Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 3 (2011) 1404–1411

Procedia Computer Science

www.elsevier.com/locate/procedia

WCIT 2010

Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches

Tamer Uçar [a], Adem Karahoca[a] *

[a] *Bahçeşehir University, Software Eng. Dept., Istanbul 34353, Turkey*

Abstract

A Correct diagnosis of tuberculosis (TB) can be only stated by applying a medical test to patient's phlegm. The result of this test is obtained after a time period of about 45 days. The purpose of this study is to develop a data mining(DM) 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.

In this research, we proposed the use of Sugeno-type "adaptive-network-based fuzzy inference system" (ANFIS) to predict the existence of mycobacterium tuberculosis. 667 different patient records which are obtained from a clinic are used in the entire process of this research. Each of the patient records consist of 30 separate input parameters. ANFIS model is generated by using 500 of those records. We also implemented a multilayer perceptron and PART model using the same data set.

The ANFIS model classifies the instances with an RMSE of 18% whereas Multilayer Perceptron does the same classification with an RMSE of % 19 and PART algorithm with an RMSE of % 20.

ANFIS is an accurate and reliable method when compared with Multilayer Perceptron and PART algorithms for classification of tuberculosis patients. This study has contribution on forecasting patients before the medical tests.

Selection and/or peer-review under responsibility of the Guest Editor.

Keywords: Tuberculosis, ANFIS, Multilayer Perceptron, PART, Data Mining

1. Introduction

Tuberculosis (TB) was believed to be almost under control; but it has once again become a serious world-wide problem. Tuberculosis disease is caused by a bacterium which is called as mycobacterium tuberculosis [20]. 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 [1].

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) and this 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

* Tamer Uçar. Tel.: +90-212-381-0575; fax: +90-212-381-0550 .
E-mail address: tamer.ucar@bahcesehir.edu.tr .

1877-0509
doi:10.1016/j.procs.2011.01.022

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 [2].

World Health Organization's Direct Observation of Therapy is the internationally accepted standard for control of tuberculosis (http://www.who.int/tb/en/). 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 [2].

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.

2. Background

Nowadays data mining techniques are extensively used in medical sciences. We may consider two different dimensions in the coverage of this study. First one is TB disease dimension and second is data mining dimension. Following sub sections will mention about recent researches on these two dimensions. In this study, we were focused on predicting the existence of mycobacterium tuberculosis on patients by using Adaptive Neuro Fuzzy Inference System (ANFIS), Multilayer Perceptron and PART Algorithm. From this view point, in order to evaluate medical data sets, firstly, we introduced the relationship between TB and DM.

2.1. Tuberculosis and Data Mining

Bakar and Febriyani (2007) applied Rough Neural Networks for classifying the tuberculosis types. 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 instances) of the data set is used for training and 30% (56 instances) 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 [3].

Sanchez and et al.(2009) implemented data mining technique to classify TB related handicaps to determine patients' sickness. This study was 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 results, age category, pulmonary tuberculosis, and extra pulmonary tuberculosis. Exhaustive CHAID is selected for generating decision trees for classes [4].

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 and each record has 4 variables. 300 of those records are used for training and 90 are used for checking the model[5].

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 [6].

Another data mining study is made on Alzheimer disease under the topic as Analysis of MEG. Background activity in Alzheimer's disease is detected by using nonlinear methods and ANFIS. This study intends to analyze magneto encephalogram background action on patients using sample entropy and Lempel-Ziv complexity [7].

Shlomi et al.(2009) studied for predicting metabolic biomarkers of human inborn errors of metabolism. The motivation of this study is to get 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 [8].

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 [9].

3. Classification in Data Mining

Classification, which can be described as analyzing of data in order to obtain models that are used to characterize data classes, is the most usual task in data mining. This task focuses on predicting the value of the decision class for an input object within set of classes which are predefined. There are many different classification approaches in the literature. Each of them are developed and proposed by various researchers. The most known techniques are decision tree based classification, neural network based classification, statistical classification, rough set based and genetic algorithm classifiers. We can divide data classification task into two phases. The first phase is called as the learning step, and the second phase is called as testing step. In the learning step, a model which defines predetermined set of classes will be constructed. This operation is made by analyzing a set of training data. For this data, each set of elements are assumed to belong a specific, predefined class. In the testing step, which is the second leg of classification, the constructed model is tested by using different set of data. In this phase, the accuracy of the classification is estimated by using one of the several proposed approaches. If the estimation of accuracy shows an adequate result, then the generated model can be used for classification of new input sets whose class labels are unknown. Before applying the generated classification model, some data preprocessing techniques can be executed in order to obtain a better accuracy and efficiency for the classification model [10].

