Combining Wavelets and Computational Intelligence Methods with applications on Multi-class Classification datasets.

Carlos C. Bracho[1]

1 Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada
Email: carloscb@ece.ualberta.ca

Abstract—*In this paper, we propose a novel algorithm for wavelet feature extraction as input to a supervised Multi-Class Classifier to improve classification performance. In particular, to select the best wavelets coefficient features, we first compute the energy-based variance distribution from wavelets coefficients at different subbands as well as the entropy-based fuzzy measures associated with the training instances. Once we get these entropy-based fuzzy measures associated with the different subsets of wavelets subbands, we apply the Möbius Transform to these entropy-based fuzzy measures to extract the Multivariate Mutual Information associated with the different subsets of wavelets subbands. The goal of these measures is twofold: assign weights (based on the wavelets information content) to all subsets of wavelets subbands and extract the independent (in terms of the multivariate mutual information) subsets of wavelets subbands. In our case, the optimal subsets of wavelets subbands as wavelets features vectors to train a Bayesian Network Model are those which provide a multivariate mutual information equal to zero. Experimental results with the multi-class SRBCT cancer dataset, show that our proposed approach achieves lower classification error in comparison with other methods proposed in the literature.*

Keywords— fuzzy measures, mutual information, wavelets.

1 Introduction

In recent years, Wavelets-based classification have gained a great relevance in different kinds of applications such as image processing [6][7], time series[8][2], bioinformatics[9][12], and pattern recognition[10][11]. The early research in wavelets-based classification was focused mainly on extracting energy values from the wavelet-subbands decomposition and using them for classification. However, in pattern recognition tasks such as classification, it is well known that proper feature selection help to improve the classification performance. Hence, a process which consists on removing irrelevant wavelets features as well as selecting a subset of wavelets from the wavelet-subbands decomposition has had a paramount importance in Wavelets-based Classification tasks.

In this paper, we propose a novel algorithm that uses entropy-based fuzzy measures and multivariate mutual information–based tests to identify the suitable subsets of wavelets features as the correct input to a supervised K2-based Bayesian Network to improve classification performance. At present time, Bayesian Networks [13[14] have became one of the most important technologies in the area of applied artificial intelligence. Roughly speaking, Bayesian Networks are graphical structures that model the probabilistic cause-effect relationships among several related variables. Particularly, the K2 Algorithm[15] is a greedy algorithm that learns the structure of Bayesian Networks from data in an efficient way given a prior order of the variables (in improper order of data will provide poor classification performance). In our case, we are proposing entropy-based fuzzy measures, its Möbius representation and multivariate mutual information to provide not only the correct order but also the correct wavelets features to a K2-based Multi-Class Bayesian Network for classifying a multi-class cancer dataset.

The remaining of this paper is organized as follows: in Section 2, we provide background on Wavelets, Bayesian Networks, Fuzzy Measures, Entropy-based Fuzzy Measures and Fuzzy Measures-based Multivariate Mutual Information. In Section 3, we detail our multi-class classification approach. In Section 4 we present the experimental results obtained using a cancer dataset with our proposed approach. Finally, in Section 5, we offer a summary and discussion of future work.

2 Background

2.1 The Wavelet Transform

The Wavelet Transform (WT)[2] is a natural context for analyzing the time-varying properties of most real-world signals. In a narrow sense, WT decomposes a signal into different time scales (wavelet- subbands decomposition) to detect features of its short-term as well as its long-term dynamics. Essentially, this transform is defined in two spaces[2]: continuous and discrete. In the continuous space, the Continuous Wavelet Transform CWT is defined by means of two basic functions (known as the mother and the father wavelet function, respectively). The father wavelet function $\varphi(t)$ depending of time t is used to capture the smooth part of the data. By contrast, the mother wavelet function $\psi(t)$ is stretched and translated to detect the rough components of the data. These two functions, to be considered as wavelet functions, must satisfy certain properties [2].

Complementary to the CWT described above, the Discrete Wavelet Transform (DWT) can be defined from an engineering point of view by means of cascading a set of couple of low and high filters [2]. These couples of filters are known in the literature as the wavelet and scaling filters. Together, this set of filters must fulfill the Quadrature Mirror Filter (QMF) property to obtain a perfect reconstruction pass-band filter.

