T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

PREDICTING THE EXISTENCE OF MYCOBACTERIUM TUBERCULOSIS ON PATIENTS BY DATA MINING APPROACH

Master Thesis

Tamer UÇAR

İSTANBUL, 2009

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

Institute of Science
Computer Engineering Graduate Program

PREDICTING THE EXISTENCE OF MYCOBACTERIUM TUBERCULOSIS ON PATIENTS BY DATA MINING APPROACH

Master Thesis

Tamer UÇAR

SUPERVISOR: ASSOC. PROF. DR. ADEM KARAHOCA

İSTANBUL, 2009

T.C
BAHÇEŞEHİR ÜNİVERSİTESİ

The Graduate School of Natural and Applied Sciences

Computer Engineering

Title of the Master's Thesis	: Predicting The Existence Of Mycobacterium Tuberculosis On Patients By Data Mining Approach
Name/Last Name of the Student	: Tamer UÇAR
Date of Thesis Defense	: 10.08.2009

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Signature

Prof. Dr. A. Bülent ÖZGÜLER
Director

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Committee Members:

Assoc. Prof. Dr. Adem KARAHOCA (Supervisor) :

Asst. Prof. Dr. Yalçın ÇEKİÇ :

Prof. Dr. Nizamettin AYDIN :

ACKNOWLEDGEMENTS

I would like to thank all people who have helped and inspired me during my study.

Especially, I offer my sincerest gratitude to my supervisor, Assoc. Prof. Dr. Adem Karahoca, who has supported me, thought-out my thesis with his experience and knowledge. It would be impossible to complete this study without his encouragement, motivation and guidance.

I would like to show my gratitude to my father, Dr. Necmettin Uçar and my brother Dr. Tolga Uçar for their professional insight. Without their support, medical basis of this thesis would not be constructed.

I owe my deepest gratitude to my mother, Nedret Uçar, for her endless love and support throughout my life. Not only in this study, but also in every moment in my life her encouragement made everything easier than it is.

Finally, I would like to thank to my fiancée, Elif Çöğürlü, for her everlasting love, endless support and encouragement in every part of my life.

ÖZET

HASTALARDA MYCOBACTERIUM TUBERCULOSIS BAKTERİSİNİN VARLIĞININ VERİ MADENCİLİĞİ YAKLAŞIMI İLE TAHMİNİ

Uçar, Tamer

Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Adem Karahoca

Ağustos 2009, 42 sayfa

Günümüzde veri madenciliği yöntemleri birçok problemin çözümünde oldukça popüler bir tekniktir. Kısaca tanımlamak gerekirse, veri madenciliği mevcut bir veri kümesinden çeşitli örüntüler elde etmeye yarayan bir mekanizmalar bütünüdür. Elde edilen bu örüntüler, mevcut olan ya da yeni toplanan verilerin yorumlanarak bu verilerden anlamlı bilgilerin elde edilmesinde kullanılır. Birçok çalışma alanında geniş ölçekli veriler ile çalışılır. Bu verilerin anlamlı bilgiye dönüştürülmesinde çok sayıda farklı algoritmalar ve yaklaşımlar uygulanmıştır.

Biyomedikal alanı veri madenciliği tekniklerinin kullanılarak verilerin anlamlı bilgilere dönüştürülebildiği alanlardan biridir. Kalp atımlarının sınıflandırılması, Alzheimer hastalığında arkaplandaki MEG (Magnetoencephalography) aktivitesinin analizi, insandaki kalıtsal metabolik bozuklukların metabolik biyomarkerlar ile öngörülmesi ve kanda Sikolosporin A seviyelerinin tahmin edilmesi gibi konu başlıkları altında birçok veri madenciliği çalışması yapılmıştır.

Bu çalışma tüberküloz hastalarının sınıflandırılması problemi üzerinde yoğunlaşmıştır. Tüberkülozun kesin tanısının konmasında hastanın balgamında bakterinin bulunup bulunmadığına dair bir testin yapılması gereklidir. Bu testin neticesi de yaklaşık olarak 45 günlük bir zaman dilimi sonunda belli olmaktadır. Bizim çalışmamızın amacı, veri madenciliği tekniğini kullanarak tüberküloz hastalığının tanısını kesin tıbbi test sonuçlarını beklemeden, mümkün olduğunca tutarlı bir şekilde koyabilen bir sistem geliştirmektir. Sistemin tutarlı bir şekilde çalışması çok önemlidir. Çünkü gerçekte tüberküloz olmayıp sistem tarafından tüberküloz olarak sınıflandırılan hastalar 45 gün boyunca güçlü ve yoğun bir antibiyotik tedavisine boşu boşuna alınacaklar ve bunun sonunda gereksiz olarak kullandıkları ilaçların yan etkilerine maruz kalacaklardır. Aynı şekilde gerçekte tüberküloz olup sistem tarafından tüberküloz dışı sınıflandırılan hastalar da 45 gün boyunca tedaviye alınmayıp uygulanması gereken tedavi programına geç başlayacaklar ve mevcut hastalıkları daha da ilerlemiş olacaktır.

Yapmış olduğumuz çalışmamızın bulguları neticesinde ANFIS metodunun tüberküloz hastalarının sınıflandırılması konusunda Bayesian Network, Multilayer Perceptron, Part, Jrip ve RSES metodlarına göre daha tutarlı ve güvenilir olduğunu gördük.

Anahtar Kelimeler: ANFIS, Biyomedikal, Hastaların Sınıflandırılması

ABSTRACT

PREDICTING THE EXISTENCE OF MYCOBACTERIUM TUBERCULOSIS ON PATIENTS BY DATA MINING APPROACH

Uçar, Tamer

The Institute of Sciences, Computer Engineering Graduate Program

Supervisor: Assoc. Prof. Dr. Adem Karahoca

August 2009, 42 pages

Data mining techniques are very popular for solving various problems. As a brief description, data mining is a mechanism for obtaining patterns from an existing data set. Those extracted patterns are used to interpret the new or existing data into useful information. In most of the areas, large scaled data is collected. To convert these data into information, many different algorithms and approaches are used.

Biomedical is one of the areas where data mining can be applied to convert data into information. Many studies are made under topics such as classification of cardiac beat, analysis of MEG (Magnetoencephalography) background activity in Alzheimer's disease, predicting metabolic biomarkers of human inborn errors of metabolism, prediction of Cyclosporine A blood levels and etc.

This study focuses on classification of tuberculosis patients. To make a correct diagnosis of tuberculosis, a medical test must be applied to patient’s phlegm. The result of this test is obtained about after a time period of 45 days. The purpose of this study is to develop a data mining solution which makes diagnosis of tuberculosis as accurate as possible and helps deciding if it is reasonable to start tuberculosis treatment on suspected patients without waiting the exact medical test results or not. It is imperative that, there must be a very accurate classification for this model. Because false positive classified patients will use strong antibiotics for 45 days for nothing and they have to deal with its side affects. And the false negative classified patients' treatment plan will be suspended for 45 days and within this untreated period their disease will get even worse than it is. Therefore, correct prediction of tuberculosis is a very important issue.

According to the findings of our study, we concluded that ANFIS is an accurate and reliable method comparing to Bayesian Network, Multilayer Perceptron, Part, Jrip and RSES methods for classification of tuberculosis patients.

Keywords: ANFIS, Biomedical, Patient Classification

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Tuberculosis, which a few years ago was considered to be almost under control, has once again become a serious world-wide problem because of AIDS. Tuberculosis disease is caused by a bacterium which is called as mycobacterium tuberculosis. This disease can spread among humans and the patients who suffer from tuberculosis might die unless they get the right treatment. This microorganism widely exists on humans, cattle, sheep and birds. All of the organs in the body can be affected by tuberculosis. But most of the tuberculosis cases are occur in lungs (Davidson 1999, pp. 347-354).

Tuberculosis disease occurs under different manifestations on adults and children. When the first encounter happens with bacillus, which is mostly happens on the childhood phase of a person, lymphatic glands that are located at the entry point of the lungs are picked by this microorganism for the first rooting point on the body. As a result of this event, those glands enlarge (hilar lymphadenopathy). This is called as primary tuberculosis. The adult type (secondary) tuberculosis is different than this scenario: In those cases, the person's lung is contaminated with the microorganism before. If the immune system is strong enough, microorganism can not cause any sickness but can keep itself alive. When the immune system of the person weakens for a reason, microorganism gets activated and begins to create sickness. Prostration, long term sicknesses, insomnia, tobacco and alcohol abuse, drug addiction, having an irregular life, malnutrition, stress, et cetera are some factors which are responsible for weakening the immune system and providing a suitable basis for illness to occur. Unlike primary tuberculosis, lesions are spread to lung parenchyma tissue in secondary tuberculosis cases. Cavities (holes) which may cause lung tissue to bleed can also be seen on advanced phases of the illness (Harrison 1999, pp. 1007-1014).

Lung tuberculosis can be seen on very wide age range. From new born babies to old people, everybody can be affected by this disease. Symptoms are: cough, fatigue, exhaustion, anorexia, night sweating, fever (which not exceeds 37.5 centigrade degree), cavities and hemoptysis on advanced cases (Özlü, Metintaş & Ardıç 2008, pp. 323-340).

To make an exact diagnosis, existence of microorganism in phlegm must be proven. But, some other microorganisms can also be flagged as mycobacterium tuberculosis under microscope observation. In order to avoid this problem, a special culture medium is prepared where only bacteria of mycobacterium tuberculosis can reproduce. The phlegm sample which is obtained from patient is planted to this medium and kept for 45 days at body temperature. At the end of this time period, the culture medium is checked for any reproduction sign of the bacteria.

In order to cure tuberculosis, 4-5 different major antituberculotic antibiotics are used for 6-12 months. Some cases may heal without any treatment plan if immune system is strong enough. After full recovery, lung wounds which are caused by tuberculosis disease still exist as calcific tissue. Unfortunately, cases which are not treated may result by death of patient (Harrison 1999).