Data Cleaning: The removal of noisy data and filling of missing data is considered in this step. There are many different approaches designed by researches in the literature for data cleaning issue [11].

Feature Selection: In the initial data set, there may be some attributes which are not related or not important for the learning step of model. The removal of inappropriate and unnecessary attributes from the data set is applied on this step of classification. After this process, the reduced data set is used to generate the classification model. For feature selection task, numerous applications are implemented by various researchers [12].

Data Discretization: The data set which will be used by classification algorithms may have some attributes that cannot be handled by the algorithm itself without applying some transformations. Such as, numerical scaled values are needed to be converted into nominal or discrete values in order to make some of the algorithms work correctly. This conversion step is considered in the data discretization part of classification. More detailed discussion about discretization techniques can be found in [10].

3.1. Preparing Tuberculosis Data Set

Data set that we used in this study contains information about 667 patients who were examined at a private clinic. Each of those records consists of 30 different variables. The full list of those variables is based on World Health Organization’s standard of Direct Observation of Therapy:

Table 1. Full list of variables

Variable Name	Variable Name	Variable Name
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

We categorized all of these parameters into 3 groups: clinical findings, medical laboratory findings and radiological findings. If we take a closer look about the meaning of these parameters we can briefly explain them as follows: In clinical findings, the gender parameter indicates whether the patient is male or female. Age group parameter indicates the age group that patient belongs to. All ages are grouped into 7 classes. These classes are "18-24", "25-32", "33-40", "41-45", "46-51", "52-57" and "58+". Weight parameter indicates the weight of the patient in kilograms. Smoke addiction parameter defines whether the patient is a smoker or not. Rates are grouped into 4 classes. "0" means the patient is not a smoker. "1" means the patient smokes less than 5 cigars per day. "2" means the patient smokes between 6 to 10 cigars per day. And "3" means the patient smokes more than 11 cigars per day. Alcohol addiction parameter indicates if the patient is addicted to any kind of alcohol or not. BCG vaccine parameter shows the whether the patient has BCG vaccine or not. Malaise, arthralgia, exhaustion, unwillingness for work, loss of appetite, loss in weight, sweating at nights, chest pain, back pain and coughing are binary valued parameters. They indicate if these parameters are positive for the patient or not. Hemoptysis means coughing up blood from the respiratory tract. This parameter identifies whether the patient has hemoptysis or not. Fever is classified into 3 categories: "0" means normal fever value which is nearly 36.5 degrees Celsius, "1" means fever value is in high ranges and "2" means subfebrile fever value which does not exceed 38.5 degrees Celsius.

In medical laboratory findings we categorized some of blood and skin tests' parameters. PPD parameter identifies whether the patient has the result of the PPD test positive (labeled as "1") or negative (labeled as "0"). Erythrocyte is the red blood cells. They are responsible for delivering oxygen to the body tissue. We grouped this parameter into 3 categories. "0" means erythrocyte level is in normal range (about 4.5 to 5 million per microliter). "1" means low (which is less than 4.5 million per microliter) and "2" means patient has high erythrocyte level (more than 5 million per microliter). Haematocrit is the ratio of the volume occupied by packed red blood cells to the volume of the whole blood. We also grouped this parameter into 3 categories. "0" means the patient has normal haematocrit percentage (about 40% to 45%). "1" means low (less than 40%) and "2" means patient has high haematocrit percentage (more than 45%). Haemoglobin is the iron-containing oxygen-transport metalloproteinase in the red blood cells. In our parameter values, "0" means the patient has normal haemoglobin values (about 14 to 16 g/dl). "1" means low (less than 14 g/dl) and "2" means the haemoglobin value is considered as high (more than 16 g/dl). Leucocytes are white blood cells. They are responsible for defending the body against infections, diseases etc. In our parameter values, "0" means the patient has normal leucocyte values (about 5000 to 10000 in $1mm^3$ blood). "1" means leucocyte values are low (less than 5000 in $1mm^3$ blood) and "2" means the leucocyte value is considered as high (10000 in $1mm^3$ blood). Number of leucocyte type parameter shows the density of leucocytes. If there is a normal density (labeled as "0") or a lymphocytic density (labeled as "1") or macrophage density (labeled as "2"). Sedimentation parameter is a measure of the settling of red blood cells in a tube of blood during one hour. It is grouped into 3 categories. "0" means normal sedimentation level (0 to 15 mm/hr), "1" means moderately high sedimentation level (16 to 40 mm/hr) and "2" means very high sedimentation level (more than 41 mm/hr).

