International Journal of Computational Intelligence and Informatics, Vol. 5: No. 2, September 2015

An Efficient Supervising Technique in Viral Hepatitis Surveillance System

S.R.Swarnalatha
Research scholar,
Manonmaniam Sundaranar University,
Tirunelveli, Tamilnadu, India.
Swarni_r@yahoo.co.in

G.M.KadharNawaz
Director
Department of Computer Applications
Sona College of Technology
Salem, Tamilnadu, India

***Abstract*- Dimension reduction is a critical data preprocessing step for many database and data mining applications, such as efficient classification of relevant and irrelevant groups in high-dimensional data. In the literature, a well-known dimension diminish algorithm is Linear Discriminant Analysis (LDA), Singular Value Decomposition (SVD). Due to the design complexity and poor efficiency, we need a design that should be able to help them to make a good decision. In this paper, we propose an LDA-based Fuzzy classifier, called Diminish Fuzzy (DF) classifier, which applied Medical Research. This method is used to reduce the risk of error in medicine field, especially Hepatitis Diagnosis. The proposed system is an intelligent system for the diagnosis of Hepatitis B virus disease. Hepatitis is one of the serious diseases which demands expensive treatment and major side effects can appear very often. The intelligent system consists of DF classifier which gives the output whether the patient is Hepatitis B positive or not and the severity of the patient. Finally we evaluate the effectiveness of the DF classifier in terms of classification error rate on the reduced dimensional space. Our experiments based on real-world data sets reveal that the classification rate achieved by the DF Classifier algorithm is better than other LDA-based algorithms.**

Keywords- Dimension reduction, Linear Discriminant Analysis, Singular Value Decomposition, Diminish Fuzzy classifier, Classifier algorithm.

I. INTRODUCTION

Data mining technique is very useful in the process of knowledge discovery in the medical field [1]. It is a process of extracting patterns from data. The process is becoming an increasingly important tool to transform this data into information [2]. It is commonly used in a wide range of profiling practices, such as surveillance, marketing, fraud detection and scientific discovery [3]. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis provides a better understanding of the data at large. There are many classification techniques available. They are decision tree Classification, Bayesian Classification, Rule-Based Classification, Classification by Back Propagation, Genetic Algorithms, Rough Set Approach and Fuzzy Set Approaches. However, many classification techniques for classification, fuzzy methods are very simpler and easier to understand.

A prediction or forecast is a statement about the way things will happen in the future, often but not always based on experience or knowledge. The term prediction is referred as both numeric prediction and class label prediction. Regression analysis is a statistical methodology that is most often used for numeric prediction. In this project, class label prediction is used. Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded. In this project, a Diminish fuzzy rule-based classification is used to classify the Hepatitis gene expression data.

Various classification techniques are used for classification. Among the classification techniques Genetic programming, Majority Voting Genetic Programming Classifier (MVGPC) and fuzzy methods are rule based classification techniques. Some classification techniques are given to understand classification techniques.

II. RELATED WORK

SVM (Support Vector Machine) is a technique; it is used for data classification. SVM models are closely related to neural networks. SVMs arose from statistical learning theory [5][15]; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step. SVMs are based on the structural risk minimization principle, closely related to regularization theory. The hyper plane can be found in the original dataset (and this is referred to as linear SVMs) [7] or it can be found in a higher-dimensional space

ISSN: 2349-6363

by transforming the dataset into a representation having more dimensions (input variables) than the original dataset (referred to as nonlinear SVMs). Mapping the dataset, in this way, into a higher dimensional space, and then reducing the problem to a linear problem, provides a simple solution.

Genetic programming is used for the classification of gene expression data. Genetic programming (GP) [4] is an extension of the genetic algorithm in which genetic population consists of computer programs. The main advantage of GP is that it can act as a classifier as well as a gene selection algorithm.

III. PROBLEM STATEMENT

To develop an intelligent system to predict Hepatitis using gene expression data and to improve the classification accuracy through fuzzy methods. The existing system for Hepatitis classification needs to enhanced to improve accuracy in decision making in hepatitis disease. The objective of this work is to classify Hepatitis gene expression data by using linear dicriminant analysis and set of fuzzy if-then rules that enable accurate nonlinear classification of input patterns.