A time period of 45 days is required in order to make a correct diagnosis. The aim of this study is to develop a data mining solution which makes diagnosis of tuberculosis as accurate as possible and helps deciding if it is reasonable to start tuberculosis treatment on suspected patients without waiting the exact test results or not. It is imperative that, there must be high sensitivity and specificity results for this model. Because false positive classified patients will use strong antibiotics for 45 days for nothing and they have to deal with its side effects. And the false negative classified patients' treatment plan will be suspended for 45 days and within this untreated period their disease will get even worse than it is. Therefore, correct prediction of tuberculosis is a very important issue.

1.2 BACKGROUND

Today, data mining techniques are used in very different areas. As mentioned earlier, this study focuses on predicting the existence of mycobacterium tuberculosis on patients by using ANFIS. Besides this study, there are two other research papers regarding this issue. In the following section, those studies will be mentioned. And after, recent researches on biomedical area using ANFIS will be referred.

1.2.1 Tuberculosis and Data Mining

Bakar and Febriyani applied Rough Neural Networks for classification of tuberculosis patients. Data set has 233 records, which has 14 attributes, firstly reduced as a result of preprocessing of data. The decisive data set is having 8 attributes which are gender, age, weight, fevers, night sweats, cough>3 weeks, blood phlegm and sputum test. 70% (131 data) of the data set is used for training and 30% (56) is used for testing. Discretization is applied on the numeric and continuous attributes using rough set application. After then, neural network is applied for training the data (Bakar & Febriyani 2007).

The second one is a chapter from the book “Data Mining and Medical Knowledge Management Cases and Applications”. On chapter XVI, mining tuberculosis data issue is considered. The aim of this study is classifying tuberculosis diagnostic categories based on given variables. Records of 1655 patients having 56 attributes are used as raw data set. Those 56 attributes are reduced into 5 attributes which are antecedents, bacteriology result, age category, pulmonary tuberculosis, and extra pulmonary tuberculosis. Exhaustive CHAID is selected for generating decision trees for classes (Sánchez, Uremovich & Acrogliano 2009).

1.2.2 Biomedical and Data Mining

Diagnosis of diabetes by using adaptive neuro fuzzy inference systems is another application of ANFIS. That study focuses on the fact that, determining the risks of diabetes is the best method for permeating it. According to this fact, the aim of this research is estimating diabetes risk depending on some variables such as age, total cholesterol, gender or shape of the body by using ANFIS. The data set has 390 patients’ records each having 4 variables. 300 of those records are used for training and 90 are used for checking (Kara & Karahoca 2009).

Another data mining approach on a biomedical topic is classification of cardiac beat using a fuzzy inference system. For training and testing data sets, MIT Arrhythmia Database and in-vivo records from cardiac voluntary patients were used. The point of this study is identifying and classifying normal versus premature ventricular contractions (PVC). Data used in this research has 34 records. Those records contain 4917 PVCs and 55508 normal beats. 2027 beat data which has 520 PVCs are used for training ANFIS (Monzon & Pisarello 2005).

Another data mining study is made on Alzheimer’s disease under the topic as Analysis of MEG background activity in Alzheimer's disease using nonlinear methods and ANFIS. This study intends to analyze magneto encephalogram background action on patients using sample entropy and Lempel-Ziv complexity (Gómez et al. 2009).

Shlomi et al. studied for predicting metabolic biomarkers of human inborn errors of metabolism. The motivation provider on this research is publication of the genome-scale network model of human metabolism. In the light of this event, researchers offer a novel computational approach for systematically predicting metabolic biomarkers in stochiometric metabolic models (Shlomi, Cabili & Ruppin 2009).

One of the latest researches about biomedical data mining is performed under the topic of Prediction of Cyclosporine A blood levels: an application of the adaptive-network-based fuzzy inference system (ANFIS) in assisting drug therapy. The aim of this study is predicting the results of the therapeutic drug monitoring (TDM) process with the help of ANFIS. Data was collected from 138 patients, each containing 20 input parameters. Both Takagi and Sugeno-type ANFIS is used to predict the concentration of Cyclosporine A in blood samples (Gören et al. 2008).

2. MATERIAL & METHODS

In this part, materials and applied methods are considered.

2.1 PREPARING TUBERCULOSIS DATA SET

Data set contains information about 503 patients who are examined at a clinic. Each of those records consists of 30 different variables. The full list of those variables is as following:

Table 2.1: Full list of variables

Gender	Loss of appetite	Erythrocyte
Age group	Loss in weight	Haematocrit
Weight	Sweating at nights	Haemoglobin
Smoke addiction	Chest pain	Leucocyte
Alcohol addiction	Back pain	Number of leucocyte types
BCG vaccine	Coughing	Active specific lung lesion
Malaise	Hemoptysis	Calcific tissue
Arthralgia	Fever	Cavity
Exhaustion	Sedimentation	Pneumonic infiltration
Unwillingness for work	PPD	Pleural effusion

Some variables contain direct values, some contain cluster values. The following table shows possible values that variables can have:

Table 2.2: List of types and acceptable values of variables

Variable Name	Data Type	Acceptable Values
Gender	Boolean	Female=0, Male=1
Age group	Integer	18-24=1, 25-32=2, 33-40=3, 41-45=4, 46-51=5, 52-57=6, 58+=7
Weight	Integer	40+

Smoking addiction	Integer	None=0, Little(<5 items)=1, Moderate(6-10 items)=2, Very Much(11+ items)=3
Alcohol addiction	Boolean	No=0, Yes=1
BCG vaccine	Boolean	No=0, Yes=1
Malaise	Boolean	No=0, Yes=1
Arthralgia	Boolean	No=0, Yes=1
Exhaustion	Boolean	No=0, Yes=1
Unwillingness for work	Boolean	No=0, Yes=1
Loss of appetite	Boolean	No=0, Yes=1
Loss in weight	Boolean	No=0, Yes=1
Sweating at nights	Boolean	No=0, Yes=1
Chest pain	Boolean	No=0, Yes=1
Back pain	Boolean	No=0, Yes=1
Coughing	Integer	No=0, Yes=1, With mucous=2
Hemoptysis	Boolean	No=0, Yes=1
Fever	Integer	Normal=0, High=1, Subfebrile=2
Sedimentation	Integer	Normal=0, Moderate=1, High=2
PPD	Boolean	Negative=0, Positive=1
Erythrocyte	Integer	Normal=0, Low=1, High=2
Haematocrit	Integer	Normal=0, Low=1, High=2
Haemoglobin	Integer	Normal=0, Low=1, High=2
Leucocyte	Integer	Normal=0, Low=1, High=2
Number of leucocyte types	Integer	Normal=0, Lymphocytic dense=1, Macrophage dense=2
Active specific lung lesion	Boolean	No=0, Yes=1
Calcific tissue	Boolean	No=0, Yes=1
Cavity	Boolean	No=0, Yes=1
Pneumonic infiltration	Boolean	No=0, Yes=1
Pleural effusion	Boolean	No=0, Yes=1

Before generating ANFIS model, attribute ranking function is applied using information gain ranking filter in WEKA (Witten & Frank 2005) platform. The following list shows the ranking result for each variable:

Table 2.3: Ranking of variables

Rank Percentage	**Variable**
0.70740	Active specific lung lesion
0.55116	Calcific tissue
0.48265	Number of leucocyte types
0.43528	Weight
0.38664	Fever
0.37107	Age group
0.35686	PPD
0.31945	Sweating at nights
0.31389	Leucocyte
0.21179	Loss in weight
0.21131	Hemoptysis
0.18745	Cavity
0.17851	Sedimentation
0.15977	Loss of appetite
0.13992	Pneumonic infiltration
0.11917	Exhaustion
0.11589	Unwillingness for work
0.11027	Haemoglobin
0.11027	Haematocrit
0.10775	Erythrocyte
0.09271	BCG vaccine
0.04021	Arthralgia
0.03608	Chest pain
0.03187	Smoking addiction
0.03029	Gender
0.02764	Malaise
0.02534	Coughing
0.01626	Back pain

0.01276	Alcohol addiction
0.00459	Pleural effusion

The variables which are ranked less than 10% were eliminated. According to this reducing on data set, BCG vaccine, arthralgia, chest pain, smoking addiction, gender, malaise, coughing, back pain, alcohol addiction and pleural effusion variables were ignored.

The distribution of patients by their age groups can be seen in the figures below.

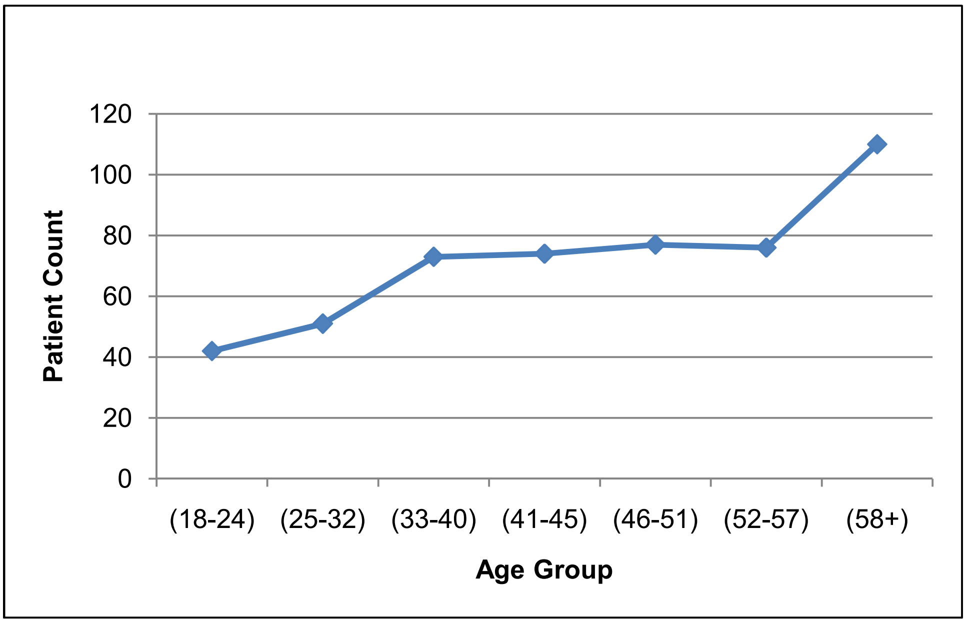

Figure 2.1: Distribution of patients by their age groups

2.2 ADAPTIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS)

ANFIS is a neural-fuzzy system which contains both neural networks and fuzzy systems. A fuzzy-logic system can be described as a non-linear mapping from the input space to the output space. This mapping is done by converting the inputs from numerical domain to fuzzy domain. To convert the inputs, firstly, fuzzy sets and fuzzifiers are used. After that process, fuzzy rules and fuzzy inference engine is applied to fuzzy domain (Jang 1992) (Jang 1993). The obtained result is then transformed back to arithmetical domain by using defuzzifiers. Gaussian functions are used for fuzzy sets and linear functions are used for rule outputs on ANFIS method. The standard deviation, mean of the membership functions and the coefficients of the output linear functions are used as network parameters of the system.

