Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis

Hardik Maniya
BVM Engineering College
Vallabh Vidhyanagar
Gujarat, India

Mosin I. Hasan
BVM Engineering College
Vallabh Vidhyanagar
Gujarat, India

Komal P. Patel
BVM Engineering College
Vallabh Vidhyanagar
Gujarat, India

ABSTRACT

Data mining is applied in medical field since long back to predict disease like diseases of the heart, lungs and various tumors based on the past data collected from the patient. In India, though the data collection of medical patient is not streamlined, we made an effort to predict the most widely spread disease in India named tuberculosis. Using data collected from various TB centers, we made an effort to fetch out hidden patterns and by learning this pattern through the collected data for tuberculosis we can diagnose and predict the disease. In the research work we are comparing naïve bayes classifier and KNN, two the most effective techniques for data classification (especially for medical diagnoses), implemented using C language and using Weka tool respectively and classify the patient affected by tuberculosis into two categories (least probable and most probable). We have used 19 symptoms of tuberculosis and collect 154 cases. We have achieved nearly 78% accuracy with low false negative.

General Terms

Pattern recognition, Medical machine learning

Keywords

Data mining, naïve bayes, KNN, pattern recognition, tuberculosis, Machine learning.

1. INTRODUCTION

Tuberculosis (TB) is one of the most infectious disease that causes thousands of deaths per year. Hence we have use a method of data mining to predict this disease. Thus we can decrease the laboratory test cost as well as time also.

1.1 Tuberculosis

Tuberculosis is a disease caused by bacteria called Mycobacterium tuberculosis. It usually attacks the lungs, but can attack any part of the body such as kidney, spine and brain. If not treated properly, this can be fatal as the bacteria attack the body and destroy tissue and create a hole in that affected part. This bacterium can spread through the air from one person to another.

1.2 Statistics of tuberculosis

As per the records of World Health Organization, globally 9.4 million new cases (including 3.3 million women) of TB is recorded out of which, 1.7 million people died from TB(including 380000 women) in the year 2009 i.e. 4700 deaths per day[1]. India has highest prevalence of TB in the World, accounting for 1/5th of the world's increased in 2011 from $50 in 2010 to $62.

Adding sputum smear was estimated to be more cost-effective (incremental cost per disability-adjusted life year [DALY] of $20 per test ($198[South Africa], $131 [Brazil], $38 (Kenya]) than a new TB diagnostic with 70% sensitivity, 95% specificity and price of $20 per test ($198 [South Africa], $275 [Brazil], $84 [Kenya]). However, compared to sputum smear, smear plus new test averted 46-49% more DALYs per 1000 TB suspects (321 vs. 215 [South Africa], 243 vs. 166 [Brazil], 790 vs. 531 [Kenya]), at an incremental cost of $170 (Kenya) to $625 (Brazil) per DALY averted [2]. Cost-effectiveness was most sensitive to the specificity and price of the new test, the baseline TB case detection rate and the discount rate.

