS MESA RESTREPO
CORPORACION FINANCIERA INTERNACIONAL IFC
INBSA
PENSIONES OBLIGATORIAS PROTECCION
KISTLER HOLDINGS LIMITED
ASTON ANDEAN PARTNERS LIMITED
LATINOAMERICANA DE INVERSIONES SURA SA
LICORSA
INTERBOLSA
VALORES INCORPORADOS SA
NARITA INVESTEMENTS
INDUFIN INTERNACIONAL SA
ASINCRO COLOMBIA
NACIONAL DE CAPITALES S
TEGLAND COMPANY
JOSE SARMIENTO VARELA
EUDORO CARVAJAL IBANEZ
CIA DE SEGUROS GENERALES CONDOR SA
JUAN MANUEL RUISECO Y CIA
INVERSORA EN ACCIONES SA IEASA
GRUPO ARGOS
INES ELENA VELEZ TRUJILLO
MANUEL GUILLERMO JARA ALBARRACIN
COMPANIA SURAMERICANA DE SEGUROS SA
MIGUEL ANGEL GAVIRIA GAVIRIA
PENSIONES HORIZONTE
VALLE CEMENT INVESTMENTS
FABRICATO
LIBIA GUTIERREZ DE GAVIRIA
SURAMERICANA DE SEGUROS DE VIDA SA
CLIENT UBS AS LONDON BRANCH A/C IPB
LILIANA TORO BETANCUR
GRUPO SURA
CEMARGOS
CONSULTORIA DE INVERSIONES SA
FIDUCIARIA ALIANZA SA
JOSE JESUS MARTINEZ MARTINEZ
SEGURIDAD FONDO MUTUO DE AHORRO E INVERSION

Fig. 1. Bipartite graph for the first quarter of 2009

C. Graph construction

Two types of graphs were used:

1) A bipartite graph having, on the one hand, owners as one group of nodes, and on the other hand the companies as the othe group. Links represent the ownership relation, whereas link weights are the percentage owned by owners in companies. The following conventions are used for these graphs:
   a) Node color is given by the in-degree value, being red the lowest value, and blue the highest one.
   b) Node size is given by the out-degree value.
   c) Node label size is given by the weighted degree; it means that the biggest the font the larger the number of connections for a particular node n_i.
   d) Link thickness is given by the ownership $\%$ of owner A in company X.
2) A projected owner graph, derived from the relationship observed on the bipartite graph. Conventions are as follows:
   a) Node color is given by the modularity class (detected community).
   b) Node size is given by the eigenvector centrality measure.
   c) Node label size and label color are given by the degree, being red the lowest degree value and blue the highest one.

In particular, we were interested in the following publicly listed companies: INTERBOLSA, FABRICATO, COLTEJER, ODINSA and BIOMAX, which were the ones that presented stronger declines in their stock prices by the time the mismanagement allegations became public. We were also interested in the following list of owners: INTERBOLSA, VALORES INCORPORADOS, INVERTATICAS, ALESANDRO CORRIDORI, HELADOS MODERNOS DE COLOMBIA, GITECO SAS, MANANTIAL SVP and

[2] http://www.elespectador.com/noticias/infografia/actuaciones-de-autoridades-el-caso-interbolsa-articulo-429718

[3] http://www.superfinanciera.gov.co

[4] One CSV file for nodes, which are ids of owners and listed companies; another one for edges, which included a tuple (owner,owned) and the ownership percentage

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

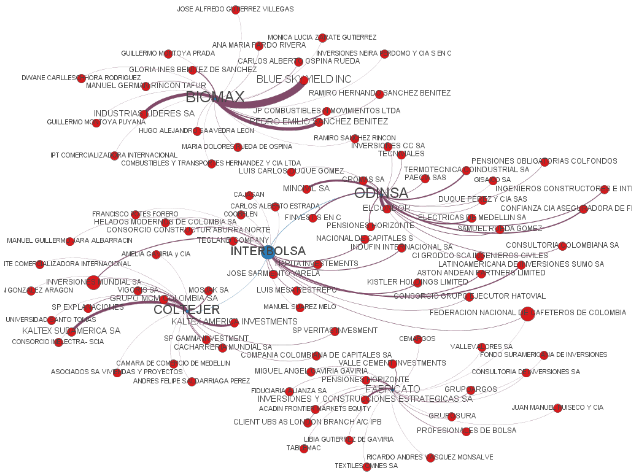

JOSE ALFREDO GUTIERREZ VILLEGAS
MONICA LUCIA ZARATE GUTIERREZ
ANA MARIA PARDO RIVERA
GUILLERMO MONTOYA PRADA
INVERSIONES NEIRA PERDOMO Y CIA S EN C
CARLOS ALBERTO OSPINA RUEDA
GLORIA INES BENITEZ DE SANCHEZ
DWANE CARLLESCHHORA RODRIGUEZ
BLUE SKY YIELD INC
MANUEL GERMAN RINCON TAFUR
RAMIRO HERNANDO SANCHEZ BENITEZ
BIOMAX
INDUSTRIAS LIDERES SA
JP COMBUSTIBLES Y MOVIMIENTOS LTDA
GUILLERMO MONTOYA PUYANA
PEDRO EMILIO SANCHEZ BENITEZ
HUGO ALEJANDRO SAAVEDRA LEON
RAMIRO SANCHEZ RINCON
INVERSIONES CC SA
MARIA DOLORES RUEDA DE OSPINA
TECNIVIALES
IPT COMERCIALIZADORA INTERNACIONAL
COMBUSTIBLES Y TRANSPORTES HERNANDEZ Y CIA LTDA
PENSIONES OBLIGATORIAS COLFONDOS
LUIS CARLOS DUQUE GOMEZ
TERMOTECNICA COINDUSTRIAL SA
CROMAS SA
PAECIA SAS
GISAICO SA
MINCIVIL SA
ODINSA
INGENIEROS CONSTRUCTORES E INTI
CAJASAN
DUQUE PEREZ Y CIA SAS
CARLOS ALBERTO ESTRADA
ELCONDOR
CONFIANZA CIA ASEGURADORA DE FI
FRANCISCO RATES FORERO
COOBELEN
FINVESTS EN C
ELECTRICAS DE MEDELLIN SA
HELADOS MODERNOS DE COLOMBIA SA
PENSIONES HORIZONTE
SAMUEL RUEDA GOMEZ
CONSORCIO CONSTRUCTOR ABURRA NORTE
NACIONAL DE CAPITALES S
MANUEL GUILLERMO JARA ALBARRACIN
TEGLAND COMPANY
INDUFIN INTERNACIONAL SA
CONSULTORIA COLOMBIANA SA
INTERBOLSA
AMELIA GAVIRIA y CIA
CI GRODCO SCA INGENIEROS CIVILES
NARITA INVESTEMENTS
LATINOAMERICANA DE INVERSIONES SUMO SA
ITE COMERCIALIZADORA INTERNACIONAL
JOSE SARMIENTO VARELA
ASTON ANDEAN PARTNERS LIMITED
INVERSIONES MUNDIAL SA
KISTLER HOLDINGS LIMITED
GONZALEZ ARAGON
VIGORIS SA
MOSDAK SA
LUIS MESA RESTREPO
CONSORCIO GRUPO EJECUTOR HATOVIAL
GRUPO MCM COLOMBIA SA
SP EXPLANACIONES
MANUEL SUAREZ MELO
FEDERACION NACIONAL DE CAFETEROS DE COLOMBIA
COLTEJER
UNIVERSIDAD SANTO TOMAS
KALTEX AMERICA INVESTMENTS
KALTEX SUDAMERICA SA
SP VERITAS INVESMENT
CONSORCIO INELECTRA- SCIA
SP GAMMA INVESTMENT
CEMARGOS
CACHARRERIA MUNDIAL SA
VALLEVALORES SA
COMPANIA COLOMBIANA DE CAPITALES SA
FONDO SURAMERICANA DE INVERSIONES
CAMARA DE COMERCIO DE MEDELLIN
VALLE CEMENT INVESTMENTS
ASOCIADOS SA VIVIENDAS Y PROYECTOS
MIGUEL ANGEL GAVIRIA GAVIRIA
CONSULTORIA DE INVERSIONES SA
ANDRES FELIPE SALDARRIAGA PEREZ
PENSIONES HORIZONTE
FIDUCIARIA ALIANZA SA
FABRICATO
GRUPO ARGOS
INVERSIONES Y CONSTRUCCIONES ESTRATEGICAS SA
ACADIN FRONTIER MARKETS EQUITY
JUAN MANUEL RUISECO Y CIA
CLIENT UBS AS LONDON BRANCH A/C IPB
GRUPO SURA
PROFESIONALES DE BOLSA
LIBIA GUTIERREZ DE GAVIRIA
TABLEMAC
RICARDO ANDRES VASQUEZ MONSALVE
TEXTILES OMNES SA

Fig. 2. Bipartite graph for the third quarter of 2009

RENTAFOLIO BURSATIL, since they were accused of collusion for trying to take ownership of FABRICATO using others investors' money, with consent of INTERBOLSA.

V. Results

Results indicate that network theory uncovers relationships difficult to identify by other means, which are different from the observed ownership dynamics of companies that did not go bankrupt.

In order to present the results, Figures 1 - 5 illustrate the bipartite graphs of the top 20 owners and companies for different quarters, over time. As the reader can observe, all the listed companies previously mentioned share the same owners. This feature is kept by all of the companies during the period of analysis.

Visualizations reveal a strong relationship among names involved in the allegations of FABRICATO takeover (i.e., INTERBOLSA, VALORES INCORPORADOS, INVERTATICAS, ALESANDRO CORRIDORI, HELADOS MODERNOS DE COLOMBIA, GITECO SAS, MANANTIAL SVP and RENTAFOLIO BURSATIL).

Link thickness represents owner's stake A in company X. As a result, visualizations confirm that involved owners had indeed an important stake in FABRICATO stock, and it was increasing over time. Figure 5 shows that all of the names previously mentioned account for a large interest in FABRICATO.

The second part of the results about ownership structure dynamics are the most interesting. Owners projected graph are analyzed. Figure 6 reveals the community structure of the involved names. For most of the time periods analyzed these

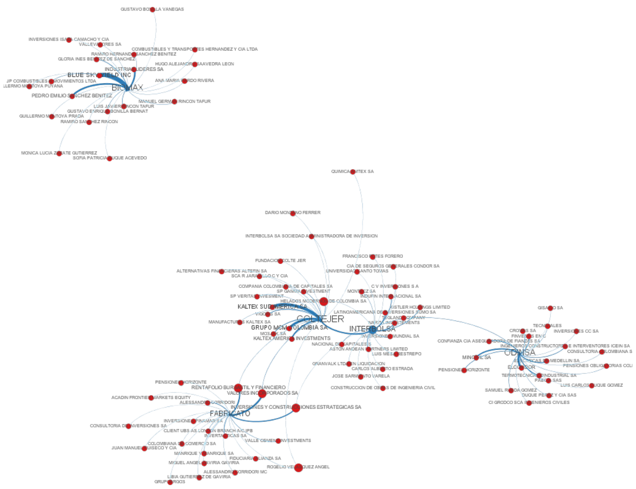


Fig. 3. Bipartite graph for the first quarter of 2010

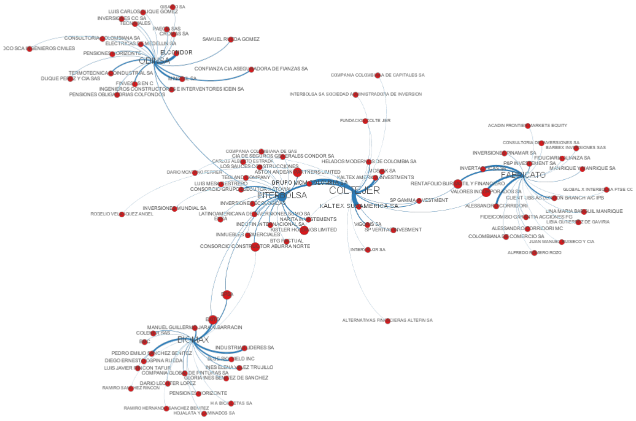


Fig. 4. Bipartite graph for the first quarter of 2011

https://doi.org/10.17562/PB-56-5

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

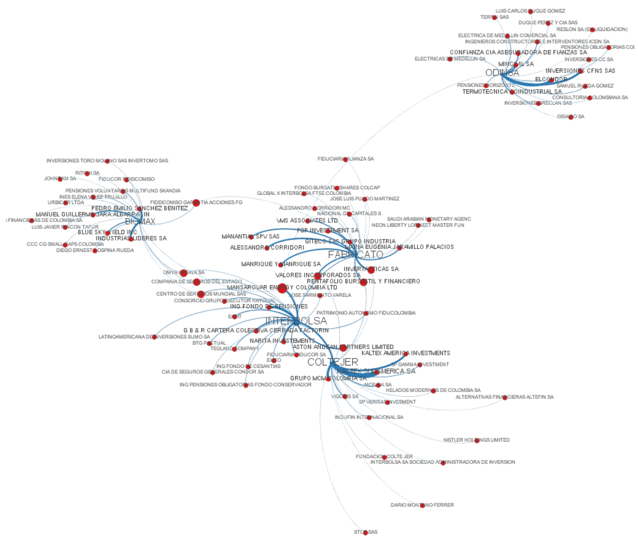


Fig. 5. Bipartite graph for the third quarter of 2012

names are grouped under a few number of communities. In fact, they are grouped under just one community by the time allegations went public (2012-III).

Analysis of eigenvector centrality measures (Figures 7 to 10) revealed that VALORES INCORPORADOS, RENTAFOLIO BURSATIL and INVERTACTICAS, were the most influential nodes within the analyzed community structure. In fact, these three companies were the most compromised in the scandal.

Ownership structure strongly changed for the three public traded companies most involved in the scandal (COLTEJER, FABRICATO and INTEROLSA). Figure 11 evidences how ownership concentration among the top 20 owners builds up over time, reaching similar levels by the time the scandal went public.

VI. Discussion

Network theory facilitates the analysis of ownership interdependencies for the case proposed. In fact, community detection and centrality measures allowed fast identification of key players. Visualizing results for each quarter revealed the influence of these owners through the whole period of analysis. Also, by considering the weights of the graph, it was possible to observe how the concentration of ownership grew over the different quarters for the three companies which shared the most owners (INTERBOLSA, FABRICATO and COLTEJER). See Figure 11. While the ownership concentration of these firms increased over time, for other companies not involved in ownership manipulation their concentration pattern proved to be completely different: some of them appeared to be almost unchanged over the same period of time.

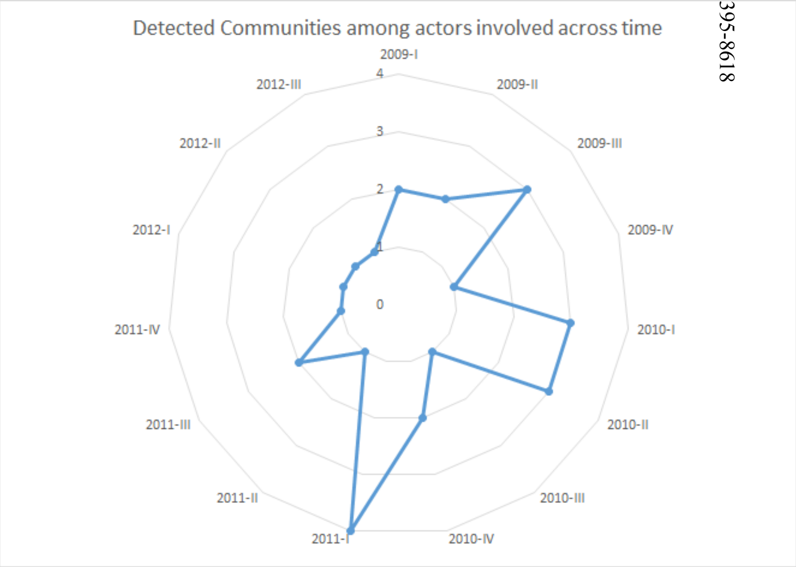


Fig. 6. Summary of number of communities detected across quarterly data

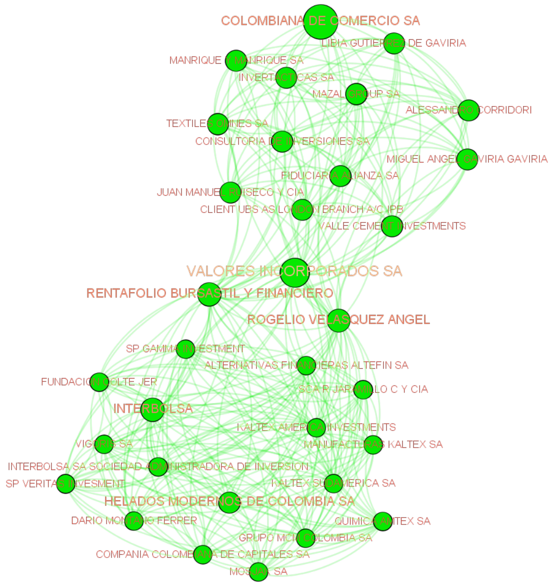


Fig. 7. Projected graph for the fourth quarter of 2009, colored by community

VII. Conclusions

Network theory concepts applied to financial data yield compelling evidence regarding the particularities of complex ownership dynamics. Data analyzed in a timely manner could serve as early warning system for market participants, particularly when reflecting salient changes in ownership structure. As we saw in the case of this paper, activities that derived into INTERBOLSA bankruptcy during the third quarter of 2012, seemed to have started back in the four quarter of 2009. As a result, it is plausible to apply network

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

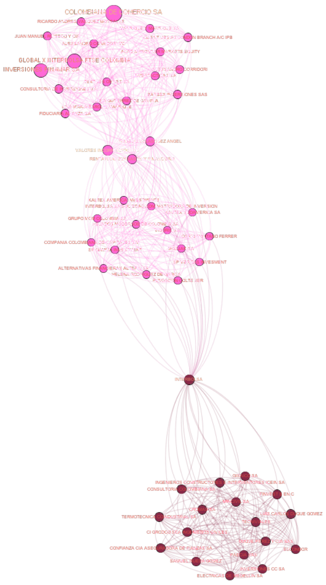


Fig. 8. Projected graph for the fourth quarter of 2010, colored by community

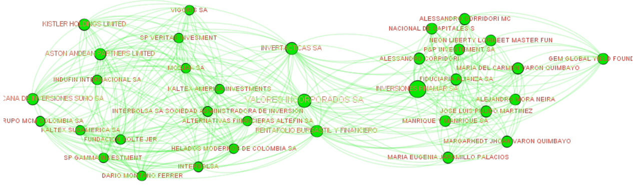

VIGIA SA
KISTLER HOLDINGS LIMITED
SP VERITAS INVESTMENT
ASTON ANDEAN PARTNERS LIMITED
INVERTCAS SA
MONICA SA
INDUFIN INTERNACIONAL SA
KALTEX AMERICA INVESTMENTS
VALORES INCORPORADOS SA
INTERBOLSA SA SOCIEDAD ADMINISTRADORA DE INVERSION
ALTERNATIVAS FINANCIERAS ALTEFIN SA
RENTAFOLIO BURSATIL Y FINANCIERO
HELADOS MODERNOS DE COLOMBIA SA
FUNDACION COLTEJER
SP GAMMA INVESTMENT
INTERBOLSA
DARIO MONTOYA FERRER
ALESSANDRO CORRIDORI MC
NACIONAL DE CAPITALES S
NEON LIBERTY LORIKEET MASTER FUN
PGP INVESTMENT SA
GEM GLOBAL YIELD FOUND
MARIA DEL CARMEN VARON QUIMBAYO
FIDUCIARIA ALIANZA SA
ALEJANDRO SORA NEIRA
JOSE LUIS PINZON MARTINEZ
MANRIQUE MANRIQUE SA
MARGARHEDT JHON VARON QUIMBAYO
MARIA EUGENIA JARAMILLO PALACIOS

Fig. 9. Projected graph for the forth quarter of 2011, colored by community

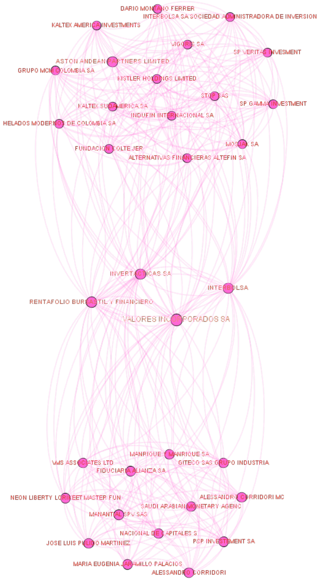


Fig. 10. Projected graph for the third quarter of 2012, colored by community

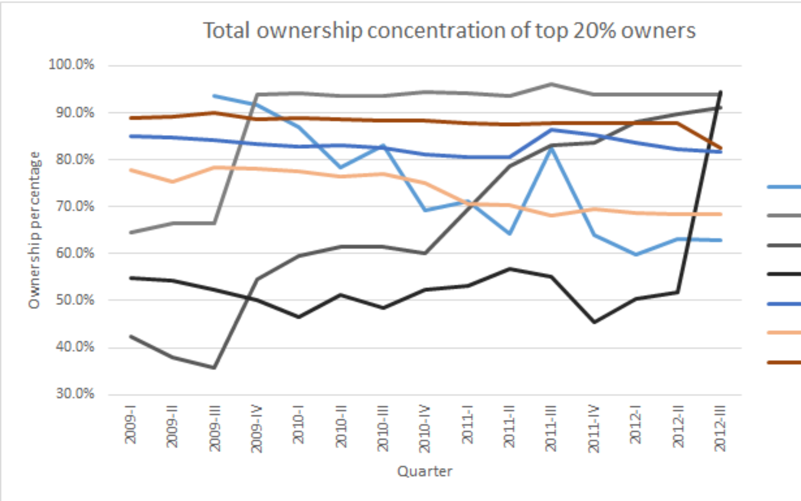


Fig. 11. Graph showing behavior of total ownership for companies involved from 2009 to 2012

theory concepts to ownership reporting of public companies to look for possible changes in communities, ownership concentration and centrality of players. This information could timely uncover undesired dynamics that might trigger negative market impacts in the future. Given the fact that network theory has been widely used to analyze contagion, approaches such as [5] could complement our case in order to analyze the stress that this bankruptcy brought to the Colombian security markets. Moreover, incorporating other sources of information such as financial indicators (Net profit, D/E ratios, etc.) could be helpful in order to complement our analyses and should be considered in future works.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

a) Acknowledgment.: The authors of the paper thank infoCapitales.com for the expertise provided.