The summation of outputs is calculated at the last node of the system. The last node is the rightmost node of a network. In Sugeno fuzzy model, fuzzy if-then rules are used (Sugeno & Kang 1988) (Takagi & Sugeno 1985). The following is a typical fuzzy rule for a Sugeno type fuzzy system:

If x is A and y is B then x = f(x, y)

In this rule, A and B are fuzzy sets in anterior. The crisp function in the resulting is z=f(x, y). This function mostly represents a polynomial. But exceptionally, it can be another kind of function which can properly fit the output of the system inside of the fuzzy region that is characterized by the anterior of the fuzzy rule. We use first-order Sugeno fuzzy model for cases which are having f(x, y) as a first-order polynomial. This model was originally proposed in (Sugeno & Kang 1988) (Takagi & Sugeno 1985). We use zero-order Sugeno fuzzy model for cases where f is constant. This can be called as a special case for Mamdani fuzzy inference system (Mamdani & Assilian 1975). In this case, a fuzzy singleton is defined for each rule's resultant. Or, this can be also called as a special case for Tsukamoto's fuzzy model (Tsukamato 1979, pp. 137-149). In this case, a membership function of a step function is defined where it is centered at the constant for each rules' consequent. Additionally, a radial basis function network under certain minor constraints is functionally correlative to a zero order Sugeno fuzzy model (Jang 1993). Let's scrutinize a first-order Sugeno fuzzy inference system having two rules:

Rule 1: If X is A_1 and Y is B_1, then $f_1 = p_1x + q_1y + r_1$
Rule 2: If X is A_2 and Y is B_2, then $f_2 = p_2x + q_2y + r_1$

In the following figure, the fuzzy reasoning system is illustrated in a shortened form (Jang 1996). In order to bypass excessive computational complexity in the process of defuzzification, only weighted averages are used.

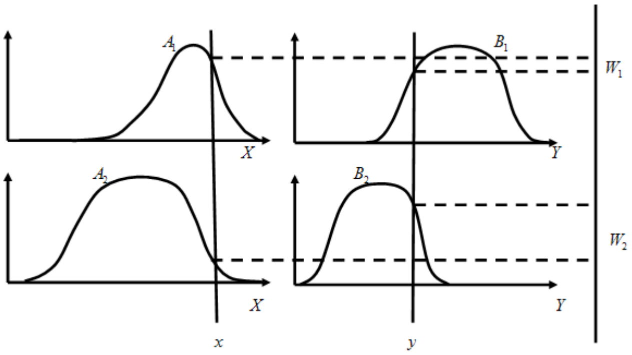

Figure 2.2: First-order Sugeno fuzzy model

$$\begin{aligned} f_1 &= p_1 x + q_1 y + r_1 \\ f_2 &= p_2 x + q_2 y + r_2 \end{aligned} \quad \Rightarrow \quad \begin{aligned} t &= \frac{w_1 + f_1 + w_2 f_2}{w_1 + w_2} \\ &= \overline{w_1}\, f_1 + \overline{w_2}\, f_2 \end{aligned} \qquad \textbf{(2.1)}$$

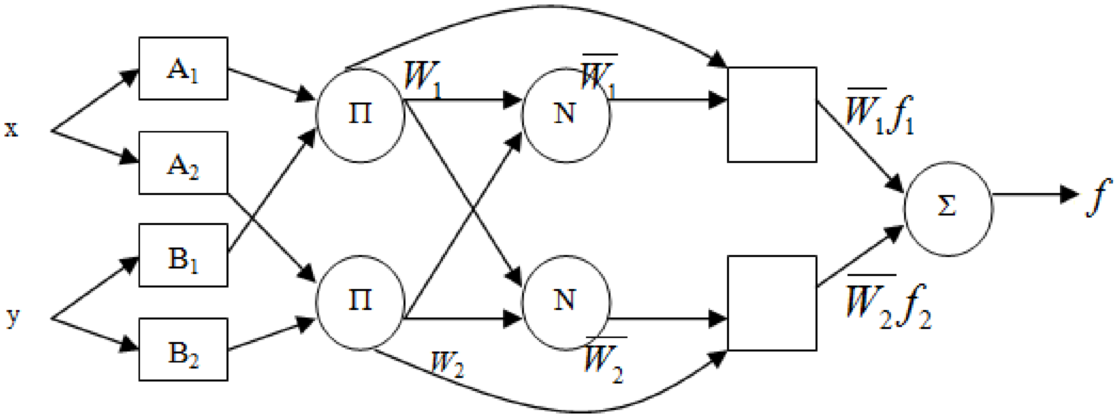

Figure 2.3: ANFIS Architecture

On the previous figure, we see a fuzzy reasoning system. This system generates an output which is shown as f. To generate this output, system accepts an input vector [x, y]. The output is calculated by computing each rule's weighted average. Those weights are achieved from the product of the membership grades in the assumption part. Using adaptive networks which are bound with the fuzzy model can compute gradient vectors. This computation is very helpful for learning of the Sugeno fuzzy model. In the next figure, we see the resultant network. This network architecture is called as ANFIS (Adaptive Neuro-Fuzzy Inference System).