In radiological findings, active specific lung lesion parameter indicates whether there is a radiological proof of a tuberculosis lung lesion on the patient or not. Calcific tissue shows that whether the patient has had tuberculosis disease before. If this parameter is positive, it indicates that the patient has had tuberculosis disease at least once. Cavity parameter states if there are opening-like lesions on the patient's lung or not. Positive value means those

kinds of lesions exist on lung. Pneumonic infiltration parameter is positive if a pneumonia-like lesion is seen on the chest x-ray of patient. Pleural effusion means the accumulation of excessive pleural fluid in pleura. This parameter is positive if such thing is seen on the chest x-ray of patient.

Before generating ANFIS model, attribute ranking function is applied using information gain ranking filter in WEKA [13] platform. By applying this function, we chose the most important parameters that will affect the fuzzy model mostly. The variables which are ranked less than 10% were eliminated. According to this reduction on the data set, BCG vaccine, arthralgia, chest pain, smoking addiction, gender, malaise, coughing, back pain, alcohol addiction and pleural effusion variables were ignored.

4. Methods

In this research, ANFIS, Multilayer Perceptron and PART methods are used for classification. The following subsections contain brief descriptions about these methods.

4.1. 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 [19]. 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[14] [15]. 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 [16][17]. The following figure shows the structure of a basic ANFIS model:

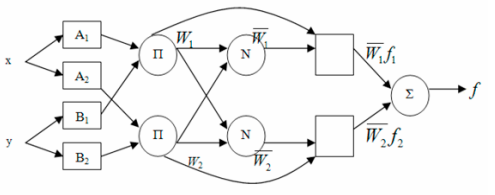

Fig. 1. ANFIS Architecture

4.2. 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 [13].

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 [18].

4.3. 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 partial tree generation contains iteration for every subset. On each step, the selected subset is expanded. This process is repeated until there is no subset left unexpanded [13].

5. Findings

In this section, the results of this research will be stated. As stated in the previous sections, different models on the data were applied which are ANFIS, Multilayer Perceptron, and PART (a Partial Decision Trees algorithm implementation). Each model generated close results to each other but in an overall point of view, ANFIS has the best scores when compared to other methods. The following table summarizes the test data benchmarking results for these methods:

Table 2. Testing result of models

Method Name	Sensitivity	Specificity	Precision	Correctness
ANFIS	0.95	0.97	0.89	0.97
Multilayer Perceptron	0.89	0.97	0.90	0.89
PART	0.85	0.96	0.87	0.85

If we consider the overall scores, ANFIS generated best results for testing data as it is clearly shown on Table 2. It is also clear that results of both Multilayer Perceptron and PART methods are very close to ANFIS. When comparing sensitivity and correctness, ANFIS has obviously better scores. The following figure shows the classification plot of ANFIS testing data.

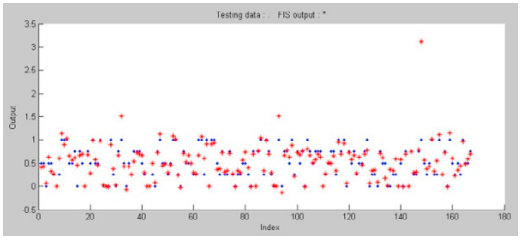

Fig 2. ANFIS classification of testing data

ANFIS, Multilayer Perceptron and PART algorithms generated RMSE values of 0.18, 0.19 and 0.20 respectively. Since a lower RMSE value shows more reliability in testing of data, it can be stated that ANFIS scored best RMSE result among other methods.