IV. PROBLEM DESCRIPTION

There are three types of Hepatitis namely Hepatitis A, Hepatitis B and Hepatitis C. These Viral diseases affect the liver and having different types of symptoms. This type of data set is downloaded from UCI repository machine learning system. The data are preprocessed according to data mining techniques. According to the problem fuzzy if-then rules are framed. The rules are validated by the expert. The validated rules can be applied to test data and then the data will be classified according the rules.

V. SYSTEM ARCHITECTURE

In this proposed system, Hepatitis B virus data is preprocessed by the data mining techniques. The preprocessed data are given to fuzzy rule extraction subsystem. The data is divided into training and test data. Data is trained by fuzzy if-then rules are extracted and then rules are validated by experts. Depending on the query, classes are classified and the maximum accuracy is predicted from the knowledge base.

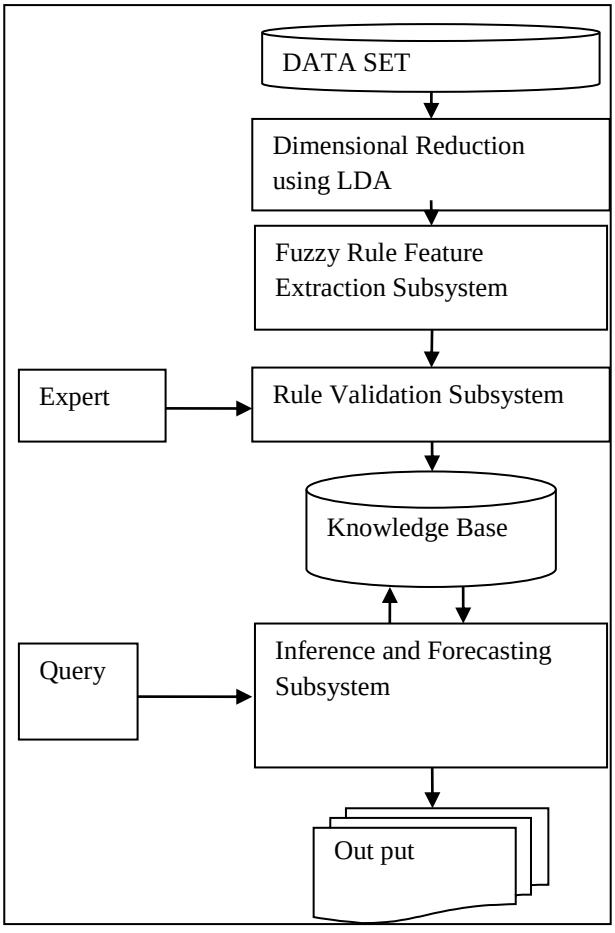

Figure 1. System for Prediction of Hepatitis Data

A. Preprocessing Subsystem

The input data is Hepatitis B virus data. It contains only numeric values. Affected liver and Normal liver are in the data set. Using LDA techniques, data can be normalized into [0 1]. The data set can be divided into two sets. One is training set and another one is called test data.

B. Fuzzy Rule extraction Subsystem

Depends upon the problem, Fuzzy if-then rules are extracted from the training set. It may be small medium and high.

C. Rule Validation Subsystem

Experts may validate the fuzzy if-then rules according to the problem. From the rules, membership function can be framed for classification and prediction.

1) Inference and Forecasting Subsystem

After rule validation data is classified according to the rules and accuracy can be calculated.

VI. DATA PREPROCESSING

Preprocessing is a method to improve the quality of the data. There are a number of data preprocessing techniques. They are

A. Data cleaning

B. Data integration

C. Data transformations

D. Data reduction

Data cleaning can be applied to remove noise and correct inconsistencies in the data. Noise is a random error or variance in a measured variable. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These data processing techniques, when applied prior to mining, can substantially improve the overall data mining results.

Usually, micro array data files contain Affymetrix's Gene-Chip software generated gene expression values in scaled average difference in units. There is a *P*, *M*, or *A* label associated with each average difference expression value, which indicates whether RNA for the gene is Present, Marginal, or Absent, respectively (as determined by the GeneChip software). Files are organized such that each column contains expression levels of different genes in a single sample, and each row contains expression levels of a single gene in different samples. These files may have many negative values that are replaced by using a threshold of θ_l and a ceiling of θ_h. If a value is less than θ_l, it is replaced with θ_l.