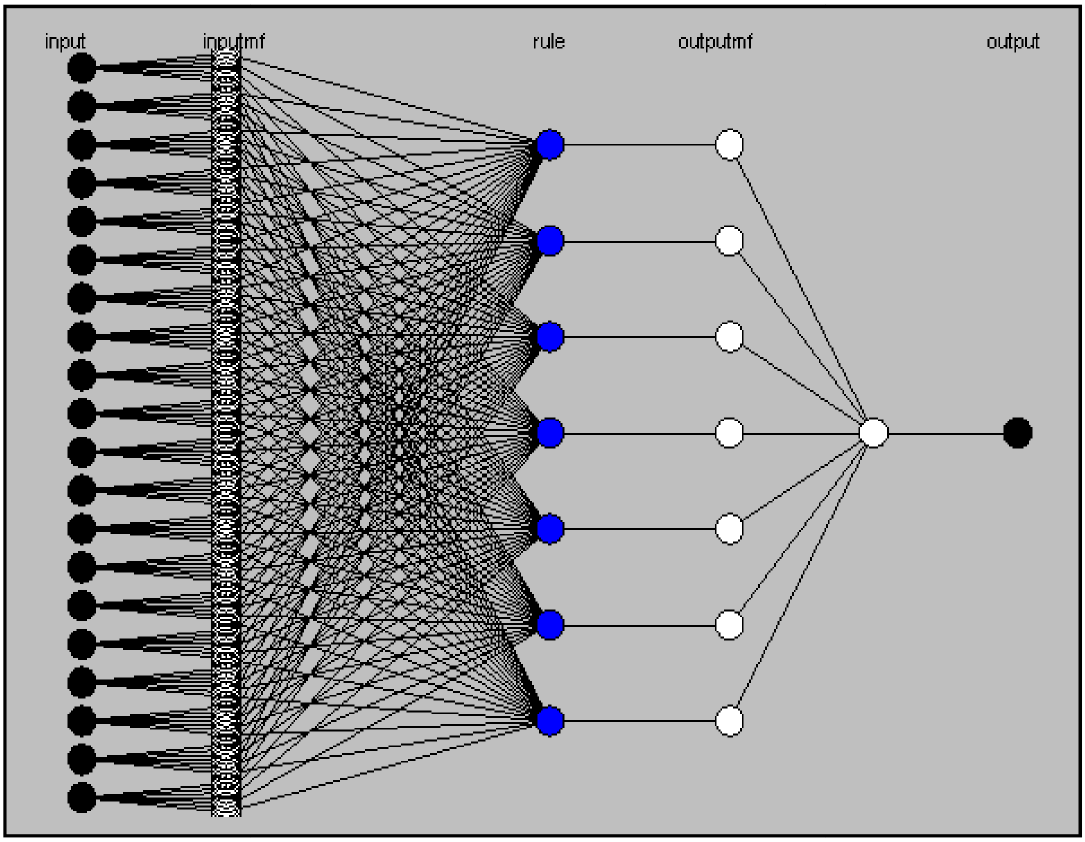

Figure 2.4: ANFIS model of fuzzy interference

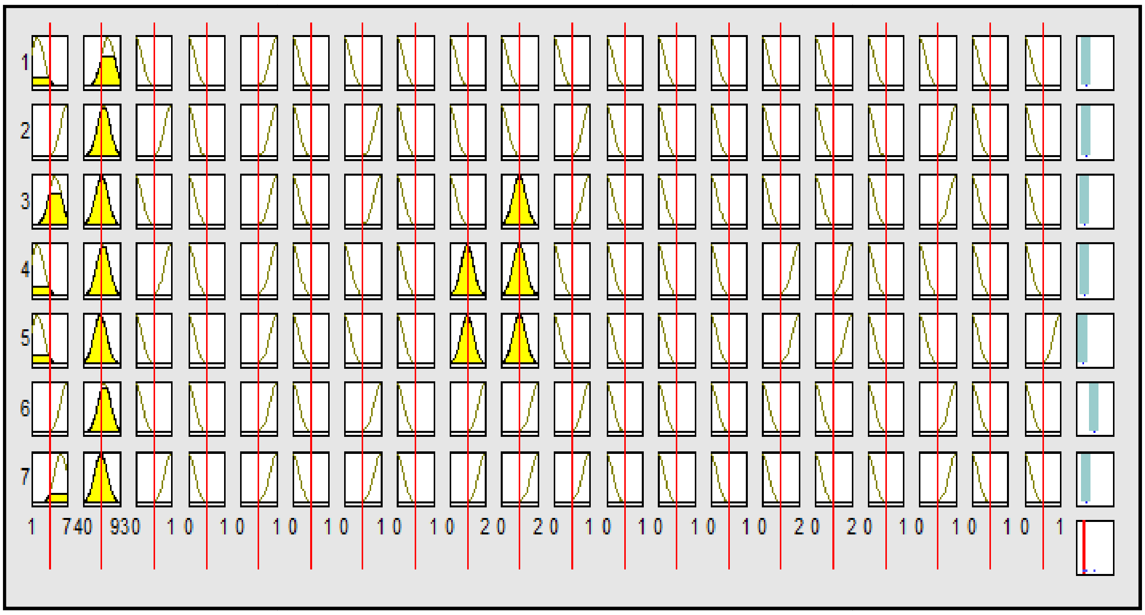

Figure 2.5: Sample rule set of an ANFIS model

The learning algorithm that ANFIS uses contains both gradient descent and the least-squares estimate. This algorithm runs over and over till an acceptable error is reached. Running process of each iteration has two phases: forward step and backward step. In forward step, linear least-squares estimate method is used for obtaining consequent parameters and precedent parameters are corrected. In backward step, fixing of consequent parameters is done. Gradient descent method is used for updating precedent parameters. And also, the output error is back-propagated through network.

It is very important that the number of training epochs, the number of membership functions and the number of fuzzy rules hold a critical position in the designing of ANFIS. Adjusting of those parameters is very crucial for the system because it may lead system to over-fit the data or will not be able to fit the data. This adjusting is made by a hybrid algorithm combining the least squares method and the gradient descent method with a mean square error method. The lesser difference between ANFIS output and the actual objective means a better (more accurate) ANFIS system. So we tend to reduce the training error in training process.

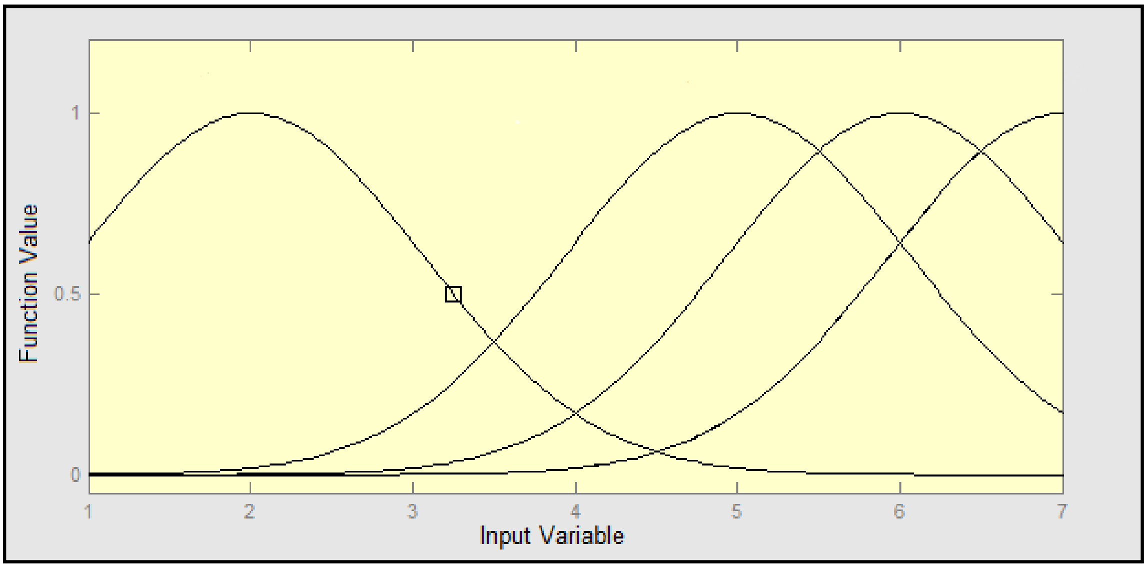

Figure 2.6: A sample membership function plot

A brief summary of 6 of the ANFIS layers algorithm can be viewed in the following table. Each layer is described and necessary formulas are stated.

Table 2.4: Layers of ANFIS Algorithm

Layer 0: It consists of plain input variable set.

Layer 1: Each node in this layer generates a membership grade of a linguistic label. For instance, the node function of the i-th node may be a generalized bell membership function:

$$\mu_{A_i}(x) = \frac{1}{1+\left[\left(\frac{x-c_i}{a_i}\right)^2\right]^{b_i}} \qquad \textbf{(2.2)}$$

Where x is the input to node *i*; A$_i$ is the linguistic label (small, large, etc.) associated with this node; and {a$_i$, b$_i$, c$_i$} is the parameter set that changes the shapes of the membership function. Parameters in this layer are referred to as the premise parameters.

Layer 2: The function is a T-norm operator that performs the firing strength of the rule, e.g., fuzzy conjunctives AND and OR. The simplest implementation just calculates the product of all incoming signals.

$$w_i = \mu A_i(x)\mu B_i(y), i = 1,2. \qquad \textbf{(2.3)}$$

Layer 3: Every node in this layer is fixed and determines a normalized firing strength. It calculates the ratio of the jth rule's firing strength to the sum of all rules firing strength.

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2. \qquad \textbf{(2.4)}$$

Layer 4: The nodes in this layer are adaptive and are connected with the input nodes (of layer 0) and the preceding node of layer 3. The result is the weighted output of the rule j.

$$\overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \qquad \textbf{(2.5)}$$

Where $\overline{W_i}$ is the output of layer 3, and {p$_i$, q$_i$, r$_i$}is the parameter set. Parameters in this layer are referred to as the consequent parameters.

Layer 5: This layer consists of one single node which computes the overall output as the summation of all incoming signals.

$$\text{Overall Output} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \qquad \textbf{(2.6)}$$

The constructed adaptive network in Figure 2 is functionally equivalent to a fuzzy inference system in Figure l. The basic learning rule of ANFIS is the back-propagation gradient descent (Werbos 1974), which calculates error signals (the derivative of the squared error with respect to each node's output) recursively from the output layer backward to the input nodes. This learning rule is exactly the same as the back-propagation learning rule used in the common feed-forward neural networks (Jang 1992) (Jang 1993) (Jang 1996) (Chiu 1997).

2.3 BAYESIAN NETWORK

Bayesian Networks produce probability estimates as network output like logistic regression models. Prediction is not produced by the system itself. The main purpose of the system is estimating the probability of an instance for each class value whether that value suits for a class or not. If we compare plain predictions with probability estimates, we see that probability estimates are more useful than plain predictions because we can rank the predictions with probability estimates. In Bayesian Networks, the conditional probability distribution of the value of a given class attribute is predicted within other class attribute (Witten & Frank 2005).

2.4 MULTILAYER PERCEPTRON

An artificial neural network is a simulation system based on mathematical models. Those systems are called as "neural networks" because their working principles are inspired from biological neural networks. Artificial neural networks are basically non-linear statistical data modeling tools. They usually have many inputs (each having different weights) and one output. A neural network has multiple layers. Those layers are mostly input layer, hidden layer and output layer. At input layer, the network gets its values from a vector of variables. At the hidden layer, each input is multiplied by their weight and the results are summed to produce a combined value. And then, this value is fed into a function which will generate the output of the network (Witten & Frank 2005).

Multilayer perceptron is an artificial neural network which has a feed forward structure. Feed forward means that the values only move through the network layers, no resultant values are fed back to any previous inner network layer. A multilayer perceptron network must have an input and an output layer. But the number of hidden layers may change due to the network architecture.

2.5 RIPPER ALGORITHM (JRIP)

RIPPER is an acronym for Repeated Incremental Pruning to Produce Error Reduction. In this algorithm, classes are considered by their sizes. Incremental reduced-error pruning is used for initial rule generation for the classes. After having individual rule sets for each class, two variants is produced for each rule. Again, reduced-error pruning is used and the instances that are covered by other class rules are removed. After that process, if one of the variants is better than the original rule, the rule is replaced by the variant (Witten & Frank 2005).

2.6 PARTIAL DECISION TREES

A partial decision tree is indifferent from conventional decision trees which are having branches to other sub-trees. To generate this kind of tree, a recursive algorithm is required to divide the instances into smaller subsets. The rules for partial decision trees are generated different than standard approach. Rule generation process is done by building a pruned decision tree for the current set of instance and the leaf which has the largest coverage is promoted as a rule. The following summarizes the partial tree generation:

```
Expand-subset(S):
  Choose a test T and use it to split the set of examples into subsets
  Sort subsets into increasing order of average entropy
  while (there is a subset X that has not yet been expanded
         AND all subsets expanded so far are leaves)
    expand-subset(X)
  if (all the subsets expanded are leaves
      AND estimated error for subtree >= estimated error for node)
    undo expansion into subsets and make node a leaf
```

As we see in the algorithm above, there is iteration for every subset. On each step, the selected subset is expanded. This process is repeated until there is no subset left unexpanded (Witten & Frank 2005).

2.7 ROUGH NEURAL NETWORKS

Rough Set Theory is firstly brought up by Zdzisław I. Pawlak who is a mathematician. This methodology is used for finding which attributes separates one class or classification from another (Pawlak 1982). To do this, rules must be generated on a training data set. Then, those generated rules should be applied to a test data set by using rough set classification methods in order to accomplish the necessary classification task. In rule generation phase of Rough Set Theory, we see that lower and upper bounds are considered in a very useful way (Pawlak 1982).

