Proceedings of the World Congress on Engineering 2016 Vol I
WCE 2016, June 29 - July 1, 2016, London, U.K.

Towards a Data Mining Class Library for Building Decision Making Applications

Juan Ramón Castro, José A. Maldonado, Ángel A. García, Manuel Castañón-Puga and Edaniel Figueroa

***Abstract*—The constant development of technologies around the globe has brought the big problem of data acumulation to different areas where data adquisition is necesarie. This is where data mining algorithms come in handly. With the past of the years the need of a tool to process and utilies efficiently data has brought data mining techniques as a valuable resource that has been incorporated under applications like information systems, decision-making systems, intelligent systems, recommendation systems, and others. In this paper, we present a data mining class library (DMCL) that is currently been develop who incorporates techniques like clustering, fuzzy logic and decision-making algorithms. This group of techniques have been used in multiple aplications found in literature making them easy to be compared with our implementation. Our main goal is to help programmers who build any type of decision making system needing the aid of a data procesing tool easy to implement.**

***Index Terms*—Data mining; Clustering, Fuzzy Inference Systems, Decision Making.**

I. Introduction

In recent years, data mining techniques had been apply to a variety of problems from particular areas with a great acceptance. Its use in this areas helps in the discovering of characteristics in the mass data collection that they handle. Some of the areas that had been incorporating data mining techniques by the pass years are Engineering, Medicine, Finance and Science. The particular tasks of data mining techniques in those areas are knowledge extraction, data classification and data filtering [1]. Data mining has been used also as a part of decision making systems to help in task such decison-making, data processing, data management and others. The main reason this type of techniques have taking attention is because of the constant grow of data and the few applications that really takes advantage of it. The data storage have brought problems as retrieving information efficiently, discovering useful data and extracting knowledge, reason why data mining comes in handy for these particular tasks. [2].

A. Data mining

Data mining is a collection of techniques used as a subprocess of knowledge discovery in data bases (KDD) its main purpose is the extraction of useful data from particular sets. This collection of techniques has been applied in recent years because of the constant grow of data acumulation that the mentioned areas are generating [7].

B. Clustering Techniques

This group of techniques have by function forming and discovering groups based in similar characteristics. There are different classification criteria depending in how the data partitions are form [3]. Some of the basic clustering techniques based on patitions form are:

- Patitional: this clustering algorithms constructs k partitions of a particular database of n objects. It aims to minimize a particular objective function, such as sum of squared distances from the mean.
- Hierarchical: creates a hierarchical decomposition of the data, either agglomerative or divisive. Agglomerative decomposition treats each data as a different cluster the first time, then by iterative merge clusters creates new groups with same distance. Divisive decomposition uses the opposite way, start with data in a single cluster and iterative splits groups into smaller ones.
- Density-based clustering: cluster generation is made by incorporating a specific density object function, which is defined as a number of objects in a neighborhood.
- Grid-Based clustering: this type of algorithms focuses in spatial data. A grid divides the data in cells used to form clusters. Tt doesnt́ depend in the distance for determinate the characteristics of the data.
- Model-Based clustering: its aim is to find models that fits the input data. In some point they can be view as a density-based clustering, but they dont́ use the same principal, they only apply a part of it to determinate if the selected model satisfy the data. They normally starts with a predefined number of clusters.
- Categorical Data Clustering: this algorithms uses the data to determinate their behaviour.

The different clustering partitions approaches enables a great diversity of clustering algorithms [8]. Some of the best known clustering algorithms are:

1) K-means: Also known as the Hard c-means algorithm, aims to find clusters based in the similarity of the data. This similarity is determinate by calculating a distance from each data to a possible center of the cluster. This algorithm uses the Euclidian distance [10].

2) Gustafson-Kessel (GK): The GK algorithm has some differences with respect of the above. One of the characteristics is that employs the Mahalanobis distance, which adapts to the data so it can be able to cover more of it. The type of clusters this algorithm generates have an ellipsoidal shape

Juan Ramón Castro, José A. Maldonado, Ángel A. García and Manuel Castañón-Puga are with Autonomous University of Baja California, Calzada Universidad 14418, Tijuana, México, 22390, e-mail: {jrcastror, jose.martinez, aagarciaramirez, puga}@uabc.edu.mx, web page: http://www.uabc.edu.mx

Edaniel Figueroa is with Tecnología en Comunicaciones e Identificaciones de México, S.A. de C.V., Del Bosque 10774, Jardines de Chapultepec, Tijuana, México, 22025, e-mail: edaniel@sdpoint.com, web page: http://www.grupotimex.com

ISBN: 978-988-19253-0-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

that helps with data dispersion. Incorporates an objective function as an stop condition [11].

