LUNG CANCER DETECTION USING A FUZZY LOGIC SYSTEM

Navid Samimi Behbahan*

Amin Samimi Behbahab*

Milad Samimi Behbahan*

Abstract

Creating a method to exploiting knowledge (information) from dataset has been a main cancer diseased of data mining researches. Data mining is a process that helps us to discover such knowledge and have used in many fields. Hence in this study using a new combination of fuzzy logic, GA and SA. A new method has developed to operate on dataset with greats features. One of the advantages of this method is parameters reduction. To the end the lung cancer datasets of UCI which has a great number of features have used. To reduce the feature dimension two methods of PCA and Conscious Selection of Effective Features by GA have used. Finally, the presented method have implemented in software .Compared with other methods a considerable improvement in the result have seen. Learning dataset was 98.33 and testing dataset was 96.15 present.

* Department of Computer Engineering, Omidiyeh Branch, Islamic Azad University, Omidiyeh, Iran

1- Introduction

According to extend dataset and its complexity, nowadays there is a need to a more effective and useful implement in order to discover a useful knowledge about them searching data is a process help human in such discovery and recently is used in widespread filed detection of different diseases in medicine is one of the full way field of searching data and recently has done very studies about it.

Janecek et al. studied relation between division and selection of features and surveyed the effect of dimension reduction by analyzing main parts in division.

Finding explains division accuracy based on PCA is depended on dataset and changes in parts which expresses a basic need to division.

Duangsai thong and Windeatt presented a way to reduce dataset in collection which has lots of features and little samples and reduce division. New datasets is created by omitting repeated and irrelevant data. The result is more accuracy and time for calculating for example, for long cancer, they has reduced 56 features and found 11 features.

Hayward et al, Surveyed application of searching data projects using regression of symbol logic and different cancers dataset. Findings expresses pre-proceed data will improve application of classifying algorithm, if the feature be chosen properly.

Azofra et al, in a research an effective features of dataset, found a subset of available features to improve operation of a learning algorithm. According to research many of ways are scheduled to improve features of evaluating submission. Reliability is a good well-known feature for a set of raters. According to reliability idea, it established a set of measuring features to evaluate features set. It is also shown how effective is this proposed measuring an improving searching process (such as selecting the best features set).

In a research presented by Jyang et al., they proposed a mixed-style of selecting feature using algorithm of non-pragmatism polarity and genetic for using of their benefits. This style consists of 2 phases. Filtering phae in which omits non-pragmatism polarity features and guides us to establish a basic group of genetic algorithm. Packing phase in which searches the extreme (goal) dataset features. Finally the effect of algorithm and different datasets is shown.

This paper is preceded to long cancer detection because it has many features. After several pre-proceed phases, for lung cancer detection, a mixed of fuzzy logic systems and improving algorithms is used. And knowing of input dataset is presented. The goal of this research is reducing the volume of great datasets using several pre-proceed phases (PCA) for selecting best features and them packing with fuzzy systems based on sets of fuzzy if-then rules to discover needed knowledge the processes of using algorithms of proposed style is shown in figure 1.