Based on Rough Set Theory, creating Rough Neural Networks are introduced by Lingras (Lingras 1996). According to Lingras, every neuron in a Rough Neural Network consists of two pairs. One pair is called as upper neuron (for upper bound value) and the other is called as lower neuron (for lower bound value). The information can be changed between those two neurons as well as other rough neurons.

The rough neurons can use rough patterns. Rough patterns are built on rough values. Basically rough values contain an upper and a lower value. In fact, this is a definition for a value range. So this type of values can be used to represent variables such as weight, age, etc. (Lingras 1998).

2.8 STATISTICAL ACCURACY METRICS

Statistical accuracy metrics are used for measuring experimental results. Widely used common metrics are mean absolute error, mean square error and root mean squared error. In this thesis study, one of the metrics that I preferred for benchmarking the methods is root mean squared error.

2.8.1 Root Mean Squared Error

Root mean squared error, also called as RMSE, is used to measure errors that is biased to weigh large errors disproportionately more heavily than small errors. RMSE is calculated by the following formula:

$$RMSE = \sqrt{E} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - r_i)^2}{N}} \qquad (2.6)$$

In this formula, p_i is the predicted values, r_i is the actual values and N is the total number of the records used in calculation.

2.9 RECEIVER OPERATING CHARACTERISTIC

ROC (which stands for Receiver Operating Characteristic) is a visualization technique for experimental results. ROC is basically a graphical plot of sensitivity and specificity values. Spackman (1989) is one of the earliest appliers of ROC graphs. He used ROC graphs in machine learning for evaluating and comparing algorithms (Spackman 1989). ROC sensitivity is an important metric for measuring system's distinguishing capability between clusters. In order to draw ROC plot, we need to calculate sensitivity and specificity values. Sensitivity and specificity values are calculated by true positive, true negative, false negative and false positive values. True positive (TP) value is the number of positive examples correctly predicted by the classification model. False negative (FN) value is the number of positive examples wrongly predicted as negative by the classification model. False positive (FP) value is the number of negative examples wrongly predicted as positive by the classification model. And true negative (TN) value is the number of negative examples correctly predicted by the classification model (Fawcett 2004).

Table 2.5: Structure of a confusion matrix

		Actual Value		
		p	**n**	
Predicted Value	**p'**	True Positives	False Positives	**P'**
	n'	False Negatives	True Negatives	**N'**
	Totals	**P**	**N**	

$$Sensitivity = \frac{TP}{P} \tag{2.7}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2.8}$$

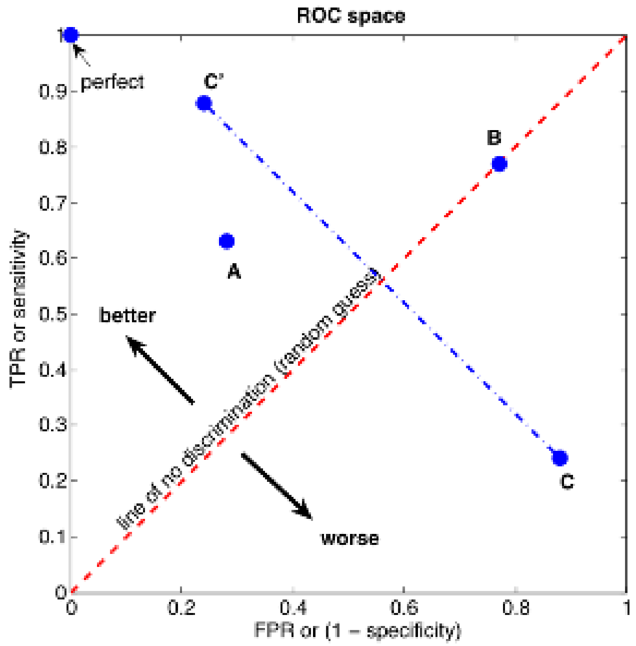

Figure 2.7: A sample ROC space plot

Source: http://en.wikipedia.org/wiki/File:ROC_space.png

3. FINDINGS

In this section, the results of the thesis will be stated. Different models on the data were applied which are Bayesian Networks, Multilayer Perceptron, JRip (a RIPPER algorithm implementation), PART (a Partial Decision Trees algorithm implementation), RSES (a Rough Set algorithm implementation) and ANFIS. As mentioned earlier, in this study, ANFIS method is mostly focused. The other methods that are used for comparison of ANFIS results. In the table below, there is a full listing of the results of the methods which are used in thesis.

Table 3.1: Benchmarking of methods

Method Name	Sensitivity (TPR)	Root Mean Squared Error
ANFIS	0.80	0.17
Bayesian Network	0.84	0.22
Multilayer Perceptron	0.84	0.23
Part	0.84	0.22
Jrip	0.79	0.25
RSES	0.69	0.37

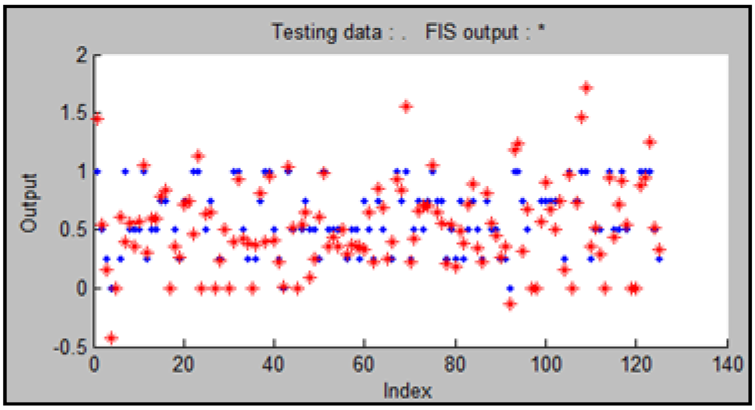

Figure 3.1: ANFIS testing error plot

As we see above, test results of the six methods are very close to each other. ANFIS has the best root mean squared error value whereas Bayesian Network, Multilayer Perceptron and Part methods have the best sensitivity (true positive rate) values. As an overall result, RSES method produced the worst values both on sensitivity and root mean squared error results.

MATLAB's Fuzzy Logic Toolbox is used for ANFIS method. For Bayesian Network, Multilayer Perceptron, Part and Jrip methods, built in algorithm implementations of WEKA (Witten & Frank 2005) is used. And for RSES algorithm, ROSETTA (Øhrn 1999) software is set in motion. WEKA and ROSETTA are both freeware data mining tools which are distributed freely (Witten & Frank 2005) (Øhrn 1999).

RSES algorithm generated 30 rules for fuzzy rough set classification. The following is the full list of the generated rough set rules:

Rule 1: Leucocyte(2) => Result(0.00) OR Result(1.00)

Rule 2: Weight([76, 95]) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 3: Weight([0, 45]) => Result(0.75) OR Result(1.00)

Rule 4: Fever(1) => Result(0.00)

Rule 5: Age Group(2) => Result(0.00) OR Result(0.75) OR Result(1.00)

Rule 6: Age Group(1) => Result(0.00) OR Result(1.00)

Rule 7: Age Group(6) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 8: Age Group(3) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 9: Age Group(7) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 10: Age Group(4) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 11: Age Group(5) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 12: Pneumonic infiltration(1) => Result(0.00)

Rule 13: Calcific tissue(0) => Result(0.00) OR Result(1.00)

Rule 14: Unwillingness for work(1) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 15: Unwillingness for work(2) => Result(1.00)

Rule 16: Sedimentation(0) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 17: Sedimentation(2) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 18: Loss in weight(1) AND Sweating at nights(0) => Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 19: Age Group(3) AND Loss in appetite(1) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75)

Rule 20: Number of leucocyte types(1) => Result(0.25) OR Result(0.50) OR

Result(0.75) OR Result(1.00)

Rule 21: Loss in weight(1) => Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 22: Erythrocyte(1) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 23: Loss in weight(1) AND Sedimentation(1) => Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 24: Age Group(6) AND Exhaustion(0) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75)

Rule 25: Age Group(4) AND Loss in weight(1) => Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 26: Age Group(7) AND Fever(2) => Result(0.00) OR Result(0.25) OR Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 27: Hemoptysis(1) => Result(0.50) OR Result(0.75) OR Result(1.00)

Rule 28: Hemoptysis(2) => Result(0.75)

Rule 29: Active specific lung lesion(1) => Result(1.00)

Rule 30: Cavity(1) => Result(1.00)

If we examine the rules above, we see possibilities for each class output for given inputs. Those inputs are mostly single variables but combinations of more than one variable are also seen. Each rule points to a result and union of multiple rules conclude the resultant classification for a given input set.

In rule 1, system states that a patient having a high leucocyte value is classified as cluster 0 or cluster 1.

In rule 2, it is told that patients who are having weight value within range of 76kg and 95kg can be classified as in every output clusters.

Rule 3 describes that patients who are less than 45kg can be classified as in cluster 0.75 or in cluster 1.

Rule 4 expresses that if a patient is having a high fever value; then he/she is cannot be suffering tuberculosis.

Rule 5, rule 6, rule 7, rule 8, rule 9, rule10 and rule 11 states possible output classes on single age group values.

Rule 12 describes that if pneumonic infiltration is positive on a patient; then it indicates that this patient is most probably not suffering from tuberculosis.

In rule 13, system states that if existence of calcific tissue is negative; then patient can be suffering from tuberculosis, or it may also indicate that patient is suffering from another disease.

Rule 14 shows that unwillingness of work parameter can be negative for all classes and in rule 15 we see that this parameter is positive for class 1.0.

In rule 16 and 17, it is expressed that, sedimentation value can be observed as high or low for each of the classes.

Rule 18 shows that having positive on loss of weight but negative on sweating at nights means that the patient is most probably in one of the positive classes such as 0.25, 0.50, 0.75 or 1.

In rule 19, we see that, having an age group value as 3 and a positive value on loss of appetite is not a solid indicator.

In rule 20, it is stated that, if lymphocytic density is positive, system can classify such a case in one of the classes: 0.25, 0.50, 0.75 or 1.

