Tuberculosis Analysis by Naive Bayesian Classification

El Mi Mi San, Nwe Nwe
Computer University, Hpan-An
el.misan1@gmail.com

Abstract

Tuberculosis (TB) is a social disease with medical aspects. By the increasing availability of bio-medical and health-care data with a wide range of characteristics, computer-based medical system is playing an increasingly relevant role in assisting both diagnosis and treatment. Base on the knowledge stored, the system will learn the patterns using Naive Bayesian classification and decides the category of TB by probabilities. It is based on the theorem of posterior probability. This system intends to develop a diagnosis system of automatic classification method for TB diagnosis based on the symptoms of the patients. The system stores the knowledge of the medical experts and the medical records of the previous case as Training database. This system also considers the missing value by filling data completely, because it needs the actual symptoms of the patient for increasing accuracy to classify. This system can give the category of TB and treatment for the patient who has TB symptoms by using Naive Bayesian classification method on the Training database. The accuracy of the system for that patient is shown by using hold-out method on testing database.

Keywords:

Machine learning system, Naive Bayesian classification, Computer-Based medical diagnosis system

1. Introduction

Artificial Intelligence (AI) is the study of intelligent machines capable of performing complex tasks that require thought and behavior normally associated with human intelligence. It stores large quantities of information and uses set rules to manipulate and access data for the purpose of providing analysis and solutions to correct the problem it detects. As such, expert systems act like doctors who cure patients by using logic and information. The goal of AI is to make computers more useful for humans[7].

Machine Learning is the study of computtationally methods improving performance by mechantising the acquisition of knowledge from experience. Machine Learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with atomic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data. The ultimate test of machine learning [5] is its ability to produce systems that are used regularly in industry, education, medicine and else-where.

Data mining is defined as the process of seeking interesting or valuable information within large data sets. Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. According to the rules of classification method in data mining, this system gives the category of TB and accuracy for user. This system uses Naive Bayesian method in classification because it is the most accurate than other classifier [2]. In classification, the derived model is based on the analysis of a set of training data (i.e., data objects whose class label is unknown). The derived model may be represented in various forms. Classification can be used for predicting the class label of data objects. This system intends to classify the disease by using Naive Bayesian classification method with the highest accuracy for using reality in nature [10]. And then, this system proves that the classification method is reality useful in nature.

2. System Architecture

This system uses the two hundreds of training data, that is two third of the total data and it is obtained from medical experts and knowledge. These training data records are used by Naive Bayesian Classification to get the rules and to result the category of disease. There are fifteen features of TB and user need to choose the attribute of each features. According to the user selecting attributes, there are five categories of TB disease. The system can also show the treatment for that category of the result. The testing database is also the TB patient records who are testing with the system.

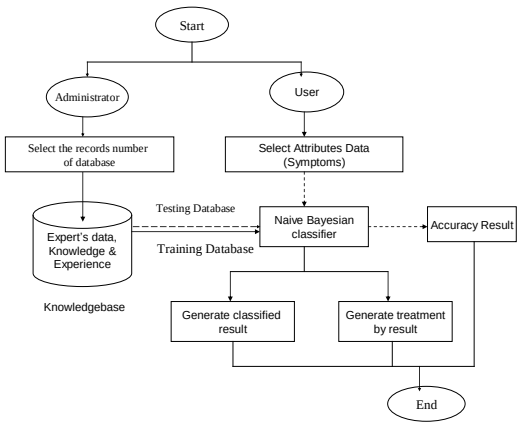

Figure 1 Design of system architecture

Accuracy can be calculated upon the one hundred of testing data**.** This method is called the hold out method that the given data are randomly partitioned into two independent sets, a training set and a test set [2]. The training set is used to derive the classifier, and whose accuracy is estimated by test set. The administrator can adjust the records numbers of two databases for system performance and system analysis. The currently user data can be saved in history database. The design of system architecture is shown in Figure 1.

User interface is one of the main component of the expert system. It provides the interactive communication with the system user. It has the user authentication to prevent the unauthorized usage of the system as described in [3][7]. The interface structure of the system is shown in Figure 2.

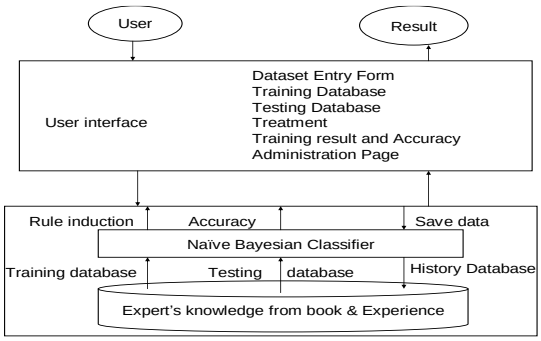

Figure 2 Interface Structure of the System

3. Tuberculosis in the System

The likely outcome of the disease (prognosis) and treatment of TB is influenced by the stage of the disease (what stage of the TB) when first diagnosed. Tuberculosis (TB) is broadly categorized in five stages:

Category 1 (Cat 1): - New smear (+) pulmonary TB.
- Severe smear (-) pulmonary TB.