Similarly, if a value is greater than θ_h, it is replaced with θ_h. Then, variation filters are applied to exclude those genes that violate $max(g) - min(g) > \Delta$ and $max(g) / min(g) > \Omega$. Different researchers have applied different values of θ_l , θ_h, Δ and Ω for preprocessing of their microarray data. Then, these values are scaled.

If y is the expression value of a gene g, its linearly scaled value in the range [a, b] will be

$$(b-a)\frac{y-minval}{maxval-minval}+a \quad (1)$$

where *minval* and *maxval* are the minimum and maximum values of gene expressions across all genes and samples. The standard normalized value of y will be

$$\frac{y-\mu}{\sigma} \quad (2)$$

where μ and σ are the mean and standard deviation of genes across all genes and samples.

VII. DATA SET AND METHODS

The hepatitis data imported from UCI machine learning repository. The main objective of dataset is to predict the presence or absence of hepatitis b virus in the given output of various medical tests. The dataset contains 155 samples, of which 32 cases belong to ‘‘die’’ class and the remaining 123 cases belong to ‘‘live’’ class. Each sample in the dataset has 20 attributes besides the label. The total attributes are listed in Table 1, in which 14 attributes with binary values and 6 attributes has discrete values.

A. DF Classifier

The proposed DF Classifier algorithm combines feature extraction and parameter optimization. The feature extraction achieved by two stages LDA. The parameter optimization is done by fuzzy rule classifier. All input data are normalized before the feature extraction to avoid attributes maxima and minima numerical changes. In addition normalization it could avoid the computational complexity. After normalization LDA performed data diminishes. The second phase of work is pattern classification. The pattern classification problem is done by if-then rule based fuzzy classifier. The fuzzy classifier divided two patterns, first one is training pattern and next one is test pattern. Third phase of the work is membership function generated. These membership functions were used to test each training pattern with optimal parameter. Final phase of the work generates predictor model that is used to classify each test pattern.

B. Algorithm –DFC

1. Load the Hepatitis dataset $X_n=\{x1,x2\ldots\ldots xn\};$
2. Select the feature space X_{ns}.
3. Normalise the featurespace $Xns \in [0,1]$
4. Data Diminsih usingTwo stage LDA
5. Obtain the Dimension Reductional Featue subset.
6. Splits subet S_i={low,medium,High}
7. Fuzzy classifier on training pattern
8. Apply Rule R_j : If x_1 is A_{j1} and . . . and x_n is A_{jn}
   then Class C_j with $CF_{j,j} = 1, 2, \ldots, N$
9. Generate membership function

$$A_i(x_i) = \exp\left(-\frac{(x_i-\mu_i)^2}{2(\sigma_i)^2}\right) \quad (3)$$

10. obtain optimal DF classifier
11. Predict and labels data subset

TABLE I. HEPATITIS DATA COLLECTION

SERIEL NO	VARIABLE	VALUES
1	Class	DIE, LIVE
2	AGE	10, 20, 30, 40, 50, 60, 70, 80
3	SEX	male, female
4	STEROID	no, yes
5	ANTIVIRALS	no, yes
6	FATIGUE	no, yes
7	MALAISE	no, yes
8	ANOREXIA	no, yes

9	LIVER BIG	no, yes
10	LIVER FIRM	no, yes
11	SPLEEN PALPABLE	no, yes
12	SPIDERS	no, yes
13	ASCITES	no, yes
14	VARICES	no, yes
15	BILIRUBIN	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16	ALK PHOSPHATE	33, 80, 120, 160, 200, 250
17	SGOT	13, 100, 200, 300, 400, 500,
18	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90
20	HISTOLOGY	no, yes

C. Data Diminish using LDA

Linear Discriminant Analysis, or simply LDA, is a well-known classification and data reduction technique that has been used successfully in many statistical pattern recognition problems. It was developed by Ronald Fisher, who was a professor of statistics at University College London and is sometimes called Fisher Discriminant Analysis (FDA). The primary purpose of LDA is to separate samples of distinct groups. We do

The first step in the LDA is finding two scatter matrices referred to as the "between class"and "within class"scatter matrices. Suppose in a given problem we have g different classes or (sample groups). Each sample group π_i has a class mean, which we denote $x_{i.}$

$$y=W^T x_i i \qquad i=1,2,\ldots,n \tag{4}$$

$$U(W)=\arg_w \max \frac{W^T S_b W}{W^T S_w W} \tag{5}$$

$$S_b = \sum_{i=1}^{c} n_i (u_i - u_i)(u_i - u)^T \tag{6}$$

$$Ui = \frac{1}{ni}\sum_{j=1}^{ni} xj, \quad u = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{ni} xj \tag{7}$$

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{ni} (xj - ui)(xj - ui)T \tag{8}$$

Where, S_b-between class Matrix and S_w-within class matrix.