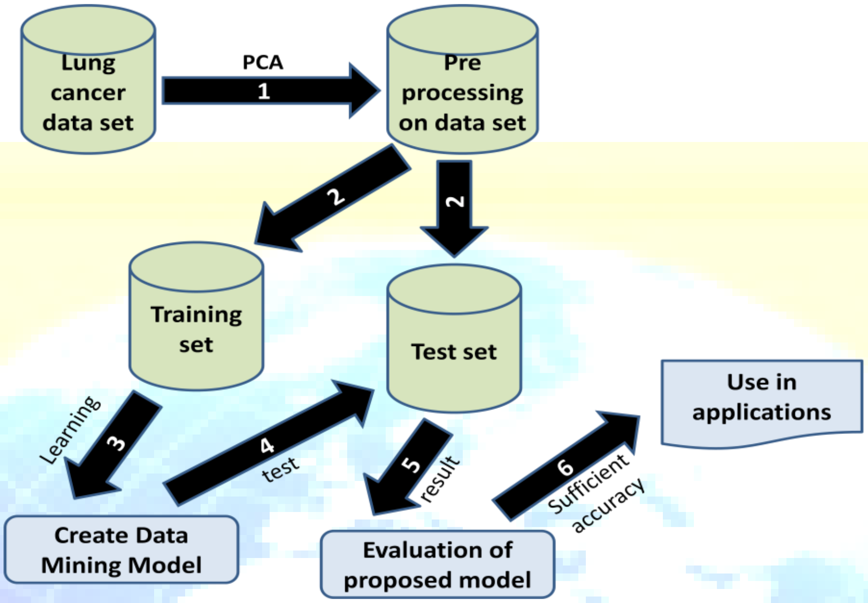

Figure 1: the processes of using algorithms of proposed style

Reduction the volume of dataset is taken into consideration because cheiring knowledge performs in much less time and more accuracy. This knowledge can be a base of fuzzy logic that will improve with recovery algorithms during searching data process. For improving the set of fuzzy logic, mixed-algorithms of genetic and annealing is used. These algorithms are based on evolutional and statistical concepts. The algorithms in co-operating each other searches set of if-then rules related to their cases to have more efficiency. Finally, presented system a lung cancer dataset from searching data store of California University is used

2. Pre-proceed using PCA algorithm

Analyzing main parts (PCA) is a useful statistical technique. This reduction is used to condense and reduce dimension and minimize the squares average errors from condensing. If there be enough data, PCA can produce the best changing. If we have to select the main modification or same of them from a set, well use PCA. PCA technique is the best way to reduce dimensions streaky. It means with eliminating non-important coefficients of this changing, lost information is less than other ways. In this way, new axes are defined based on new axes. First axis should be at the side in which data's variance is maximum (at the side data dispersion is maximum) second axis must be such pillar on the first axis that data variance be maximum and so on other axes must be such pillar on previous axes that data have the most dispersion. PCA is a best streaky changing as bellow.

$$\mathrm{Y}=W_x \qquad (1)$$

In which it is a change of n dimension data model with real amount ($\mathrm{X}\in R^n$) to m dimension axis ($\mathrm{y} \in R^m$) .streaky changing matrix $\mathrm{W}\in R^{m*n}$ is the best due to achieving maximum information and is accounted as bellow:

$$\mu=\mathrm{E}[X] = [E[x_1], E[x_2], \dots, E[x_n]\,]^T \qquad (2)$$

And co-variance matrix will be a square (n*n) equals to:

$$R_{xx} = E[(X - M)(X - M)^T] \qquad (3)$$

In which E[.], shows math hope agent, M is average axis of a data set such X and co-variance matrix R_{xx} with real amount, describes correlation between different features of dataset X.

Co-variance matrix R_{xx} is n*n and it has symmetrical and real amount. ($R_{xx} \in R^{n*n}$)Especial amounts of covariance matrix R_{xx} are arranged descending ($\lambda_1 = \lambda_{\max}, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$). we suppose special pillar parallel vectors from a pillar matrix n*n (these vectors length is 1).

$$\mathrm{E}= [e^1, e^2, \dots, e^n] \qquad (4)$$

It can form matrix W by selecting m special vector more important than matrix E (selecting special vectors with larger amounts) and change their places.

$$w = \begin{Bmatrix} (e^1)^T \\ (e^1)^T \\ . \\ . \\ . \\ (e^m)^T \end{Bmatrix} \qquad (5)$$

PCA can be in consideration as a learning without observer of data's In fact , PCA not only use of available information about a related class with a model, but also it discover correlation between models, their elements and principle sides in which data's model change in (with maximum variance) and it is shown in figure 2.

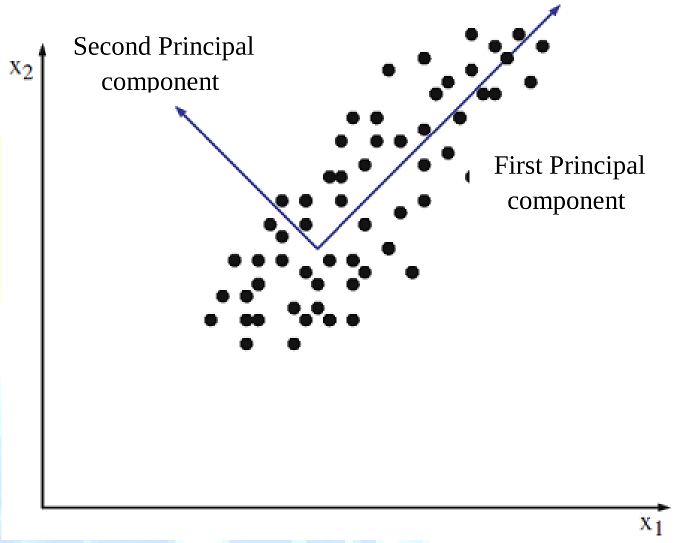

Figure 2: Principal components extracted by PCA

3. Fuzzy legislation system

This part expresses how to show knowledge using fuzzy rules and models classification. Then a present the formation of basic logic rules, classification and positiveness of each rule and continues by how classifying a new model (sample) and surveys the used fuzzy reason.