6. Conclusion

As we mentioned in Section 1, 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. From this view point, we focused on three different data mining methods. In Section 3, we mentioned that a data set of 667 patients each having 30 input parameters is used in this study. A reduction on data set is performed according to the result of attribute ranking function which is applied using information gain ranking filter. This operation reduced the data set into 20 input variables. This final data set is used when generating data mining models.

According to our experiment results, ANFIS is an accurate and reliable method when compared with Multilayer Perceptron and PART algorithms for classification of tuberculosis patients.

References

[1] Davidson S. Davidson's Principles and Practice of Medicine. Churchill Livingstone; 1999.

[2] Harrison TR. Harrison's Principles of Internal Medicine. McGraw-Hill Education; 1999.

[3] Bakar AA, Febriyani F. Rough Neural Network Model for Tuberculosis Patient Categorization. In: Proceedings of the International Conference on Electrical Engineering and Informatics; 2007; Indonesia. p. 765-768.

[4] Sánchez MA, Uremovich S, Acrogliano P. Mining Tuberculosis Data. In: Berka P, Rauch J, Zighed DA, editors. Data Mining and Medical Knowledge Management: Cases and Applications. New York: Medical Information Science Reference; 2009.

[5] Kara A, Karahoca A. Diagnosis of Diabetes by using Adaptive Neuro Fuzzy Inference Systems. In: ICSCCW; 2009; Famagusta.

[6] Monzon JE, Pisarello MI. Cardiac Beat Classification using a Fuzzy Inference System. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference; 2005; Shanghai. p. 5582-5584.

[7] Gómez C, Hornero R, Abásolo D, Fernández A, Escudero J. Analysis of MEG Background Activity in Alzheimer's Disease Using Nonlinear Methods and ANFIS. Annals of Biomedical Engineering. 2009;37(3):586-594.

[8] Shlomi T, Cabili MN, Ruppin E. Predicting metabolic biomarkers of human inborn errors of metabolism. Haifa: EMBO and Macmillan Publishers Limited; 2009 Israel Institute of Technology.

[9] Gören S, Karahoca A, Onat FY, Gören Z. Prediction of cyclosporine A blood levels: an application of the adaptive-network-based fuzzy inference system (ANFIS) in assisting drug therapy. Springer-Verlag. 2008;64:807-814.

[10] Han J, Kamber M. Data Mining: Concepts and Techniques. 1st ed. Morgan Kaufmann Publishers; 2000.

[11] Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering. 2000;23(4).

[12] Zhou J. Feature Selection in Data Mining - Approaches Based on Information Theory. VDM Verlag; 2007.

[13] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. San Fransisco: Morgan Kaufmann Publishers; 2005.

[14] Jang JS. Self-learning fuzzy controllers based on temporal back propagation. IEEE Trans Neural Networks. 1992;3(5):714-723.

[15] Jang JS. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybernet. 1993;23(3):665-685.

[16] Sugeno M, Kang GT. Sturcture identification of fuzzy model. Fuzzy Sets and Systems. 1988;28(1):15-33.

[17] Takagi T, Sugeno M. Fuzzy identification of systems and its application to modeling and control. IEEE Trans. On Systems, Man & Cybernetics. 1985;15:116-132.

[18] Jitgarun, K., & Tongsakul, A. (2009). Virtual-based training and critical thinking in higher-level education. Cypriot Journal Of Educational Sciences, 4(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/cjes/article/view/68

[19] Dagli, G., & Uzunboylu, H. (2009). Competence of School Principals Regarding Knowledge Management in Elementary Schools. Cypriot Journal Of Educational Sciences, 2(2). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/cjes/article/view/16

[20] Tuncay, N., & Uzunboylu, H. (2010). Trend of Distance Education in the last three Decades. World Journal On Educational Technology, 2(1). Retrieved November 15, 2010, from http://www.world-education-center.org/index.php/wjet/article/view/183