3) Gath-Geva (GG): Is based in the GK algorithm. Its difference resides in the size and density of the cluster which is evaluated in this algorithm, the objective function is not used in this algorithm [9].

4) Fuzzy C-means (FCM): Is one of the most popular algorithms with a considerable amount of modifications reported in the literature. The main difference that has with the others algorithms is the use of a parameter to assign fuzziness to a determinate data. This helps in the group formation allowing that a single data could belong to multiple groups [12].

5) Subtractive: This algorithm uses ratios of acceptance and rejection to determinate if some data belong to a certain cluster. Its based in the mountain algorithm which uses grid-based clusters [13].

II. Decision Making

A. Decision Tree

Decision trees have been used in many areas as a powerful tool in decision making systems, as a way to help the expert take the best decision that not only solves the problem but also solve it efficiently [5] [14]. Decision tree algorithms creates a collection of nodes strategy evaluated by a predictor attribute that finds the best ramification in order to separate data into more homogeneos subgroups, it iterates until every data is classify. The whole process starts dividing the data into two sets. One is used for training and a second one for testing. Once the training data is given to the tree, it selects a root attribute; used to obtain the nodes that best classify the data until they hit the best attribute that matches the data. One of the most utilized decision tree algorithm is the iterative dichotomiser 3 (ID3) introduced by Quinlan in 1986. Years later the ID3 was upgraded to C4.5. There are different variations implemented of this two algorithms. These algorithms uses the information gain calculation in any single attribute in order to build decision tree.

1) ID3: This algorithm generates the node employing a top-down, greedy search of the data. The information gain calculation in this algorithm determinates the most valuable attribute that classifies better the given data. This function helps minimizing the tree levels [5].

2) C4.5: This algorithm was also develop by Ross Quinlan, who proposed it to overcome some limitations of the ID3 like the sensivility to sets with large number of values. In order to solve this, C4.5 uses a gain information ratio to evaluate the splitting attribute of each node.

B. Fuzzy Logic System

A Fuzzy Logic System (FLS) is a system that tries to handle data with some sort of uncertainty. As we could observe in the real world there does not exist a simple logic representation like 0 and 1 or true and false. For example in the real world when someone express about the weather, they could say l̈ess coldör m̈uch coolerẗhese two expressions are know as linguistic variables, and there not express a precise value, instead we have a degree of pertinence between 0 and 1. To change between a linguistic variable to a numeric value, we need some components called rules, so it can evaluate and process the input to determinate a single value for that linguistic variable. According to a set of rules, a fuzzifier could transform the linguistic variabes to a range between 0 and 1. An inference engine that will make the operations needed to obtain a single value for the system and finally an output process that will transform that fuzzy value into a real one [6] [15].

The basic FLS stucture process the crisp input transforming the data to fuzzy then process by the inference machine which has a set of rules. Once it ends processing, it sends the data to the output module where defussifies into a crisp output, the FLS structure can be observed in figure 1.

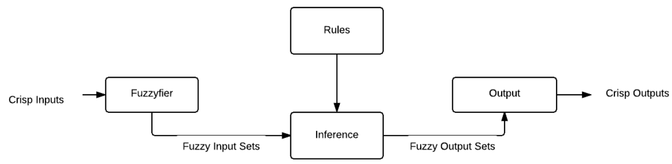

Fig. 1. FLS Structure

There are two fuzzy model that we explore in the literature:

1) Mamdani: Was first introduced to control a steam engine and boiler. The input was a set of linguistic data obtained from humans expers. In this fuzzy model the inputs and outputs are all fuzzy sets. A typical fuzzy rule has the form:

if x is A and y is B then z *is* C

2) Takagi-Sugeno-Kang (TSK): Proposed by Takagi, Sugeno and Kand in order to use fuzzy sets as inputs and a crisp function in the output. A typical fuzzy rule has the form:

if x is A and y is B then $z = f(x, y)$

III. Data Mining Class Library

DMCL is a library currently being develop as a solution in Microsoft C# language. The main purpose its to provide tools for data processing. The uses of this tools could be patter discovering, decision-making and classification. The current content of the library incorporates clustering techniques, FIS and decision making algorithms. Under clustering techniques we have K-means, GK, GG, FCM and subtractive, those are the more used and known techniques. For FIS we have Mamdani model and TSK model and for the decision-making we have incorporated the ID3 algorithm, which is one of the first decision tree techniques develop by Ross Quinlan. In the table I appears the techniques incorporated in this library and some of their application found in literature.

A. DMCL structure

The DMCL has a package structure, where it contains clases separated by type of algorithm. We have the stadistics package who contains the decision tree classes 4. Also we

ISBN: 978-988-19253-0-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
COLLECTION OF DATA MINING ALGORITHMS AND TASKS CONDUCTED FOUND IN LITERATURE

Algorithm	Task
K-means	Classification of EEG signals
Gustafson-Kessel	Fuzzy rule extraction
Fuzzy C-Means	Fuzzy Partition
Subtractive	Preprocessing Data
Decision Trees (ID3 and C4.5)	Decision Making
Fuzzy Logic System (Mamdani and TSK)	Evaluation

have the FIS package where we contain the Mamdani and TSK fuzzy models 3 and we have other package in where it lies the hard and fuzzy package of clustering techniques 2. The library uses objected oriented characteristics: heritage, encapsulation and polymorphism.

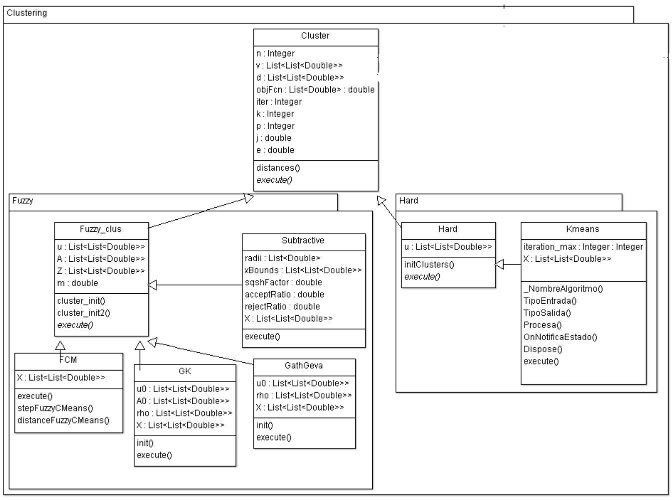

Fig. 2. Clustering Package Diagram

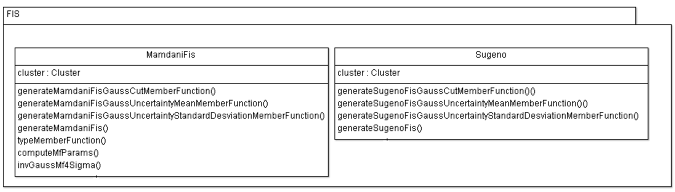

Fig. 3. FIS Package Diagram

As we can see in the image above; we have for the clustering package a super class called Cluster and each technique inherits from it. The classes that inherits directly from it are fuzzy_clust and hard, each one in its own subpackage. The fuzzy_clust and hard are also super classes of the different type of algorithms implemented. We have a FIS package in where we have two clases each one implementing a fuzzy model. One implements the Mamdani fuzzy model and the second one implements TSK. Both uses the cluster class because this algorithms use clustering techniques in order to determinate the appropied rules to work with the data provided. The fuzzy classes are based on a Java Library called JT2FIS [4]. And we also have the statics subpackage in which we have the decision tree implementation the ID3 called DecisionTreeID3 and the clases TreeNode and Attribute which are used by the ID3 implementation.