Rule 21 is similar to rule 20 in structural way. If loss in weight value is positive, it means that system can classify such a case in one of the classes: 0.25, 0.50, 0.75 or 1.

Rule 22 shows us that if erythrocyte value is low, this kind of patient can be belonging to any class. So this rule does not provide us a solid indication for classification.

In rule 23, we see that, if loss in weight is positive and sedimentation is moderately high, then the patient can be a member of the classes: 0.25, 0.50, 0.75 or 1.

Rule 24 shows us that if age group parameter is 6 and exhaustion is negative, this kind of patient can be belonging to any class. So this rule does not provide us a solid indication for classification.

Rule 25 expresses that if age group parameter is 4 and loss in weight parameter is true; then it indicates that such a case can be classified as 0.50, 0.75 or 1.

Rule 26 covers a case for age group is 7 and a subfebrile typed fever. This case can be suitable for all of the classes.

In rules 27 and 28 we see that if hemoptysis is positive, the patient can be classified as 0.75.

In rule 29, if active specific lung lesion is positive; then this case is most probably classified as class 1.

And rule 30 suggests us that there is a relation between tuberculosis disease and cavity parameter. But this is not very likely in most cases.

Table 3.2: Confusion matrix of Rough Set test data

		Actual Classes				
		0	**0.25**	**0.50**	**0.75**	**1**
Predicted Classes	**0**	27	7	0	0	1
	0.25	0	1	2	0	0
	0.50	0	4	0	0	0
	0.75	1	10	18	22	0
	1	1	1	1	0	30

Table 3.3: MATLAB code of generating and training FIS

```
% Load data files
% ---------------

load c:\data\train_rnd_in.txt
load c:\data\train_rnd_out.txt
load c:\data\train_rnd.txt
load c:\data\check_rnd.txt
load c:\data\check_rnd_in.txt
load c:\data\check_rnd_out.txt
load c:\data\test_rnd.txt
load c:\data\test_rnd_in.txt
load c:\data\test_rnd_out.txt


% Generate initial fis
% --------------------

fismat = genfis2(train_rnd_in, train_rnd_out, 0.5);


% Train fis for 3 epoches
% -----------------------

for ct=1:3
     [fismat,error] = anfis(train_rnd, fismat,2, NaN, check_rnd, 1);
end;


% Evaluation of generated fis
% ---------------------------

predicted_train_output = evalfis(train_rnd_in, fismat);
predicted_test_output  = evalfis(test_rnd_in, fismat);
predicted_check_output = evalfis(check_rnd_in, fismat);


% RMSE calculation
% ----------------

N=125; % number of records
difference = normalize(predicted_test_output) - test_rnd_out;
rmse = power(sum(power(difference,2))/N, 1/2);
disp(rmse);
```

ANFIS generated 20 rules in order to obtain the values above in Table 3.1. Those rules can be expressed as follows:

Rule 1: [4 54 1 1 0 1 1 1 2 2 1 1 1 1 0 1 1 1 0 0] [1]
If Age Group=4 and Weight=54 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=1 and Fever=2 and Sedimentation=2 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=1 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 1

Rule 2: [4 54 1 1 0 1 1 1 2 2 1 1 1 1 0 0 1 1 0 0] [1]
If Age Group=4 and Weight=54 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=1 and Fever=2 and Sedimentation=2 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=1 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 1

Rule 3: [4 54 1 1 0 1 1 1 2 2 1 1 1 1 0 1 1 0 0 0] [1]
If Age Group=4 and Weight=54 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=1 and Fever=2 and Sedimentation=2 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=1 and Calcific Tissue=0 and Cavity=0 and Pneumonic Infiltration=0 then Output is 1

Rule 4: [4 54 1 1 0 1 1 1 0 2 1 1 1 1 0 1 1 1 0 0] [1]

If Age Group=4 and Weight=54 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=1 and Fever=0 and Sedimentation=2 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=1 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 1

Rule 5: [7 68 0 0 0 1 1 1 2 2 0 0 0 0 0 0 0 1 0 0] [0.75]

If Age Group=7 and Weight=68 and Exhaustion=0 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=1 and Fever=2 and Sedimentation=2 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.75

Rule 6: [7 70 1 0 0 1 1 1 2 2 0 1 1 1 0 0 0 1 0 0] [0.75]

If Age Group=7 and Weight=70 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=1 and Fever=2 and Sedimentation=2 and PPD=0 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.75

Rule 7: [6 49 1 1 0 1 0 0 2 1 1 1 1 1 0 1 0 1 0 0] [0.75]

If Age Group=6 and Weight=49 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=0 and Hemoptysis=0 and Fever=2 and Sedimentation=1 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and

Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.75

Rule 8: [4 63 1 0 0 0 1 0 2 1 1 1 1 1 0 1 0 1 0 0] [0.75]

If Age Group=4 and Weight=63 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=0 and Sweating at Nights=1 and Hemoptysis=0 and Fever=2 and Sedimentation=1 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.75

Rule 9: [3 67 1 0 0 1 1 0 0 1 1 1 1 1 0 1 0 1 0 0] [0.50]

If Age Group=3 and Weight=67 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=0 and Fever=0 and Sedimentation=1 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.50

Rule 10: [5 72 1 0 1 0 1 0 2 1 1 0 0 0 0 1 0 1 0 0] [0.50]

If Age Group=5 and Weight=72 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=0 and Loss in Weight=0 and Sweating at Nights=1 and Hemoptysis=0 and Fever=2 and Sedimentation=1 and PPD=1 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.50

Rule 11: [3 57 1 0 0 1 1 0 0 1 1 1 1 1 0 1 0 1 0 0] [0.50]

If Age Group=3 and Weight=57 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=0 and Fever=0 and Sedimentation=1 and PPD=1 and Erythrocyte=1 and Haematocrit=1 and Haemoglobin=1 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.50

Rule 12: [5 47 1 0 0 1 1 0 2 0 1 0 0 0 0 1 0 1 0 0] [0.50]

If Age Group=5 and Weight=47 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=0 and Fever=2 and Sedimentation=0 and PPD=1 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.50

Rule 13: [4 64 1 0 0 1 1 0 0 0 1 0 0 0 0 0 0 1 0 0] [0.25]

If Age Group=4 and Weight=64 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=0 and Fever=0 and Sedimentation=0 and PPD=1 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.25

Rule 14: [6 78 1 1 0 1 1 0 0 0 1 0 0 0 0 1 0 1 0 0] [0.25]

If Age Group=6 and Weight=78 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=1 and Sweating at Nights=1 and Hemoptysis=0 and Fever=0 and Sedimentation=0 and PPD=1 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=1 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.25

Rule 15: [7 78 1 0 1 0 0 0 2 0 0 0 0 0 0 0 0 1 0 0] [0.25]

If Age Group=7 and Weight=78 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=0 and Loss in Weight=0 and Sweating at Nights=0 and Hemoptysis=0 and Fever=2 and Sedimentation=0 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.25

Rule 16: [5 54 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0] [0.25]

If Age Group=5 and Weight=54 and Exhaustion=1 and Unwillingness for Work=0 and Loss of Appetite=0 and Loss in Weight=0 and Sweating at Nights=0 and Hemoptysis=0 and Fever=0 and Sedimentation=1 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=0 and Calcific Tissue=1 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0.25

Rule 17: [2 69 1 1 0 0 0 0 1 1 0 0 0 0 2 2 0 0 0 1] [0]

If Age Group=2 and Weight=69 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=0 and Sweating at Nights=0 and Hemoptysis=0 and Fever=1 and Sedimentation=1 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=2 and Number of Leucocyte Types=2 and Active Specific Lung Lesion=0 and Calcific Tissue=0 and Cavity=0 and Pneumonic Infiltration=1 then Output is 0

Rule 18: [1 71 1 1 0 0 0 0 1 2 0 0 0 0 2 2 0 0 0 0] [0]

If Age Group=1 and Weight=71 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=0 and Sweating at Nights=0 and Hemoptysis=0 and Fever=1 and Sedimentation=2 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=2 and Number of Leucocyte Types=2 and Active Specific Lung Lesion=0 and Calcific Tissue=0 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0

Rule 19: [2 68 1 1 1 0 0 0 1 2 0 0 0 0 2 2 0 0 0 1] [0]

If Age Group=2 and Weight=68 and Exhaustion=1 and Unwillingness for Work=1 and Loss of Appetite=1 and Loss in Weight=0 and Sweating at Nights=0 and Hemoptysis=0 and Fever=1 and Sedimentation=2 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=2 and Number of Leucocyte Types=2 and Active Specific Lung Lesion=0 and Calcific Tissue=0 and Cavity=0 and Pneumonic Infiltration=1 then Output is 0

Rule 20: [3 55 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0] [0]

If Age Group=3 and Weight=55 and Exhaustion=0 and Unwillingness for Work=0 and Loss of Appetite=0 and Loss in Weight=0 and Sweating at Nights=0 and Hemoptysis=0 and Fever=0 and Sedimentation=0 and PPD=0 and Erythrocyte=0 and Haematocrit=0 and Haemoglobin=0 and Leucocyte=0 and Number of Leucocyte Types=0 and Active Specific Lung Lesion=0 and Calcific Tissue=0 and Cavity=0 and Pneumonic Infiltration=0 then Output is 0

The rules above indicate different cases for the trained ANFIS model. Each rule is represented by a vector which consists input values for the system. As stated in the data preparation section, each input variable has a different ranking on affecting the output. For instance; Active specific lung lesion parameter has a ranking of 70%, calcific tissue existence parameter has a ranking of 55%, number of leucocyte types parameter has a ranking of 48% and weight of the patient parameter has a ranking value of 43%. Those four parameters are the strongest ones among other parameters. Especially active specific lung lesion parameter shows great importance. In the following surface diagram, two of the strongest input parameters which are active specific lung lesion and calcific tissue existence are plotted versus output.