Category 2 (Cat 2): - Relapses (smear +).
- Treatment after failure Cat 1 or 3.

Category 3 (Cat 3): - New smear (-) pulmonary TB or other than Cat 1.

Category 4 (Cat 4): - Chronic cases fail after Cat 2.
- Multi-Drug Resistant TB.

Category 0 (Cat 0): - No TB.

The result for the system user is one of the above categories. This system will classify TB disease by the following symptoms as features: Family contact {Present, Absent}, Weight loss {Yes, No}, Low grade fever {Yes, No}, Night sweating {Yes, No}, Loss of appetite {Yes, No}, Breathlessness {Yes, No}, Cough for 3 weeks {Yes, No}, Haemoptysis {Yes, No}, Chest Pain {Yes, No}, Spectum smear {Positive, Negative}, Chest x-ray {Positive, Negative}, Taking previous treatment {Present, Absent}, Treatment after failure {Present, Absent}, Treatment of Cat 2 failure {Present, Absent}, HIV {Positive, Negative} that are shown in dataset entry form.

The treatment category of regimens for Tuberculosis disease [9] is classified four types according to the category of the disease. The kinds of drugs are Rifampicin(R), Isoniazid(H), Pyrazinamide (Z), Streptomycin(S), Ethambutol(E). The TB patient must take 6 months of drug. For Cat 1, Isoniazid, Rifampicin, Pyrazinamide, Ethambutol must be taken 2 months and Isoniazid, Rifampicin must be taken 4 months. For Cat 2, Streptomycin, Isoniazid, Pyrazinamide, Ethambutol must be taken 2 months, Isoniazid, Rifampicin, Pyrazinamide, Ethambutol must be taken 1 months and Isoniazid, Rifampicin, Ethambutol are taken 5 months. For Cat 3, Isoniazid, Rifampicin, Pyrazinamide must be taken 2 months and Isoniazid, Rifampicin must be taken 4 months. For Cat 4, the patient for Cat 4 needs to be taken to the special hospital of TB, doctors and TB center. This system produces treatment for the user according to the category of TB.

4. Theory Background

4.1 Naive Bayesian Classification

The naive Bayesian classifier, or simple classifier, that is used in the system to classify the disease works as follows. According to the rule, each data sample is presented by an n-dimensional feature vector, $X=(x_1,x_2,x_3,...,x_n)$, that is fifteen features in this system as shown in dataset. Each feature has n attributes $A_1,A_2,\ldots,A_n$, respectively. That attributes are the user inputs of each features. Suppose that there are m classes, $C_1,C_2,\ldots,C_m$,that is the five categories of TB in the system. Given an unknown data sample X, it is user's input, the classifier will predict the class of X. That is the naive Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(C_i\backslash X) > P(C_j\backslash X) \quad \text{for } 1 \le j \le m, j \ne 1 \tag{4.1}$$

Thus $P(C_i\backslash X)$ was maximized. The class C_i for $P(C_i\backslash X)$ that is maximized, is called the maximum posterior hypothesis. That hypothesis is the data sample X belongs to the specific class C_i. By Bayes theorem,

$$P(C_i\backslash X) = \frac{P(X|C_i)\,P(C_i)}{P(X)} \tag{4.2}$$

Given the data sets with many attributes, it would be extremely computationally expensive to compute $P(X\backslash C_i)$. In order to reduce computation in evaluating $P(X\backslash C_i)$, the naive assumption of class conditional independence is made. Giving the class label of the sample, that is, there are no dependence relationships among the attributes. In order to classify an unknown sample X, $P(X\backslash C_i)P(C_i)$ is evaluated for each class C_i. Sample X is then assigned to the class C_i if and only if

$$P(X\backslash C_i)P(C_i) > P(X\backslash C_j)P(C_j) \quad \text{for } 1 \le j \le m\,, j \ne i. \tag{4.3}$$

In other words, it is assigned to the class C_i for which $P(X\backslash C_i)$ is the maximum. Bayesian

classifiers have the minimum error rate in comparison to all other classifiers [2]. Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes Theorem [2][1].