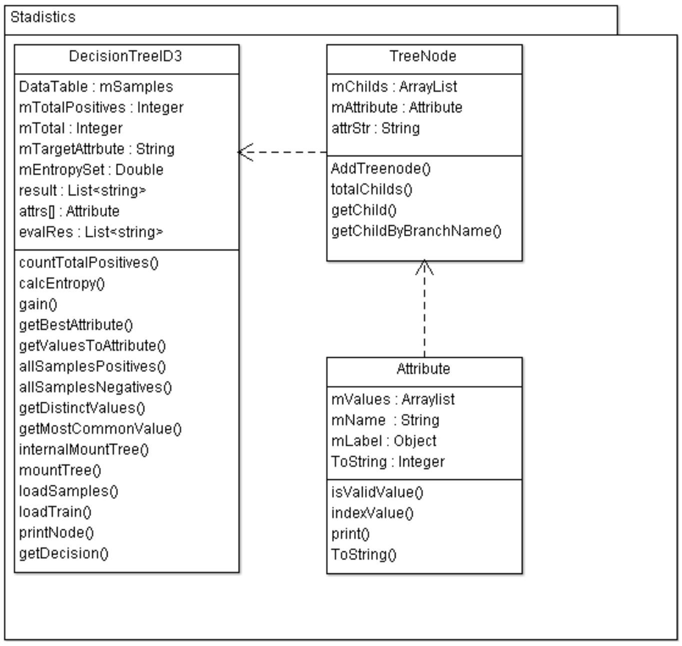

Fig. 4. Stadistics Package Diagram

The DMCL is designed to be accesible for developers as a web service in which the data woud be loaded to the specified algorithms and processed in the server side. The server will send the results back to the program who called the service 5.

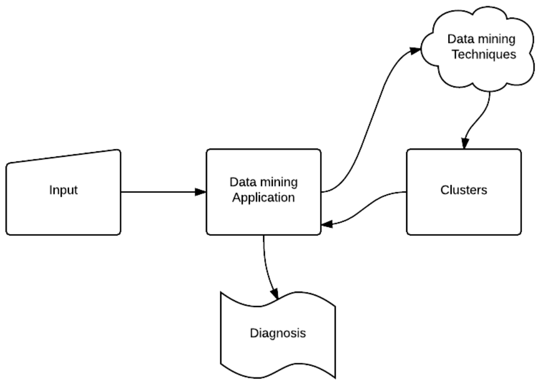

Fig. 5. Web service process.

Since we are using an objective oriented programming language, we can instance every algorithm that we would want to use in an application. Each algorithm has some particular parameters that can be change depending on the result we want to achieve. A code example of the algorithm K-means can be viewed in the listing 1.

Listing 1. Object-oriented coding example

```
//parameters
int K = 4;
X = matrix;
// K-means Algorithm object
Kmeans Km = new Kmeans(X, K);
Km.execute();  // Algorithm start
Km.V // getClusters
```

ISBN: 978-988-19253-0-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Another object included in DMCL is the FIS who can be instanciatedas shown in the listing 2.

Listing 2. Object-oriented coding example of Fuzzy Inference System
```
//parameters
Input inputList = new Input();
Outpur outputList = new Outpur();
inputList.Add(inputs); // data vector
outputList.Add(outputs); // data vector
double uncertainty = 0.0;
//Clustering technique to determinated
//the model rules
FuzzyCMeans m = new FuzzyCMeans();
MamdaniFis genFis = new MamdaniFis(m);
    // Fuzzy Model

//FIS instance
Mamdani fis =
genFis.generateMamdaniFisGaussUncertainty
MeanMemberFunction(inputList,
    outputList, uncertainty);
```

For the decision-making, we implemented the ID3 decision tree,aides from the data in order to learn, so it can be able to solve decision making problems. For this algorithm the sintaxis is 3. Since we are using recursivity in the principal method of the class we used the class TreeNode to a groupe the different ramification the tree will generate. The dataset used in the below example was the Iris dataset.