1.3 Medical practitioner's approach

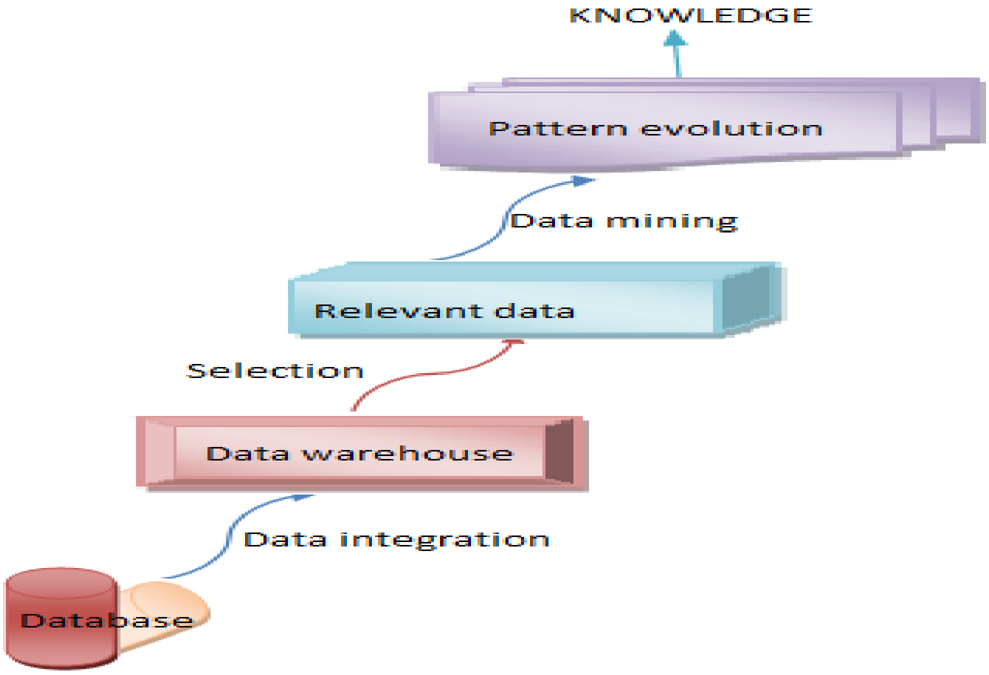
Disease diagnosis is often done on the basis of doctor's experience and personal opinion rather than the data pattern hidden in the database. This approach leads to errors and increase the medical costs which affects the quality of services provided to the patients [3]. According to medical practitioners, the diagnosis of TB is only possible through pathological tests (sputum test, x-ray and skin test). These laboratory tests are unnecessary for cases least suspected for TB. There are data mining methods that can be applied to biomedical field to extract knowledge from existing data set and using these extracted pattern we can diagnose the disease very well[3].

2. RESEARCH OBJECTIVE

The healthcare industry is facing the major challenge is to provide quality services at affordable costs. The quality service implies to diagnose patients correctly and provide them treatment effectively [3]. In diagnosis poor clinical decisions can lead to disastrous result which should not happened. In records of hospitals there is a large amount of unanalyzed data that can be turned into useful information. Firstly, medical diagnosis is depends on physician (his experience, intuition and bases, his psycho-physiological conditions). Secondly, the most of data that should be analyzed to make a good prediction is usually huge and unmanageable [4]. In this context, machine learning can be used to automatically infer diagnostic rules from descriptions of successfully treated patients and can help specialists to make the diagnostic process more objective and reliable. The main objective of this research is to develop a novel technique to categorize TB into two categories Yes & No through a data mining technique called Naïve Bayesian classification. This technique can discover and extract hidden patterns associated with TB from the past patient data.

3. EXTRACTION TECHNIQUES

According to Fayyad ‘data mining is extraction of interesting (non-trival, implicit, previously unknown and potentially unknown) information or patterns from data in large database’ [5]. The approach of data mining is to use past data to help identify similar instances.



Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large database. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (E.g. k-means clustering is unsupervised) [3].

3.1 NAÏVE BAYES CLASSIFIER

NB's main strength is its efficiency; it combines efficiency with good accuracy it is often used as a baseline in text classification research.

3.1.1 Classifier Module

Naive Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the variable values necessary for classification. Classification problem is given to the classifier as combination of different values of chosen variables with related value of class variable; the classifier then returns a posterior probability distribution over the class variable.[3]

Probability that a training pattern with attribute array A belongs to class C_K, also known as the posterior probability is given Bayes theorem of probability,

$$P\ (C_k|A) = P\ (C_k)*P\ (A|C_k)/P\ (A)$$

Where,

A is an array of M >= 1 attributes A_1 ,A_2 ,….,A_M for the patterns of a training set.

P (C_K) is the probability that a training pattern belongs to class A, also called prior probability.

P (C_K|A) is also called the posterior probability because it is probability of training pattern with attribute array *B* belongs to class C_k.

P (A|C_K) is the conditional probability of *B* given *A*. It is also called the likelihood shows probability of class A has attribute array B.

