International Journal of Advanced and Innovative Research (2278-7844) /

Volume 7 Issue 2

A Mining Approach for Detection and Classification Techniques of Tuberculosis Diseases

N. Suresh[#1], D. Arulanandam[#2]

1#Research Scholar, Dept. of Computer Applications, Govt.Thirumagal Mills College, Gudiyattam, India.

2#Research Supervisor, Dept. of Computer Applications, Govt.Thirumagal Mills College, Gudiyattam, India.

Abstract - A Correct diagnosis of Tuberculosis (TB) can be only stated by applying a medical test to patient's. The result of this test is obtained after a time period of about few days. The need 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. In this research, we proposed the classification techniques to predict the existence of tuberculosis. we propose efficient Decision Tree algorithm technique approach for Tuberculosis prediction. Today medical field have come a long way to treat patients with various kind of diseases. Among the most threatening one is the Tuberculosis which cannot be observed with a naked eye and comes instantly when its limitations are reached. Bad clinical decisions would cause death of a patient which cannot be afforded by any hospital. The Decision Tree algorithm technique classifies the most usual types are: (i) Latent Tuberculosis,(ii) Active Tuberculosis. is an accurate and reliable method of tuberculosis patients. This study has contribution on forecasting patients before the medical tests.

Keywords-Tuberculosis, Decision Tree, Latent, Data Mining, WEKA

1. INTRODUCTION

1.1 DATA MINING

Data Mining is the abstraction of useful data from concealed knowledge of valuable information from large databases. It is the relevant method of searching legitimate, novel, potentially beneficial and in the end understandable templates in records. It uses a variance of technique to identify chunks of information for determination in the available dataset and deriving these in such way that they can be used in various areas. The term is contradicted because the target is to extract the knowledge from enormous amounts of data. It can be applied to any form of information processing system as well as any decision support system.

The real mining venture is the partially automatic or computerized evaluation of huge quantities of information to extract unknown, exciting styles consisting of agencies of records, unusual statistics and dependencies. This generally includes using database strategies. The output patterns are a kind of summary of the input data, and may be used in future analysis. The statistics mining step may become aware of multiple groups within the records that could be used to attain greater correct prediction results with the assistance of a support machine.

The related terms "Data degrading, Data fishing and Data snooping" refers to the use of data mining statistics to a large populated data. These methods can be used in creating new axioms to test against the larger data. This testing finds a model which explains the data by creating a hypothesis and then tests the hypothesis against the data. It is usually verified by searching the data sample. As the dataset has grown in size and complexity manual process becomes complicated and ineffective. So the implementation of automated process aided by the mining techniques is put forth.

1.2 TUBERCULOSIS

Tuberculosis is a tremendous infectious disease caused by "Myco-bacterium tuberculosis". Humans who have active TB usually spreads the ailment via the air even while coughing, spiting, talking and sneezing.

Facts of Tuberculosis

- The 'World Health Organization' reports nine million people get affected with TB in one year.
- Women of age fifteen to forty four gets affected by TB and it includes in the top three causes of death.

- TB symptoms may be dull for several months, and the infected people spreads TB up to ten to fifteen to other.
- It is an 'Airborne Pathogen'- spread through air.

Tuberculosis conditions

If the bacteria enter the body the following three things may happen.

- Body kills bacteria and no harm.
- The bacteria remain silent inside and are called 'Latent TB'.
- The bacteria make the body ill and is called 'Active TB'.

Tuberculosis in other parts

In bones -- leads to joint failure and spinal pain.
In brain -- leads to meningitis.
In kidney and liver -- leads to bleeding in the urine.
In heart -- leads to cardiac arrest.

Types

The most usual types are: (i) Latent Tuberculosis,
(ii) Active Tuberculosis.

Latent TB

The germ a sleep inside in an idle state. It has no symptoms and not transmissible. But still they can form as active. To control, it should be identified and treat properly which is generally carried out for several months.

Risk Factors

The risk factors include:

- HIV infection,
- Recent contact with an infectious people,
- Under treatment of 'Antitumor necrosis factor (TNF)',
- Undergoing dialysis,
- Organ Transplantation,
- Silicosis,
- Immigrant from highly affected TB burden countries,
- Illicit drug user.

Active TB

It can be spread to others. If the body resistance is minimum the bacteria lead to cause active tuberculosis. This germ begins to increase in number and cause active tuberculosis. It destroys the tissue. If the lungs get affected then it actually creates a gap in the lung. Based on the immunity power the people are affected soon or later. Normally young children have weakly immune systems so they can easily get affected.