3.1. Classification of models

Classification of models is a matter with n dimensions, c classes and m educational models. ($X_p = (X_{p^1}, X_{p^2},\ldots, X_{p^m})$, P= 1,2,…,m). Without lasting totality of matter, each educational models feature is normalized. It means models space for each feature is a number between 0, 1 independently. In this research, we use of if – then fuzzy rules based on classification systems.

$$\text{Rule } j = \text{ if } X_1 \text{ is } A_{j_1} \text{ and } \ldots \text{ and } X_n \text{ is } A_{j_n}$$
$$\text{Then class } C_j \text{ with } CF_j \text{ , } j=1,\ 2,\ \ldots\ ,\ n \qquad (6)$$

R_j: J is trustee of if-then fuzzy rule.

$A_{j_1},\dots,A_{j_1}$: Prior language amount like large, medium; small that each ones range is between 0,1 and overcast each other.

C_j: achieved class for each model

N: number of if-then fuzzy rules

CF_j: positiveness degree of R_j rule.

They use as basic fuzzy collections from triangle fuzzy collections.

3.2. Used fuzzy model

As show in figure 3, the used fuzzy collection is a 5 parts fuzzy collection. For coding each part, a number is taken into consideration.

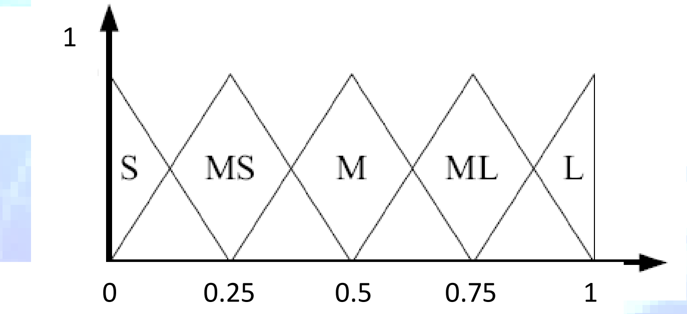

Figure 3: Fuzzy model have been used

S (small) =1

MS (Medium Small) = 2

M (Medium) = 3

ML (Medium Large) = 4

L (Large) =5

3.3. Establishing Basic Fuzzy Rules:

Each rule is coded by number between 1 to 5. At first a collection is random established consists of 100 rules in which each one is a branch with long of datasets features and surveyed their recognitions scales.

In many of similar works, basic rules are established independently from educational data's. Although this way in view of time complexity is suitable but because these rules are not with good education, they have high recognition an only educational data's. So it is better to establish basic rules randomly and always try to improve them.

The way of accounting related class and positiveness degree of each if-then fuzzy rule for the best rules in fuzzy classification system according to rule is as bellow. For example for determining the class(C) and positiveness degree (CF_j), there is bellow processes:

Stage 1: calculating $B_h(R_j)$ for h class (h=1,…, c)

$$B_h(R_j) = \sum M_{j1}\ (X_{p1}) * \ldots * M_{jn}\ (X_{pn})$$
$$H= 1,\ 2, \ldots,\ c \qquad (7)$$

Stage 2: finding (h^) class in which it has maximum amount of $B_h(R_j)$

$$B_h(R_j) = \max \{\ B_1(R_j),\ B_2(R_j),\ \ldots\ B_c(R_j)\ \} \qquad (8)$$

If more than one class has maximum amount, R_j cant merely specialized to C_j class. In this condition, we take it empty (C_j=0). If only one class have has maximum amount. Then C_j=class h.

Stage 3: if only one class has maximum amount of $B_h(R_j)$, then positiveness degree for j rule (CF_j) will identify by formula 9.

$$CF_j = \frac{\beta_{h^\wedge}(R_j) - \beta\square}{\beta_h(R_j)}$$

$$\beta\square = \frac{\sum_{h \neq h^\wedge} h \neq h^\wedge(R_j)}{C-1} \qquad (9)$$

3.4. Fuzzy Deductive

With helping the way of producing a rule in previous part, we can produce randomly, N fuzzy rules.