P (A) is the probability that a training pattern has attribute array B, regardless of the class to which the pattern belongs.

3.1.2 Learning Module

A set of 50 cases was taken and the program was trained with these data sets such that the probabilities of all the classes with all the conditions were calculated. Result was stored in database and when the test data was given we got the probabilities for the various classes for the given symptom values on the basis of which we inferred that the patient fell into the class with the highest probability. This is what is called the Naïve Bayes’ classification. This is a very powerful technique that is instrumental in helping us predict the category a patient falls into.

3.2 NEAREST NEIGHBOUR CLASSIFIER

In pattern recognition, the *k*-nearest neighbor algorithm (*K*NN) is a method for classifying objects based on closest training examples in the feature space.

3.2.1 Introduction of KNN

K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. It is called lazy because it does not have any training phase or minimal training phase. All the training data is needed during the testing phase and it uses all the training data so if we have large number of data set then we need special method to work on part of data which is heuristic approach. Although classification remains the primary application of KNN, we can use it to do density estimation also. The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. K-nearest-neighbor classification was developed from the need to perform discriminate analysis when reliable parametric estimates of probability densities are unknown or difficult to determine.[9]

When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution this is pretty useful because real word data does not obey the typical theoretical assumptions made. Since KNN is non parametric, it can do estimation for arbitrary distributions. [7][10]

One of their striking results is to obtain a fairly tight error bound to the Nearest Neighbor rule.

$$P' \leq P \leq P^{'}(2 - \frac{c}{1-c}P')$$

The bound is where the Bayes error rate, c is the number of classes and P is the error rate of Nearest Neighbor [9]. The result is indeed very striking because it says that if the number of points is fairly large then the error rate of Nearest Neighbor is less than twice the Bayes error rate [8].

3.2.2 Execution Model for KNN

This algorithm is very similar to our daily life like poor localities have poor people and rich localities have the elite class. There are very few instances that rich people live in a poor locality. These are called outliers. In our case, the classes of the diseases are the localities and the characteristics of people are the various symptoms. So, on the basis of the symptoms and the distance of the particular symptom from the various classes, we classify the patient as having a certain disease or not. The samples are trained first and then the testing data is worked upon according to the Hamming distance calculated for various symptoms. [5][6]

As the data collected by us is non numeric, we have used weighted Hamming distance algorithm to recognize the pattern and classify the diseases. The Hamming distance between two patterns is defined to be the number of attributes for which the two patterns have different values. The smaller the Hamming distance between a recall pattern and a training pattern, the nearer the two patterns are. [5][6]

Steps for the classification of recall pattern:

Calculate the Hamming distance between the recall pattern and each of the training patterns.

Assign the recall pattern to the class of the training pattern (Neighbor) nearest to it.

In k-Nearest Neighbor Rule if we consider k = 5 and there are 3 instances of C1 and 2 instances of C2. In this case, KNN says that new point has to label as C1 as it forms the majority. We follow a similar argument when there are multiple classes. If we assume that the points are d dimensional, then the straight forward implementation of finding k Nearest Neighbor takes O(dn) time. In our case d is 2 and hence 0(n). For large N this is not acceptable and hence there are some efficient data structures like KD-Tree which can reduce the time complexity but they do it at the cost of increased training Time and Complexity [6].

Clearly if *K* becomes very large, then the classifications will become all the same {simply classify each x as the most numerous class. We can argue therefore that there is some sense in making *K* > 1, but certainly little sense in making *K* = *P* (*P* is the number of training points). This suggests that there is some optimal intermediate setting of K[6]. We can expect a nearest neighbor classifier to perform well when there are a lot of training patterns. However, the more the training patterns, there are, the more distances have to be calculated, and consequently the computation required increases, thus slowing down the process of classification [6].

After calculating the Hamming distance for all the training patterns, we need to consider the minimum distance among all the training patterns. For these minimum distances, the corresponding classes need to be found out and then amongst these the class with maximum occurrence is recognized as the final class and the pattern is said to be recognized.