Of particular interest in this paper, is the application of a modified version of DWT known as the Maximal Overlap

ISBN: 978-989-95079-6-8

Wavelet Transform (MODWT)[2], to provide wavelets features to a K2-based Bayesian Network. Roughly speaking, MODWT is a rescaled version of the wavelet and scaling filter used in the common DWT but with the following advantages:

- Wavelet coefficients are aligned with the original signal values.
- MODWT provides a suitable format for the traditional tabular framework used in machine learning approaches.
- MODWT provides more redundancy at each wavelet-subband.

Definition 1
Let *j*, *t* be the decomposition level (wavelet subband), and time index, respectively, and X_t a real-valued uniformly sampled signal whose size is *N*. The j^{th} level MODWT wavelet and scaling coefficient $W_{j,t}$ and $V_{j,t}$, respectively, are defined by convolving the impulse response of the wavelet and scaling filters with the signal X_t in the following way [2]:

$$W_{j,t} = \sum_l h_{j,l} * X_{t-l \bmod N} \quad (1)$$

$$V_{j,t} = \sum_l g_{j,l} * X_{t-l \bmod N} \quad (2)$$

Where *l* is the length of both the wavelet filter impulse response *h* and the scaling impulse response *g*. In this paper, MODWT was implemented by means of a Pass-Band Wavelet Filter (using the WMTSA matlab toolbox [2]). In particular, we use a Daubechies Extremal Phase Filter with *l=6* (DB6). One fundamental property of this filter is that it transforms a signal in terms of difference of several order of averages at different wavelet-subbands.

2.2.1 MODWT Energy-based variance distribution
Another important property of WT is its ability to preserve the energy of a signal. Roughly speaking, the WT decomposes the total energy of a signal at different wavelet-subbands defining an Energy Density Distribution (EDD). In particular, this EDD is computed by means of wavelets and scaling coefficients of MODWT expressed as follows[2]:

$$\sum_t X_t^2 = \sum_j \sum_{n=0}^{N_j-1} W^2{}_{j,t} + \sum_{n=0}^{j-1} V^2{}_n \quad (3)$$

A closely related concept to the wavelet-based energy decomposition is the wavelet variance which can be associated with a probability distribution and used in this paper for purposes of computing the entropy-based fuzzy measures. Mathematically, this wavelet variance is defined as follows [2]:

$$\sigma_X^2 = \frac{1}{N} \sum_{j=1}^{MaxLevel} \left\| W_j \right\|^2 + \frac{1}{N} \left\| V_{MaxLevel} \right\|^2 - \overline{X}^2 \quad (4)$$

Where:

N: Number of observations in the signal.
$MaxLevel$: Maximum level of decomposition .

W_j : Vector of wavelet coefficients defined at subband j .

$V_{MaxLevel}$: Vector of scaling coefficients defined at subband j .

$\overline{X}$: Sample mean of signal X_t.

2.2 Bayesian Networks
Bayesian Networks [16][13] have become one of the most important technologies in the area of applied artificial intelligence. They have shown to offer reliable methods for prediction, decision making, classification, and data mining in different areas such as medicine, image processing, marketing, banking, finance, etc. Roughly speaking, Bayesian Networks are graphical structures that model the probabilistic cause-effect relationships among several related variables. Mathematically, a Bayesian Network is modelled by means of a Directed Acyclic Graph (representing the above cause-effect relationships) and a set of Conditional Probabilities (one conditional probability for each node given its parent set in the graph).

Definition 2
A Bayesian Network $B = (\chi, G, P)$ consists of :

- A Direct Acyclic Graph $G = (V, E)$ with nodes $V = (V_1, V_2, .., V_n)$.
- A set of discrete random variables χ modeling the nodes of G.
- A set of conditional probability distributions $P(X_v \mid X_{parent(v)})$ for each random variable $X_v \in \chi$ where :

$$P(\chi) = {}_{v \in V} P(X_v \mid X_{pa(v)}) \quad (5)$$

The last property specifies a joint probability distribution over X known as the Bayesian Network chain rule.