4.2 Missing values

There are many tuples for several attributes. Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning routines attempt to fill value in missing data, smooth out noise and correct inconsistencies in the data [6].To get high accuracy of the result, it need to be consider for the missing value among in six types of missing value. There are many methods for missing value. They are " Ignoring the tuple", "Filling in the missing value manually", "Using the attribute mean to fill in the missing value", "Using the attributes mean for all samples belonging to the same class as given tuple" and "Using the most probable value to fill in the missing value"[1][4]. This system uses the filling data in the missing value manually because the tuples for the attributes of the system is not much for time consuming and for bioinformatics. The system will alert the message box for the missing value and incorrect data to fill necessary. When the user need to fill missing value in dataset entry form to get high accuracy for the system, the message box for the missing value will give the user until the system is completely filled.

4.3 Classifier Accuracy

Estimating classifier accuracy is important since it determines to evaluate how accurately a given classifier will label future data, data on which the classifier has not been trained. The following classi-fication features are used to train and test the classifier.
Given : a collection of labeled records (training set) with features (attributes), and the true class (label).
Find : a model for the class as a function of the values of the features.
Goal : previously unseen records should be assigned a class as accurately as possible.

A test set is used to determine the accuracy of the model. The given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

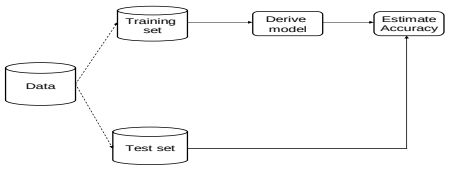

Figure 3 Estimating accuracy with the hold out method

The *Sensitivity* and *Specificity* measures can be used to determine the accuracy measures. *Precision* may also used to access the percentage of samples labeled. These measures are defined as

$$\text{Sensitivity} = \frac{t\_pos}{pos} \quad (4.4)$$

$$\text{Specificity} = \frac{t\_neg}{neg} \quad (4.5)$$

$$\text{Precision} = \frac{t\_pos}{(t\_pos + f\_pos)} \quad (4.6)$$

Where,
t_pos = the number of category of result that is the attributes of that category are same with the attributes of training database of the category of that result
pos = the number of category of result in training database
t _neg = the number of category of result that is the attributes of these category are not same with the attributes of training database of that result
neg = the number of all categories of disease that do not include the category of result
f_pos = the number of category in database that is the same attributes with the result but not the category of result

$$\text{accuracy} = \text{sensitivity}\frac{pos}{(pos+neg)} + \text{specificity}\frac{neg}{(pos+neg)} \quad (4.7)$$

In this system, the accuracy will be shown by the message box for the user.

4.4 Experimental Result

Diagnosing the patient can be done through the patient diagnosis panel, where entry of user's symptoms can be filled to get the result. Patient's medical records are stored in the history records database. There are the symptoms with attributes of disease that are use in the system. They are: family contact {Present, Absent}, weight loss {Yes, No}, low grade fever {Yes, No}, night sweating {Yes, No}, loss of appetite {Yes, No}, breathlessness {Yes, No}, cough for 3 week {Yes, No}, haemoptysis {Yes, No}, Chest pain {Yes, No}, spectum smears { Positive, Negative}, chest X-rays{ Positive, Negative}, taking previous treatment{Present, Absent}, treatment after failure{Present, Absent}, cat 2 failure {Present, Absent}, HIV {Positive, Negative}. Figure 4 shows the result for the user according the filling data items using Naive Bayesian classification method. For example, the user input is X = ("Present", "Yes", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Positive", "Positive", "Absent", "Absent", "Absent", "Positive") as the fifteen attributes. Using the classification rule, the system will result "Cat 1" as the result for the user. This can be shown in Figure 4.

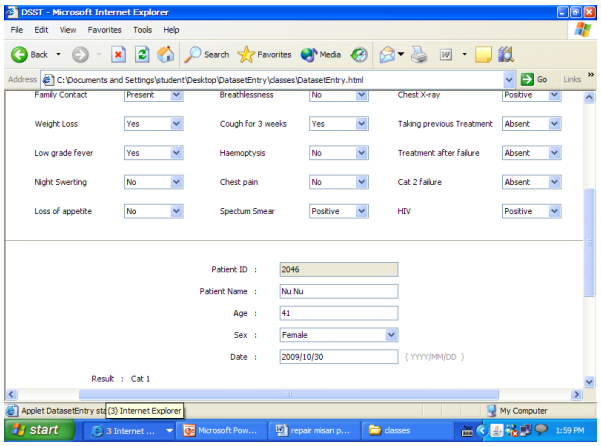

Figure 4 Classification result for the user

This system also gives the treatment according the category of the result. The treatment is also shown with message box. According to the example, this system will result the treatment as shown in Figure 5.