LDA finds directions on which the data samples of different classes are far from each other while requiring data samples of the same class to be close to each other thus achieving the maximum class discrimination.

D. Fuzzy – Rule Based Classification

Fuzzy-rule based systems are mainly used for pattern classification problems. Pattern classification is a supervised process. Fuzzy Systems based on fuzzy if–then rules are successfully applied to various control problems. Fuzzy if-then rules are derived by human experts.

One advantage of such fuzzy systems is their comprehensibility. That is, human users can easily understand each fuzzy if–then rule through linguistic interpretation because its antecedent and consequent fuzzy sets are

related to linguistic labels, such as "small" and "large." The meaning of each linguistic label is specified by its membership function.

Recently, several approaches are proposed for automatically generating fuzzy if-then rules from numerical data. Self-learning methods also proposed for adjusting membership functions of fuzzy sets in fuzzy if-then rules.

A fuzzy set A in X is characterized by a membership function which is easily implemented by fuzzy conditional statements. For example, if the antecedent is true to some degree of membership, then the consequent is also true to that same degree.

If <antecedent> **Then** <consequent>

For example

Rule: If variable1 is low and variable2 is high

Then output is benign **Else** output is malignant.

A fuzzy system is characterized by a set of linguistic statements based on expert knowledge. The expert knowledge is usually in the form of "if-then" rules.

E. Pattern Classification Problem

Let us assume that our pattern classification problem is an n-dimensional problem with C classes (in microarray analysis, C is often 2) and m given training patterns $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn}), p = 1, 2, \ldots, m$. Each attribute of the given training patterns is normalized into the unit interval [0, 1]; i. e., the pattern space is an n-dimensional unit hypercube $[0, 1]^n$. In this study, fuzzy if-then rules of the following type as a base of our fuzzy rule-based classification systems

$$\text{Rule } Rj : \text{If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \text{ then Class } C_j \text{ with } CF_{j,j} = 1, 2, \ldots, N \quad (9)$$

where R_j is the label of the j^{th} fuzzy if-then rule, $A_{j1}, \ldots, A_{jn}$ are antecedent fuzzy sets on the unit interval [0, 1], Cj is the consequent class (i.e., one of the C given classes), and

CF_j is the grade of certainty of the fuzzy if-then rule R_j

Fuzzy-rule-based classification system consists of N linguistic rules each of which has a form as in (5). There are two steps in the generation of fuzzy if-then rules: specification of antecedent part and determination of consequent class C_j and the grade of certainty CF_j. The antecedent part of fuzzy if-then rules is specified manually. Then, the consequent part (i.e., consequent class and the grade of certainty) is determined from the given training patterns. The use of the grade of certainty in fuzzy if-then rules allows us to generate comprehensible fuzzy-rule-based classification systems with high classification performance.

1) Rule Generation Based on Mean and Standard Deviation of Attribute Values

In this approach, a single fuzzy if-then rule is generated for each class. The fuzzy if-then rule for the k^{th} class can be written as

If x_1 is A_1 and ... and x_n is A_n, then Class k

where A_i is an antecedent fuzzy set for the i^{th} attribute.

The membership function of A_i is given by

$$A_i(x_i) = \exp\left(-\frac{(x_i - \mu_i)^2}{2(\sigma_i)^2}\right) \quad (10)$$

A_i is the mean of the i^{th} attribute values x_{pi} of Class k patterns, and sigma is the standard deviation. Data set is divided into two categories, namely training and test data. Membership function can be generated using (6^{th}). There are only three categories in membership function. They are small medium and high. Then depends on the three categories, the performance is tested on the test data of gene expression Hepatitis B data set.