References

[1] Barabasi, A.: Network Science. Cambridge University Press, Cambridge (2015)

[2] Bezsudnov, I., Snarskii, A.: From the time series to the complex networks: The parametric natural visibility graph. Physica A: Statistical Mechanics and its Applications, vol 414, pp 53-60 (2014)

[3] Bohlin, Ludvig: Network Analysis of the share ownership structure on the Swedish market. Umea University, Faculty of Science and Technology (2012)

[4] Easley, D., Kelinberg, J.: Networks, Crowds and Markets: Reasoning about a Highly Connected World. Cambridge University Press, Cambridge (2010)

[5] Elliot, M., Golup, B., Jackson, M.: Financial Networks and Contagion. American Economic Review, Vol 104, issue 10, pp 3115-3153 (2014)

[6] Bouveret, A., Guillaumie, C., Roqueiro, C. A., Winkler, C., Nauhaus, S.: High-frequency trading activity in EU equity markets. ESMA Report on Trends, Risks and Vulnerabilities, issue 1, pp 41-47 (2014)

[7] Fortunato, S.: Community detection in graphs. Physics Reports Vol 486, issue 3, pp 75-174 (2010)

[8] Geetha, N., Sekar, P.: Graph Theory Matrix Approach A Review. Indian Journal of Science and Technology, vol 9, issue 16, pp 1-4, (2016)

[9] Vitali, S., Glattfelder, J. B., Battiston, S.: The network of global corporate control. PloS one, vol 6, issue 10, e25995 (2011)

[10] Luca, C.: Technical Analysis Applications in the Global Currency Markets, 2nd Edition, New York Institute of Finance, New York (2000).

[11] Naish, J.: Managing chaos, Nursing management, vol 2, issue 1, p 3 (1995).

[12] Newman, M.: The mathematics of networks. The new palgrave encyclopedia of economics, Vol 2, pp 1-12 (2008)

[13] Ohlsson, S.: Deep Learning: How the Mind Overrides Experience. Cambridge University Press, New York (2011).

[14] Ortega, L.: A neuro-wavelet method for the forecasting of financial time series. Proceedings of the World Congress on Engineering and Computer Science, vol. I, pp 24-26 (2012)

[15] Ruhnau, B.: Eigenvector-centrality: a node centrality? Social Networks, vol 22, issue 4, pp 357-365 (2000)

[16] Stasinakis, C., Sermpinis, G.: Financial forecasting and trading strategies: a survey. Computational Intelligence Techniques and Trading and Investment (2014)

[17] Tay, F., Cao, L.: Application of support vector machines in financial time series forecasting. Omega, vol 29, pp 309317 (2001)

[18] Treleaven, P., Galas, M., Lalchand. V.: Algorithmic trading review. Commununications of the ACM, vol 56, issue 11, pp 76-85 (2013)

[19] Yan, S., Wang, D.: Time series analysis based on visibility graph theory. 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Vol 2, pp 311-314 (2015)

[20] Yang, Y., Wang, J., Yang, H., Mang, J.: Visibility graph approach to exchange rate series. Physica A: Statistical Mechanics and its Applications, vol 388, issue 20, pp 4431-4437 (2009)

[21] Yang, Y., Yang, H.: Complex network-based time series analysis. Physica A: Statistical Mechanics and its Applications, vol 387, issue 5, pp 1381-1386 (2008)

[22] Zhuang, E., Small, M., Feng, G.: Time series analysis of the developed financial markets' integration using visibility graphs. Physica A: Statistical Mechanics and its Applications, Vol 410, pp 483-495 (2014)

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Raster data implemented in a FPGA device

J. Sandoval-Gutierrez, J.A. Alvarez-Cedillo, J.C. Herrera-Lozada, T. Alvarez-Sanchez and M. Olguin-Carbajal

***Abstract*—The instrumentation for image processing sends the information through a communication interface, and the applications are developed by two general methods: a Software Development Kit as a PC-based programming language and a hardware implementation. In the first method, the users are limited by other processess that are being executed at the same time, energy consumption, size of the system, mobility, and the visual interface. Conversely, an embedded system provides more efficient features than general purpose software. To confirm this idea, in this paper VGA display is used as the output reference to compare the results applying a set of raster data as portable pixmap (*.ppm), graymap (*.pgm) and bitmap (*.pbm). Two tests in a different field of study are comparing: pre-processing of an image format with four operations: original, grayscale, binary and inverted; the other test is a laser triangulation measurement system. In the tests: RPLIDAR A1M1-R1 Development Kit, ImageJ, GIMP and a Spartan 3E FPGA hardware 12 bits RGB output image was used as a reference at 640x480 pixels in a conventional computer monitor. The method proposed as an image processing was compared with a conventional computer, and the results in the visualization were similar in both cases, but with less energy consumption, less size and capacity for mobile systems.**

***Index Terms*—FPGA, Imaging processing, Lidar, Netpbm format, VGA.**

I. Introduction

IMAGE processing is a part of signal processing that uses some segmentation that researchers are using in many fields, such as measuring [1] [2] laser scaning [42] [43] [44] [45], food [3], surgery[4], corrosion [5], industrial [6] [7], particles [8] and others. A general framework image processing according to [9] [10] [11] is:

- Image acquisition
- Pre-processing
- Segmentation
- Representation
- Classification

A. Image acquisition and pre-processing

There are two ways to obtain image data either cases (Electronic device or software) the result is a digital image storage in an array of bits, within a memory using a particular format file. In a review of various applications aimed at image processing, the characteristics of an image were found in different disciplines. Specifically an overview of the major file formats currently used in medical imaging, define universal concepts to all file formats such as pixel depth, photometric interpretation, metadata and pixel data [12]. A particular software package for image processing of electron, micrographs, interpretation of reconstructions, molecular modeling and general image processing generate a text file [13].

Some image file format provided by GIMP software are: Animation .flic, Animation .mng, PostScript .ps, Icon .ico, Digital Imaging and Comunications in Medicine .dcm .dicom, BMP Image .bmp, Photoshop .psd, Encapsulated PostScript .eps, GIF .gif, IRIS de Silicon Ghraphics .sgi, JPEG .jpg, PBM .pbm, PGM .pgm, PIX .pix, PNG .png, PNM .pnm, PPM .ppm, SUN .im1, im8 .im24 im32, TarGa .tga, TIFF .tif, X BitMap .xbm,X pixMao .xpm, Zsoft PCX .pcx, KISS CEL .cel, OpenRaster .ora, GIMP. pat, PDF .pdf and Flexible image .fit. For example JPEG 2000 standard (Joint Photographic Experts Group) file format is used widely on the internet, color facsimile, printing, scanning, digital photography, remote sensing, mobile, and others. It is processed with the block tiles to produce a JPEG file [14] such as occurs with BMP, PNG, TIFF, among others. All formats have implicit features as image nature, resolution, number of colors [15] [16] even a Holographic Data System applies a similar storage [17].

This paper focused on three file formats as mentioned above: PPM, PGM and PBM [18] [19] [20] in order to share the data with other devices. The Netpbm is a toolkit for the manipulation of graphic images including conversion of images from a variety of different formats. Also it is portable to Unix-based systems, Windows, Mac OS X, VMS and Amiga OS. Netpbm was developed to be a single source for all the primitive graphics utilities [21] and in this paper on hardware applications.

B. LiDAR

LiDAR is a distance sensor [42] that allows showing environmental visual information through a grid map or point-cloud. Normally it is mounted on mobile systems such as vehicles [42] [44], UGV [41], coordinate motion [45] and static environment [43]. An RPLIDAR A1M1-R1 module tested with the SDK was connected to USB from a conventional computer and a set of points over a radar background is shown on a screen as in Figure 6. The raw data sent by the LiDAR is an array of values representing a distance and orientation in digital bits. The hardware implementation proposed avoids connecting the device to a computer, but allows drawing the color of each pixel as an image pre-processing.

Manuscript received Month DD, YYYY; revised Month DD, YYYY.

We would like to thank the Cátedras CONACYT for the financial support of this work

J. Sandoval-Gutierrez is with the Universidad Autónoma Metropolitana at Lerma, J.A. Alvarez-Cedillo, J.C. Herrera-Lozada and M. Olguin-Carbajal are with the Centro de Innovación y Desarrollo Tecnológico en Cómputo (CIDETEC), Instituto Politécnico Nacional (IPN), Juan de Dios Bátiz s/n, C.P. 07700 D.F., México (e-mail: jacobosandoval@hotmail.com; jaalvarez,jlozada@ipn.mx) T. Alvarez-Sanchez is with the Centro de Investigación y Desarrollo de Tecnología Digital(CITEDI), Instituto Politécnico Nacional (IPN), Av. del Parque No. 1310, Mesa de Otay, Tijuana, Baja California, México

https://doi.org/10.17562/PB-56-6

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

(a) PGM Image

```
P2
# test.pgm
19 7
15
0 0  0  0  0  0 0  0  0  0  0  0 0 0 0 0 0 0  0
0 15 15 15 15 0 0  11 11 11 11 0 0 5 5 0 5 5  0
0 15 0  0  15 0 0  11 0  0  0  0 0 5 0 5 0 5  0
0 15 15 15 15 0 0  11 0  11 11 0 0 5 0 5 0 5  0
0 15 0  0  0  0 0  11 0  0  11 0 0 5 0 5 0 5  0
0 15 0  0  0  0 0  11 11 11 11 0 0 5 0 5 0 5  0
0 0  0  0  0  0 0  0  0  0  0  0 0 0 0 0 0 0  0
```

(b) PGM code

Fig. 1: PGM image representation

C. Visualization

After the digital image has been stored in any electronic device by any format, the image is displayed using a screen electronic device without specific software [6] [7] [28], MATLAB® [1] [2] [3] [5] [8] [43] [25] [26] , JAVA ® [1], C language [26] [42] [44], ROS [43], Qt SDK [41] and also alternative methods for MPI CUDA in HPC [27].

This process is a common task and known as visualization. However, when a device with embedded screen (display size, resolution, and color) shows the raw data in real time, there is no possibility of knowing if the file format shared is the same as the original. Since the process is a Black-box for the users. An alternative solution is a visual direct manipulation as a software [22]. In this paper, an FPGA-based implementation is shown using a VGA display.

D. Netpbm kernel

PPM is a raw ASCII image format and is a suitable string representation of an image in a file. Each pixel contains ASCII information in an arbitrary size. In the first line, a P3 tag(color file format) is used, in the second line the columns and rows number must be added, in the third line an RGB maximum number value, and in the other lines the rest of the data.

Another format is PBM, where each pixel is represented with 0 or 1 (black and white); white space in the raster section is ignored and the heading in the first line is P1 instead of P3 used by PPM [29].

PGM is a format consisting of four lines, providing a maximum of 256 gray scale levels or 8 bit data per pixel [30] [32].

A sample code of PGM file is shown with a P2 indicating a gray level from 0 up to 15 values, 19 columns, seven rows and ASCII information of one character is equal to one pixel. In Figure 1 the result of this code is shown (test.pgm), and the file was generated by ImageJ and GIMP Software.

E. VGA Display

The general considerations for VGA display controller have been referenced by development in Verilog Hardware Description Language [33] and VHDL [34] [36] [37].

In Table I VGA signal 640 x 480 @ 60 Hz Industry standard timing is shown.

TABLE I: Timing

Horizontal timing (line)

Scanline part	Pixels	Time (μs)
Visible area	640	25.422045680238
Front porch	16	0.63555114200596
Sync pulse	96	3.8133068520357
Back porch	48	1.9066534260179
Whole line	800	31.777557100298

Vertical timing (frame)

Frame part	Lines	Time (μs)
Visible area	480	15.253227408143
Front porch	10	0.31777557100298
Sync pulse	2	0.063555114200596
Back porch	33	1.0486593843098
Whole frame	525	16.683217477656

The Spartan-3A FPGA Starter Kit board includes a VGA display port via a DB15 connector with a red, green, and blue signal. VGA display port provides 4-bit RED, 4-bit GREEN, 4-bit BLUE, (444 color), or 4,096 possible colors. In (1) the color output is described.

$$\text{color}_{\text{out}} = \frac{\text{vga}[3:0]}{15} \times \text{color} \tag{1}$$

II. Test design

This section specifies the characteristics utilized to produce a VGA output on FPGA. The first step is to read a file with the raw data saved in the RAM memory block. The RAM memory has three parameters to set: the address vector (depth), width vector (value) and writing in an enable signal.

Read after write is used to compute three functions through a processing module: inverted function (2) gray (3) and binary (4).

$$\text{Inverted}_{\text{out}} = 2^{\text{ vga}[3:0]} - \ \text{color}_{\text{out}} \tag{2}$$

$$\text{Gray}_{\text{out}} = \frac{\sum\limits_{RGB} (\text{color}_{\text{out}})}{3} \tag{3}$$

$$\text{Binary}_{\text{out}} \begin{cases} 1 & \text{if } (\frac{2^{\text{ vga}[3:0]}}{2} > \text{color}_{\text{out}}) \\ 0 & \text{else} \end{cases} \tag{4}$$

This module reads all the addresses of the raw data, computing (4), (3) and (2), and writing in a new RAM memory section.

The VGA controller with 25 MHZ clock (clk2) reads all the memory block and creates a synchronization with the data and the addresses. The data must be written in the vector before the horizontal and vertical requires it.

Finally, VGA output port receives four images of VGA controller and shows the results on the computer screen. A diagram of the design is shown in Figure 2.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

A. Reading a COE file

A memory coefficient (COE file) loaded in the initialization with a single port A, 12-bit width, (25 600 deep RAM). The syntax is:
memory_initialization_radix = 16;
memory_initialization_vector = 100, 200, . . . 100, 200;

B. VGA controller

A set of six signals selected is: address, data, clock 2, synchrony, horizontal and vertical value for VGA controller. The process begins when the horizontal value is greater than 144 and less than 784, and the vertical value is greater than 31 and less than 511 while another parallel process reads the data with its corresponding address. The relation (5) describes its values by section with two clocks (clkdiv = 25MHz and clk = 50 Mhz).

$$[\mathrm{h}][\mathrm{v}]_{\mathrm{value}} \times \mathrm{clkdiv} = (\mathrm{data}_{\mathrm{address}} \times \mathrm{clk}) \times \mathrm{sync} \qquad (5)$$

C. Processing module

The data of RAM memory is divided into four sections. The processing module reads and computes (4), (3) and (2) address by address in a parallel process to the VGA controller. The algorithm uses a single instruction multiple data streams (SIMD) [38] [39].

D. VGA output port

The Spartan® 3A FPGA Starter Kit board, includes an HD-DB15 female connector with the horizontal sync signal (row), the vertical sync signal (column); these two continuously running counters from the address into a video display buffer (RGB Values) [40].

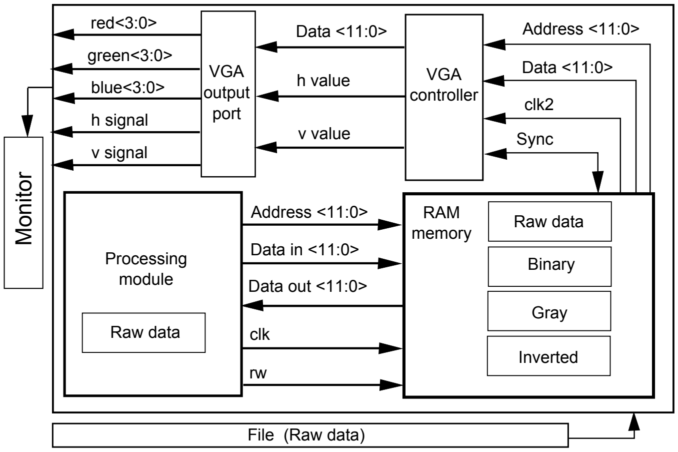
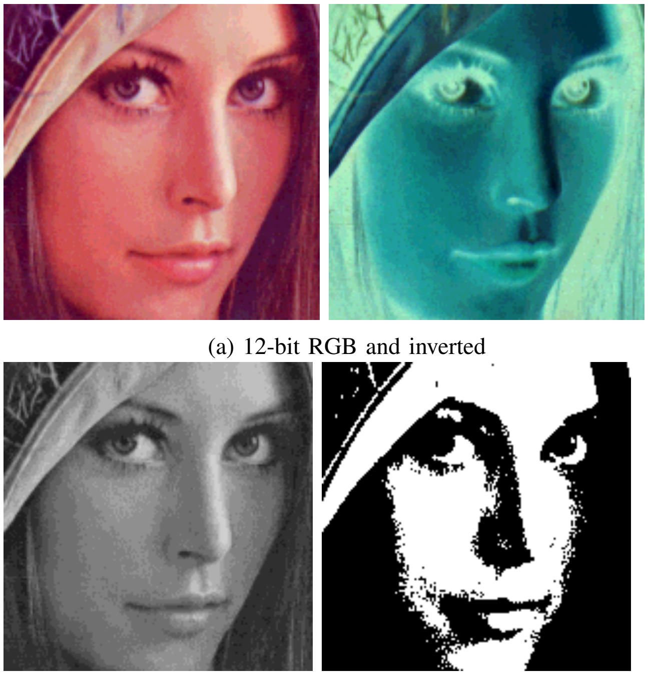


Fig. 2: Test design of FPGA implementation

III. Testing

A. Image Proccesing Software

A Lena image has been tested and the result of ImageJ software is shown in Figure 3a 12-bit RGB (1), 3a inverted (2), 3b gray (3) and 3b binary (4).



(a) 12-bit RGB and inverted

(b) Gray and bin

Fig. 3: Lena image processing using ImageJ software

B. Image LiDAR

The original image has a set of 360 distances and angles that are display in the demo application developed by the SDK of the manufacturer. The on-line data is shown in Figure 4, but it is not clearly visible to the human eye.

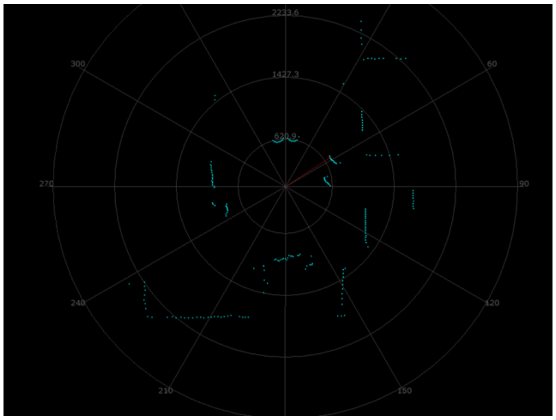


Fig. 4: LiDAR image by SDK's manufacturer

The final tests are shown in Figures 5 and 6, the outputs by software is shown on the left and the hardware output is shown on the right.

IV. Analysis and Discussion

In the first test a comparative table in II was filled out with five features: hardware type, software or file format, energy consumption, mobility and finally the system size. The four cases were included in the proposed design having an FPGA implementation with Mif file [33] and Hexadecimal file [35]. It

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

means that an embedded application requires a more suitable format in order to be manipulated, but there is no problem with the format file in PC-based processing. The computer uses 1000 % more energy than FPGAs implementation, and this wasted energy avoids reducing the size of the system and consistently only static applications could be developed.