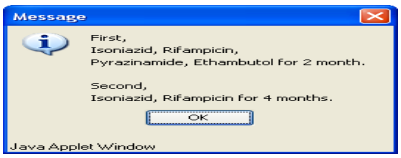

Figure 5 Treatment according result type

In this system, accuracy is 98.599994122 % as example input that is higher than other classifiers, is obtained by using hold out method.

4.5 Analysis of the System

For the analysis of the system, it can be used to get higher accuracy by analyzing the number of testing data according to the hold-out method. It can be used two hundreds of training data records and one hundred of testing data records by analyzing by administrator in administration page that can be shown as table in Figure 6 and Figure 7 for runtime performance.

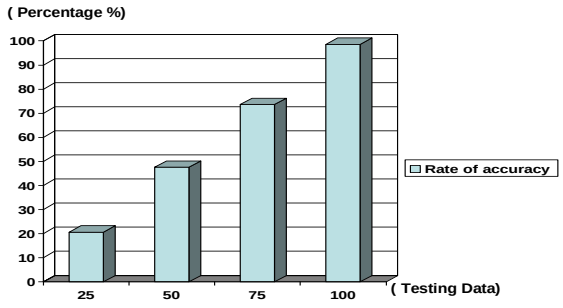

Figure 6 Comparison of accuracy upon testing data

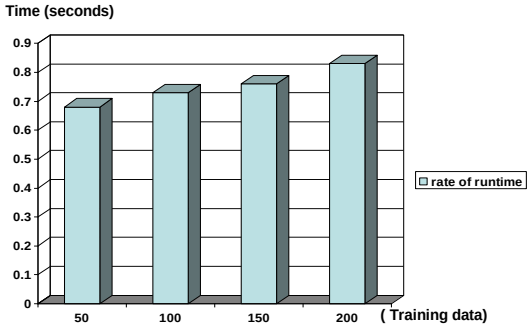

Figure 7 Comparison of runtime upon training data

5. Conclusion

Computer-based diagnostic decision support systems will play an increasingly important role in health care. They may improve the quality of the diagnostic process in accuracy and efficiency while cost and burden of patients may be reduced. The system helps a person to improve their diagnostic skill to solve problems concerned with Tuberculosis when experience medical one is not available. Moreover, an authorized person can construct the new rules easily by just clicking the button without the need of the knowledge engineer [8]. Naive Bayesian methods are computionally intractable, as they can provide standard of optimal decision making against which other methods can be measured.

This system helps users in classifying Tuberculosis (TB) diagnosis based on the symptoms of the patients. People can easily test themselves at home or on-line by entering data in TB diagnosis. If a man who is coughing and doubts on TB, they can get treatment early-before it spreads other people and they can lead their long and healthy lives by testing these system. In the end, without doubt, this system significantly enhance the insights into the various facets, both patient specific and patient unspecific of the process, and may result in new diagnostic, therapeutic and prognostic methods by entering data in TB diagnosis. It can also prove that the methods used in the system are reality useful in nature for people and this system gives the highest accuracy.

Bibliographical Reference:

[1] R. Duda & Hart, *Pattern classification and Scence analysis*, New York: John Wiley & Sons, 1973.

[2] Jiawei Han, Micheline Kamber, *Data Mining Concepts an Techniques*, Academic Press.

[3] W . S . Jeffrey , *Data Mining and Homeland Security* ,2006.

[4] D . A . Keim , *Knowledge Discovery and Data Mining* , Newport Beach, USA, 1997.

[5] P. Langley, *Elements of Machine Learning* , San Francisco: Morgan Kaufmann, 1996.

[6] T. S. Lim, W-Y.Loh, and .-S. Shih. *A comparison of predictive accuracy, complexity, and training time of thirty-three old and new classification algorithms, MachineLearning* ,39,2000.

[7] K. I. Modesitt, *The Four W 's of Expert Knowledge Based Systems : Why, What, When, And Why-not* , Proc. Nat'l Computer Graphics Conf., NCGA, Fairfax,1987

[8] R .Z. O' Osmar, *Principles of Knowledge Discovery in Databases* , Introduction to Data Mining, 1999.

[9] S.Phyu, R.Jureen, T.Ti, U.R.Dahle , HM.Grewal , *Heterogeneity of Mycobacterium tuberculosis Isolates in Yangon, Myanmar*, J.Clin Microbiol, 2003.

[10] Shortliffe, H.Edward , *Computer-BasedMedical Consultations : MYCIN* , American Elsevier , New York, 1976.