VIII. RESULT AND DISCUSSSION

To evaluate the effectiveness, the proposed method is applied in hepatitis database. The highest classification accuracy of 98.77% has been achieved for the high value partition data in training–testing partition.

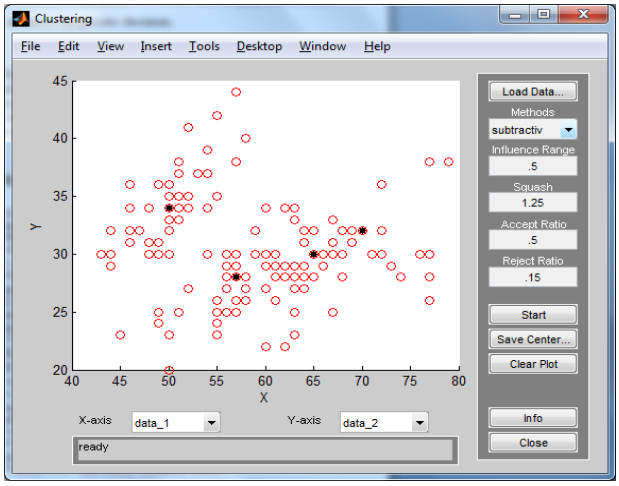

Figure 2. Hepatitis B dataset

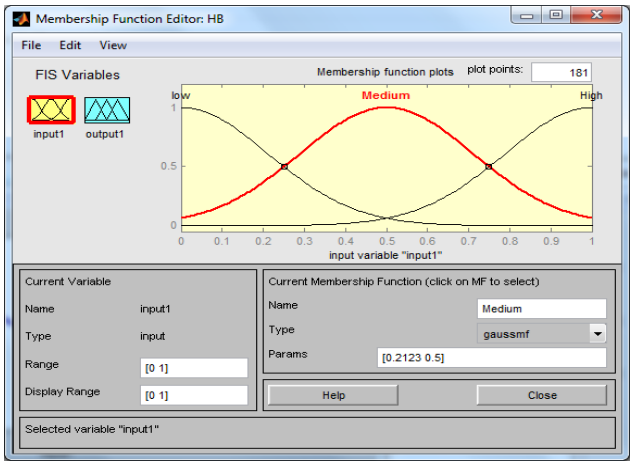

Figure 3. Membership Function Generation

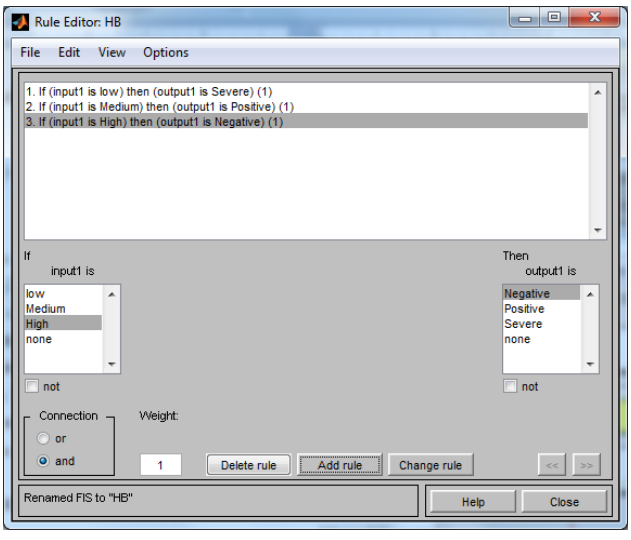

Figure 4. Fuzzy if then Rule classifier

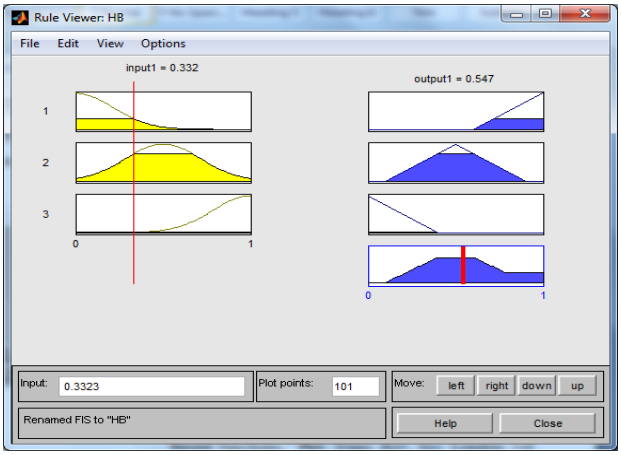

Figure 5. Fuzzy if then Rule Viewer

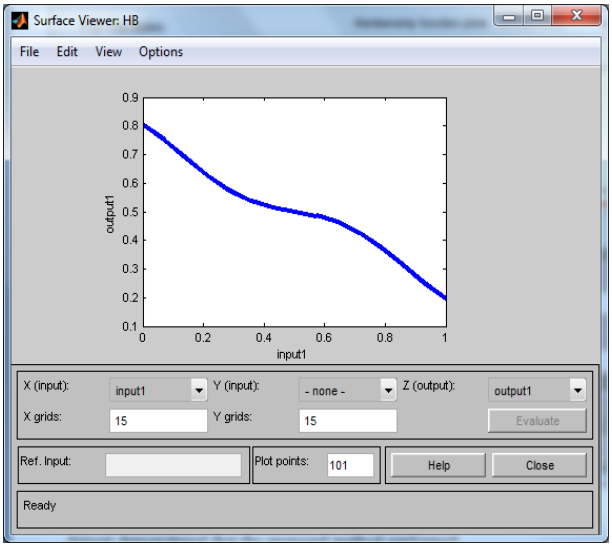

Figure 6. Fuzzy if then surface Viewer

TABLE II. HEPATITIS DATA COLLECTION

	ATTRIBUTES	CLASS	PATTERN	CLASSIFICATION
LOW	19	2	3	97.2%
MEDIUM	16	2	3	98.5%
HIGH	20	2	3	98.8%

IX. CONCLUSION

In this work, we have developed a new medical diagnostic method, DF classifier, for addressing hepatitis diagnosis problem. Experiments on different portions of the hepatitis dataset demonstrated that the proposed method performed significantly well in distinguishing the live liver from the dead one. It was observed that DFC achieved the best classification accuracies (98.77% for 80–20% training–testing partition) for a reduced feature subset that contained two features. Meanwhile, comparative study was conducted on the methods of the PCA_SVM, the FDA_SVM and the SVM. The experimental results showed that the DFC performed advantageously over the other three methods in terms of the classification accuracy. We believe the promising results demonstrated by the DFC can ensure that the physicians make very accurate diagnostic decision.

REFERENCES

[1] K..Thearling, Information About Data Mining and Analytic Technologies, http://www.thearling.com/text/dmwhite/dmwhite.htm, accessed July, 2009.