4. RESEARCH WORK

The various cases of TB is collected and that data was fed into database. Then, the program was trained with these data sets. Once it was ready, the test data was fed (i.e. the symptoms were entered) and the result was obtained.

4.1 Data Source

There have been many cases of TB in India especially this part of Gujarat and so it was easy to collect required number of TB cases for training and testing. We consulted a number of medical practitioners from various hospitals and were able to categorize the symptoms and ranges according to which we have generated a few cases and authenticated the same with the practitioners. We took expert help of medical practitioner to choose best symptoms for classification. A total of 154 records with 19 medical attributes (factors) were collected and validated from Sardar Gopaldas TB Center, Anand and the attributes are listed. The records were split equally into two datasets: training dataset and testing dataset. To avoid bias, the records for each set were selected randomly. For the sake of consistency, only categorical attributes were used. All the non-categorical medical attributes were transformed to categorical data. The attribute “Medical test” was identified as the predictable attribute. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

Following are the TB symptoms:

Age: Numerical values
Gender: male-0, female-1
Fever: high-0, low-1, not-2
Chills: yes-0, no-1
Sweating: extreme-0, little-1, no-2
Night sweats: yes-0, no-1
No appetite: yes-0, no-1
Weakness: yes-0, no-1
Cough: persistent-0, little-1, no-2
Chest pain: yes-0, no-1
Sputum: yes-0, no-1
Shortness of breath: yes-0, no-1
Duration: Numeric values (in days)
Haemoptysis: yes-0, no-1
Dyspnoea: yes-0, no-1
Palpitation: yes-0, no-1
Change in voice: yes-0, no-1
Weight: Numeric values, not available-0
Occupation: labor-0, job-1, farming-2, driving-3, other-4
Given and the program learns and gets trained with the data given to it.

4.2 Implementation

The program has been implemented in the C language which accesses data from the database. These are the steps that the program follows:

The symptom set is fed into the program for detection like fever, chills, cough etc.

The probability of the various categories is calculated and stored into the database. The Bayesian probability can be calculated as follows-

$$P\ (Ck) = (|Ck|+1)/(m+ \Sigma|Cj|)$$

Where m is the number of classes and j=1 to m which will be 2 in our case as we have 2 cases (A, B). And P (Ck) is the probability of a class Ck. If Ai is the set of all the attributes (in our case, symptoms i.e. fever, chills, breathlessness, cough etc) then the probability of a category Ck can be calculated as follows-

$$P(Ck)=P(A1|Ck)*P\ (A2|Ck)*P\ (A3|Ck)*\ldots\ldots.*P\ (AM|Ck)$$

In our case, M will be 19 as we have 19 symptoms and A1, A2, A3….will be the various symptoms. This formula will give the probability of the test symptoms to fall in one category. Similarly we'll have to calculate for the other category.

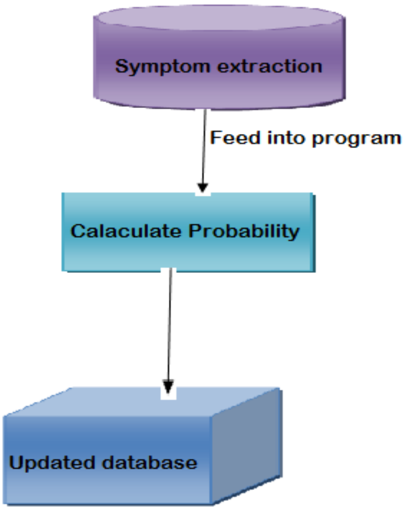
The result is inferred from the probabilities that is, the category with the highest probability is taken as the current status of the patient with the given symptoms. Suppose the probabilities of the various classes are as follows-

P (Category A) =0.12

P (Category B) =0.88

Then it is clear that the patient with the symptoms specified falls into the category B.