Then we define which rule is related to which class and positiveness degree for all N *if-then* fuzzy rules the class of a new model(X) is defined as below:

Stage 1: accounting $\alpha_h(x)$ for class h and h=1, 2,…, c . So

$$\alpha_h(x) = \max \{M_j(x) * CF_j | C_j\text{=class h , j=1,2,…N} \}\text{ , h= 1,2,…C}$$

$$M_j(x) = M_{j1}(x_1) * \ldots * M_{jn}(x_n) \qquad (10)$$

Stage 2: finding a class for model (h_p^*) in which it has maximum amount of $\alpha_h(x)$

$$\alpha_{h_p^*}(x) = \max\,(\alpha_1(x), \ldots, \alpha_c(x)) \qquad (11)$$

If more than one class had max amount, then model X doesn't classify. Otherwise, class h_p^* specializes to model X.

In this way, only for achieving new rule, we uses of genetic algorithm and we ever try to produce new Rules using modern rules later; we will present the genetic algorithm such as used in this model.

4. New Rules and Improving Them

Here we use a mix of genetic algorithm and simulated annealing to produce new rules for improving them using genetic algorithm, available rules will change and produce new rules. With helping of simulated annealing we try to abscond (slip) of local extremisms.

4.1. Genetic Algorithm

Genetic algorithm can be considered as a way to do best randomly and objectively which move to the best point. In comparison with other doing best ways, we can say that genetic algorithm can be used for every matters without any limitation and with no information about the matter and its benefits in finding relative best is proved. The ability if this way is in solving the complicated matters of doing best in which we can't use classic ways or they aren't reliable.

4.1.1. The Structure of Genetic Algorithms

Genetic algorithms consist of:

* Chromosome: In genetic algorithm, each chromosome is a symbol of a point in searching space and a possible solution for the matter. Chromosomes themselves consist if fixed number of gene. For showing them, dependent on gene numbers (features), N is "between" 1 to 5.

* Population: A collection of chromosomes mace a population with effecting genetic factor (agent) on each population, new population with the same number is 100.

*Competence function: In order to solve each matter using genetic algorithms, at first there should be created a competence function. These functions return non-negative number for each chromosome in which shows that chromosome' ability .costiveness degree of each rule is as competence function.

4.1.2. Genetic Agents (Factors)

Genetic operators are described as bellow:

* Selection: Selecting 2 good rules and 2 bad rules from modern rules' collection based on their competence.

* Exchanging: Applying monotonous exchanging agent on 2 rules in order to shift the amounts of selected rules.

* Leaping: Applying agent on 2 suitable selected rules and changing these' amounts randomly in order to produce 2 new rules.

* Displacement: Displacing 2 produced new rules and changing rule with 2 other rules in which selected based on their less competence.

In selection agent, we select 20 good rules by leaping rule with 2 other rules from 100 rules using random function. Using leaping and exchanging agents, high competence rules will change 2 to 2 and produce new rules.

4.1.2.1. Exchange Agent

This exactly act such that confronted genes of 2 selected chromosomes (parents) will shift randomly from several points , and produce 2 new rules and displaced by 2 rules with less competence.

4.1.2.2. Leaping Agent

Difference of this agent with exchange agent is that it replaces randomly new amounts" between" 1 to 5 instead of previous amounts in selected genes of 2 chromosomes (parents) and 2 new chromosome will exchange with 2 leas competence chromosomes.

The main problem of genetic algorithm in solving matter, is falling in tap named relative maximum condition .It sometimes Coues we can't find best answer (abstract max) to the matter whereas sometimes for finding algorithm always try to improve the space of solving matter whereas sometimes for finding the best answer it is needed to find some worse answers and with improving them, find the best one (abstract max). so in this way we don't use of improving genetic algorithm way in which it always try to find better answers more than modern one instead we use of simulated annealing algorithm.

4.2. Simulated Annealing Algorithm

This algorithm is excerpted of read from world heating and melting metals. It's a simulated proceeding algorithm in which cold and solid the melt metal slowly and solids' structure features dependent the amount of cold. If fluid, be cold slowly, it will shape large crystals. So, if it cold rapidly, the crystals will be deficient.