TABLE II: First test comparative features with the proposed design

Application	Hardware	Software / File Format	Watts	Mobility	Size
Proposed design	FPGA		+5 W	Yes	Small
Image [33]	FPGA	mif file	+5 W	Yes	Small
Image [34]	FPGA	—	+5 W	Yes	Small
Image [35]	FPGA	Hex File	+5 W	Yes	Small
Image [3]	PC	Matlab	+65 W	No	Normal
Image [4]	PC	Matlab	+65 W	No	Normal
Image [5]	PC	Matlab	+65 W	No	Normal
Image [6]	PC	SDK	+65 W	No	Normal
Image [8]	PC	MatLab	+65 W	No	Normal
Image [13]	PC	Bsoft	+65 W	No	Normal
Image [22]	PC	Palimpses	+65 W	No	Normal
Image [43]	PC	Matlab	+65 W	No	Normal
Image [24]	PC	Java	+65 W	No	Normal
Image [25]	PC	Matlab	+65 W	No	Normal
Image [26]	PC	Matlab	+65 W	No	Normal
Image [30]	PC	Java	+65 W	No	Normal
Image [31]	PC	Matlab	+65 W	No	Normal

In the second test a comparative Table in III as in TableII was compared. The most similar applications are developing with a low consumption energy technology [41] where the software provides sufficent resources. The other applications are a typical PC-based SDK with high wasted energy and normal size that is not efficient for a mobile robot applications.

TABLE III: Second test comparative features with the proposed design

Application	Hardware	Software	Watts	Mobility	Size
Proposed design	FPGA		+5 W	Yes	Small
Robot [41]	Intel Atom	Qt SDK	+9 W	Yes	Small
Measure [1]	Pc	Kinect	+65 W	No	Normal
Measure [2]	PC	MatLab	+65 W	No	Normal
Measure [42]	PC i7	Open CV	+95 W	No	Normal
Measure [43]	PC	ROS	+60 W	No	Normal
Measure [44]	PC i5	C++	+73 W	No	Normal
Measure [45]	Pc core2	unknown	+65 W	No	Normal

V. Conclusion

Raster data as Netpbm is a compatible file format that could be implemented in embedded systems such as the FPGA proposed design and other similar cited papers. The compressed algorithm used by JPGE, PGN and others is not a suitable format for the hardware applications. While a 444 RGB and 160 x 160 pixels *.jpg and *.ppm file use a variable size from 10 KB up to 346 KB in the hard disk. The memory in the FPGA uses a fixed size of 38.4 KB. The most common applications in image processing are developed using an SDK tool on the computer, but the problem is that the energy consumption is more than a 100 times the embedded application. The design proposed has capacity to be implemented in a mobile robot platform, because it satisfies three necessary conditions. A small size, less than $2{\times}10^{-3}m^3$, low consumption around $5W$ (consequently avoids a cooler system) and the electronic supports vibration. Only a smart computer has similar characteristics, but this requires an OS sharing the resources and a heat sink that avoids damaging the components. Figure 5 and 6 compare the final VGA output with their counterpart (personal computer). Another characteristic is that it only takes a few seconds to boot the embedded application; conversely the PC lost time booting the OS.

Appendix A
PPM, PGM and PBM file format tested

PPM file.

```
P3
160 160
5 1 4 . . .
```

PGM file.

```
P2
160 160
16
3 3 3 . . .
```

PBM file.

```
P1
160 160
1 1 1 . . .
```

Appendix B
Test code for 8-bit VGA image and RAM black out

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;
use IEEE.NUMERIC_STD.ALL;
use IEEE.STD_LOGIC_ARITH.ALL;

entity vga is
 port(
 sw        : IN STD_LOGIC_VECTOR(1 downto 0);
 Led       : INOUT STD_LOGIC_VECTOR(0 DOWNTO 0);
 wea1      : INOUT STD_LOGIC_VECTOR(0 DOWNTO 0);
 addra1    : INOUT STD_LOGIC_VECTOR (14 downto 0);
 dina1     : INOUT STD_LOGIC_VECTOR(7 DOWNTO 0);
 douta1    : INOUT STD_LOGIC_VECTOR(7 DOWNTO 0);
 clk       : IN STD_LOGIC;
 red_out   : OUT STD_LOGIC_VECTOR(2 downto 0) ;
 green_out: OUT STD_LOGIC_VECTOR(2 downto 0) ;
 blue_out : OUT STD_LOGIC_VECTOR(1 downto 0) ;
 hs_out    : OUT STD_LOGIC;
 vs_out    : OUT STD_LOGIC_VECTOR
 );
end vga;

architecture Behavioral of vga is

COMPONENT ram
  PORT (
    clka  : IN STD_LOGIC;
    wea   : INOUT STD_LOGIC_VECTOR(0 DOWNTO 0);
    addra : IN STD_LOGIC_VECTOR(14 DOWNTO 0);
    dina  : IN STD_LOGIC_VECTOR(7 DOWNTO 0);
    douta : OUT STD_LOGIC_VECTOR(7 DOWNTO 0)
```

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

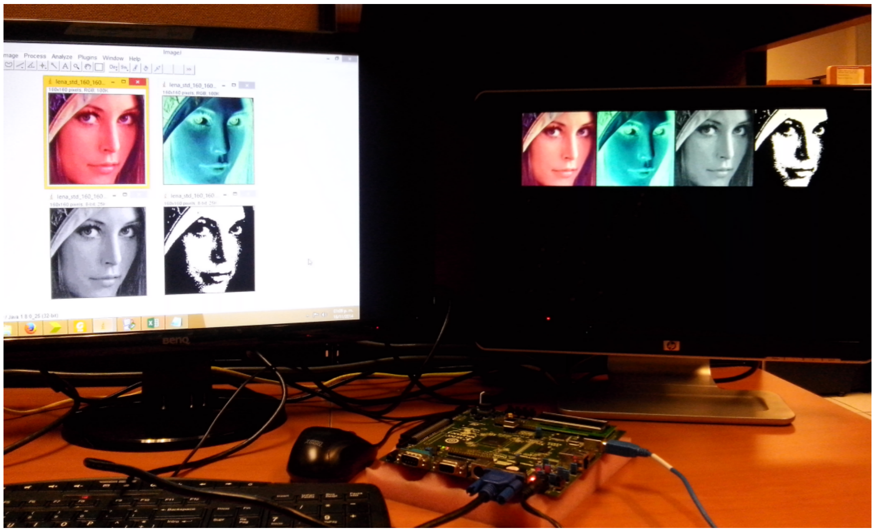


Fig. 5: Image processing ImageJ software (left monitor) and FPGA (right monitor)

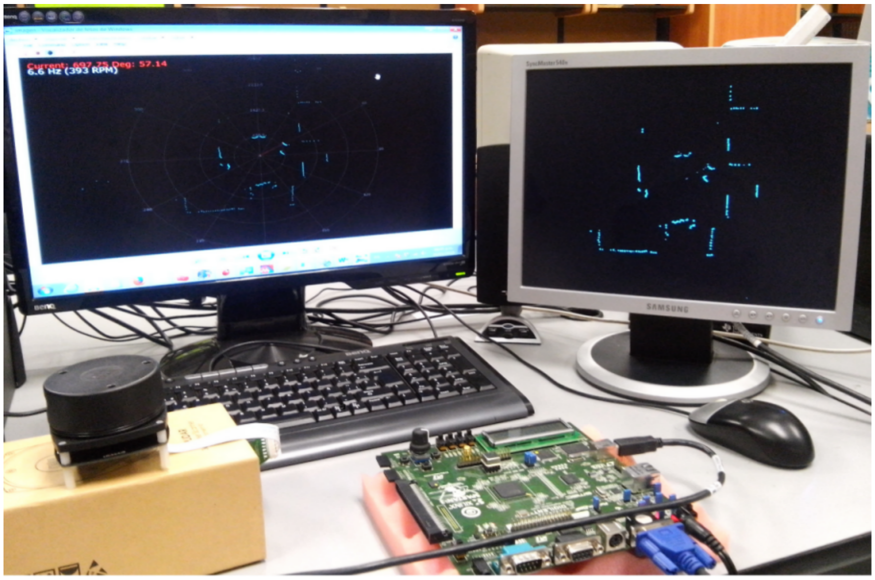


Fig. 6: LiDAR SDK (left monitor) and FPGA (right monitor)

```
  );
END COMPONENT;

 signal clkdiv   : std_logic := '0';
 signal clkdiv2  : std_logic := '0';
 constant hsyn   : integer :=  800;
 constant vsyn   : integer :=  521;
 constant pwh    : integer :=  96;
 constant pwv    : integer :=  2;
 constant bph    : integer :=  48;
 constant fph    : integer :=  16;
 constant bpv    : integer :=  29;
 constant fpv    : integer :=  10;
 constant x0     : integer :=  320;
 constant y0     : integer :=  240;
 signal hc       : integer range 0 to 1024;
 signal vc       : integer range 0 to 1024;
 signal hc0      : integer range 0 to 1024;
 signal vc0      : integer range 0 to 1024;
 signal hvc0     : integer range 0 to 32768;

begin
```

```
process (clk)
begin
        if clk' event and clk = '1' then
                clkdiv <= not clkdiv;
        end if;
end process;

process (clkdiv)

begin

 if clkdiv' event and clkdiv = '1' then
hc <= hc + 1;
if (hc = hsyn) then
vc <= vc + 1;
    hc <= 0;
    end if;

    if (vc = vsyn) then
    vc <= 0;
    end if;

          if (hc > pwh )
```

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

```
        then            hs_out <= ’1’;

if (vc > pwv )
      then vs_out <= ’1’;
    else vs_out <= ’0’;
end if;

If (hc > (bph+pwh) ) and (hc < (hsyn - fph) )
and (vc >= (pwv+bpv)) and (vc < (vsyn - fpv))

        then
 if (hc < (bph+pwh+180) ) and (vc < (pwv+bpv)+180)
 then

        hc0 <= hc-144;
        vc0 <= vc-31;
        hvc0 <= vc0*180+hc0;
        addra1 <= conv_std_logic_vector(hvc0,15);
        red_out   <= douta1(7 downto 5);
        green_out <= douta1(4 downto 2);
        blue_out <= douta1(1 downto 0);
        else     red_out   <= "111";
        green_out <= "000";
        blue_out  <= "11";
 end if;
 else
                 red_out   <= "000";
                green_out <= "000";
            blue_out  <= "00";
     end if;
 end if;

end process;

your_instance_name : ram
  PORT MAP (
    clka => clk,
    wea => wea1,
    addra => addra1,
    dina => dina1,
    douta => douta1
  );

with sw select

 wea1<= "1" when "11,
       "0" when others;

led <= wea1;