ISBN: 978-989-95079-6-8

2.2.1 K2-based Bayesian Network learning

Generally speaking, the structure learning process under the K2 algorithm is made by means of an optimization process in which a quality measure Q of a Bayesian Network structure given a training dataset X is maximized; that is:

$$\max Q(B_S \mid X) \tag{6}$$

K2 is a greedy search algorithm that learns from data presenting a good performance when a prior order of the data exists. This order refers to have first in a sequence the parents and then the children. Assuming that the attributes have an order, the algorithm starts setting the parent of node x_1 to the empty set $\varnothing$. Then, the algorithm visists each subsequent node following the order in the sequence order adding $Parents_i$ to the set of parents of node attribute x_i only if the inclusion of the $Parents_i$ to the node attribute x_i maximize the network-structure posterior probability distribution. At the same time, it includes the relationship between the parent and the children in question by adding arcs between them.

2.2.2 Fuzzy Measures

Fuzzy measures are mainly known in the context of Fuzzy Integrals[3][5]. Grabisch[4], presents the usage of fuzzy measures and fuzzy integral in supervised classification and feature extraction problems. In these papers the author proposes an optimization approach which is based on a heuristic least mean squares (HLMS) algorithm, to compute the fuzzy measures as well as the Choquet Integral in classification tasks obtaining remarkable results.

Definition 3

Let the triple $K = (\Omega, A, g)$ a fuzzy measure space containing a family A of subsets defined on Ω. A set-valued function $g : A \to 0,1]$ is called a fuzzy measure if:

$$g(\varnothing) = 0 \tag{7}$$

$$g(S) = 1 \tag{8}$$

$$A_i \subset A_j \to g(A_i) \le g(A_j). \forall A_i, A_j \in S \tag{9}$$

Any set-valued function $g : A \to [0,1]$ can be uniquely expressed in terms of its MöbiusTransform represented by:

$$g(T) = \sum_{S \in T} m^g(S), \ \forall T \subseteq A, \tag{10}$$

where the set function $m^g : A \to [0,1]$ is called the Möbius Inversion Transform of g and is given by[17]:

$$m^g(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} g(T), \ \forall S \subseteq A, \tag{11}$$

2.2.2.1Entropy-based fuzzy measures

In comparison with the Grabish approach, in this paper we are interested in applying and extending a novel unsupervised approach to compute entropy-based fuzzy measures proposed by Kojadinovic[1]. The main characteristic of this approach is the replacement of the notion of attributes importance with the notion of information content in attributes by means of the entropy and the mutual information concepts.

Definition 5

Let 2^S denote the power set of $S = \{W_1, W_2, .., W_m\}$. An entropy-based fuzzy measure is a set-valued function $g_H : 2^S \to [0,1]$ defined as follows[1]:

$$g_H(S) = \begin{cases} 0 & S = \varnothing \\ (-1)^{|S|+1} H(P_{(\overleftarrow{W}_1,..)}) & S = \{\overleftarrow{W}_1,.\} \end{cases} \tag{12}$$

where :

$H(P_{(\overleftarrow{W}_1,..)})$: Entropy associated with the joint probability distribution $P_{(\overleftarrow{W}_1,..)}$). (In our case W_m means the maximum wavelet-subband, $P_{(\overleftarrow{W}_1,..)}$ refers to the joint probability distribution among the different subset of wavelet-subbands, and $H(P_{(\overleftarrow{W}_1,..)})$ to their associated entropy).

Definition 6

Any entropy-based fuzzy measure $g^H : 2^S \to [0,1]$ can be estimated by:

$$\tilde{g}^H = \frac{H(P_{(W_{i1},..,W_{is})})}{H(P_{(W_{i1},..,W_{iM})})} \tag{13}$$

Definition 7

Any entropy-based fuzzy measure $g^H : 2^S \to [0,1]$ can be uniquely expressed in terms of its Möbius Transform represented by:

$$g^H(S) = \begin{cases} 0 & S = \varnothing \\ (-1)^{|S|+1} I(\overleftarrow{W}_1;..) & S = \{\overleftarrow{W}_1;..\} \end{cases} \tag{14}$$

where :

I : *Multivariate Mutual Information*

ISBN: 978-989-95079-6-8

Proposition 1
A set of wavelets subbands is called independent if its Möbius Transform associated with its entropy-based fuzzy measure is equal to zero.

Proof. From equation 7, clearly $(-1)^{|S|+1} \neq 0$ for all S. Hence, $I(\overleftarrow{X}_1;..\overleftarrow{X}_r) = 0$ implies $g^H(S) = 0$ ■.