Listing 3. Object-oriented coding example of Decision Tree
```
//parameters
//clases
String[] columns = new string[] {
    "Sepal_len", "Sepal_wid",
    "Petal_len", "Petal_wid", "class" };
// Possible outputs
String[] answer = new String[] { "1",
    "2", "3" };
DecisionTreeID3()  id3 = new
    DecisionTreeID3();
// Data to be evaluated
DecisionTree.Attribute[] attribute =
    new DecisionTree.Attribute[] { sl,
    sw, pl, pw };
// Data to train the Tree
DataTable samples =
    id3.loadSamples(columns, data);
// Passing the posible outputs of the
    Tree
id3.Answers = answer;
// Root generated for algorithm
TreeNode root = id3.start(samples,
    "class", attribute);
// Print Tree
id3.printNode(root, "");
```

IV. Test Cases and Results

We tested the DMCL using weel know benchmark datasets, to ensure that our results were correct. We compare the library algorithms with the Matlab® implementations. We also wanted to see how efficient was our implementations comparing execution times with Matlab®.

A. *Clustering Comparations from Data Mining Library and Matlab®*

The dataset used to test the K-means implementation was the cholesterol, which contains an input matrix of 21 spectral measurement of 264 blood samples. The clusters obtained from both implementations are shown in table II. Also we used the cholesterol dataset to test Fuzzy C-means and Subtractive algorithms, each can be viewed in table III and IV.

TABLE II
K-means cluster comparation between Data mining class library and Matlab®.

	Cluster #1	Cluster #2	Cluster #3
Matlab®	0.3322	0.5235	0.2159
	0.3820	0.6057	0.2454
	0.3725	0.5903	0.2389
	0.3118	0.4896	0.2015
	0.2612	0.3926	0.1736
DMCL	0.3322	0.5234	0.2158
	0.3820	0.6057	0.2454
	0.3725	0.5903	0.2389
	0.3117	0.4896	0.2014
	0.2611	0.3926	0.1735

TABLE III
Fuzzy C-means cluster comparation between data mining class library and Matlab®

	Cluster #1	Cluster #2	Cluster #3
Matlab®	0.3149	0.4434	0.2126
	0.3617	0.5120	0.2416
	0.3525	0.4988	0.2352
	0.2952	0.4150	0.1984
	0.2489	0.3396	0.1708
DMCL	0.3148	0.4431	0.2125
	0.3615	0.5117	0.2415
	0.3524	0.4985	0.2351
	0.2950	0.4148	0.1983
	0.2488	0.3394	0.1707

TABLE IV
Subractive cluster comparation between data mining class library and Matlab®

	Cluster #1	Cluster #2
Matlab®	-0.3592	0.3399
	0.8262	0.5588
	1.0030	0.1464
	0.1098	-0.0744
	1.2785	0.6150
DMCL	-0.3592	0.8262
	0.3398	0.5588
	1.003	0.1463
	0.1098	-0.0743
	1.2785	0.6150

The decision tree was evaluated with the Iris dataset which contains data of iris flower classification, the dataset contains two matrices: the first one contains the 4 different attributes of the flower, sepal length and width and petal width and length of 150 flowers, the second one is a matrix of the 3 different classifications of the 150 flowers; the input data was

ISBN: 978-988-19253-0-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

separeated into two matrices one of 4*100 to train the tree, and a 4*4 to query the tree in order to get the classification, we compared the result with the target matrix provided by the dataset. The result can be observed in table V and the data used to query and their classification can be observed in table VI.

TABLE V
COMPARING ID3 IMPLEMENTATION BETWEEN DATA MINING CLASS LIBRARY AND MATLAB®

	Class #1	Class #2	Class #3	Class #4
Matlab®	1	2	2	2
DMCL	1	2	2	2

TABLE VI
BENCHMARK DATASET EXTRACT TO TEST THE DECISION TREE WITH CALSSIFICATION

Sepal lenght	Sepal width	Petal length	Petal width	Class
4.7	1.6	4.4	4.6	1
5	4.8	5.5	6.1	2
3.2	3.1	2.6	3	2
1.4	0.2	1.2	1.4	2

B. Data mining class library and Matlab® time comparition

Respecting to time comparitions, we also used the Matlab® implementation to determinate the eficiency of DMCL as a development language for data mining algorithms. The dataset used to test was the simplecluster dataset incorporated in Matlab®. We expand this dataset multiplying it with a matrix of ones making the size of the dataset 1000*50. In the table VII times can be observed by some of the clustering algorithms.