end Behavioral;
```

References

[1] Omar Rodríguez Zalapa, Antonio Hernández Zavala y Jorge Adalberto Huerta Ruelas. Sistema de medición de distancia mediante imágenes para determinar la posición de una esfera utilizando el sensor Kinect XBOX, Revista Polibits, Vol. 49, 2014, pp. 59–67.

[2] Hofer D. and Zagar B.G., Image processing for calibrating a coordinate measurement set-up, Measurement Science and Technology, Vol. 25, No. 11, 2014, pp. 115003-115017

[3] Hosseinpour Soleiman, Rafiee Shahin, Aghbashlo Mortaza and Mohtasebi Seyed Saeid, A novel image processing approach for in-line monitoring of visual texture during shrimp drying, JOURNAL OF FOOD ENGINEERING, Vol. 143, 2014, pp. 154-166.

[4] Lee Sang Hee, Lee Minho and Kim Hee Jin, Anatomy-based image processing analysis of the running pattern of the perioral artery for minimally invasive surgery BRITISH JOURNAL OF ORAL & MAXILLOFACIAL SURGERY, Vol. 52, No. 8, 2014, pp. 688-692.

[5] Gamarra Acosta, Margarita R., Velez Diaz Juan C., Schettini Castro Norelli, An innovative image-processing model for rust detection using Perlin Noise to simulate oxide textures, CORROSION SCIENCE, Vol. 88, 2014, pp. 141-151.

[6] Deyong You, Xiangdong Gao and Katayama, S. Monitoring of high-power laser welding using high-speed photographing and image processing, Mechanical Systems and Signal Processing, Vol. 49, No. 1, 2014, pp. 39-52.

[7] Lopez F., Maldague X., and Ibarra-Castanedo, Enhanced image processing for infrared non-destructive testing, OPTO-ELECTRONICS REVIEW, Vol. 22. No. 4, 2014, pp. 245-251.

[8] Charonko John J, Antoine Elizabeth and Vlachos Pavlos P., Multispectral processing for color particle image velocimetry, MICROFLUIDICS AND NANOFLUIDICS, Vol. 17, No. 4, 2014, pp. 729-743.

[9] Russ John C., The Image Processing Handbook, Sixth Edition, CRC Press 2011.

[10] Pinoli Jean-Charles, Mathematical Foundations of Image Processing and Analysis 1, John Wiley & Sons, Inc. 2014

[11] Bernd Jähne, Practical Handbook on Image Processing for Scientific and Technical Applications, Second Edition CRC Press 2004.

[12] Larobina Michele and Murino Loredana, Medical Image File Formats, JOURNAL OF DIGITAL IMAGING, Vol. 27, No. 2, 2014, 200-206

[13] Heymann J. Bernard and Belnap David M., Bsoft: Image processing and molecular modeling for electron microscopy, JOURNAL OF STRUCTURAL BIOLOGY, Vol. 157 No. 1, 2007, pp. 3-18.

[14] Skodras A, Christopoulos C. and Ebrahimi, T, The JPEG 2000 still image compression standard IEEE SIGNAL PROCESSING MAGAZINE, Vol. 18, No. 5, 2001, pp. 36-58.

[15] Wiggins RH, Davidson HC, Harnsberger HR, Lauman JR and Goede PA, Image file formats: Past, present, and future RADIOGRAPHICS, Vol. 21, No. 3, 2001, pp. 789-798.

[16] Lins RD and Machado DSA, Comparative study of file formats for image storage and transmission, JOURNAL OF ELECTRONIC IMAGING, Vol. 13, No. 1, 2004, pp. 175-181.

[17] Kim Do-Hyung, Jeon Sungbin, Park No-Cheol and Park, Kyoung-Su, Iterative design method for an image filter to improve the bit error rate in holographic data storage systems, MICROSYSTEM TECHNOLOGIES-MICRO-AND NANOSYSTEMS-INFORMATION STORAGE AND PROCESSING SYSTEMS, Vol. 28, No. 8-9,2014, pp. 1661-1669.

[18] Nadal J, Keeping the bits in place: A case study of raster image migration, SOC IMAGING SCI & TECHNOL, Final Program and Proceedings, 2005, pp. 249-252

[19] ZAMA C M S, System for converting word file into other format e.g. JPEG file format using e.g. HTML software, has CPU to convert individual characters from scanned image into comprehensible code in other format using optical character recognition, patent: ZA200803391-A.

[20] DARGELAS A M, Waveform image e.g. bitmap file, generating method, involves providing localized rendering of temporally organized waveform data based on user input and waveform viewer resolution, and loading waveform images into waveform viewer. patent: US2011234600-A1.

[21] Home page for Netpbm: http://netpbm.sourceforge.net/

[22] Blackwell Alan F., Palimpsest: A layered language for exploratory image processing, JOURNAL OF VISUAL LANGUAGES AND COMPUTING, Vol. 2, No. 5, 2014, 545-571.

[23] Wang Z. and Bovik AC. A universal image quality index, IEEE SIGNAL PROCESSING LETTERS, Vol. 9, No. 3, 2002, pp. 81-84.

[24] Pavel Surynek and Ivana Lukšová, Automated Classification of Bitmap Images using Decision Trees, Revista Polibits, Vol. 44, 2011, pp. 11–18.

[25] Minh N. Do and Martin Vetterli, The contourlet transform: An efficient directional multiresolution image representation, IEEE TRANSACTIONS ON IMAGE PROCESSING,Vol. 14, No. 12, 2005, 2091-2106.

[26] Manjunath BS and Ma WY, Texture features for browsing and retrieval of image data IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 18,Issue. 8, 1996, pp. 837-842.

[27] Galizia, Antonella, D’Agostino, Daniele and Clematis, A Clematis, Andrea, An MPI-CUDA library for image processing on HPC architectures, JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS, Vol. 273, 2015, pp. 414-427.

[28] Ishigami Yuta, Waskitoaji Wihatmoko, Yoneda Masakazu, Takada Kenji, Hyakutake Tsuyoshi, Suga Takeo, Uchida Makoto, Nagumo Yuzo, Inukai Junji and Nishide Hiroyuki, Oxygen partial pressures on gas-diffusion layer surface and gas-flow channel wall in polymer electrolyte fuel cell during power generation studied by visualization technique combined with numerical simulation, JOURNAL OF POWER SOURCES, Vol. 269, 2014, pp. 556-564.

[29] Pakhira M.K. and Dutta A., Computing approximate value of the PBM index for counting number of clusters using genetic algorithm, 2011 International Conference on Recent Trends in Information Systems (ReTIS), 2011,mpp. 241-5

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

[30] Shiva Shankar R., Mnssvkr Gupta V, Murthy K.V.S. and Someswararao C., Object Oriented Fuzzy Filter for Noise Reduction of PGM Images, Proceedings of the 2012 8th International Conference on Information Science and Digital Content Technology (ICIS and IDCTA), Vol. 3, 2012, pp. 776-82.

[31] Philippot E., Belaid A. and Belaid Y., Use of PGM for Form Recognition, Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS 2012), pp. 374-378

[32] Abdul-Jabbar I.A.-A, Jieqang Tan and Zhengfeng Hou, Face Recognition Enhancement Based on Image File Formats and Wavelet De-noising, International MultiConference of Computer Scientists (IMEC 2014). Proceedings, Vol. 1, pp. 441-445

[33] Radi H.R., Caleb W. W. K., M.N.Shah Zainudin and M.Muzafar Ismail, The Design and Implementation of VGA Controller on FPGA International Journal of Electrical & Computer Sciences, IJENS Vol. 12, No. 05, 2012, pp. 56-60.

[34] Ashish B. Pasaya and Kiritkumar R. Bhatt, Implementing VGA Application on FPGA using an Innovative Algorithm with the help of NIOS-II, International Journal Of Computational Engineering, Vol. 2, No.3, 771-775.

[35] Guohui Wang, Yong Guan and Yan Zhang, Designing of VGA Character String Display Module Base on FPGA, 2009 International Symposium on Intelligent Ubiquitous Computing and Education, IEEE, 2009, pp. 499-502.

[36] Ioan, A.D., Designing an optimal single chip FPGA video interface for embedded systems, Electrical and Electronics Engineering (ISEEE), 2010 3rd International Symposium on, pp. 58-63.

[37] Van-Huan Tran and Xuan-Tu Tran, An efficient architecture design for VGA monitor controller, Consumer Electronics, Communications and Networks (CECNet), 2011 International Conference on, pp. 3917-3921.

[38] Elliott D.G., Stumm M., Snelgrove W.M., Cojocaru C. and McKenzie R., Computational RAM: Implementing processors in memory IEEE DESIGN & TEST OF COMPUTERS, Vol. 16, No. 1, 1999, pp. 32-41

[39] Tessier Russell, Betz Vaughn, Neto David, Egier Aaron and Gopalsamy Thiagaraja, Power-efficient RAM mapping algorithms for FPGA embedded memory blocks, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, Vol. 26, No. 2, 2007, pp.278-290.

[40] Spartan-3A/3AN FPGA Starter Kit Board User Guide 11 UG334 (v1.1) June 19, 2008.

[41] T. Jian, C. Yuwei, A. Jaakkola, L. Jinbing, J. Hyyppa, and H. Hyyppa, NAVIS-An UGV Indoor Positioning System Using Laser Scan Matching for Large-Area Real-Time Applications, Sensors, vol. 14, no. 7, pp. 11805-11824, July, 2014.

[42] Y. Li, and Y. Ruichek, Occupancy Grid Mapping in Urban Environments from a Moving On-Board Stereo-Vision System, Sensors, vol. 14, no. 6, pp. 10454-10478, Jun, 2014.

[43] R. Wang, X. Li, and S. Wang, A laser scanning data acquisition and display system based on ROS, Proceedings of the 33rd Chinese Control Conference, pp. 8433-8437, 2014.

[44] Y. Yongtao, J. Li, G. Haiyan, W. Cheng, and Y. Jun, Semiautomated Extraction of Street Light Poles from Mobile LiDAR Point-Clouds, IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 3, pp. 1374-1386, March, 2015.

[45] H. Yuqing, and M. Yuangang, An efficient registration algorithm based on spin image for LiDAR 3D point cloud models, Neurocomputing, vol. 151, pp. 354-363, 3, 2015.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

A Bootstrapping approach for Entity Linking from Biomedical Literature

ISSN 2395-8618

Kanimozhi U
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
kanimozhiu.03@gmail.com

Manjula D
Department of Computer Science and Engineering
College of Engineering Guindy, Anna University
Chennai, India
manju@annauniv.edu

***Abstract* - Entity Linking (EL) is a task of aligning literal of a named-entity from an unstructured document to appropriate entities in a knowledge base. The main objective of EL in biomedical domain stems on construction of efficient computational models. This paper proposes a bootstrap approach based on uniformity perception and similarity computation to link entities from unstructured biomedical texts to ontologies. A rich semantic information and structures in ontologies are influenced by the proposed approach for similarity computation and entity ranking. The proposed approach address the Entity Linking in biomedical domain. The experiments show that our approach outperforms the existing state-of-the-art algorithms in terms of linkage accuracy.**

Keywords - Entity Linking, Bootstrapping, Biomedical Literature.

I. BACKGROUND

Over the past years, there is an emergence of enormous amount of unexplained abbreviations and terminologies that leverages a major bottleneck in understanding scientific literature. Mining and linking significant facts/information from biomedical literature have great impact on knowledge discovery in biomedical domain. It is also very challenging even for domain experts to keep up with the large number of articles published [2]. For instance, supporting the modeling task by means of identifying the key proteins, and their behaviors and interactions. Hence, there is a need for advancements in methodologies for making sense of large amount of unstructured textual data which is explosively increasing. In order to facilitate it, specific way of analysis can be enabled, where phrases comprising of a distinct term or sequence of terms are automatically linked to entries in a knowledge base.

Here, the focus is on the task of Entity Linking (EL) from biomedical literature. Entity linking is a process that links different entities that refer to the same source of data. Such entities exists in many other fields such as semantic web, multimedia, personal profiling, publication, geography, etc. Our main aim is to automatically identify the prominent entity mentions from unstructured texts and linking them to terms described in a Knowledge Base (KB) and define in an ontology in order to enrich the text documents. These knowledge base and ontology terms are also referred to as reference entities. EL can helps human end user navigate biomedical literature and improve many other Natural Language Processing (NLP) tasks such as gene-disease association, gene-gene and protein-protein interaction event extraction [3, 4]. Entities enable semantic exploration of biomedical mentions, numerous information prerequisites can be encountered by recurring a list of entities, their properties, and/or their relations. Those entities can be utilized to identify unforeseen relations or functions and link the gap between unstructured and structured data.

Some recent works have been done on improving linkage performance using Machine Learning techniques [5,6]. However, it is laborious and effortful in building large-scale high-quality training set. Hence we propose a bootstrapping approach for entity linkage by utilizing semantics-based and similarity-based methods. For a given entity, our approach initially infers a set of semantically co-referent entities and then, iteratively expands this entity set using distinct classes. In order to improve the performance of the classifier bootstrapping [7] technique is used which is suitable for entity linkage due to the abundant uncertain entities. A publicly accessible ontologies in biomedical domain known as BioPortal [8] is utilized here. These ontologies consist of rich structures with declaratively defined semantic relations, along with comprehensive text descriptions provided by domain experts. We assume that multiple entities are semantically related in unstructured texts (i.e., they co-occur in the same sentence, are linked through dependency paths, or play certain semantic roles in the same event, etc.). Thereby address entity linking by means of uniformity perception by leveraging the global topical coherence and linking a set of relevant mentions simultaneously and generated labeled EL data through bootstrap approach.

Generally, there are two categories of EL algorithms namely collective inference and non-collective methods respectively [1]. Collective inference approaches influences concept mentions through supervised or graph based re-ranking methods. Besides they discourse the linking problem through exploiting the agreement between the mention document's text and the context of the entities of the knowledge base. Graph based re-ranking models typically collects linking agreement information from training data and propagates to other nodes. Non-collective methods usually rely on prior knowledge and context similarity with supervised models. Ranking scores for each concept mentions are computed individually. Whereas, both these approaches requires large amount of manually labelled entity mentions in order to achieve a reasonable linking accuracy.

https://doi.org/10.17562/PB-56-7

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

This paper presents a study on identifying prominent links between entities and label gene/protein-disease relations in PubMed and MEDLINE abstracts by using bootstrapping approach. Beginning with PubMed and MEDLINE abstracts, we first recognized gene/protein and disease entities using existing Natural Language Processing (NLP) tools such as Regex NER. Then we extract candidate gene/protein-disease pairs by mapping it with the existing ontology and knowledge base based on different levels of co-occurrence such as, abstract level, sentence level, phrase level and paragraph level. In order to find the most linked gene/protein-disease relations, we finally rank candidate gene/protein-disease pairs using Information Gain (IG). The evaluation using a manually annotated data set from Gene Ontology (GO) indicated that the Bootstrap method outperformed others. To our best knowledge, this is the first attempt that applied bootstrap approach to rank gene/protein-disease entity linking from biomedical literature.

II. METHODOLOGY

A. PROPOSED FRAMEWORK

A bootstrapping approach is proposed, in which it accepts the candidate entity as an input. The following section describes the methodological steps for our approach depicted in Fig. 1. The major goal of the paper is to link the identified entities to the concepts in the knowledge base. For a given biomedical text document as input, we extract the entity mentions and automatically construct a kernel, consisting of semantically co-referent entities, via uniformity perception on several gene/protein/disease vocabularies from ontologies. Then the kernel is expanded iteratively using distinct classes to probe different co-referent entities. The distinguishability of each class is learned with a statistical measure, reveling the importance of the class characterizing the co-referent entities and match the class by comparing the functions of those entities. Furthermore, frequent class combinations (i.e., the functions often used together) are mined to enhance entity labeling criterion in bootstrapping, so that the linkage accuracy can be improved. Entity mapping for an entity mention is assigned by measuring the popularity of an entity among all other candidate entities.

i. Input Documents

The development corpus is a subset of PubMed and Medline abstracts dealing with Huntington Disease and its genes. It was annotated with disease and gene relations, based on “etiology” and “clinical biomarker”. Beginning with PubMed and MEDLINE abstract collection. The initial step is the pre-processing which is done to determine entity boundaries in a text by sentence splitting and tokenization. Natural language processing incurs creation of a set of patterns to match the possible linguistic realizations of the individual facts. Due to this complexity, the preprocessing on structural input requires assigning parts-of-speech and features to words and idiomatic phrases. Annotated corpus drive construction of training data for machine learning that will filter out false positives from the dictionary-based results. These data are used for training and testing purposes. The input corpus consists of text related to Huntington Disease, gene names with their functions and all words related to neurogenetic disorders. The input corpus which is manually curated has 8998 sentences and 140481 words. Let, entity mentions $u \in \mathrm{E}$ are prominent phrases in the input biomedical text document. All classes, properties and individuals described in the ontologies $r \in R$ are considered to be the reference entities. Relations based on sentence level and paragraph level are extracted based on co-occurrence are extracted. A list of candidate entities X are located from the biomedical dictionaries for each entity in the Context Graph CG. Then we compute the imporatance score and link them by the bootstrap approach. Finally, we compute similarity scores for each entity/candidate pairs $< u, x >$ and select the candidate with the highest score as the appropriate entity for linking.

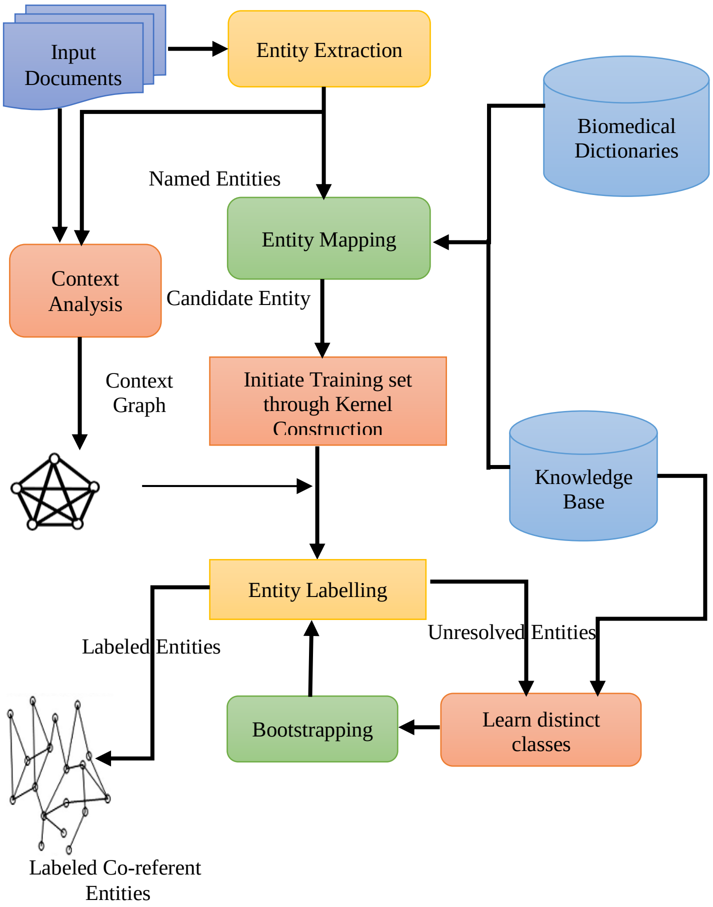


Figure 1 Work flow of the Bootstrap Entity Linkage

ii. Entity Extraction

We apply the publicly available natural language processing tools for identifying prominent biomedical entities from unstructured texts to recognize the entity and ascribe it to a class or entity type. The occurrences of gene/disease entities in a text automatically identified by Gene/disease NER. Initially, a name tagger [9] is used to extract entity mentions. Regular expressions are used to join named entities that might have been considered separate by looking for intervening

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

prepositions, articles, and punctuation marks. After that, a shallow parser [10] is used to add noun phrase chunks to the list of entities. A parameter controls the minimum and maximum number of chunks per entity, in which by default one and four are considered and whether overlapping entities are allowed. The entity normalization process is characterized by representing entities' names to their canonical names and by associating them with unique representations so as to help in solving issues resulting from variations in the synonym terms as well as the ambiguous abbreviations.

iii. Knowledge Base

A comprehensive Knowledge Base (KB) is developed based on the classes, characters and functions/properties present in the aggregated ontologies. Graph-based approach is used to construct the KB. We create a document for Triplet Construction in which each entity e is entity is described as as a set of triples $\mathrm{t} \in \mathrm{T}$. The knowledge base which is constructed from 300 biomedical ontologies from BioPortal [8], consist of the Triplets in which each entity is connected to other entities via a set of triples T. And these connections are regarded as edges of CG_{KB}. where, $\mathrm{CG_{KB}}$ is the Context Graph with respect to the Knowledge Base.

iv. Entity Mapping and Candidate Retrieval

We perform entity matching for all entity types based on regular expressions [11] using Regex NER. It defines cascaded patterns over token sequences. Set of rules are defined for each entity type that expresses some patterns of entity mentions by exploring the corpora, and BIO labels are assigned to those patterns. Also, triples describing the entities are analyzed based on the properties such as: labels and names (e.g. rdfs:label), synonyms (e.g. synonym from gene ontology), aliases, and symbols (e.g. from Orphanet ontology). Thus providing more than 160 properties to map with its respective entity. Then we retrieve all the entities that are similar to the mentions in the ontologies and knowledge base and consider them as candidate entities.

v. Kernel Construction

A set of semantically co-referent entities u, mentioned as kernel of u is automatically inferred by using the functional aspects of gene and disease mined from biomedical dictionaries. We use Human Metabolome Database (HMDB) [12], Gene Ontology [13] and UniProt [14] as gene dictionary; Medical Subject Heading (MeSH) produced by US National Librart of Medicine, and KEGG Disease [15] as disease dictionary. The training set is initialized by combining the candidate entity with the co-referent entities based on functional property, partial match and full match of the elements. Beside we assume that the correct entities are infers in the kernel. Yet, error accumulation in the bootstrapping process can be encountered due noisy data.

vi. Entity Labeling

Linking entities refers to the description of functions or target genes through which it is associated with the disease. Entity labeling is a task that deliberated for context graph in our experiments. Assume that the classification component in a given context graph is given a set of labels. The problem is simplified by initially constructing a network with only a single type of entity from the context graph. So that we introduce a link between two entities if they are connected to the same function and having introduced these entity-entity links delete all the functional nodes and the links originating from them from the context graph.

vii. Learning Distinct Classes

This iterative step is based on the hypothesis that co-referent entities share some similar functional aspects and a few functions are more essential for linking entities. For a given set of candidate entities with respect to u, we estimated a set of co-referent and non-coreferent entities together establish the training set of u. A pair of matched functions (partial/full) are chosen to hold the maximum distinguishability and is measured in terms of Information Gain [16]. Then assigned a unique value to separate function in that class. Since, functional relations are involved in the iteration. Functional relations are extracted and compared with a string matching algorithm [17] for the entities given in the training set. Since, each entity is described in the dictionary that contains all phrases matching the string. If the similarity between the values are larger than a threshold, the related two functions are matched. The highest computational cost in the Boosting process is incurred due to functional aspect comparison. The learned functional classes reveal important characteristics of the mined biomedical literatures and enable to find new co-referent entities holding the same function. Additionally, we employ Apriori algorithm to find the frequent grouping of functions and refine them using heuristic rules beforehand. In each iteration, when a class is chosen and it belongs to some frequent class cluster, its counterpart in the group. Finally, the classes in the group with their associated value would be used together to obtain new links.

viii. Bootstrapping Algorithm

The proposed entity linkage algorithm is a kind of semi-supervised learning and is depicted in Algorithm 1. Given the kernel **K**(e) of an entity e, and a set x of (uncertain) entities from the input document D, the goal is to incrementally learn the most distinct classes (steps 3-5) and use them to continue linking entities in ***X*** by retraining itself on an explained training set (steps 7-8).

Algorithm 1: Bootstrapping Biomedical Entity Linkage

Input: The kernel **K** of an entity e, a set **X** of Candidate entities in a set **D** of input documents.

Output: A set **E** of labelled coreferent entities for e.

```
1. Initialize two empty lists $L_P$ and $L_V$, and print $\mathbf{K}$ to $\mathbf{E}$;
2. Estimate a set $\mathbf{N}$ of non-coreferent entities for e
   Such that $\mathbf{N} \subseteq \mathbf{X}, |\mathbf{N}| \approx |\mathbf{E}|$;
3. The most distinct classes are selected $(\mathrm{f_i}, \mathrm{f_j}) \notin \mathrm{LP}$ by
   $\mathrm{Distinct}(\mathrm{f_i}, \mathrm{f_j}) = \mathrm{IG}(\mathrm{f_i}, \mathrm{f_j}) = \mathrm{E(T)} - \mathrm{E}(\mathrm{T}_{(\mathrm{f_i}, \mathrm{f_j})})$, such
   that
```

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