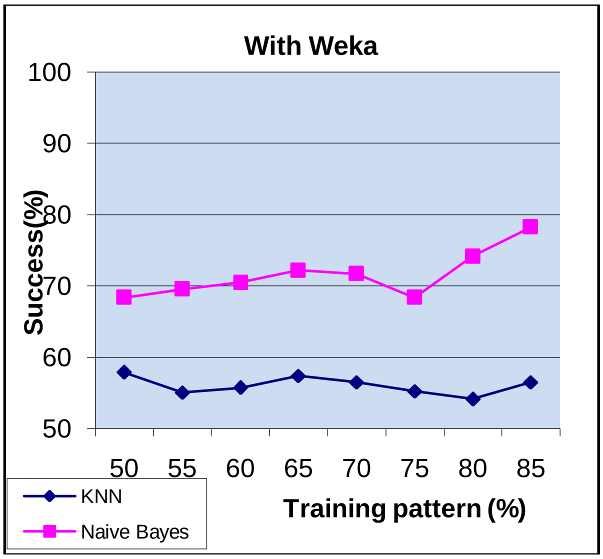
If the patient falls under Category A, he is absolutely normal Category B, it is likely that the patient is suffering from tuberculosis and tests should be carried out immediately to confirm the supposition.

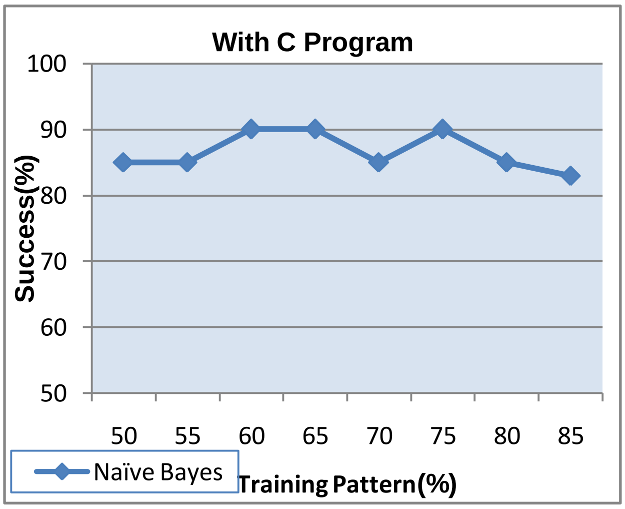


This is the flow of execution of the test module. The program retrieves data from the database which already has the probabilities stored in it. And a request from the program calculates the probabilities as per the above mentioned procedure.

4.3 Statistics

After training the program with the data sets, the test symptoms were given to the program and the execution of the same gave the following results-





5. CONCLUSION

This algorithm extracts hidden patterns from available TB database. Naïve Bayes could identify all the significant medical predictors. The prototype can further be improved by incorporating various other attributes and increasing the number of cases for training and testing. The efficiency of results using KNN can be further improved by increasing the number of data sets and for Naïve Bayesian classifier by increasing attributes or by selecting weighted features.

6. ACKNOWLEDGEMENT

We would like to thank medical practitioners at Sardar Gopaldas TB Center, Anand for providing us with valuable thoughts and useful data for our research work. Our hearty thanks also to our worthy institute BVM Engineering College and all those that we have inadvertently forgotten for their immense support and encouragement.

7. REFERENCES

[1] www.tbevidence.org/documents/dxres/models/tb_diagnostics.pdf World Health Organization

[2] Intelligent Heart Disease Prediction System Using Data Mining Techniques by Sellappan Palaniappan and Rafiah Awang IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

[3] Machine Learning for Medical Diagnosis: History, State of the Art and Perspective by Igor Kononenko

[4] An Introduction to Data Mining by Prof. S. Sudarshan CSE Dept, IIT Bombay

[5] SCHOLARPEDIA, available at http://www.scholarpedia.org/article/K-nearest_neighbor

[6] Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier Artcom international conference, 978-1-4244-8093-7

[7] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

[8] Pattern Classification (2nd. Edition) by R. O. Duda, P. E. Hart and D. Stork, Wiley 2002

[9] Inductive and Bayesian Learning in Medical Diagnosis Igor Kononenko University of Ljubljana