3 Proposed Approach

The proposed approach aims to identify the suitable wavelet information as the correct input for the the Bayesian Network. By contrast with other wavelet feature extraction approaches reported in the literature (in which most of them deal at the granularity of wavelets coefficients), this approach presents a novel approach which deals with sets of wavelets-subbands using entropy-based fuzzy measures, its Möbius Transform and multivariate mutual information. To compute the entropy-based fuzzy measures associated with each set of wavelets-subbands, we first consider the energy-based variance distribution at different wavelet-subbands (this Energy Distribution is discretized to fulfill the monotonicity property of fuzzy measure) as random vectors. This set of energy-based variance distributions are used to estimate the entropy-based fuzzy measures associated with the different sets of wavelets subbands. Once we have defined the set of entropy-based fuzzy measures, we apply the Möbius Inversion Transform to these entropybased fuzzy measures and then we select those sets of wavelets subbands whose Möbius Inversion Transform is equal to zero (i.e. set of independent wavelets-subbands in all cancer classes whose multivariate mutual information is equal to zero). This set of wavelets subbands are the wavelet-training and wavelet-testing to our supervised Bayesian Network Model (see Algorithm I and II).

Algorithm I: Training Phase

1.- Data Preprocessing withWavelet Transform

- Apply Daubechies (DB6) Extremal Phase Filter to the training dataset.

2.- Energy-based Variance Distribution Estimation and discretization

- Compute equations (3) and (4).

3.- Estimate entropy-based Fuzzy Measures

- Compute the Wavelet-subbands Join Probability Distributions with the maximum likelihood estimator[1].
- Compute the entropy associated with the above Wavelet-subbands Joint Probability Distributions.
- Compute entropy-based fuzzy measures using equation (13).

4.- Apply the Möbius Transform to the entropy-based Fuzzy Measures obtained in step 3 to obtain their associated Multivariate Mutual Information

- Compute equations (14).

5.- Wavelet Feature Vector Extraction to train Bayesian Network.

- Select sets of wavelet-subbands whose Möbius Transform is equal (or aprox.) to zero (Proposition 1).
- Concatenate wavelets coefficients of different wavelet-subbands that fulfill the above condition. These wavelets features are the input to train a Bayesian Network. (If exists several sets of wavelet-subbands that fulfill the above condition, you might select the subset with less wavelet coefficients)

End Algorithm I

Algorithm II: Testing Phase

1.- Data Preprocessing withWavelet Transform

- Apply Daubechies (DB6) Extremal Phase Filter to the testing dataset.

2.- Select the same set of wavelet-subbands obtained in step 5 of algorithm I.

- Apply 10-fold cross validation to verify performance of the trained Bayesian Network with the selected wavelet coefficients.

End Algorithm II

4 Experimental results

Given the great interest in the scientific community in exploring DNA microarray which in general consist of thousands of genes, to evaluate the performance of our proposed approach, we use a highly and noisy multidimensional multi-class microarray dataset. In particular, we use a cancer microarray dataset known as SRBCT[9] consisting of 2,308 genes and 63 samples from four types of cancers: neuroblastoma (NB), rhabdomyosacoma (RMS), Burkitt lymphomas (BL) and Ewing family of tumors(EWS). In our case, we consider a total of 32 instances for the training phase partitioned uniformly with respect to each kind of cancer (8 samples for each type of cancer) and 31 instances as the testing dataset (eight samples from each type of cancer except one with seven instances as testing corresponding to the Burkitt lymphomas).

ISBN: 978-989-95079-6-8

In the training phase after applying the first four steps, we found that the wavelet-subband 5 (containing 2308 wavelet coefficients) provide us a multivariate mutual information equal to zero fulfilling the condition established in proposition 1. Once we apply step 5 of algorithm 1, we use a 10-fold cross-validation method and consider an input SRBCT-training dataset to the Multi- Class Bayesian Network (which is implemented. in the data mining environment known as weka [19]) in the following way:

$$T = \left\{ \left(\overleftarrow{W}_i, y_i \right) : \overleftarrow{W}_i \in \Re^{2308} \right\} \tag{14}$$

where:

$$\overleftarrow{W}_i \in \Re^{2308} \tag{15}$$

$$y_i \in \{NB, RMS, BL, EWS\} \tag{16}$$

Once trained, we test our Bayesian Network classifier obtaining the following results reported in table I (in comparison with other results reported in the literature, in this table we can see that there are two methods providing 100% of accuracy partitioning the multi-class problem as a collection of N binary sub-problems. In our case, we solve the multi-class problem without using this artefact):