```
    $\mathrm{f_i} \in \bigcup_{s \in \mathbf{E} \cup \mathbf{N}} \mathrm{Pred}(\mathbf{D}, s), \mathrm{f_j} \in \bigcup_{t \in \mathbf{E} \cup \mathbf{N}, \mathbf{t \neq s}} \mathrm{Pred}(\mathbf{D}, t)$;
4.  If $(\mathrm{f_i}, \mathrm{f_j}) = NULL$ then break;
5.  Assign the maximum score values $(\mathrm{v_i}, v_\mathrm{j})$ to $(\mathrm{f_i}, \mathrm{f_j})$
    respectively, based on occurrence given by
    $(\mathrm{v_i}, \mathrm{v_j}) = \mathrm{argmax}|\{(s, s') \in \mathbf{E} \times \mathbf{E}|\ \mathrm{sub}(v, v') \geq$
    $\delta, \langle s, \mathrm{f_i}, v \rangle \in \mathbf{D}, \langle s', \mathrm{f_j}, \mathbf{v'} \rangle \in \mathbf{D}\}|$, such that
    $\mathrm{v_i} \in \bigcup_{s \in \mathbf{E}} \mathrm{Obj}(\mathbf{D}, s, \mathrm{f_i})$,  $\mathrm{v_j} \in \bigcup_{t \in \mathbf{E}} \mathrm{Obj}(\mathbf{D}, t, \mathrm{f_j})$,  while
    $(\mathrm{f_i}, \mathrm{v_i}) \notin L_V$ or $(\mathrm{f_j}, \mathrm{v_j}) \notin L_V$;
6.  If $o_i = NULL$ and $o_j = NULL$ then Push $(\mathrm{f_i}, \mathrm{f_j})$ in $L_P$;
    Go to step 3;
7.  If $o_i \neq NULL$ then
    Use $(\mathrm{f_i}, v_\mathrm{i})$ to fetch out a set **U** of candidate entities,
    such that $\mathrm{U} \longleftarrow \{u \in \mathbf{X}|\langle u, \mathrm{f_i}, \mathrm{v_i} \rangle \in \mathbf{D}\}$;
    Else if $(\mathrm{f_i}, v_\mathrm{i})$ is distinct by the following equation,
    $\mathrm{Distinct}(\mathrm{f_i}, \mathrm{v_i}) = \frac{|\{s \in \mathbf{E}|\langle s, \mathrm{f_i}, \mathrm{v_i} \rangle \in \mathbf{D}\}|}{|\{s \in \mathbf{X}|\langle s, \mathrm{f_i}, \mathrm{v_i} \rangle \in \mathbf{D}\}|}$, then
    Add **U** to **E**, and eliminate **U** from **X;**
8.  If $o_j \neq NULL$ then
    Use $(\mathrm{f_j}, v_\mathrm{i})$ to draw a set **W** of candidate entities, such
    that $\mathrm{W} \longleftarrow \{u \in \mathbf{X}|\langle w, \mathrm{f_j}, \mathrm{v_j} \rangle \in \mathbf{D}\}$;
    Else if $(\mathrm{f_i}, v_\mathrm{i})$ is distinct by the following equation,
    $\mathrm{Distinct}(\mathrm{f_i}, \mathrm{v_i}) = \frac{|\{s \in \mathbf{E}|\langle s, \mathrm{f_i}, \mathrm{v_i} \rangle \in \mathbf{D}\}|}{|\{s \in \mathbf{X}|\langle s, \mathrm{f_i}, \mathrm{v_i} \rangle \in \mathbf{D}\}|}$, then
    Add **W** to **E**, and eliminate **W** from **X;**
9.  Push $(\mathrm{f_i}, v_\mathrm{i})$ and $(\mathrm{f_i}, v_\mathrm{i})$ in $L_V$;
10. Continue iteration until $iteration\_times > \tau$; finally,
    return **E** with the set of labels**.**
```

The distinctiveness of a class is measured w.r.t. the amount of potentially co-referent entities that can be found by using the class. Let $(\mathrm{f_j}, v_\mathrm{i})$ be the distinct function and value selected from a set **D** of documents respectively. The distinctiveness of a class is computed as mentioned in step 7 and 8, where **E**, **X** are the co-referent and uncertain entity sets in **D**, respectively. The algorithm terminates when all distinct functions have been checked (step 7) or the iteration time exceeds the threshold value τ (step 10). A subset of E is randomly sampled by the algorithm to reduce the computational cost. The sample size is denoted by N is set to set to 240 based on the computational capability of our system. The time complexity of the algorithm $\mathrm{O}(\tau * \mathrm{N})^2$, since in an iteration at most $\mathrm{O}(\mathrm{N})$ entities need to be compared, which is the most time-consuming step in the algorithm and the main issue of our approach.

III. EXPERIMENTAL RESULTS

A. *Evaluation Measures*

The assessment of our entity linking systems is performed in terms of evaluation measures, such as precision, recall, $\mathrm{F_1}$-measure and accuracy. The precision of an entity linking system is computed as the portion of correctly linked entity mentions that are generated by the system:

$$precision = \frac{|\{correctly\ linked\ entity\ mentions\}|}{|\{linked\ mentions\ generated\ by\ system\}|} \quad (1)$$

Precision takes into account all entity mentions that are linked by the system and determines how correct entity mentions linked by the entity linking system are. Precision is usually used with the measure recall, the portion of correctly linked entity mentions that should be linked:

$$recall = \frac{|\{correctly\ linked\ entity\ mentions\}|}{|\{entity\ mentions\ that\ should\ be\ linked\}|} \quad (2)$$

Recall takes into account all entity mentions that should be linked and determines how correct linked entity mentions are with regard to total entity mentions that should be linked. These two measures are used together in $F_1 - measure$ to provide a single measurement for a system. $F_1 - measure$ is defined as the harmonic mean of precision and recall:

$$F_1 = \frac{2.precision.recall}{precision + recall} \quad (3)$$

Accuracy is calculated as the number of correctly linked entity mentions divided by the total number of all entity mentions. Therefore, here $precision = recall = F_1 = accuracy$.

B. *Reults*

The experimental evaluation were performed on a personal desktop with an Intel Core i5 3.1GHz CPU, 4 GB memory, Ubundu 11.10 and Java 7. The datasets were stored on a server with two Xeon Quad 2.4 GHz CPUs, 64 GB memory, CentOS 6.4 and MySQL 5.6. We conducted our experiment using the evaluation dataset created by Zheng et al. (2015) which contains 208 linkable mentions extracted from several biomedical publications. Among all of the ontologies, there are more than 2 million entities and more than 50 million factual statements. We observed that for each mention, the candidate entity types are not as diverse. The kernel achieved the highest precision but the lowest relative recall, because some coreferent entities cannot simply be identified via uniformity perception. During bootstrapping, our approach estimated non-coreferent entities to measure distinctiveness and employed frequent combination of functions/relations between entities to enhance the selection criterion of functions/relations. The candidate *"Neural Nucleus"* a non-coreferent entity indirectly links to *"Nerve Impulse"*, due to frequent combination of relations it links to *"Neural Nucleus"* from candidates of coreferent entities enables the candidate entity *"Cell Nucleus"* to obtain the correct label and rank to link. Both of their contribution increased the overall accuracy of the proposed system. It is observed that, 64% of the correct links were inferred from the kernel, and 36% correct links were established through bootstrapping, in which 4% were of frequent combinations of relations.

The entity linking is to map an entity mentioned in an input text to the Knowledge Base, which consist of articles from PubMed and MEDLINE. The Bootstrap track gives a sample entity set which consists of 416 entities for developing. The test set consists of 3904 entities. 2229 of these entities cannot be mapped to Knowledge Base, for which the systems should

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

return NIL links. The remaining 1675 entities all can be aligned to Knowledge Base. We will firstly analyze the ranking methods with those non-NIL entities, and then with an additional validation module, we train and test with all entities including NIL link entities. We have shown our results for Biomedical Entity Linking system before and after bootstrapping of the biomedical entities in Table 2.

Table 1 Performance of the Entity Linking Systems

EL System	**Correct Links**	**Total Links**	**Linkage Accuracy**
(Chan and Roth, 2013)	84	113	74.34%
(Zheng et al, 2015)	173	208	83.17%
Bootstrap approach	192	208	92.30%

Table 2 Results for Bootstrapping Biomedical Entity Linking System

	Precision (%)	**Recall (%)**	**F-Score (%)**
Without Bootstrap	86.44	87.23	86.83
With Bootstrap	92.83	83.12	92.19

For example, given a sentence *"The effects of the **MEK** inhibitor on total **HER2**, **HER3** and on phosphorylated **pHER3** were dose dependent."* it can link "**HER2**" to *"ERBB2"* in BioPortal and extract the class *'Proto-Oncogenes→Oncogenes→Genes→Genome Components→Genome→Phenomena and Processes'* as the class and label the co-referent entities for this entity mention.

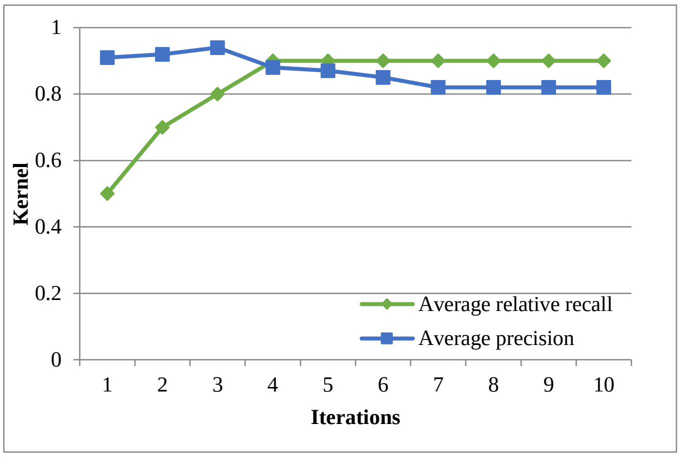


Figure 2 Precision and relative recall w.r.t number of iteration

Figure 2 depicts the average precision and relative recall on the 50 testing entities with respect to the number of iterations, where the relative recall continuously rises up at the beginning and ascends slowly later. The result suggests that a small amount of distinct functions is accurate enough for entity linkage. If bootstrapping continues, some non-distinct functions would be chosen and cause decrease in precision. Based on the figure, we set the maximum number of iteration $\tau = 4$.

The empirical comparison has been made between the proposed Bootstrap approach and two other systems, which have several variations and it is hard to cover them all in our test. The Indexing + Similarity computation approach [18] leverages indexing techniques on a few important relations/functions to locate the candidate entities, and then combines various matchers to compute similarities between these candidates. In our evaluation, we indexed the labels and mention names of all the entities in our input corpus and used the TF_IDF model to compute the similarities between the descriptions of entities. The similarity threshold was set to 0.24 based on the best accuracy in our test. Class based learning approach identifies distinct functions/relations statistically w.r.t. different classes, and matches other entities under the same classes using the learned function/relations. We chose [19] in the test, which conducted uniformity perception to create a training set and ranked entities w.r.t. the information gain in different classes. Value similarities from top-5 function/relations were linearly aggregated with equal weighting, and the threshold was fixed to 0.14 according to the best accuracy.

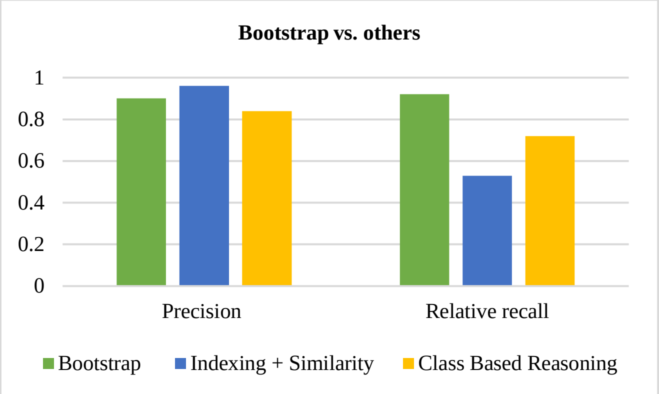


Figure 3 Precision and relative recall comparison

Fig. 3 shows the precision and relative recall comparison between the proposed Bootstrap approach and the other two systems. It is observed that the Bootstrap approach achieved the best overall accuracy, while the class based learning largely depend on the sufficiency of the training sets, causing its accuracy varied between testing entities. The system based on Indexing + Similarity computation performed worse than the others, because it gen`erated too many candidate entities with diverse functions/relation-values, and failed to decide a uniform threshold to eliminate wrong links.

V. CONCLUSION

We proposed a bootstrapping approach to entity linkage on biomedical domain. It automatically extract and link prominent entities from unstructured biomedical literatures to ontologies. The proposed Bootstrap approach is based on uniformity perception and similarity computation to link entities from unstructured biomedical texts to ontologies. Also bridges the gap between semantically coreferent entities and potential candidates. The experimental results show that our approach achieved superior precision and recall by comparing with the existing state-of-the art algorithms with improved linkage accuracy. In future, we look forward to designing other semi-supervised learning approaches for entity linkage over biomedical domain.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

REFERENCES

[1] Jin G Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler,Heng Ji, BMC Med Inform Decis Mak 2015 20;15 Suppl 1:S4. Epub 2015 May 20.

[2] Hunter L, Cohen KB: Biomedical language processing: Perspective what’s beyond PubMed? Molecular cell 2006, 21(5):589-594, doi:10.1016/j. molcel.2006.02.012

[3] Miwa M, Sætre R, Miyao Y, Tsujii J: A rich feature vector for proteinprotein interaction extraction from multiple corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore; 2009, 121-130.

[4] Liu B, Qian L, Wang H, Zhou G: Dependency-driven feature-based learning for extracting protein-protein interactions from biomedical text. Proceedings of the 23rd International Conference on Computational Linguistics: Posters Association for Computational Linguistics, Beijing, China; 2010, 757-765

[5] R. Isele, C. Bizer, Active learning of expressive linkage rules using genetic programming, J. Web Semant. 23 (2013) 2–15.

[6] W. Hu, R. Yang, Y. Qu, Automatically generating data linkages using class-based discriminative properties, Data Knowl. Eng. 91 (2014) 34–51.

[7] S. Abney, Bootstrapping, in: Proc. Annual Meeting on Association for Computational Linguistics, ACL’02, 2002, pp. 360–367

[8] National Center for Biomedical Ontology: BioPortal. 2014.

[9] Ratinov L, Roth D: Design challenges and misconceptions in named entity recognition. Proceedings of the Thirteenth Conference on Computational Natural Language Learning Association for Computational Linguistics, Boulder, CO; 2009, 147-155.

[10] Punyakanok V, Roth D: The use of classifiers in sequential inference. NIPS, Vancouver British Columbia, Canada; 2001.

[11] A.X. Chang, C.D. Manning, TokensRegex: defining cascaded regular expressions over tokens, Technical Report CSTR 2014-02, Department of Computer Science, Stanford University, 2014.

[12] Wishart,D.S., Jewison,T., Guo,A.C. et al. (2013) HMDB 3.0– The Human Metabolome Database in 2013. Nucleic Acids Res., 41, D801–D807

[13] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:259.

[14] UniPort Consortium. The universal protein resource (UniProt). Nucleic Acids Res 2008;36: D190–5.

[15] National Library of Medicine. MeSH. http://www.ncbi.nlm.nih.gov/mesh.

[16] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[17] G. Stoilos, G. Stamou, S. Kollias, A string metric for ontology alignment, in: Proc. International Semantic Web Conference, ISWC’05, 2005, pp. 623–637.

[18] S. Rong, X. Niu, E. Xiang, H. Wang, Q. Yang, Y. Yu, A machine learning approach for instance matching based on similarity metrics, in: Proc. International Semantic Web Conference, ISWC’12, 2012, pp. 460–475.

[19] W. Hu, R. Yang, Y. Qu, Automatically generating data linkages using class-based discriminative properties, Data Knowl. Eng. 91 (2014) 34–51.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Identification of axiomatic relations from unstructured texts using named entity recognition

Ana B. Rios-Alvarado, Ivan Lopez-Arevalo, and Edgar Tello-Leal

***Abstract*—Domain ontologies facilitate the organization, sharing and reuse of domain knowledge. The construction of ontologies from text deals with the extraction of concepts and relations from a text collection. A huge challenge is the learning of more expressive ontologies which includes relations such as disjointness or equivalence between classes. In this paper, we propose a method for recognition of named entities, which operates on the levels of instance and class. Firstly, at the instance level, using a named entity recognition tool named entities from unstructured texts are extracted. In addition, the type and subtype of the extracted named entity are identified. Secondly, at the class level, for each class a set of instances that allow characterizing the class is associated. Then, using the type and the set of instances of each class, the proposed method can identify the axiomatic relation. The different axiomatic relations that approach identifies can be *subClassOf*, *disjointWith*, and *equivalentClass*. The evaluation of the method for named entity recognition proposed was performed using a data set of 3542 English text documents.**

***Index Terms*—Knowledge acquisition, ontologies, text processing**

I. Introduction

The use of information and communication technologies have motivated an exponential growth in the available information. This growth is not only present on web resources, but it also can be seen in organizations. For example, in an organization, documents represent a significant source of collective expertise (*know-how*) and the most of the data are in unstructured text format. For instance, the number of business emails sent and received per user per day totals 122 emails per day[1]. In order to store, retrieve, or infer knowledge from this information, it is necessary to represent it using a conceptual schema. This can be achieved by means ontologies. Ontologies are formal vocabularies of terms, often shared by a community of users [1]. Ontologies facilitate the organization, sharing, and reuse of domain knowledge, they also are one of the key technologies for the Semantic Web and its current success.

Ontology learning from text consists in deriving high-level concepts and relations on the basis of the words appearing in the text [2]. To carry out this process, textual documents are an important source of knowledge. Moreover, in the recent years, the availability of unstructured textual information has increased, which can serve to extract useful knowledge. In many areas, such as medicine, bioinformatics, and finance, the main benefits of using ontologies for knowledge modeling is the ability to infer new knowledge that allows the development of more realistic applications, which requires the inclusion of more expressive elements, such as disjointness or equivalence relations. Axioms involving semantic features that can provide expressivity to ontologies [3]. Consequently, the addition of such relations allows the implementation of applications based on reasoning tasks, such as ontology classification and query answering.

In the context of languages for Semantic Web (for example OWL-DL), an axiom is an assertion in a logical form. All axioms together comprise the overall theory that the ontology describes in its domain of application. Taking into account the elements of the ontology, there are three types of axioms: 1) *class expression axioms*, which refer to general restrictions between classes, for example, the *subClassOf* relation between the *SoccerClub* and *SportTeam* classes, or *disjointWith* relation between the *City* and *SoccerClub* classes; 2) *properties* allow to define the attributes or facts associated with the members of classes or specific instances, for example, the relation *birthPlace* between *Place* and *Person* classes or the relation *birthYear* between *Person* class and `xsd:integer`; and 3) *assertions* on individuals commonly called *facts*, for example, the relation between individuals with the same characteristics establishes a particular property between them, such as *Ronaldo* `owl:sameAs` *Ronaldo Luís Nazário de Lima*. In particular, OWL-DL gives the formal syntax to represent the axioms above described in the ontology. The disjointness of classes can be expressed using the `owl:disjointWith` constructor. This relation guarantees that an individual, as member of one class, cannot be simultaneously an instance of a specified other class. Similarly, the constructor `owl:equivalentClass` is used to indicate that two classes have precisely the same instances. The obtaining of instances for each class is a key step in the identification of subsumption, disjointness or equivalence relations.

This paper presents a method based on named entity recognition from unstructured text to identify class expression axioms. A named entity is an information unit such as the name of a person, an organization, a location, a brand, a product, or a numeric expression (time, date, money, and percent) as can be found in text. The presented approach starts with the detection of named entities. Subsequently, at the class level, for each class a set of instances that allow characterizing the class are identified and associated. In a complementary way, the sentences where the instances and their corresponding type of class appear are analyzed. Consequently, the context relation and the *instanceOf* relations based on entity extraction task, determines one of the following relations between classes: *subClassOf*, *disjointWith*, or *equivalentClass*. This is

[1] The Radicati Group, Inc, Email Statistics Report, 2015-2019 www.radicati.com

https://doi.org/10.17562/PB-56-8

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

possible due to the use of schema types from AlchemyAPI or OpenCalais have also been collected in an ontology called NERD (Named Entity Recognition Disambiguation)[2]. Finally, the evaluation of the method was performed using a data set of 3542 English documents in the Football domain, allowing evaluate the identification of the *instanceOf* relation, and evaluate the learning axioms. In [4] has been reported the results for a set of documents in Tourist domain.

The rest of the paper is structured as follows. In Section 2, a brief description of the work related to generation of axioms is presented. Next, in Section 3 the method to identify class expression axioms is described. In Section 4, the experiments carried out are presented and discussed. Finally, in Section 5, we provide some conclusions.

II. Related work

In order to provide a higher level of expressiveness to learned ontologies, several approaches have been proposed for extending logical properties of the modeled knowledge in an unsupervised or automatic way. According to the the type of axioms, works such as [5], [6], and [7] are focused on class expression axioms. The tool named LEDA [5] permits the automated generation of disjointness axioms based on machine learning classification. The classifier, which determines disjointness for any given pair of classes, is trained based on a gold standard baseline of disjoint axioms manually created. Zhang *et al.* [6] proposed an unsupervised method for minning equivalent relations from Linked Data. It consists of two components: 1) a measure of equivalency between pairs of relations of a concept and 2) a clustering process to group equivalent relations. Ma *et al.* [7] introduced an approach to discover disjointness between two concepts. In this work, the task of association rule minning is to generate patterns like the form $A \rightarrow \neg B$, and then transform them to disjointness axiom "A `owl:disjointWith` B". On the other hand, Sánchez *et al.* [8] presented an approach for discovering object properties. Their method is based on natural language processing techniques, linguistic patterns and statistical analyses performed at a Web-scale to extract and evaluate semantic evidences from textual resources. In [9] and [10] the approaches are related work to assertions or inference rules acquisition. Völker *et al.* [9] presented the methodology named LExO. The first step of the methodology is analyzing the syntactic structure of an input sentence. The resulting dependency tree is transformed into a set of OWL axioms (concept inclusion, transitivity, role inclusion, role assertions, concept assertions, and individual equalities) by means of manually engineered transformation rules. Li and Sima [10] proposed an ontology mining approach, where the ontology axioms are obtained through statistical measures by running SPARQL queries on Linked Data.

The above approaches do not examine how to determine what classes are relevant in an automatic way for getting axioms neither do they consider the individuals as part of the extensional definition of a class. In order to get axioms, by taking into account the evidence of named entities in domain-specific text, we propose to resolve the following question: Does the *instanceOf(named entity, class)* relation provide evidence for an axiomatic relation? To address this question, the named entities have been identified by a Named Entity Recognition (NER) tool and subsequently, *subClassOf*, *disjointWith*, and *equivalentClass* relations are established. The NER aims to identify meaningful segments in input text and categorize them into pre-defined semantic classes such as the names of people, locations and organizations.

We assumed that a taxonomy structure exists and it represents the domain of the texts. Following a method from specific to general, the approach involves identifying individuals, which are instances of some class. Such classes belong to a taxonomic structure, which is at the core of the ontology. Figure 1 shows that the instance level corresponds to the leaves in a taxonomic tree structure and the class level to the branches. The difference between one class and another is that its set of leaves is different and therefore it can be characterized as a separate (disjoint) class, otherwise if the set of leaves is very similar, then it can be characterized as an equivalent class. For example, in the instance level, the set of leaves for *Country* class includes *Brazil*, *Germany*, and *Denmark* as members, but the set of leaves for *SoccerFederation* class contains *FIFA*, *CONMEBOL*, and *UEFA* members. Then, *Country* class and *SoccerFederation* class are disjoint. Thus, the collection of named entities provides the members for a specific class, and defines a class in an extensional manner.

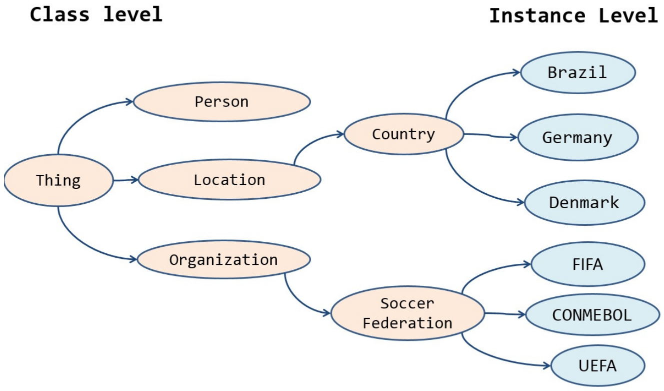


Fig. 1. Example of ontology for Sports domain

III. A method for acquisition of axioms

The proposed method starts at the instance level, where an NER tool extracts the named entities from input text. Later, at the class level, each class has a set of instances associated with it that characterize it. The NER tool provides a set of types (type/subtype) associated to each named entity. Using the type and the linguistic context of each class, an axiomatic relation is identified. Figure 2 shows the general overview of the proposed steps to extract axioms. This method consists of a bottom-up approach and it follows the next steps:

1) Identification of instances: An NER tool obtains the named entities from text. The named entities can correspond to one of the following types (defined by the tool):

[2] Available: http://nerd.eurecom.fr/ontology

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

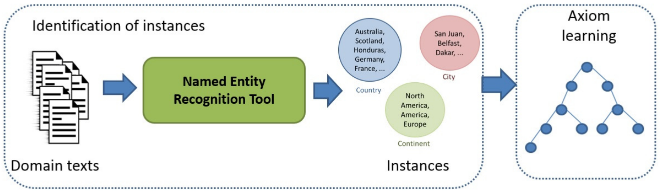

Identification of instances
Domain texts
Named Entity Recognition Tool
Australia, Scotland, Honduras, Germany, France, ...
Country
San Juan, Belfast, Dakar, ...
City
North America, America, Europe
Continent
Instances
Axiom learning

Fig. 2. The proposed method for identification of axioms

Person, Organization, Location, Country, or Quantity among others. The NER tools exploit the Linked Data[3] principles, which consists of a unique global identifier defines an entity. Such referenced identifier provides useful information about the corresponding resources and links to other relevant identifiers. Later, the relations of type *instanceOf(named entity, class)* between a named entity and a class are obtained by two methods: 1) the given type from the NER tool and 2) the context where the named entity and its class co-occurs.

2) Axiom learning. The sentences where a set of instances and its corresponding class occur are grouped to determine if there exists a relation between the contexts of two classes. A part-of-speech (POS) tagger and a syntactic parser are used to get the linguistic context (i.e., representative elements such as nouns, verbs, or adjectives and their grammatical relations). The linguistic context supports the identification of relations based on entities used to derive one of the following axioms: *disjointWith* or *equivalentClass*.

   At the class level, the *subClassOf* relation represents one of the main axioms, which structures the set of classes into a taxonomy where a higher class is more general than a lower class. We propose the use of NER and linguistic context as an additional approach for identifying *subClassOf* relations in text.

IV. Experiments and Results

For our experiments, we used the Smart Web Football dataset used by Jiang and Tan [11], which consists of 3,542 English documents. It covers a list of 2295 classes, 1459 individuals, and 633 taxonomy relations. The measures used for the evaluation are precision, recall, and F-measure.