Conditions for weakly immune system

- Substance abuse,
- Diabetes mellitus,
- Silicosis,
- Cancer,
- Leukaemia ,
- Kidney disease,
- Less body weight,
- Medical treatments ,
- Specialized treatment for rheumatoid arthritis or Crohn's disease.

Symptoms of TB

The symptoms include:

- Coughing with blood,
- Chills,
- Fatigue,
- Fever,
- Loss of weight,
- Loss of appetite,
- Night sweats.

Treatment

The disease may be cured with right medication and administration. The antibiotic treatment have factors of age, health, drugs, type of TB and the place of infection. Patient with latent TB may need one kind of antibiotics, whereas patient with active will require a prescription of numerous drugs for a long time. The course is about 6 months.

Prevention

A following are used to prevent the disease,
(a) TB vaccination ("BCG"),
(b) Finish the medication completely while the patient in latent TB.

Differences between the types

Latent TB	Active TB
Bacteria are a sleep in the body	TB bacteria are awake and makes ill.
Have no symptoms and well	Have symptoms that make feel unwell.
Cannot spread TB to others	Can pass TB to others if it is in lungs.
Detected by a skin test or blood test.	Detected by a chest x-ray if the TB is in the lungs.
Treated with a small amount of medicines over three to six months.	Treated with high amount medicines over at least six months.

Table 1.1 Difference between latent and active TB

2. METHODOLOGY

Classification

Classification aims to assign a label to a collection of points to aide more accurate predictions and analysis. It

comes under machine learning that uses known value to fix how the new object should be put into existing categories. It can be also used for large sets in a effective manner. The analysis will be start with a training set that contains a attributes and its likely outcome.

For Example, Weather forecasting might make use of this technique to order the day will be sunny, cloudy or rainy. The medical will use this to analyze fitness conditions to predict. Another classification technique classifies the E-Mail into junk and non-junk.

Working Principle

Classification is carried out as follows,

Step: 1

Model Construction: It has a pre - determined classes. Each observation is will belong to that pre-defined class and is determined by the label. The observations used for constructing the model are known as training set. The model is characterised by the rules.

Step: 2

Model usage: This step is used to analyse unknown objects. This model measures the correctness. And the set is known as test-set. This is classified on the basis of training set and it is independent of the case.

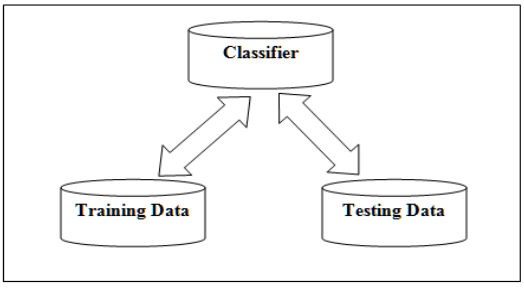


Figure 2.1(i) Classification Techniques

Figure 2.1 shows the classification technique.

Procedure of Classification

- Generate Sample data,
- Create Training set from data,
- Convert it to vectors,
- Train the vectors,
- Test the vectors.
- Assign the category.

Techniques used for Classification

Decision Tree

The tree is a systematic, 'tree-shaped diagram' used to analyse a action or to visualize a statistics. Each branch of the tree denotes a possible action, co-occurrence or decision. The aim is to show how the choice lead to the further level and each option is mutually exclusive. It may range from simple to a complex. All decision trees begin with a particular action. This is illustrated using a small square.

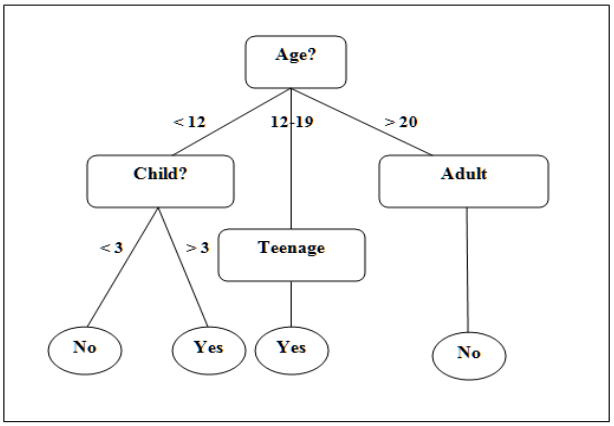


Figure 2.2(ii) Decision tree

Lines are drawn outward from the box, representing each available option. At the end the results is analysed. If the result finals in a new decision, a box is sketched at the stop of the line, and new lines are drawn from there. If the result is not clear, a circle is sketched at the end of the line denoting a potential risk. If the results are the solution to the decision, the line is left blank. The line moves from left to right, and it can extend to get a proper conclusion.