[2] A. S. Koyuncugil, "Fuzzy Data Mining and Its Application to Capital Markets", PHD. Thesis, Ankara University, 2006.

[3] S. J. Lee, and K. Siau, "A Review of Data Mining Techniques," Industrial Management & Data Systems, vol.101,no.1,pp. 41-46,2001.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol. 17, pp. 37-54, 1996.

[5] B. Thuraisingham, "A Primer for Understanding and Applying Data Mining", IT Professional, pp: 28-31, 2000.

[6] Osmar R. Zaiane, "Chapter I: Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1999.

[7] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery: an overview, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, Menlo Park, CA, AAAI/MIT Press, pp: 1-36, 1996.

[8] C. Romero, S. Ventura, P. G. Espejo, and C. Hervas, "Data Mining Algorithms to Classify Students", proceedings of the 1st Int'l conference on educational data mining, Canada, pp: 8-17, 2008.

[9] J. Zhang, I. Mani, KNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction, In Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II, August 21, 2003.

[10] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufman Publishers, 1993.

[11] M. Grabisch. 1996, "The representation of importance and interaction of features by fuzzy measures", Pattern Recognition Letters, Vol. 17:567-575.

[12] M. Grabisch, F. Dispot. 1992, "A comparison of some methods of fuzzy classification on real data", Proc. of 2nd Intl. Conf. on Fuzzy Logic and Neural Networks, 659-662.

[13] M. Grabisch, and J.-M. Nicolas. 1994, "Classification by fuzzy integral: performance and tests", Fuzzy Sets and Systems, Vol. 65, No. 2/3:255-271.

[14] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Fifth annual workshop on computational learning theory.

[15] Pittsburgh: ACM. Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. Volume 2 Issue 3, April 2011.