A. Identification of instances

In this stage, the objective was to evaluate the identification of the *instanceOf* relation using AlchemyAPI and OpenCalais tools. These tools execute the named entity recognition task and define a taxonomy of types. The comparison was made on 185 *instanceOf* relations that were manually annotated. According to the evaluation, AlchemyAPI had better precision than OpenCalais in this task. More in detail, Table I presents the performance of AlchemyAPI and OpenCalais for the identification of instances belonging to these classes: *Country*, *Person*, *City*, and *Company*. The obtained results were compared manually with 70 *instanceOf* relations from the test dataset manually annotated. In most cases, AlchemyAPI showed the best precision.

TABLE I
Performance NER tools - identified instances by class

Class	Tool	Precision	Recall	F Measure
Country	AlchemyAPI	0.4529	0.4900	0.4707
	OpenCalais	0.4000	0.5000	0.4444
Person	AlchemyAPI	0.7331	0.8582	0.7907
	OpenCalais	0.6500	0.7000	0.6740
City	AlchemyAPI	0.5678	0.4000	0.4693
	OpenCalais	0.5234	0.3550	0.4230
Company	AlchemyAPI	0.3333	0.2667	0.2963
	OpenCalais	0.2480	0.3000	0.2715

TABLE II
Examples of sentences where *instanceOf* relation ocurrs

Sentence	Lexical Pattern
Messi is an Argentine professional *footballer* who plays as a forward for Spanish club Barcelona and the Argentina national team.	<NE> is a <NP>
I have actually wanted to be a proffessional *goalkeeper, like Iker Casillas* from Spain.	<NP> like <NE>
Eight *players including Brian McBride*, *Claudio Rayner*, and *Brad Friedel*	<NP> including <NE> {, <NE>, ... and <NE>}

In addition, using the context, we can see that instances of different classes appear in the same sentence, i.e. they co-occur. For extracting relations, the linguistic context for each of the extracted named entity was analyzed. The Table II shows examples of sentences with patterns that identify the *instanceOf* relation, where <NE> is a named entity and <NP> is a noun phrase. In the first example, *Messi* is an instance of the *footballer* class and the pattern associated is <NE> is a <NP>. In the second example, *Iker Casillas* is an instance of the *goalkepper* class. In this case, the pattern associated is <NP> like <NE>. For the third example, the instances are *Brian McBride*, *Claudio Rayner*, and *Brad Friedel* for the class called *player*.

B. Axiom learning

In this section, we present a description on the experiments to identify *subClassOf*, *disjointWith*, and *equivalentClass* relation.

The NER tool used for this was AlchemyAPI because it shows the best precision in obtaining instances. AlchemyAPI obtains 16 types of classes and 62 subtypes on a sample corpus with 541 files from the Smart Web Football corpus. A human team was asked to evaluate all extracted subtype relation, which gave a precision of 73.58% for the extracted relations based on AlchemyAPI identified subtypes-types representing the football domain. The Table III shows some examples of relations correctly identified.

A *disjointWith* relation states that one class has not an instance member in common with another class. For learning the disjoint relationship between two classes, we consider named

[3] http://www.w3.org/DesignIssues/LinkedData.html

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

TABLE III
EXAMPLES OF LEARNED *types-subtypes* RELATIONS

Type	**Subtype**
Organization	*SoccerClub*, *FootballTeam*, *FootballOrganization*
Company	*FootballAssociation*, *SportsAssociation*
Person	*FootballPlayer*, *FootballManager*
Sport	*AwardDiscipline*
Region	*Location, Country*

TABLE IV
EXAMPLES OF LEARNED *disjointClass* RELATIONS AND ITS NAMED ENTITIES

$class1$/$class2$	$class1$**'s NE**	$class2$**'s NE**
SportingEvent/ Organization	*World Cup*, *Nations Cup*, *Olympics*	*Arsenal*, *FIFA*, *Champions League*, *Glasgow Rangers*, *East Asian Football Federation*
Country/Organization	*Italy*, *Japan*, *Iraq*, *Germany*, *United States*, *France*, ...	*Arsenal*, *FIFA*, *Champions League*, *Glasgow Rangers*, *East Asian Football Federation*
City/SportingEvent	*Kuwaits*, *Cologne*, *Aruba*, *Liverpool*, *Caracas*, *Miami*, *Madrid*, ...	*World Cup*, *Nations Cup*, *Olympics*
City/Person	*Kuwaits*, *Cologne*, *Aruba*, *Liverpool*, *Caracas*, *Miami*, *Madrid*, ...	*Jacques Santini*, *Patrick Mboma*, *Edwin van der Sar*, *Hidetoshi Nakata*, ...

entities that co-occur in the same context. For each NER ($class_1$, $class_2$) duple, the list of instances was compared. If there is not a common named entity between the two classes then the *disjointWith(*$class_1$, $class_2$*)* relation is established. To illustrate the evaluation of *disjointWith* relation extraction, it was used a sample corpus with 541 files. A number of 120 duple ($class_1$, $class_2$) were obtained. According to the evaluation of the human team, where it was evaluated if obtained duple has disjoint relation between $class_1$ and $class_2$, 102 of the relationships correspond correctly to *disjointWith (*$class_1$, $class_2$*)* and the rest of them (18) have some other relation. As a result, the precision was 85.00% for the learned disjoint relations. Some examples of learned disjoint relations between classes are the *SportingEvent* and *Organization* classes as well as the *Country* and *Organization* classes, *City* and *SportingEvent* classes, and the *City* and *Person* classes. However, the *Organization* and *Company* are not necessary disjoint classes. Even although according to NER tool results, the set of instances were very different between *Organization* and *Company*, according to human expert the classes meet in a *subClassOf* relation. The Table IV shows some disjoint relations learned and their corresponding named entities where it is clear that their set of named entities is disjoint.

In a particular case, the sets of named entities associated with *City* class and *SoccerClub* class are very similar, but these class are disjoint although they share elements.

The *equivalentClass* relation is established between two classes when the class descriptions include the same set of individuals. It is important to mention that class equality means that the classes have the same intensional meaning i.e. denote the same concept. For learning *equivalentClass* relation, two ontologies were considered and for each ontology class its set of instances obtained by two different NER tools were compared, if the set of instances between two different classes is highly similar then an *equivalentClass(*$class_1$, $class_2$*)* relation can be established. Highly similar means that almost the total of named entities detected by the NER tool is the same in both classes, that is because the identification of instances depends on the precision of the NER tool. In this case, using the same sample corpus with 541 files, the AlchemyAPI and OpenCalais tools identify 16 and 17 classes, respectively. However, only 32 duple (AlchemyAPI : $class_1$, OpenCalais : $class_2$) of the total (272) have overlap between their set of instances. For example, *AlchemyAPI : Organization / OpenCalais:Organization* and *AlchemyAPI : Country / OpenCalais : Country* can clearly be determined a equivalence relationship between them. In contrast, the classes *AlchemyAPI : Organization / OpenCalais : Company* or *AlchemyAPI : Person / OpenCalais : Holiday* which have similar individuals but they are not equivalent. According to the evaluation of the human team, 24 of the relationships correspond correctly to *equivalentClass(*$class_1$, $class_2$*)* and the rest of them have some other relation. As a result, the precision was 75.00% for the learned *equivalentClass* relations. The Table V shows some examples of learned duples, where *AAPI* and *OC* correspond to *AlchemyAPI* ontology and *OpenCalais* ontology, respectively.

TABLE V
EXAMPLES OF EQUIVALENT CLASSES

$class1$	$class2$	*equivalent class*	**other relation**
AAPI:Organization	*OC:Organization*	*	
AAPI:Country	*OC:Country*	*	
AAPI:Sports	*OC:SportsGame*	*	
AAPI:Health Condition	*OC:Medical Condition*	*	
AAPI:Organization	*OC:Company*		*

V. CONCLUSIONS

The approach described in this paper is based on identifying named entities as class' members and comparing their set of instances to establish axiomatic relations *subClassOf*, *disjointWith* and *equivalentClass*. Our approach is unsupervised and the identified relationships can enrich ontologies lack of expressiveness. New technologies in NER tools based on Linked Data can be useful in the process of extracting axioms.

According to the experiments, we observed that the identified instances that belong to a specific class could be considered as the extensional definition of this class and then it is described by the named entities associated to it. However, the method must take into account the fact that the incorrect identification of instances can derive erroneous axiomatic relations. For example, other relations such as *subClassOf* and *partOf* were learned instead as a *disjointWith* relation, or as *equivalentClass* relation. One of the main difficulties lies with

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

resolving ambiguity in named entities. In such case, other tools could be exploited for named entity disambiguation task.

In the experiments, one of the main dificulties lies with ambiguity. Further work will be focus on more experiments for adding other resources and evaluating the similarity of classes. Also, new experiments will consider a comparison with other approaches.

REFERENCES

[1] I. Horrocks, "Tool support for ontology engineering," in *Foundations for the Web of Information and Services*, 2011, pp. 103–112.

[2] P. Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3] J. Völker and S. Rudolph, "Lexico-logical acquisition of owl - dl axioms," in *Formal Concept Analysis*, ser. Lecture Notes in Computer Science, R. Medina and S. Obiedkov, Eds. Springer Berlin / Heidelberg, 2008, vol. 4933, pp. 62–77. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-78137-0_5

[4] A. Rios-Alvarado, I. Lopez-Arevalo, and E. Tello-Leal, "The acquisition of axioms for ontology learning using named entities," *IEEE Latin America Transactions*, vol. 14, no. 5, pp. 2498–2503, 2016.

[5] J. Völker, D. Vrandečić, Y. Sure, and A. Hotho, "Learning disjointness," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, E. Franconi, M. Kifer, and W. May, Eds. Springer Berlin Heidelberg, 2007, vol. 4519, pp. 175–189. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72667-8_14

[6] Z. Zhang, E. Blomqvist, I. Augenstein, F. Ciravegna, and A. L. Gentile, "Mining equivalent relations from linked data," *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 289–293, 2013.

[7] Y. Ma, H. Gao, T. Wu, and G. Qi, "Learning disjointness axioms with association rule mining and its application to inconsistency detection of linked data," in *The Semantic Web and Web Science*. Springer, 2014, pp. 29–41.

[8] D. Sánchez, A. Moreno, and L. Del Vasto-Terrientes, "Learning relation axioms from text: An automatic Web-based approach," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5792–5805, 2012.

[9] J. Völker, P. Hitzler, and P. Cimiano, "Acquisition of owl dl axioms from lexical resources," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, E. Franconi, M. Kifer, and W. May, Eds. Springer Berlin / Heidelberg, 2007, vol. 4519, pp. 670–685. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-72667-8_47

[10] H. Li and Q. Sima, "Parallel mining of OWL 2 EL ontology from large linked datasets," *Knowledge-Based Systems*, vol. 84, pp. 10–17, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2015.03.023

[11] X. Jiang and A.-H. Tan, "Crctol: A semantic-based domain ontology learning system," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 150–168, Jan. 2010.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Meta-heuristic Coefficient for Indexing Geometric Objects in Hierarchical Spatial Data-Structures

Gerdys E. Jiménez Moya and Jairo Rojas Delgado

***Abstract*—Hierarchical spatial data structures are usually employed for indexing geometric objects and are characterized by recursively decomposing the space. Due to this recursion process is necessary to define a decision criteria to determine when to stop the process of spatial decomposition. In this paper, a new recursion threshold for indexing hierarchical spatial data structures that is independent of the nature of the data is introduced. The objective is to reduce the execution time of space searches that arise in various applications of modern computing systems such as mining, solid modelling, simulation and others. Results indicate that the proposed recursion threshold reduces the execution time of space searches respect to others criteria of general purpose reported in the literature, however, RAM consumption is increased considerably.**

***Index Terms*—execution time, recursion threshold, spatial data-structure.**

I. Introduction

MODERN software systems are increasingly important to the development of human activity. Nowadays these systems implements several complex operations from a computational point of view related to spatial and time consumption indicators. Among these operations are the diminution of the computational cost when processing a geological block model[1], the estimation of mineral resources and the calculation of the mineral concentration located within two surfaces which are common operations of any mining software system.

In the problem of reducing the computational cost of processing a block model [1], a calculation of the mineral concentration and economic cost for each block is made. This is done as part of an optimization process to design an open pit mine. Here the goal is to maximize profits and to select the proper blocks for mineral extraction while the physical boundaries of the pit mine, usually with the shape of a regular inverted cone, are defined according to geometric constraints as illustrated in Figure 1a.

The time complexity required for finding the best combination of blocks using brute force is 2^n where n is the number of blocks and since there is a spatial relation between blocks and a geometric constraint in the physical boundaries of the mine pit, this problem can be seen as a spatial search problem.

In the estimation of mineral resources operations, Figure 1b, there is an interest in calculate the value of the mineral concentration in a spatial region from which there is not any explicit measure. In order to obtain the mentioned value, a spatial search is performed around the region of interest and all measurements are considered as inputs for a spatial interpolation method that makes an estimation according to the values of the k-nearest neighbours. According to [2] the time complexity required to perform a k-nearest neighbours search using a naive algorithm for n query point is in the order of n^2 which is expensive.

Another example of the operations performed in modern mining software is showed in Figure 1c and is a special case of a spatial range search. In the figure, a calculation of the value of the mineral concentration within two surfaces represented by polygonal meshes is required. In order to do this, mining systems computes the intersection between meshes components and a search volume. This operation has a linear time complexity but still expensive since the intersection tests usually require a lot of computational effort.

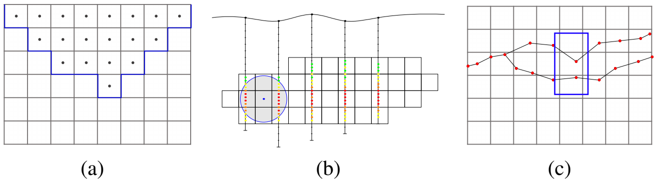


Fig. 1: Common spatial operations performed in the actual mining software.

In addition to the previous mentioned issues, there are other operations performed by modern software systems focused in tridimensional visualization that can be treated as spatial search problems. In order to achieve a higher performance of software and hardware, several methodologies have been developed, but usually the great amount of data and the complexity of the implemented operations leads to computational burden [3]. An open issue in this research field can be identified and it states as follows: how to reduce the execution time of spatial queries?

II. Spatial search operations

Spatial queries can be classified according to its basement: query based in phenomena and query based in location [4].

G. E. Jiménez was with the University of Informatics Sciences, Geoinformatics and Digital Signals Center, La Habana, Cuba. e-mail: (gejimenez@uci.cu).

J. Rojas was with the University of Informatics Sciences, Geoinformatics and Digital Signals Center, e-mail: (jrdelgado@uci.cu).

[1] Block model: structure for modelling mine pits that represents the space through a set of regular blocks.

https://doi.org/10.17562/PB-56-9

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Spatial queries based in phenomenon makes use of a semantic criterion associated to data to satisfy some constraint, while spatial queries based in location only makes use of geometric descriptions of the spatial data. In the latest years, the effort related to the diminution of the time complexity of spatial queries based in location have been intense because this kind of spatial query are context-independent[2]. In recent investigations two fields can be identified: parallel computing and spatial data-structures.

Despite parallel computing has showed good results in many application fields, there are several limitations related to the mandatory requirement of specific hardware and the difficulty of writing concurrent code. Until now, several spatial data-structures with the objective of performing spatial queries have been proposed: the Binary Space Partitioning trees (BSP), the Kd-trees, the R-tree and the quadtrees.

III. Hierarchical spatial data-structures

A spatial data-structure organizes geometric objects located in a dimensional space M according to some geometric and spatial constraint. According to [5] the organization of a spatial data-structure is usually hierarchical and its construction is based on recursive decomposition of space. Recursive decomposition of space refers to a systematic subdivision of a space in two or more subsets M_i according to Equation 1.

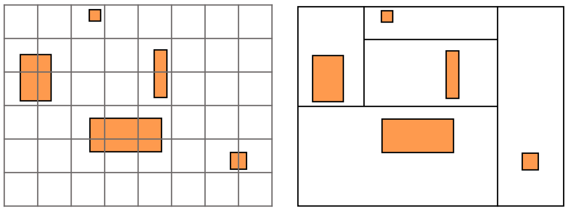
$$M = M_1 \cup ... \cup M_n i \neq j, 0 < i, j \leq n \qquad (1)$$

Given the previous definition, a set of geometric objects can be partitioned in a set of discrete adjacent elements. These elements are usually primitives and can have different shape, size, position and orientation. For tridimensional spaces, the most common of these primitives are boxes and when are used, it's a regular decomposition process because boxes are uniform in shape, size and orientation [6]. Figure 2a shows a bidimensional space regularly partitioned through squares.

An example of non-regular space partitioning process is the resulting from the indexation of a BSP tree like Figure 2b. On its simplest form, a BSP tree use a line to divide the space in two subregions and then reorganize the geometric objects according to its position respect that line [4]. The process is repeated recursively until a condition is met. The previously mentioned condition is also known as stopping criteria or indexation criteria of the recursive space decomposition process and determines when to continuing subdividing the space or not.

The recursive space partitioning process defined by a Quadtree is an example of a regular one. With a Quadtree , space is subdivided by squares in four subregions of equal size and then geometric objects are reorganized according to its position respect to the corresponding portion of space [7]. Like BSP trees, Quadtrees recursively continue subdividing the space until a condition is met.

As can be expected, there are more complex variations for BSP trees [8], [9] and Quadtrees [10], [11], [12]. BSP trees can be used for dimensional spaces higher than two by using planes in R^3 and hyperplanes in R^k for $k > 3$ also known as Kd-trees. Additionally, BSP trees can be Axis-Aligned and Polygon-Aligned when the subdivision line or plane is aligned to the axes of the coordinate system or not respectively. Quadtrees also have several variations used in R^3 named Octree and Hyperoctree in R^k for $k > 3$. In all cases a sort of recursion threshold or indexing criteria must be defined and according to [13] the setup of this recursion threshold has a direct impact in the ability of these spatial data-structure for solving specific problems.



(a) (b)

Fig. 2: Regular and no regular recursive space particioning process.

For precise mathematical description, an Octree exhaustive definition will be given and in the following section, common studied recursion thresholds for Octrees will be enunciated although it can be generalized for BSP, kd-trees and others hierarchical spatial data-structures. Octrees were proposed for first time by [14] and extensively studied by [15]. On it's general form, an Octree is a tridimensional approximation of an object by a set of boxes. According to [16], given an Octree O, a set of geometric objects S contained in a restriction volume V associated to O and $\gamma(\{O, S, V\}) \rightarrow 0, 1$ a function that determines if the Octree O should be partitioned or not; the formal definition of an Octree is as enunciated in Equation 2.

$$\begin{array}{lcc} 1 & O & \text{if } |S| = 0 \\ 2 & O & \text{if } \gamma(O, S, V) = 1 \\ 3 & \{\{O_1, S_1, V_1\}, ..., \{O_8, S_8, V_8\}\} & \text{in other case} \end{array} \qquad (2)$$

In Equation 2 the construction process of an Octree O stops if their restriction volume does not contain any geometric object or if the threshold function have been satisfied. In any other case space is recursively partitioned into eight subregions O_i calculating V_i and S_i.

A. *Recursion threshold for hierarchical spatial data-structures*

According to [13] the setup of the recursion threshold has a direct impact in the ability of spatial data-structure for solving specific problems. The recursion threshold has a direct influence in the amount of nodes generated in the hierarchical structure that at the same time determines the amount of recursions needed to process a spatial query. The behaviour for the time complexity in these structures is usually logarithmic while the search executes in the branches of the

[2] Context-independent means that is not necessary to use semantic information associated to the spatial data, but only their geometric description

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

hierarchical structure and is linear when search is inside a specific node. Despite the importance of this configuration parameter, according to [13] very few have been written about this. Some applications can be seen in [17], [18].

In [19] for example a stopping criterion for an Octree construction is defined based on a priory knowledge in radiative transfer simulation but this stopping criteria can not be generalized to other fields. In [20] an Artificial Neural Network (ANN) is proposed to calculate an *information value* index that evaluates if to continue subdividing an Octree box given a specified threshold. However, very little information is given about this process, or about how to train the ANN or what kind of test were executed with this novel methodology. A review of the state of the art about recursion threshold functions reveal that there are three main criteria:

- **Recursion level**(α). The generation of nodes will continue until the level of a node O in the hierarchical structure gets higher than α.
- **Volume of the region**(β). The generation of nodes will continue until the volume of a region assosiated to a node O gets under β.
- **Number of geometric objects**(δ). The generation of nodes will continue until the number of geometric objects assosiated to a node O gets under δ.

Each of the previously mentioned indexation criteria has their own advantages and disadvantages and their adjust is context-dependent. This means that depending of the structure of the data to index, the spatial data-structure will be sensitive to formation of clusters and over generation of nodes. In the following section a novel indexation criteria will be introduced.

B. Meta-heuristic coefficient for indexing geometric objects in hierarchical spatial data-structures

Meta-heuristic refers to the use of some sort of experience or prior knowledge to approximate an acceptable solution to a given problem. Actual knowledge allows deducing that decision about when to stop the space partitioning is related to the size of the region to subdivide, the amount contained geometric objects and to the times that the space has been subdivided. These factors match with indexation criteria seen in Section III-A.

In Equation 3 function γ from Section III is defined as a new indexation criteria that establish a direct relation between the volume of a region associated to Octree O and number of geometric objects associated to O. At the same time the new indexation criteria establish an inverse relation with the recursion level of O.

$$\gamma(O,S,V) = \begin{cases} 1 & \text{if} \quad \epsilon < \phi \times \frac{|S| \times \sigma(V)}{\theta(O)} \\ \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In Equation 3, $\sigma(V)$ calculates the volume value for the restriction volume V associated to Octree O, S is the set of geometric objects contained in V. Index ϵ represents the recursion threshold and its value depends of factors: restriction volume, amount of geometric objects and depth of the node O.

Small values of ϵ increments the chance of subdivision of a node. When a node is subdivided, its children have less volume, less number of geometric objects and a higher value for their level in the hierarchical tree structure. Because of that, the chance of subdivision is smaller. The values for ϵ factor can be normalized between $0 <= \epsilon <= 1$ with Equation 4 with $0 < \epsilon < 1$. In Equation 4 V_{root} is the restriction volume associated to the root node and S_{root} is the set of geometric objects contained in V_{root}.

$$f(\epsilon) = |S_{root}| \times \sigma(V_{root}) \times \epsilon \tag{4}$$

Factor ϕ controls the influence of factors $|S| \times \sigma(V)$ and $\theta(O)$. For values of $0 < \phi < 1$ importance to factor $\theta(O)$ is granted, causing that the node has fewer chances of subdivision and the generation of children gets decreased. For values of $\phi > 1$ factor $\theta(O)$ gets less importance and the generation of children is bigger. Factor ϕ can be used to balance the effects of time and spatial complexity.

IV. Results and discusion

The following section describes a set of tests conducted to measure the time and memory consumption of spatial queries while using an Octree and a specific indexation criteria. Input data consist in point clouds randomly generated using a normal distribution with mean of 500 thousand and typical deviation of 200 thousand. A point cloud is a set of tridimensional points in the space, defined by X, Y, and Z coordinates, and often are intended to represent the external surface of an object. The generated point clouds has different sizes, starting with 100 thousand points until 2 millions points.

An Octree pointer variation data-structure was implemented in C++11 programming language [21]. Tests were executed in Xubuntu 14.04 operating system using a *Core i3* microprocessor and 4 Gb of RAM. The implemented Octree can construct an index given a point cloud using the indexation criteria specified in Section (III-A, III-B). An orthogonal spatial search was also implemented as described in Algorithm1 using a box as V parameter.

Algorithm 1 Orthogonal spatial search.