A tree is used to get an answer to a complicated problem. The design allows the users to have more solutions for a problem and visualize it by a simple format that displays the relation between distinct events. The farthest nodes on the tree represent the final results and it assigns a risk and weight. The conclusion is arrived with the highest total value.

Advantages

- **Easeful to Understand:** It can be understood by users with non-analytical background. And it does not need knowledge to interpret them. The users can simply with the graphical representation.
- **Useful in finding significant variable:** If the problem has enormous variables it is used to get most significant variable. Also the user can make other variables and feature to get target value.
- **Less cleaning required:** It is not affected by outliers and missed values leading to a less cleaning.
- **No constraint in data type:** It can handle both numerical and categorical values.
- **Non Parametric Method:** It has no assumptions on the distribution and the structure.

Disadvantages

- **Over fitting:** Pruning is carried to solve the problem.

Not fit for continuous variables: It loses information when it orders the variables in different categories.

3. IMPLEMENTATION

Proposed Work

The work classifies the tuberculosis dataset for high risk and low risk and clusters the patients on the category of tuberculosis.

Data set

The tuberculosis dataset consists of 1250 details collected from a medical practitioner in city hospital.

Flow of Work

The work is evaluated in three phases.

Phase I – Pre-Processing,

Phase II – Classification,

Phase-I → Pre-Processing

Pre processing involves changing ambiguous info into a clear format that can be made understandable. The file available in real may be unfinished, inconsistent or lacking some important certain characteristics and contain many errors. This technique has many methods to clear up such issues. It transforms the it for further processing.

Major steps involved in data pre processing are:

(a) Cleaning,
(b) Integration,
(c) Reduction and
(d) Transformation.

(a) Cleaning

The dataset may be noisy, incomplete and inconsistent. Data cleaning procedures make a attempt to fill the lost values, clear the noisy data while finding outliers and rectify the discrepancies.

Methods used in Cleaning

(i) Missing Values

The file can have more missing entries and it make way to misclassification.

1. **Ignoring the tuple:** This is taken when the label is missing. While ignoring the tuple, the remaining attributes is also not used. But this not effective, if the tuple does not contains several attributes with lost values. It is considered bad when the percentage of unseen values per attribute varies.

2. **Filling the value:** The values are entered manually but this is time consuming.

3. **Replace the value:** Replace all values by a unique constant. This method is easy but it is perfect.
4. **Use a mean or median to fill value:** which pick the "middle" value of a series to fill the value. The symmetric distributions require the mean value, while skewed distribution requires median to fill or replace.
5. **Use the most possible value to fill the value:** Using a "Bayesian method or decision tree" induction this may be handled.

(ii) Noisy Data

Noise is said as a "random error or variance" in a measured variable.

The followings are used to recover from noise.

Binning: These methods ease sorted facts based on the values around it. Then the values are allotted into a number of "bins". It is a variation that each value is exchanged with the mean.

Regression: Data smoothing can also done by regression. It assigns the values to a function. Linear regression finds the "best" two attributes so that one can be used to foresee the other. Multiple linear is a variation in which multi attributes are selected and are distributed in a multi-dimensional surface.

Outlier analysis: It is analysed by clustering. The unique values are formed as a unit in which the most unfit one forms a outlier.

(b) Integration

It combines the information from multiple sources. There exist redundancies and inconsistencies after merging. If it is avoided it can help to improve the efficiency and speed of the process. The well known implementation is building a warehouse. This enables to perform analyses in the warehouse.

Integration Techniques

- Manual Integration – User operate with the relevant information.
- Application Integration- A particular application do all the process.
- Middleware Integration–The integration process is transferred to the middleware.
- Virtual Integration –Data resides in the system and have a set of Unified view for accessing.
- Physical Integration – Creates a advanced system and have a Xerox from the base system.

(c) Reduction

It is a reduced form that is much smaller from the normal size but maintains the efficacy of the original. Mining

on the reduced one should reflect the same analytical results as the originality. It increases storage efficiency and reduce costs.

Reduction Strategies

It includes,

- Dimensionality reduction,
- Numerosity reduction,
- Data compression.

Dimensionality reduction

It lessen the number of variables or attributes on some consideration.

And it includes,

- Wavelet transforms – Transforms the initial set onto a smaller space.
- Principal Components Analytics – Selects the important attributes.
- Attribute subset formation - redundant attributes are removed.