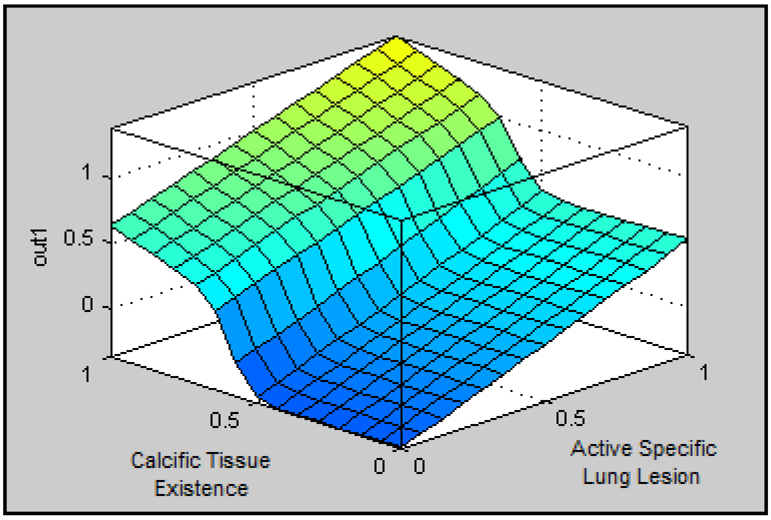

Figure 3.2: Surface plot of active specific lung lesion and calcific tissue existence parameters versus output

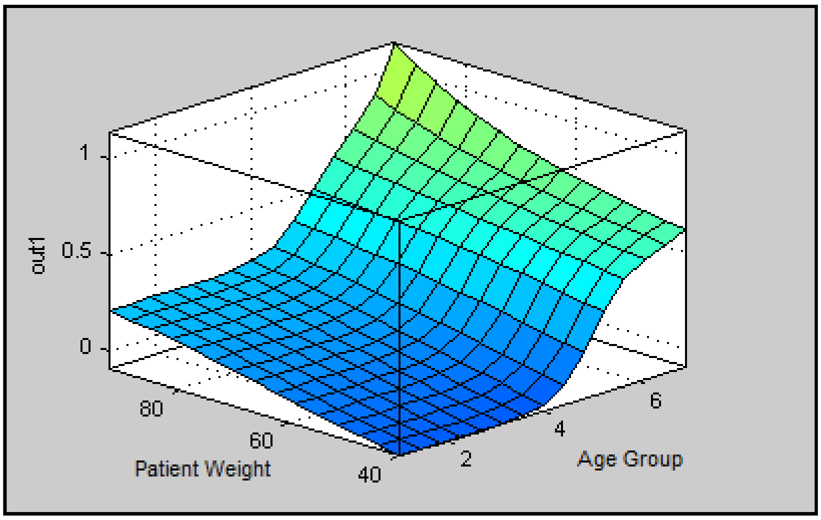

Figure 3.3: Surface plot of patient weight and age group parameters versus output

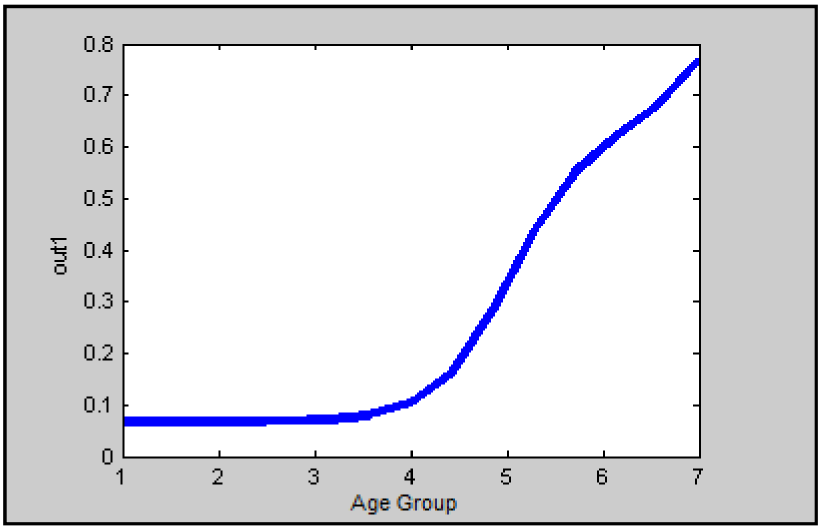

Figure 3.4: Plot of age group versus output

For the inputs for rule 1, system states that this patient belongs to cluster 1 which means the patient is suffering from tuberculosis disease. If we take a closer look to the parameters, we see that active specific lung lesion and existence of calcific tissue parameters are both positive. Calcific tissue existence is a proof that patient has had tuberculosis disease at least once before. Sedimentation value is high, subfebrile fever and PPD test result is also positive. Number of leucocyte types parameters shows that there is a lymphocytic density. And patient is suffering from hemoptysis. Exhaustion, unwillingness for work, sweating at nights and loss in weight parameters also support this output.

Rule 2 is similar to rule 1. The strongest parameters are positive. The supportive parameters such as exhaustion, unwillingness for work, sweating at nights and loss in weight is positive too. The only change in this case is number of leucocyte type's parameters. These parameters have a value within normal range. But the other parameters still affect the output as 1 which means the patient is suffering from tuberculosis disease.

In rule 3, the parameters are mostly similar to rule 1 and rule 2. The only difference is existence of calcific tissue is negative in this rule. This means that this patient is not suffered from tuberculosis disease before. But the rest of the parameters support that patient is suffering from tuberculosis now. So the output is stated as cluster 1 by system.

Rule 4 covers cases where patient has a fever value within normal ranges. This is a normal possibility in real life too. When we look into other parameters, we see that, active specific lung lesion and existence of calcific tissue parameters are both positive. PPD is positive and patient has a high level of sedimentation. Patient is also having hemoptysis. And the supportive parameters such as exhaustion, unwillingness for work, sweating at nights and loss in weight is positive. So system classifies this kind of patient to cluster 1 which means there is an active tuberculosis disease.

Input parameters for rule 5 are different from the rules before. This time, patient does not have active specific lung lesion. But he/she has calcific tissue, has subfebrile typed fever and high sedimentation. Patient also has loss in weight, positive sign of hemoptysis. According to those parameters, system classifies this input into cluster 0.75 which means that patient is having tuberculosis disease for a probability of 75%.

Rule 6 is similar to rule 5. This time, there is no exhaustion but erythrocyte, haematocrit and haemoglobin values are within low range. The rest of the parameters indicate that this patient is in cluster 0.75.

Rule 7 shows little changes in parameters. There is exhaustion, unwillingness for work, subfebrile typed fever, moderately high sedimentation, positive PPD, low erythrocyte, haematocrit and haemoglobin values. And there is also lymphocytic dense in number of leucocyte types. Existence of calcific tissue is positive but there is no sweating at nights, no sign of hemoptysis and no active specific lung lesion positivity. According to these values, system classifies this input as cluster 0.75.

In rule 8, parameter values are like rule 7. Major parameters are mostly same. There are only some changes such as unwillingness for work is negative, no loss in weight, no sweating at nights. In the light of these parameters such a condition is classified as 0.75.

Rule 9 covers a different case for cluster 0.50. In this case, existence of calcific tissue is positive whereas active specific lung lesion parameter is negative. In number of leucocyte types, there is lymphocytic density. Erythrocyte, haematocrit and haemoglobin values are within low range. Sedimentation is moderately high, PPD is positive. Fever value is in normal ranges and some minor parameters such as exhaustion and sweating at nights is positive. But, a strong parameter, Hemoptysis, is negative. So these parameters indicate that such a case is classified by 0.50.

When we look into rule 10 we see that erythrocyte, haematocrit and haemoglobin values are within normal bounds. PPD value is positive, lymphocytic density is seen in number of leucocyte types. Existence of calcific tissue is positive but there is no sign of active specific lung lesion. Leucocyte level is also normal. Sedimentation is moderately high and patient is having subfebrile typed fever. Hemoptysis is negative and minor supportive parameters such as exhaustion, sweating at nights is positive. All of these parameters indicate that these types of inputs are classified as cluster 0.50.

Rule 11 is similar to the pervious rule. The only major change is erythrocyte, haematocrit and haemoglobin values are in low range. And the patient is not having high level or subfebrile typed fever. With these input parameters, this rule classifies such a patient as cluster 0.50.

Rule 12 and rule 11 have very common parameters. The difference between these two rules is sedimentation value, loss of appetite and loss in weight. System states these two rules both as a member of cluster 0.50.

Rule 13 covers a case for 25% possibility of tuberculosis disease which is classified as cluster 0.25. The input parameters are mostly in normal ranges. But patient is having exhaustion, sweating at nights, loss of appetite and loss in weight. PPD value is also positive. And calcific tissue existence is positive too. So these parameters indicate a small chance which is around 25% that the patient can be suffering tuberculosis disease.

Rule 14 is a close replica of the previous rule. In this rule, the patient is having the same values as rule 13 except he/she is having unwillingness to work. And the patient’s age group is older than the previous one. This is also classified as cluster 0.25.

If we look into rule 15, we see that PPD parameter is negative but patient is having subfebrile typed fever. The rest of the parameters are mostly same with minor changes. System classifies such an input vector as 25% possibility of being mycobacterium tuberculosis positive.

Rule 16 shows us positive exhaustion, and subfebrile typed fever. Existence of calcific tissue is also positive like the previous rules. System classifies this kind of input vector as a member of 0.25 cluster's.

Rule 17 has an output of 0 which means that patient is not suffering from tuberculosis disease. Exhaustion and unwillingness for work is positive. Fever is in high values. Sedimentation is moderately high, leucocyte value is high and macrophage density is spotted in number of leucocyte types. Pneumonic infiltration is also positive. These parameters indicate that patient is not having tuberculosis disease. He/she is most probably having another disease such as pneumonia.

In rule 18, parameters are like rule 17. But pneumonic infiltration is not positive. This indicates that patient is not having tuberculosis disease but he/she is most probably suffering from acute bronchitis.

Rule 19 is similar to rule 17. In this case there are only some minor changes that sedimentation is very high. This patient is in cluster 0 and he/she is probably having another disease such as pneumonia.