```
function SEARCH(V: Search volume, O : Octree)
    R: Geometric Objects List;
    if INTERSECTION(V, O.V) = $True$ then
        if O.ISLEAF = $True$ then
            $R \leftarrow R+$ INTERSECTION(O.S, V);
        else
            for i = 1..8 do
                $R \leftarrow R+$ SEARCH(V, $O.O_i$);
            end for
        end if
    end if
    return R;
end function
```

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Algorithm 1 takes two parameters, the search volume V and an Octree O. Procedure first determines if the restriction volume associated to O intersects with V. If this condition is not met then no other computation is performed and there is not other geometric object contained in V. Otherwise, if O is leaf, procedure computes the intersection of each geometric object with V, if O is not leaf, spatial search procedure is called for each node of O.

First thing to do is to configure the threshold for indexation. The time consumption for a spatial search using Octree has a direct relation with the indexation criteria and the volume of data, so this parameter must be adjusted to each specific data volume. Using this adjustment, a calculation of the time consumption of an orthogonal search for each indexation criteria is made in different scenarios. All measures are calculated as the average of 100 thousand orthogonal searches. At the same time, the number of generated nodes is calculated as a measure of RAM consumption for each indexation.

The setup of the recursion threshold is made using indexation criteria previously defined. For each indexation criteria the threshold value is considered as follows: $10^{14} < \beta < 10^{16}$ with step of 2×10^{14} volumetric units, $0 < \delta < 50000$ with step of 500 geometric objects and $0 < \epsilon < 1$ with step 0.01 and $\phi = 100$.

For example, using a point cloud of 100 thousand objects, the best scenario for δ is $\delta = 500$ geometric objects and the worst scenario is $\delta = 12000$ as can be seen in the red line of Figure 3. However, this configuration does not stand for 2 millions objects where the best scenario for δ is $\delta = 1000$ and the worst scenario is $\delta = 43500$ as can be seen in the blue line of Figure 4.

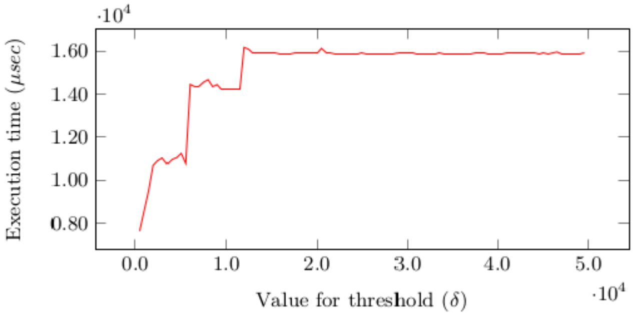


Fig. 3: Time consumption for different threshold (δ) using 100 thousand points (red)

A. *Performance for best scenario*

The objective of the following test is to measure the time and spatial consumption for best scenario of defined recursion thresholds. Figure 5 shows the time consumption for threshold (β) red, threshold (δ) blue and threshold (ϵ) black for each volume of data.

As can be seen in Figure 5 and 6 threshold ϵ performs better than other indexation criteria for each data volume tested in the best possible scenario. Considering threshold β threshold ϵ decreases the time consumption in 1.6×10^4 microseconds as average and 8.7×10^3 microseconds considering threshold δ . Figure 5 also shows the number of generated childs for each indexation criteria for different data volumes in the best scenario. Results indicate that the use of the new indexation criteria in the best scenario needs more memory space in RAM than others recursion thresholds.

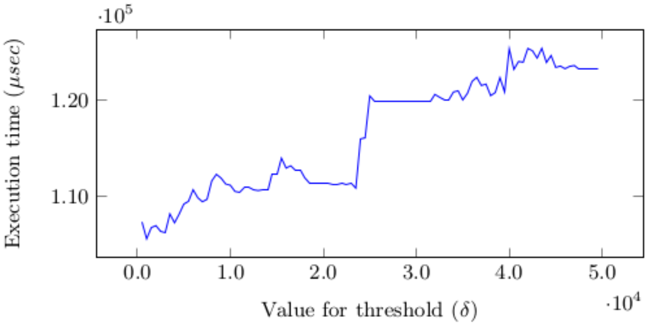


Fig. 4: Time consumption for different threshold (δ) using 2 millions points (blue).

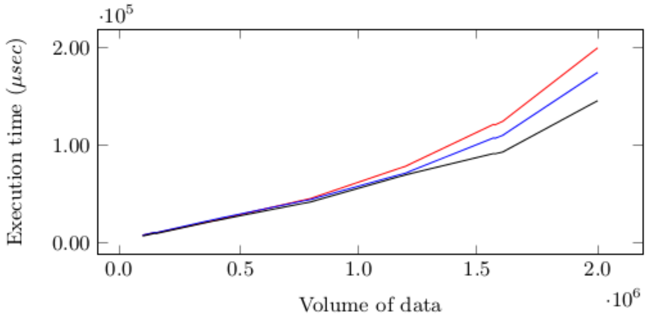


Fig. 5: Time consumption and children generation for different data volume using threshold (β) red, threshold (δ) blue and threshold (ϵ) black.

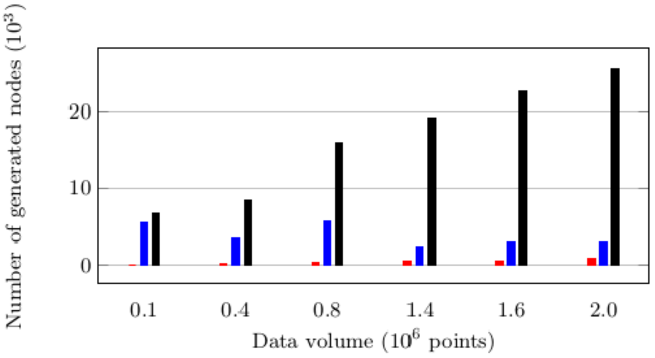


Fig. 6: Another view of time consumption and children generation for different data volume using threshold (β) red, threshold (δ) blue and threshold (ϵ) black.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

B. Performance for worst scenario

The objective of the following test is to measure the time and spatial consumption for worst scenario of defined recursion thresholds. Figure 7 shows the time consumption for threshold (β) red, threshold (δ) blue and threshold (ϵ) black for each volume of data.

As can be seen in Figure 7 threshold ϵ performs better than other indexation criteria for each data volume tested in the worst possible scenario. Considering threshold β threshold ϵ decreases the time consumption in 2.99×10^4 microseconds as average and 18.49×10^3 microseconds considering threshold δ.

In Figure 7 and 8 also shows the number of generated nodes for each indexation criteria for different data volumes in worst scenario. Results indicate that the use of the new indexation criteria in the worst scenario also needs more memory space in RAM than others recursion thresholds.

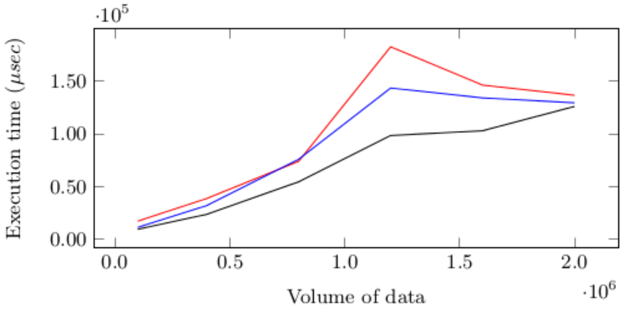


Fig. 7: Time consumption and child generation for different data volume using threshold (β) red, threshold (δ) blue and threshold (ϵ) black.

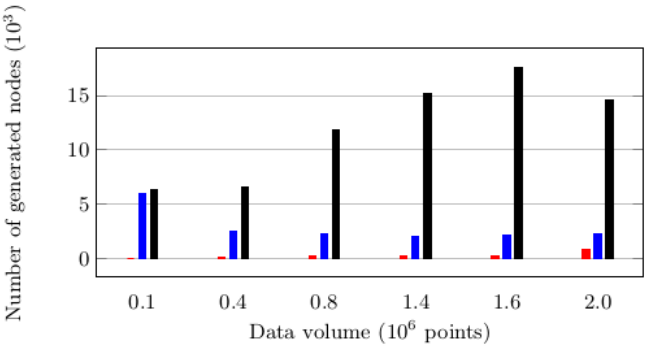


Fig. 8: Another view of time consumption and child generation for different data volume using threshold (β) red, threshold (δ) blue and threshold (ϵ) black.

V. Conclusion and Future Work

In this paper a new recursion threshold is proposed for indexing geometric objects using hierarchical spatial data-structures. The main contribution is a novel heuristic based threshold that performs better than other indexation criteria proposed for general data. Experiments show that considering threshold β threshold ϵ decreases the time consumption in 1.6×10^4 microseconds as average and 8.7×10^3 microseconds considering threshold δ when configuring parameters to the best possible scenario. The results remain positive even when the worst scenario is considered: considering threshold β threshold ϵ decreases the time consumption in 2.99×10^4 microseconds as average and 18.49×10^3 microseconds considering threshold δ. However the proposed threshold tends to generate more nodes in the hierarchical structure.

As future work, an investigation of the ϕ factor is required in different scenarios with the objective of reducing the space complexity of the proposed indexation criteria and more experiments have to be carried out in order to define new scenarios. We also pretend to investigate the effect of the re-indexation when considering our proposed method.

Acknowledgment

We thank colleagues Signal Processing and Geoinformation Research Group. This work was partially supported by, the University Cooperation Network "*Strengthening of ICT role in Cuban Universities for development of the society*", funded by VLIR-UOS.

References

[1] C. Quintana, "Algoritmo para la creación de volúmenes de restricción para búsqueda mínima en modelos de bloques geológicos," Centro Geoinformática y Señales Digitales, Universidad de las Ciencias Informáticas, Tech. Rep., 2013.

[2] J. Chen, "Computational geometry: methods and applications," *Computer Science Department Texas University*, 1996.

[3] Y. Valle Martínez and J. Rojas Ortiz, "Sistemas de información. representación de superficies de terrenos para su visualización en tres dimensiones," *Ciencias de la Información*, vol. 42, no. 3, pp. 57–64, 2011.

[4] D. P. Mehta and S. Sahni, *Handbook of data structures and applications*. CRC Press, 2004.

[5] T. Akenine-Möller, E. Haines, and N. Hoffman, *Real-time rendering*. CRC Press, 2008.

[6] C. M. Hoffmann, *Geometric and solid modeling*. Morgan Kaufmann, 1989.

[7] J. Bai, X. Zhao, and J. Chen, "Indexing of the discrete global grid using linear quadtree," *Proceedings of the XXXVIth ISPRS*, pp. 267–270, 2005.

[8] B. F. Naylor, "Binary space partitioning trees," *Handbook of Data Structures and Applications*, pp. 20–1, 2005.

[9] K. Zhou, Q. Hou, R. Wang, and B. Guo, "Real-time kd-tree construction on graphics hardware," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 5, p. 126, 2008.

[10] S. F. Frisken and R. N. Perry, "Simple and efficient traversal methods for quadtrees and octrees," *Journal of Graphics Tools*, vol. 7, no. 3, pp. 1–11, 2002.

[11] H. Sundar, R. S. Sampath, and G. Biros, "Bottom-up construction and 2: 1 balance refinement of linear octrees in parallel," *SIAM Journal on Scientific Computing*, vol. 30, no. 5, pp. 2675–2708, 2008.

[12] T. Lewiner, V. Mello, A. Peixoto, S. Pesco, and H. Lopes, "Fast generation of pointerless octree duals," in *Computer Graphics Forum*, vol. 29, no. 5. Wiley Online Library, 2010, pp. 1661–1669.

[13] B. Aronov, H. Bronnimann, A. Y. Chang, and Y.-J. Chiang, "Cost-driven octree construction schemes: an experimental study," in *Proceedings of the nineteenth annual symposium on Computational geometry*. ACM, 2003, pp. 227–236.

[14] C. L. Jackins and S. L. Tanimoto, "Octrees and their use in representing three-dimensional objects," *Computer Graphics and Image Processing*, vol. 14, pp. 249–270, 1980.

[15] D. Meagher, "Geometric modeling using octree encoding," *Computer graphics and image processing*, vol. 19, no. 2, pp. 129–147, 1982.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

[16] K. Yamaguchi, T. Kunii, K. Fujimura, and H. Toriya, "Octree-related data structures and algorithms," *IEEE Computer Graphics and Applications*, vol. 4, no. 1, pp. 53–59, 1984.

[17] T. Boubekeur, W. Heidrich, X. Granier, and C. Schlick, "Volume-surface trees," in *Computer Graphics Forum*, vol. 25, no. 3. Wiley Online Library, 2006, pp. 399–406.

[18] W. Saftly, P. Camps, M. Baes, K. Gordon, S. Vandewoude, A. Rahimi, and M. Stalevski, "Using hierarchical octrees in monte carlo radiative transfer simulations," *Astronomy & Astrophysics*, vol. 554, p. A10, 2013.

[19] W. Saftly, M. Baes, and P. Camps, "Hierarchical octree and kd tree grids for 3d radiative transfer simulations," *Astronomy & Astrophysics*, vol. 561, p. A77, 2014.

[20] P. K. Gupta, "Using neural networks for octree generation," *Journal of Multi Disciplinary Engineering Technologies Volume*, vol. 8, no. 1, p. 28, 2014.

[21] N. M. Josuttis, *The C++ standard library: a tutorial and reference*. Addison-Wesley, 2012.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Fejer-Korovkin Wavelet Based MIMO Model For Multi-step-ahead Forecasting of Monthly Fishes Catches

Nibaldo Rodriguez and Lida Barba

Abstract—This paper proposes a Multiples Input-Multiples Ouput Autoregressive (MIMO-AR) model based on two stages to improve monthly anchovy catches forecasting of the coastal zone of Chile for periods from January 1958 to December 2011. In the first stage, the stationary wavelet transform (SWT) based on Fejer-Korovkin (FK) wavelet filter is used to separate the raw time series into a high frequency (HF) component and a low frequency (LF) component. In the second stage, both the HF and LF components are the inputs into a FK+MIMO-AR model to predict the original time series. The performance of the FK-MIMO-AR model is evaluated by comparing its prediction with MIMO-AR model based on SWT with Daubechies (Db) wavelet filter (Db+MIMO-AR). Results show that the FK+MIMO-AR model outperforms the Db+MIMO-AR model in terms of root mean square error, modified Nash-Sutcliffe efficiency and coefficient of determination for 15-month-ahead anchovy catches forecasting.

Index Terms—Wavelet analysis, MIMO model, Forecasting model

1 Introduction

Citizens of fishing countries today (Chile, Peru, China, Japan, New Zealand, Mexico, etc.) are demanding that their governments develop new sustainable policies for the exploitation of fishing resources. However, the development of such policies requires an understanding of the variability of abundance of certain species in the marine ecosystem. The development of models to aid in understanding and predicting fluctuations in abundance of fishing resources is a complex task due to the dynamics underlying the marine ecosystem.

In recent years, linear regression models [1]–[3] and artificial neural network models have been proposed for 1-step-monthly time series forecasts of pelagic species [4], [5]. The disadvantage of linear regression models is the assumption that time series of pelagic species abundance are stationary. Although artificial neural networks can model the non-linear behavior of a time series, they also have some disadvantages due to the learning algorithm based on a descending gradient, as this type of algorithm shows rapid convergence to local minimums during the learning process. Gutierrez et al. [4], [5] proposed multi-layer neural network models to forecast catches in the following month (1-step-ahead) for anchoveta and sardine in northern Chile. The results obtained from the use of a neural network achieved a variance of 87%. In order to better understand the underlying dynamics of fishing resource abundance in Chile, it is necessary to develop new models to explain and predict the oscillatory behavior of pelagic resources along the Chilean coastline.

In recent decades some researchers in order to improve non-stationary time series forecasting models have used the wavelet analysis. The advantage of wavelet analysis is its ability to detect and separate high frequency and low frequency components from a non-stationary time series. After separation, each component is more regular than the original time series,

- *Nibaldo Rodriguez is with the School of Computer Engineering at the Pontificia Universidad Católica de Valparaíso, Av. Brasil 2241,Chile. e-mail: nibaldo.rodriguez@ucv.cl*
- *Lida Barba is with the School of Computer Engineering at the Universidad Nacional de Chimborazo, Av. Antonio Jose de Sucre,Riobamba, Ecuador, lbarba@unach.edu.ec*

https://doi.org/10.17562/PB-56-10

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

which may help improve the forecasting performance [6], [7]. Wavelet analysis has also been evaluated successfully in $one-step-ahead$ forecast models in different areas, such as the electricity market [6], [7], the finance market [8]-[9], smoothing methods [10]-[11] and in ecological time series modeling [12], [13]. In addition, wavelet analysis at different timescales has also been used to show that climate oscillations such as the El Nino-Southern Oscillation significantly affect marine species abundance [14]-[15].

In this paper, a multi-step-ahead forecasting model of monthly anchovy catches is proposed. Our proposed forecasting model is based on two phase. In the first phase, the stationary wavelet transform (SWT) based on Fejer-Korovkin wavelet (FK) filter is used to extract a high frequency (HF) component of intra-annual periodicity and a low frequency (LF) component of inter-annual periodicity. In the second stage, both the HF and LF components are the inputs into a MIMO-AR model to predict the original time series. Besides, the proposed MIMO-AR model is compared with a MIMO-AR model based on SWT with Daubechies wavelet filter [16], [17] denoted as Db+MIMO-AR.

This paper is organized as follows. In the next section, we present hybrid multi-step-ahead forecasting model. The simulation results are presented in Section 3 followed by conclusions in Section 4.

2 Proposed Multi-step-ahead forecasting

In order to predict the future values of time series $x(n)$, we can separate the raw time series $x(n)$ into two components by using SWT. The first extracted component x_H of the time series is characterized by fast dynamics, whereas the second component x_L is characterized by low dynamics. Therefore, in our forecasting model a time series is considered as a functional relationship of several past observations of the components x_L and x_H as follows:

$$\hat{x}(n+h) = F(z(n)) \tag{1}$$

where the h value represents forecasting horizon, m denotes lagged values of both the LF and HF components and $z(n) = [x_L(n), \dots x_L(n-m), x_H(n), \dots x_H(n-m)]$ denotes regressor vector. Besides, the functional relationship $F(\cdot)$ in this paper is estimated by using a MIMO-AR model. The proposed MIMO-AR model calibrates only one MIMO-AR model to predict the h future values. The following equation is used to represent the linear MIMO-AR model:

$$[\hat{x}(n+1), \hat{x}(n+2) \dots, \hat{x}(n+h)] = F[z(n)] + e(n) \tag{2}$$

The MIMO-AR model is used to estimate the function $\hat{F}(\cdot)$. Given a set of training data $z_i, d_i, i = 1, \dots, N$, with $z_i \in R^{2m}$ and $d_i \in R^h$, then the output forecasting in matrix form is obtained as

$$Y = ZA \tag{3}$$

where Y is the matrix dependent variables of N rows by h columns, Z is the regressor matrix of N rows by $2m$ columns and A is the parameters matrix of $2m$ rows by h columns. In order to estimate the parameters A the linear least squares method is used, which is given as

$$A = Z^{\dagger} Y \tag{4}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudoinverse [18].

2.1 Stationary wavelet transform