TABLE VII
DATA MINING LIBRARY AND MATLAB® TIME COMPARATION

Algorithm Name	Matlab® time (seconds)	DMCL time (seconds)
K-means	0.535613	0.1015944
Fuzzy C-means	0.29	0.48
Subtractive	0.11	1.74

V. CONCLUSION

We presented the DMCL developed in C# language as an alternative to help on a variety of intelligent aplications like patterns discovering, knowledge discovering, decision-making, data filtering and many others. We believe that by making the library accesible as a group of web services, the applications will only need to handle the inputs and the result of the invoked algorithm. We tested the DMCL using benchmarks dataset comparing our result aginst Matlab® which had the algorithms already incorporated. From the comparations, we found that our results were basically the same, showing that the implementation is correct.

As a future work, we want to add more algorithms to the library, to perfect some mathematical implementations in order to apply them to the algorithms. Use other benchmark data sets to test the algorithms.

ACKNOWLEDGMENT

The authors would like to thank to *Tecnología en comunicaciones e identificaciones de México, S.A. de C.V.* and the *Programa de estímulo a la investigación, desarrollo tecnológico e innovación* of CONACYT (PROINNOVA) of México to support the project *BUZÓN MÉDICO: Software como Servicio para expedientes clínicos electrónicos en la Nube* by grant no. 220590.

REFERENCES

[1] Sriraam, N., V. Natasha and H. Kaur. Data Mining Techniques and Medical Decision Making for Urological Dysfunction. *In Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, ed. John Wang, 2506-2516 (2008).
[2] Nada Labrac, Selected Techniques for Data Mining in Medicine,*Artificial Intelligence in Medicine* 16, 3-23 (1999).
[3] Periklis Andritsos, Data Clustering Techniques, *University of Toronto: Deparment of Computer Science*, March 11, 2002.
[4] Manuel Casta nón-Puga, Juan Ramón Castro, Josué Miguel Flores-Parra, Carelia Guadalupe Gaxiola-Pacheco, Luis Guillermo Martínez-Méndez, Luis Enrique Palafox-Maestre, JT2FIS A java Type-2 Fuzzy Inference System Class Library for Building Object-Oriented Intelligent Applications, *Universidad Autonoma de Baja California*.
[5] Wei Peng, Juhua Chen and Haiping Zhou, An Implementation of ID3 Decision Tree Learning Algorithm, *University of New South Wales: School of Computer Science and Engineering*.
[6] Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani, Neuro-Fuzzy and Soft Computing A computational Approach to Learning and Machine Intelligence, *Prentice-Hall, Inc*, 1997.
[7] Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Datasets, 2010.
[8] Khaled Hammouda, Prof. Fakhreddine Karray, A Comparative Study of Data Clustering Techniques,*Cite Seer X*, pp. 1-21J.
[9] Gath and A. B. Geva, Unsupervised Optimal Fuzzy Clustering,*IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol 11, No. 7, pp 773-781.
[10] Nikita Gurudath, H. Bryan Riley, 2014, Drowsy Driving Detection by EEG Analysis Using Wavelet Transform and K-Means Clustering, *Procedia Computer Science*, Vol. 34, pp. 400-409.
[11] H. Verma and R. K. Agrawal, 2002, Intuitionist Gustafson-Kessel Algorithm for Segmentation of MRI Brian Image,*Advances in Intelligent and Soft Computing*, Vol. 131, pp. 133-144.
[12] M. N. Ahmed, S. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, A Modified Fuzzy C-Means Algorithms for Bias Field Estimation and Segmentation of MR Date,*IEEE Transactions on Medical Imaging*, Vol. 21, No. 3, pp. 193-199.
[13] Chao Ma, Jihong Ouyang, Hui-Ling Chen and Xue-Hu Zhao, 2014, An Efficient Diagnosis Systm for Parkinsons Disease Using Kernel-Based Extreme Learning Machine with Subtractive Clustering Features Weighting Approach,*Computational and Mathematical Methods in Medicine*, Vol. 2014.
[14] Vili Podgorelec, Peter Kokol, Bruno Stiglic, Ivan Rozman, 2002, Decision Trees: an Overview and their use in medicine,*Journal of Medical Systems*, Vol. 26, No. 5, pp. 445-463.
[15] Jerry M. Mendel, 2001, Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions, *Prentice Hall PTR*, University of Sothern California Los Angeles, CA.