Specialized movement because of heat dispersion in situation will be like random rise and down. But annealing algorithm used in our way is as bellow:

The rules produce recently using genetic algorithm, if have had better efficiency than the previous rules, will select certainly otherwise they will select probably .Of course more higher temperature, more will select bad rules collection. Because the movement of randomly parts is more often in higher temperatures

Generally, we always save 2 collections of rules:

1) The best rules collection that there was the most recognition of models because of it.
2) The rules collection that is probably one of the worst states and they use for escaping from local maximums. The changes are always applied on this rules collection(Current)

It is obvious that in high temperatures, the selection chance of a bad rule collection for current rule collection is more than the low temperatures. Because the chance of bad changes in low temperatures is less

Therefore if the new produced rule collection be better than the previous, they certainly will be in Swrrent. If they be worse than the previous, they maybe save as current rule collections of system depend on current temperature and random function of probability.

In fact, we try to produce new rules collection using genetic algorithm and there is no selecting of rules collection in it. Acceptance or rejection of new rules collection is done using simulated annealing algorithm.

// s_{init} is the initial set of rules

// s_{best} is the best set of rules

// EF_{best} is Evaluation Fitness for best set of rules

// $EF_{current}$ is Evaluation Fitness for current set of rules

// $T_{\max}$ is initial temperature

// $T_{\min}$ is the final temperature

```
// $\alpha$ is the cooling rate
// $\beta$ is a constant
// $Time$ is the time spent for the annealing process so far
// $k$ is the number of calls of metropolis at each temperature
Begin
    $T = T_{\max}$;
    $S_{current} = T_{init}$;
    $S_{best} = S_{current}$; // $S_{best}$ is the best set of rules
                                   Soon so far
    $EF_{current} = NNCP(S_{current})$;
    $EF_{best} = NNCP(S_{best})$;
    $Time = 0$;
    Repeat
        For i = 1 to k
            Call Metropolis (
                $S_{current}, EF_{current}, S_{best}, EF_{best}, T$ )
        $Time = Time + k;$
        $k = \beta \times k;$
        $T = \alpha \times T;$
    Until ( $T \geq T_{\min}$ );
    Return ( $S_{best}$ );
End. //Genetic-Simulated Annealing
```

```
Procedure Metropolis (
    $S_{current}, EF_{current}, S_{best}, EF_{best}, T$ )
    // $S_{new}$ is the new set of rules
Begin
```

$$S_{new} = \begin{cases} Selection(S_{current}); \\ Crossover(S_{current}); \\ Mutation(S_{current}); \end{cases}$$

```
        $EF_{new} = NNCP(S_{new});$
        $\Delta EF = (EF_{new} - EF_{current});$
        If ( $\Delta EF < 0$ ) Then
$S_{current} = S_{new};$
            If $EF_{new} < EF_{best}$ Then
                $S_{best} = S_{new};$
 End If
        Else If ( $random[0,1] < e^{-\Delta EF / T}$ ) Then
            $S_{current} = S_{new};$
        End If
End. //Metropolis
```

Figure 4: Semi-Code for Combining Genetic and Simulated Annealing Algorithms

5. Selecting Effective Features Using Genetic Algorithm

As described, each sample of disease contains some features. In the model of this research, the suitable features for classifying samples are selected using genetic algorithm from all features. For this features, Binary's chromosome will describe with length of dataset features. If a gene =1it means it uses and if it =0 it means it doesn't use in the classification. (Figure 5)

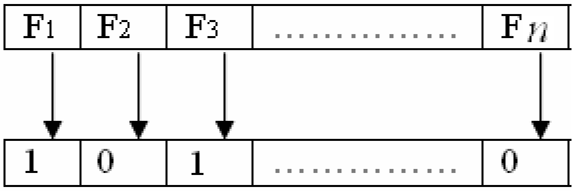

Figure 5: A chromosome and features selection

For producing initial population of genetic algorithm, Binary's chromosome produces randomly. After educating and testing each chromosome, the percentage of testing data's recognition is calculated and then its calculation way will define.

Exchange agent by producing a random number between 1 to n (n = the number of data set features) and displacing the amount of 2 chromosomes, produce new chromosomes. Leaping agent perform by producing a random number between 1 to n and changing gene amount from 1 to zero or vice versa. Producing new generation perform by selecting 50 percent randomly from previous race and 50 percent by associating parents (randomly) from previous population I which they have has more comely. Leaping agent happen with 15 percent exchange agent with 85 percent

With genetic algorithm in several races and selecting suitable chromosomes, the effective features will achieve.

The way performing algorithm especial for this goal is very time consuming. After producing chromosome, its recognition degree must be calculated.