Table I. Classification Performance

Method	Accuracy %
Our Method	**92%**
KNN	30%
Naive Bayes	60%
lvs1 SVM/Wavelet[9]	100%
SVM-OVA	80%
Decision Tree	75%

5 Conclusions

We have reported experimental results using a novel wavelet-computational intelligence approach for improving classification of multi-class datasets. In particular, we propose a novel approach for extracting entropy-based fuzzy measures from the wavelet-subband decomposition as well as their multivariate mutual information (by means of the Möbius Inversion Transform). With this information, we select the most suitable wavelets (based on the concept of independence given in terms of multivariate mutual information) to train and evaluate a multiclass Bayesian Network to classify cancer types obtaining satisfactory results. In conclusion, the propose approach provides a general framework to deal with data which are inherently noisy and high dimensional such as microarray gene expression data. Future work in this research will proceed in refining and validating the proposed approach in other applications.

Acknowledgment

This work was partially supported by CONACYT Mexico. The author also likes to thank all the reviewers of this paper for their valuable comments.

References

[1] I. Kojadinovic, Unsupervised aggregation of commensurate correlated attributes by means of the Choquet Integral and Entropy Functionals. *International Journal of Intelligent Systems*. Vol 23, pp. 128-154. 2008.

[2] Donald B. Percival, A. T.Walden,*Wavelet Method for Time Series Analysis*. Cambridge Series in Statistical and Probabilistics Mathematics. 2000.

[3] G. Choquet, Theory of Capacities. *Annales de l'Institut Fourier*, Vol 5, pp. 131-295. 1953.

[4] M. Grabisch*, Fuzzy integral for classification and feature extraction.* Fuzzy measures and integrals Theory and applications. Physica Verlag, pp. 135-148. 2000.

[5] G. Beliakov, *Aggregation Functions: A guide for Practitioners*. Springer Studies in Fuzziness and Soft Computing. 2007.

[6] K. Huang and S. Aviyente. Information-theoretic wavelet packet subband selection for texture classification.*Signal Processing* 86, 1410-1420, 2006.

[7] M. Shoshany.Wavelet decomposition for reducing flux density effects on hyperspectral classification. *IEEE Geoscience and Remote Sensing Letters*, v6, No. 1, 38-41, January 2009.

[8] T. Qian. *Wavelet Analysis and Applications*. Birkhauser, First Edition, January 2007.

[9] S. Prabakaran, R. Sahu, S. Verma. Classification of multi class dataset using wavelet power spectrum. *Data Mining Knowledge Discovery*, v15, 297-319, 2007.

[10] D. Li,W. Pedrycz, N. Pizzi. FuzzyWavelet packet based feature extraction method and its application to biomedical signal classification. *IEEE Transactions on Biomedical Eng*, v52-6, 1132- 1139, June 2005.

[11] T. Celik, and T. Tjahjadi. Multiscale texture classification using dual-tree complex wavelet transform. *Pattern Recognition Letters*, v30-3, 331-339, February 2009.

[12] C. Aggarwal, and P. Bradley. On the use of Wavelet Decomposition for String Decomposition. *Data Mining Knowledge Discovery*, v10-2, 117-139, 2005.

[13] U. Kjaerulff, and A. Madsen. *Bayesian Networks and Influence Diagrams*. Information Science and Statistics, Springer, 2008.

[14] X. Chen, G. Anantha, X. Lin. Improving Bayesian Network Structure Learning with Mutual Information-Based Node Ordering in the K2 Algorithm. *IEEE Transactions on Knowledge and Data Eng*, v20-5, 1-13, May, 2008.

[15] G. Cooper and E. Herskovits. A Bayesian Method for the induction of Probabilistic Networks from Data. *Machine Learning*, v9, 309-347, 1992.

[16] R. Neapolitan and X. Jian. *Probabilistics methods for financial and marketing informatics*. Morgan and Kauffman publishers, 2007.

[17] H. Nguyen and E. Walker. *A first course in Fuzzy logic*. Chapman & Hall/CRC , 2006.

[18] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[19] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, Morgan Kauffman, 2005.

ISBN: 978-989-95079-6-8