If we look to the input parameters of the last rule, Rule 20, it is obvious that this case is not suffering from tuberculosis disease. All of the major parameters that indicate tuberculosis are having normal values.

In the following table, the confusion matrix of ANFIS test data result is seen.

Table 3.4: Confusion matrix of ANFIS test data

		Actual Classes				
		0	**0.25**	**0.50**	**0.75**	**1**
Predicted Classes	**0**	16.0	0.0	0.0	0.0	1.0
	0.25	0.0	20.0	9.0	0.0	0.0
	0.50	0.0	3.0	21.0	7.0	1.0
	0.75	0.0	1.0	7.0	16.0	5.0
	1	0.0	0.0	0.0	1.0	17.0

The following figure is the ROC plot of ANFIS test data. In this graph, true positive rate is plotted against false positive rate.

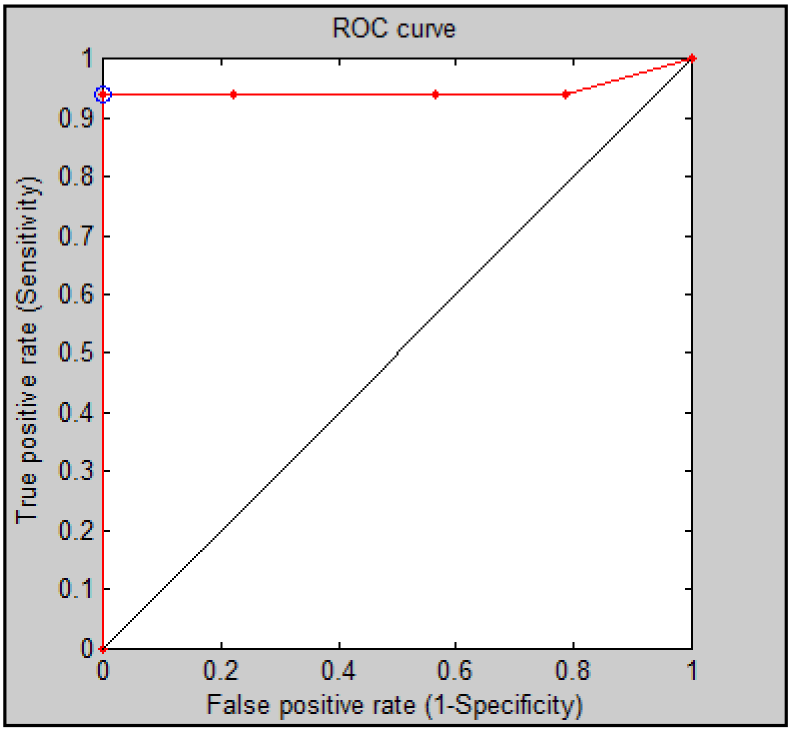

Figure 3.5: ROC plot of ANFIS test data

4. CONCLUSION AND FUTURE PLANS

Data mining techniques are used widely in biomedical area. There are many studies performed using different techniques. Each of these techniques has pros and cons. The main aim of this study is to apply ANFIS on a medical problem.

In this study, the medical problem that is chosen is predicting existence of mycobacterium tuberculosis bacteria on patients. A dataset of 503 records each having 30 attributes was used. After applying a ranking algorithm (InfoGainAttributeEval with Ranker) on the dataset, 10 attributes were removed. The removed attributes were the ones which were ranked less than 0.10. So, they were not having much importance on the dataset at all.

Each of the attributes on the dataset represent a value for the patient such as gender, loss of appetite, age group of patient, loss in weight, total weight of patient, smoke addiction level, chest pain level, etc. The model that was developed distinguishes the probability class of the patient using these attributes whether he/she is in one of the following classes:

Table 4.1: Predicted classes and output codes

Class	Output
1 (bacteria existence 0%)	0
2 (bacteria existence 25%)	0.25
3 (bacteria existence 50%)	0.50
4 (bacteria existence 75%)	0.75
5 (bacteria existence 100%)	1

ANFIS model classifies the patients with an RMSE of 17%. In order to compare the success of this result with other methods, 5 different data mining techniques are applied. We can name them as: Bayesian Network, Multilayer Perceptron, Part, Jrip and RSES. These methods had an RMSE of 22%, 23%, 22%, 25% and 37% respectively.

Benchmarking values indicate that the ANFIS model that was developed classifies the instances with a RMSE of 17% which is a very acceptable result. ANFIS's generated rules' integrity and consistency is checked by comparing each rule with real case inputs and outputs. If we compare the generated rules of ANFIS and RSES algorithms, we see that ANFIS generated more generalized rules for cases. The rules which are generated by RSES algorithm are much more specific and mostly focused on single input parameters. So this reduces accuracy of RSES. On the other hand, our findings indicate that Bayesian Network, Multilayer Perceptron and Part algorithms are having RMSE results within acceptable range. But still ANFIS has the best RMSE value.

According to the findings of this study, ANFIS is an accurate and reliable method comparing to Bayesian Network, Multilayer Perceptron, Part, Jrip and RSES methods for classification of tuberculosis patients.

REFERENCES

Bakar, A. A. & Febriyani, F. 2007, 'Rough Neural Network Model for Tuberculosis Patient Categorization', *Proceedings of the International Conference on Electrical Engineering and Informatics*, Indonesia.

Chiu, S. L. 1997, 'Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification', in D. Dubois, H. Prade, R. Yager (eds.), *Fuzzy Information Engineering: A Guided Tour of Applications*, Wiley.

Davidson, S. 1999, *Davidson's Principles and Practice of Medicine*, Churchill Livingstone.

Fawcett, T. 2004, 'ROC Graphs: Notes and Practical Considerations for Researchers', Technical report, HP Laboratories, Kluwer Academic Publishers, Palo Alto.

Gómez, C., Hornero, R., Abásolo, D., Fernández, A. & Escudero, J. 2009, 'Analysis of MEG Background Activity in Alzheimer's Disease Using Nonlinear Methods and ANFIS', *Annals of Biomedical Engineering*, vol 37, no. 3, pp. 586-594.

Gören, S., Karahoca, A., Onat, F.Y. & Gören, Z. 2008, 'Prediction of cyclosporine A blood levels: an application of the adaptive-network-based fuzzy inference system (ANFIS) in assisting drug therapy', *Springer-Verlag*, vol 64, pp. 807-814.

Harrison, T.R. 1999, *Harrison's Principles of Internal Medicine*, McGraw-Hill Education.

Jang, J-S. 1992, 'Self-learning fuzzy controllers based on temporal back propagation', *IEEE Trans Neural Networks*, vol 3, no. 5, pp. 714-723.

Jang, J-S. 1993, 'ANFIS: adaptive-network-based fuzzy inference system', *IEEE Trans. Syst. Man Cybernet*, vol 23, no. 3, pp. 665-685.

Jang, J-S. 1996, 'Input Selection for ANFIS Learning', *Proceedings of the IEEE International Conference on Fuzzy Systems*, New Orleans.

Kara, A. & Karahoca, A. 2009, 'Diagnosis of Diabetes by using Adaptive Neuro Fuzzy Inference Systems', *ICSCCW*, Famagusta.

Lingras, P. 1996, 'Rough neural networks', *Proceeding of the 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, Granada.

Lingras, P. 1998, 'Comparison of neofuzzy and rough neural networks', *Information Sciences*, vol 110, pp. 207-215.

Mamdani, E.H. & Assilian, S. 1975, 'An experiment in linguistic synthesis with a fuzzy logic controller', *International Journal of Man-Machine Studies*, vol 7, no. 1, pp. 1-13.

Monzon, J.E. & Pisarello, M.I. 2005, 'Cardiac Beat Classification using a Fuzzy Inference System', *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai.

Øhrn, A. 1999, 'Discernibility and Rough Sets in Medicine: Tools and Applications', PhD Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, ISBN 82-7984-014-1, Trondheim.

Özlü, T., Metintaş, M. & Ardıç, S. 2008, *Akciğer Hastalıkları Temel Bilgiler*, Poyraz Tıbbi Yayıncılık, Ankara.

Pawlak, Z. 1982, 'Rough classification', *International Journal of Information* , vol 11, pp. 145-172.

Sánchez, M.A., Uremovich, S. & Acrogliano, P. 2009, 'Mining Tuberculosis Data', in P. Berka, J. Rauch, D.A. Zighed (eds.), *Data Mining and Medical Knowledge Management: Cases and Applications*, Medical Information Science Reference, New York.

Shlomi, T., Cabili, M.N. & Ruppin, E. 2009, 'Predicting metabolic biomarkers of human inborn errors of metabolism', Department of Computer Science, Israel Institute of Technology, EMBO and Macmillan Publishers Limited, Haifa.

Spackman, K.A. 1989, 'Signal detection theory: Valuable tools for evaluating inductive learning', *Proceedings of the Sixth International Workshop on Machine Learning*, Morgan Kaufmann, San Mateo.

Sugeno, M. & Kang, G.T. 1988, 'Sturcture identification of fuzzy model', *Fuzzy Sets and Systems*, vol 28, no. 1, pp. 15-33.

Takagi, T. & Sugeno, M. 1985, 'Fuzzy identification of systems and its application to modeling and control', *IEEE Trans. On Systems, Man & Cybernetics*, vol 15, pp. 116-132.

Tsukamato, Y. 1979, 'An approach to fuzzy reasoning method', in M.M. Gupta, R.K. Ragade, R.R. Yager (eds.), *Advances in Fuzzy Set Theory and Applications*, Elsevier Science Ltd, Amsterdam.

Werbos, P. 1974, 'Beyond regression, new tools for prediction and analysis in the behavioural sciences', PhD Thesis, Harvard University.

Witten, I.H. & Frank, E. 2005, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, San Fransisco.

VITA

Tamer Uçar was born in İstanbul. He received his under graduate degree in Computer Engineering from Bahçeşehir University, Istanbul, in 2006. He has been working as a research assistant in Bahçeşehir University, Software Engineering Department. His areas of interests are data mining applications, database design and enterprise web programming.