Let $x(n)$ denote the value of a time series at time n, then $x(n)$ can be represented at multiple resolutions by decomposing the signal on a family of wavelets and scaling functions [10]. The approximation (scaled) signals are computed by projecting the original signal on a set of orthogonal scaling functions of the form:

$$\phi_{jk}(t) = \sqrt{2^{-j}}\phi(2^{-j}t - k) \tag{5}$$

or equivalently by filtering the signal using a low pass filter of length r, $h = [h_1, h_2, ..., h_r]$, derived from the scaling functions. On the other hand, the detail signals are computed by projecting the signal on a set of wavelet basis functions of the form

$$\psi_{jk}(t) = \sqrt{2^{-j}}\psi(2^{-j}t - k) \tag{6}$$

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

or equivalently by filtering the signal using a high pass filter of length r, $g = [g_1, g_2, ..., g_r]$, derived from the wavelet basis functions. Finally, repeating the decomposing process on any scale J, the original signal can be represented as the sum of all detail coefficients and the last approximation coefficient.

In time series analysis, discrete wavelet transform (DWT) often suffers from a lack of translation invariance. This problem can be tackled by means of the un-decimated stationary wavelet transform (SWT). The SWT is similar to the DWT in that the high-pass and low-pass filters are applied to the input signal at each level, but the output signal is never decimated. Instead, the filters are up-sampled at each level.

Consider the following discrete signal $x(n)$ of length N where $N = 2^J$ for some integer J. At the first level of SWT, the input signal $x(n)$ is convolved with the $h_1(n)$ filter to obtain the approximation coefficients $a_1(n)$ and with the $g_1(n)$ filter to obtain the detail coefficients $d_1(n)$, so that:

$$a_1(n) = \sum_k h_1(n-k)x(k) \tag{7a}$$

$$d_1(n) = \sum_k g_1(n-k)x(k) \tag{7b}$$

because no sub-sampling is performed, $a_1(n)$ and $d_1(n)$ are of length N instead of $N/2$ as in the DWT case. At the next level of the SWT, $a_1(n)$ is split into two parts by using the same scheme, but with modified filters h_2 and g_2 obtained by dyadically up-sampling h_1 and g_1.

The general process of the SWT is continued recursively for $j = 1, ..., J$ and is given as:

$$a_{j+1}(n) = \sum_k h_{j+1}(n-k)a_j(k) \tag{8a}$$

$$d_{j+1}(n) = \sum_k g_{j+1}(n-k)a_j(k) \tag{8b}$$

where h_{j+1} and g_{j+1} are obtained by the up-sampling operator inserts a zero between every adjacent pair of elements of h_j and g_j; respectively.

Therefore, the output of the SWT is then the approximation coefficients a_J and the detail coefficients $d_1, d_2, ..., d_J$, whereas the original signal $x(n)$ is represented as a superposition of the form:

$$x(n) = a_J(n) + \sum_{j=1}^{J} d_j(n) \tag{9}$$

The wavelet decomposition method is fully defined by the choice of a pair of low and high pass filters and the number of decomposition steps J.

2.2 Measures of accuracy applied in the model performance

In this study, three criteria of forecasting accuracy called root mean square error (RMSE), modified Nash-Sutcliffe efficiency (mNSE) and coefficient of determination (R2) were used to evaluate the forecasting capabilities of the proposed forecasting models, which are defined as

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\big(x(i) - \hat{x}(i)\big)^2} \tag{10}$$

$$mNSE = 1 - \frac{\sum_{i=1}^{L}\big|x(i) - \hat{x}(i)\big|}{\sum_{i=1}^{L}\big|x(i) - \bar{x}\big|} \tag{11}$$

$$R2 = 1 - \frac{\sum_{i=1}^{L}\big(x(i) - \hat{x}(i)\big)^2}{\sum_{i=1}^{L}\big(x(i) - \bar{x}\big)^2} \tag{12}$$

where $x(i)$ is the actual value at time i, $\hat{x}(i)$ is the forecasted value at time i, $\bar{x}$ is the mean of observed data and L is the number of forecasts.

3 Experiments and Results

In this section, we apply the proposed wavelet MIMO-AR model for multi-step-ahead forecasting. The data set used corresponded to landing of anchovy in the south of Chile. These samples were collected monthly from 1 January 1958 to 31 December 2011 by the National Fishery Service of Chile (www.sernapesca.cl). The raw anchovy data set have been normalized to the range from 0 to 1 by simply dividing the real value by the maximum of the appropriate set. On the other hand, the original data set was also divided into two subsets. In the first subset the 80% of the time series were chosen for

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

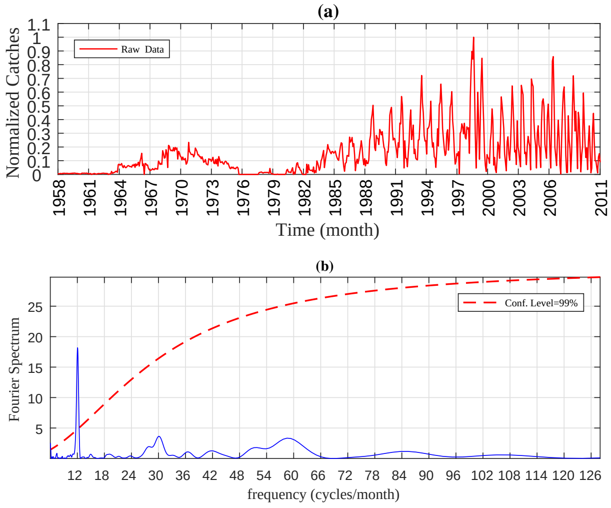


Fig. 1. Monthly anchovy catches

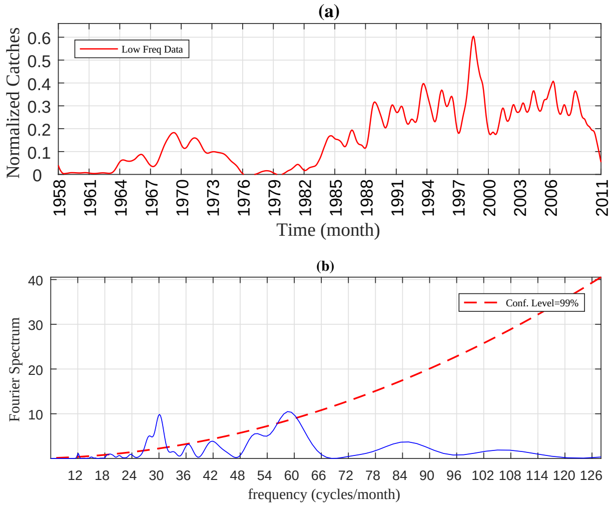


Fig. 2. Low frequency monthly anchovy catches

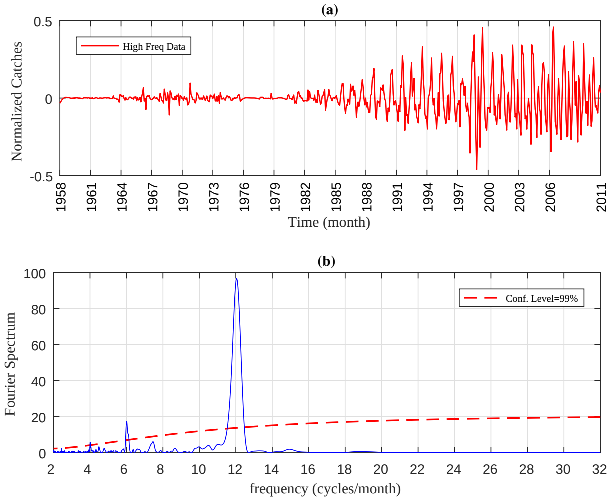


Fig. 3. High frequency monthly anchovy catches

the calibration phase (parameters estimation), whereas the remaining data set were used for the testing phase.
The normalized raw time series and the Fourier power spectrum are present in the Figure 1(a) and 1(b); respectively. The red thick line in Figure 1(b) designates the confidence level against red noise spectrum. From Figure 1(b) it can be observed that there are one peak of significant power, whose peak has periodicity of 12 months. After we applied the Fourier power spectrum to the raw time series, we decided to use 3-level SWT due to the significative peak of 12 months. Both the HF and LF times series are presented in Figures 2 and 3; respectively, whereas the power spectrum of both time series are illustrated in Figure 2(b) and 3(b); respectively. Find the order of the MIMO-AR model is a complex task, but here we will use 30 months due to significant period of the low frequency component.

The multi-step-ahead forecasting methodology used in this paper is based on SWT combined with MIMO-AR model and in order to evaluate the contribution of modeling the monthly anchovy catches using FK+MIMO-AR model, the latter is compared to Db+MIMO-AR model. The SWT is implemented by evaluating a wavelet families: (i) FK4, (ii) FK6, (iii) Daubechies db2 (iv) Daubechies db3. The results of the forecasting performance of different wavelets are reported in Figure 4.From Figure 4 it is observed that the accuracy decreases as the time horizon increases; therefore the best accuracy was obtained for the nearest months, and the lowest accuracy was obtained for the farthest months. Also, from Figure 4 it is seen that the FK4-wavelet (also Db2-wavelet) seems to perform better than other wavelets due to their good localization ability. The MIMO-AR model using FK4-wavelet has a mNSE equal to 90.45% whereas the models based on db2-wavelet and Db3-wavelet yielded results with low mNSE values equal to 80.98% and 67.46%,respectively. The Figures 5 and 6 show the results obtained with the MIMO-AR(30) for 15-month-ahead anchovy catches forecasting during the testing phase. Figure 5(a) provides

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

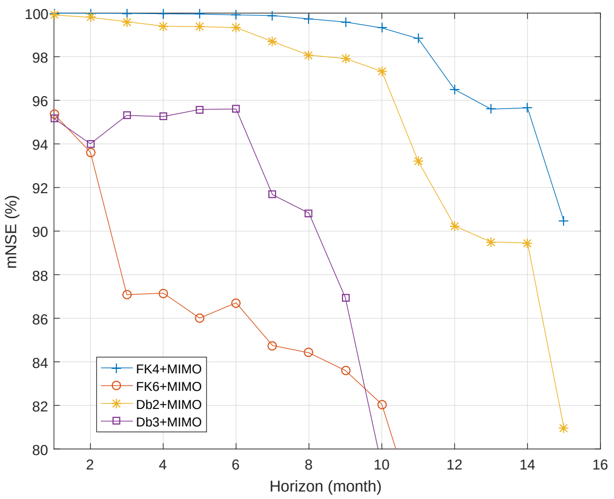


Fig. 4. FK+MIMO-AR versus Db+MIMO-AR

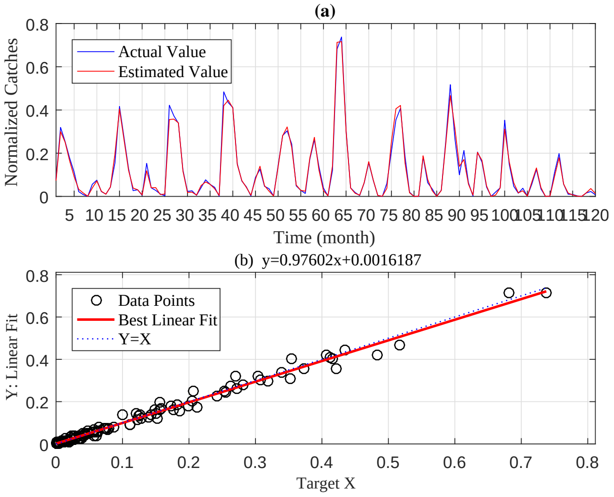


Fig. 5. FK4+MIMO-AR: Fifteen-month-ahead forecasting for test data set

data on observed monthly anchovy catches versus forecasted catches;this forecasting behavior is very accurate for testing data with a RMSE 0.017 and a mNSE of 90.45%. On the other hand, from Figure 5(b) it can be observed a good fit to a linear curve with a coefficient of determination of 98.63%.

Figures 6(a) and 6(b) show the results obtained with the Db2+MIMO-AR(30) forecasting model during the testing phase. Figure 6(a) illustrates the observed data set versus forecasted data set, which obtains a RMSE and a mNSE of 0.034 and 81%; respectively. On the other hand, Figure 6(b) shows the scatter curve between observed values and forecasted values with a R2 of 94.75%.

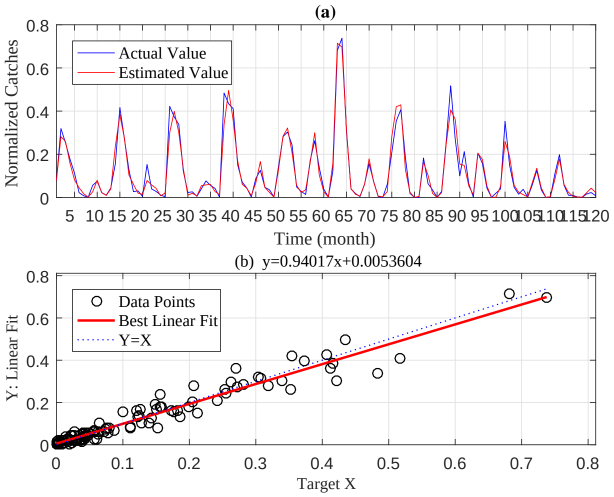


Fig. 6. Db2+MIMO-AR: Fifteen-month-ahead forecasting for test data set

4 Conclusions

In this paper was proposed a multi-step-ahead forecasting model to improve prediction accuracy based on stationary wavelet decomposition combined with MIMO-AR model. The reason of the improvement in forecasting accuracy was due to use Fejer-Korovkin wavelet filter to separate both the LF and HF components of the raw time series, since the behavior of each component is more smoothing than raw data set. It was show that the proposed FK4+MIMO-AR model achieves a mNSE of 90.45% and a R2 of 98% for 15-month-ahead anchovy catches forecasting. Besides, the experimental results demonstrated a better performance of the proposed model when compared with a Db2+MIMO-AR prediction model. Finally, hybrid forecasting model can be suitable as a very promising methodology to any other pelagic species.

Acknowledgments

This research was partially supported by the project DI-Regular 2016 of the Pontificia Universidad Católica de Valparaíso.

References

[1] K. Stergiou, E. Christou, and G. Petrakis, "Modelling and forecasting monthly fisheries catches: comparison of regression, univariate and multivariate time series methods," *Fisheries Research*, vol. 29, no. 1, pp. 55-95, 1997.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

[2] K. Stergiou and E. Christou, “Modelling and forecasting annual fisheries catches: comparison of regression, univariate and multivariate time series methods,” *Fisheries Research*, vol. 25, pp. 105-138, 1996.

[3] K. Stergiou, “Short-term fisheries forecasting: comparison of smoothing, arima and regression techniques,” *Journal of Applied Ichthyology*, vol. 7, pp. 193-204, 1991.

[4] J. Gutirrez-Estrada, C. Silva, E. Yanez, N. Rodrguez, and I. Pulido-Calvo, “Monthly catch forecasting of anchovy engraulis ringens in the north area of chile: Non-linear univariate approach,” *Fisheries Research*, vol. 86, no. 23, pp. 188-200, 2007.

[5] E. Yanez, F. Plaza, J. G. ., N. Rodriguez, and et al., “Anchovy (engraulis ringens) and sardine (sardinops sagax) abundance forecast of northern chile: A multivariate ecosystemic neural network approach,” *Progress in Oceanography*, vol. 87, no. 12, pp. 242-250, 2010.

[6] N. Shrivastava and B. Panigrahi, “A hybrid wavelet-elm based short term price forecasting for electricity markets,” *International Journal of Electrical Power Energy Systems*, vol. 55, pp. 41-50, 2014.

[7] A. Nima and K. Farshid, “Day ahead price forecasting of electricity markets by a mixed data model and hybrid forecast method,” *nternational Journal of Electrical Power Energy Systems*, vol. 30, no. 9, pp. 533-546, 2008.

[8] S. Lahmiri, “Forecasting direction of the s-p500 movement using wavelet transform and support vector machines,” *Int. J.Strateg. Decis. Sci*, vol. 4, pp. 78-88, 2013.

[9] Z. Bai-Ling, C. Richard, M. Jabri, D. Dersch, and B. Flower, “Multiresolution forecasting for futures trading using wavelet decompositions,” *IEEE Transaction on neural networks*, vol. 12, no. 4, pp. 765-775, 2001.

[10] G. Nason and B. Silverman, “The stationary wavelet transform and some statistical applications,” *Lecture Notes in Statistics: Wavelets and Statistics*, pp. 281-299, 1995.

[11] D. Percival and A. Walden, “Wavelet methods for time series analysis,” *Cambridge, England: Cambridge University Press*, 2000.

[12] M. Hidalgo, T. Rouyer, J. Molinero, E. Massut, and et al., “Synergistic effects of fishing-induced demographic changes and climate variation on fish population dynamics,” *Marine Ecology Progress Series*, vol. 426, pp. 1-12, 2011.

[13] B. Cazelles, M. Chavez, D. Berteaux, F. Mnard, and et al., “Wavelet analysis of ecological time series,” *Oecologia*, vol. 156, pp. 287-304, 2008.

[14] T. Rouyer, J. Fromentin, N. Stenseth, and B. Cazelles, “Analysing multiple time series and extending significance testing in wavelet analysis,” *Marine ecology progress series*, vol. 359, pp. 11-23, 2008.

[15] S. Lluch-Cota, A. Pars-Sierrab, and et al., “Changing climate in the gulf of california,” *Progress in Oceanography*, vol. 87, no. 14, pp. 114-126, 2010.

[16] I. Daubechies and S. Maes, “A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models. wavelets,” *Wavelets in Medicine and Biology*, pp. 527-546, 1996.

[17] I. Daubechies, “Ten lectures on wavelets,” in *SIAM monographs, Philadelphia, PA:SIAM*, 1992.

[18] D. Serre, *Matrices:Theory and Applications*. Springer, NewYork, NY, USA, 2002.

IMPORTANT: This is a pre-print version as provided by the authors, not yet processed by the journal staff. This file will be replaced when formatting is finished.

ISSN 2395-8618

Journal Information and Instructions for Authors

I. Journal Information

Polibits is a half-yearly open-access research journal published since 1989 by the *Centro de Innovación y Desarrollo Tecnológico en Cómputo* (CIDETEC: Center of Innovation and Technological Development in Computing) of the *Instituto Politécnico Nacional* (IPN: National Polytechnic Institute), Mexico City, Mexico.

The journal has double-blind review procedure. It publishes papers in English and Spanish (with abstract in English). Publication has no cost for the authors.

A. Main Topics of Interest

The journal publishes research papers in all areas of computer science and computer engineering, with emphasis on applied research. The main topics of interest include, but are not limited to, the following:

- Artificial Intelligence
- Natural Language Processing
- Fuzzy Logic
- Computer Vision
- Multiagent Systems
- Bioinformatics
- Neural Networks
- Evolutionary Algorithms
- Knowledge Representation
- Expert Systems
- Intelligent Interfaces
- Multimedia and Virtual Reality
- Machine Learning
- Pattern Recognition
- Intelligent Tutoring Systems
- Semantic Web
- Robotics
- Geo-processing
- Database Systems
- Data Mining
- Software Engineering
- Web Design
- Compilers
- Formal Languages
- Operating Systems
- Distributed Systems
- Parallelism
- Real Time Systems
- Algorithm Theory
- Scientific Computing
- High-Performance Computing
- Networks and Connectivity
- Cryptography
- Informatics Security
- Digital Systems Design
- Digital Signal Processing
- Control Systems
- Virtual Instrumentation
- Computer Architectures

B. Indexing

The journal is listed in the list of excellence of the CONACYT (Mexican Ministry of Science) and indexed in the following international indices: Web of Science (via SciELO citation index), LatIndex, SciELO, Redalyc, Periódica, e-revistas, and Cabell's Directories.

There are currently only two Mexican computer science journals recognized by the CONACYT in its list of excellence, *Polibits* being one of them.

II. Instructions for Authors

A. Submission

Papers ready for peer review are received through the Web submission system on www.easychair.org/conferences/?conf=polibits1; see also updated information on the web page of the journal, www.cidetec.ipn.mx/polibits.

The papers can be written in English or Spanish. In case of Spanish, author names, abstract, and keywords must be provided in both Spanish and English; in recent issues of the journal you can find examples of how they are formatted.

The papers should be structures in a way traditional for scientific paper. Only full papers are reviewed; abstracts are not considered as submissions. The review procedure is double-blind. Therefore, papers should be submitted without names and affiliations of the authors and without any other data that reveal the authors' identity.

For review, a PDF file is to be submitted. In case of acceptance, the authors will need to upload the source code of the paper, either Microsoft Word or LaTeX with all supplementary files necessary for compilation. Upon acceptance notification, the authors receive further instructions on uploading the camera-ready source files.

Papers can be submitted at any moment; if accepted, the paper will be scheduled for inclusion in one of forthcoming issues, according to availability and the size of backlog.

See more detailed information at the website of the journal.

B. Format

The papers should be submitted in the format of the IEEE Transactions 8x11 2-column format, see http://www.ieee.org/publications_standards/publications/authors/author_templates.html. (while the journal uses this format for submissions, it is in no way affiliated with, or endorsed by, IEEE). The actual publication format differs from the one mentioned above; the papers will be adjusted by the editorial team.

There is no specific page limit: we welcome both short and long papers, provided that the quality and novelty of the paper adequately justifies its length. Usually the papers are between 10 and 20 pages; much shorter papers often do not offer sufficient detail to justify publication.

The editors keep the right to copyedit or modify the format and style of the final version of the paper if necessary.

See more detailed information at the website of the journal.