For calculating the recognition degree, system has to educate for each chromosome and at last calculate its recognition degree, and this amount is selected as the amount of chromosomes' recognition degree. This project has to repeat for all chromosomes of a race and it's obvious it's time consuming. But after producing several race, features are lessen and dimension reduce, the classification speed in testing stage increase and it's mistakes reduce remarkably. In fact, how more time and cost spend education, it compensate in testing and system application.

Genetic algorithm guarantee that select the most effective features and offer the best situation to the user. Of course it's very slow and time –consuming.

6. Reeducation of available rules number

As it is obvious, how number of recognizing disease be less, the speed of class recognition for a new experimental sample is higher. So, it is necessary to eliminate some rules. The used way is such that after performing the program completely and producing 100 rules existing from educational data's, each rule will eliminate temporarily. If there isn't any change in recognition of experimental data's or it don't better, then the rule will eliminate permanently. Otherwise we overlook of eliminating that rule and test the next rule. We do that for all 100 rules and in this way we can reduce rules.

7. Result

The presented way was tested on lung cancer dataset with 32 samples, 56 features and 3 classes in which there is 9, 13 and 10 samples in order. The parameters quantity is shown in table 1 and features reduction (using PCA) is from 56 to 6. We use of 10 – fold cross validation for performing experiment. Each feature normalized separately between 0, 1.Wffective features for increasing system accuracy exploited using genetic algorithm. With suitable genetic algorithm and simulated annealing algorithm, it search for best rules in matters' situation and try to escape from local best point. Finally reduce rules using an evaluation way.

Table 1: Initialize the parameters of proposed method

parameters	value
Number of initial rules	100
initial temperature	10
final temperature	0.00001
coefficient of temperature decrease	0.95
Number of iterations of simulated cooling function at each temperature	15
Number of chromosomes to select effective features	20
Probability of cross over in effective features	85
Probability of mutation in effective features	15
Probability of cross over in set of rules	80
Probability of mutation in set of rules	20
Rate of Replacement in set of rules	20

The proposed way implemented using software and amount of correct recognition in lung cancer dataset is 99.66 percent and in experimental data's is 96.67 percent that indicates the efficiency of proposed algorithm collection.

With consideration to high accuracy of proposed model, we can use it in different diseases with high features, in processing pictures, speeches and generally in all fields in which we face to high features (dimensions).

REFRENCES

[1] Janecek, A.G.K., Gansterer, W.N., Demel, M.A., Ecker, G.F.: "On the relationship between feature selection and classification accuracy". Journal of Machine Learning and Research. JMLR: Workshop and Conference Proceedings 4. pp 90–105 (2008).

[2] Rakkrit Duangsoithong and Terry Windeatt: "Relevance and Redundancy Analysis for Ensemble Classifiers", Springer-Verlag Berlin Heidelberg. 2009.

[3] Hayward, J., Alvarez, S., Ruiz, C., Sullivan, M., Tseng, J. Whalen, G.: "Knowledge discovery in clinical performance of cancer patients". IEEE International Conference on Bioinformatics and Biomedicine, USA, pp. 51–58 (2008)

[4] Antonio Arauzo-Azofra, Jose Manuel Benitez and Juan Luis Castro: "A feature set measure based on Relief", RASC. 2004.

[5] Bai-Ning Jiang Xiang-Qian Ding Lin-Tao Ma, Ying He, Tao Wang,& Wei-Wei Xie. "A Hybrid Feature Selection Algorithm: Combination of Symmetrical Uncertainty and Genetic Algorithms". The Second International Symposium on Optimization and Systems Biology. pp. 152–157. Lijiang, China. October 31– November 3, 2008.

[6] http://archive.ics.uci.edu/ml/machine-learning-databases/lung-cancer/lung-cancer.data

[7] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski & Lukasz A. Kurgan: "Data Mining A Knowledge Discovery Approach". Springer Science, New York, 2007.

[8] Tomoharu nakashima and hisao ishibuchi: "Using Boosting Techniques To Improve The Performance Of Fuzzy Classification Systems" In "Classification And Clustering For Knowledge Discovery". Studies In Computational Intelligence. Vol 4. pp.146-157 . Springer, Netserlands. 2005.

[9] Ajith Abraham and ravi jain: "Soft computing models for network intrusion detection system". In "Classification And Clustering For Knowledge Discovery". Studies In Computational Intelligence. Vol 4. pp.190-207 . Springer, Netserlands. 2005.