Numerosity reduction

This will replace the initial volume with the other forms of data representation.

It may be,

- Parametric - Data parameters are stored instead of actual data.
  Ex: "Regression & Log-linear models".
- Non-parametric - Stores simplified forms.
  Ex: "Histograms, Clustering, Sampling & Data cube aggregation".

Compression

The transformations are put to get a "reduced or compressed" form of the prior.

It may be,

Lossless – Primary info can be recreated without loss.

Lossy – Cannot recreate the primary one, only an approximation will get.

There are numerous lossless algorithms for compression, but they allow only limited data manipulation. "Dimensionality reduction and numerosity reduction" is also considered one form of compression.

(d) Transformation and Discretization

The record is changed into a form which is apt for the mining process.

Transformation Strategies

(i) Smoothing – It removes noise.
Techniques include 'binning, regression, and clustering'.

(ii) Attribute construction - New attributes are created and added.

(iii) Aggregation - Constructs a data cube for analysis.

(iv) Normalization - Attributes are scaled with a mini range.

(v) Discretization - Normal values are redirected by intervals.

(vi) Hierarchy generation – Attributes are modified from higher to lower level.

Transformation by Normalization

It assigns an equal numbers to all the attributes. It is particularly used for classification which includes 'neural networks, nearest-neighbour classification and clustering'.

Phase –II Classification

It aims to assign data item from a collection to the intended categories. The goal is to correctly assume the specified class for each case. The task starts with a set in which the labels are known. They are discrete, real, floating-point values are denoted by numerical rather than categorical value. A numerical predictive model uses a regression algorithm. The simplest form is binary classifier. In this, the attribute has only two possible out forms either high or low. There are also multiclass targets which have three as low, medium and high.

This procedure finds relations between the observations and the target.

Process

The set is as "building the model and testing the model". The built model is used to build the model and the balance is for testing.

Steps in the Process

Building the Model

This is referred as the "learning step or phase". It is created from the training set from the database tuples with their labels. Each is referred to as a "category or class".

- **Using the Classifier**
  The test data is used to measure the efficiency and property in place of the training set.

Training and Test set

It is an important part in the models to differentiate the set into training and testing. While separating most of the series is allotted for training, and only a smaller portion is for testing. After a model has been created using a training set, the balance set is tested by making predictions. If similar thing is used for training and test set the discrepancy is minimized.

The algorithms includes,

- Linear classifiers,
- Fisher's linear discriminator,
- Logistic regression,

- Naïve Bayes,
- Perception,
- Support vector machine,
- Least squares vector,
- Quadratic classifiers,
- Kernel estimation,
- k-nearest neighbour,
- Boosting,
- Decision trees,
- Random forests,
- Neural networks,
- FMM Neural Networks ,
- Vector quantization.

Method Used

Random Forest

"Random forests or Random decision forests" is an ensemble learning procedure used for classification, regression in models which is operated by creating a multitude of decision trees at the sequence of training and output the class which is the mode of the classes for classifying and mean prediction for regression.

It was first created by "Tin Kam Ho" by using the random subspace method where his formulation is used. "Stochastic Discrimination" is also implemented with this proposed by "Eugene Kleinberg". An extension was developed by "Leo Breiman & Adele Cutler", and they named as "Random Forests" as the brand name. This extension combines prior "bagging" method with "Random selection of features".

In this classifier, the higher the number of trees gives the high certainty. Averages the multiple decision trees, trained on completely variant components to mini the variance.

Features

- It is applicable for both 'classification and the regression' problem.
- It can handle the missing values automatically and handle categorical values.
- It won't over fit the model even the forest extends larger.

Terminologies used

Bagging

Each tutoring set picks a sample of instances with replacement and is referred as "Bootstrap Sample" from data set. By "Sampling with Replacement", the instances can be replaced in each set.

So, N models are created using the bootstrap samples and finally it is combined by taking average of the votes.

For Training - $2/3^{rd}$ of the instances (63.2%) is used.

For Testing - $1/3^{rd}$ of the instances (36.8%) is used.

Trees are extended only for the training.

Out-of-Bag Error

It is similar to the validation data or test data. There is no separation for validating the result.

It is estimated during the process as follows,

The testing set is not used in building that tree and it is mentioned as "Out of bag error estimation".

Bootstrap Sample

This is chosen by "random with replacement sampling" method. The sample which is selected for training is again put again into the set. So there is a chance to pick up again and referred as "Random with Replacement".

Proximity

Random Forest calculates proximity between the observations.

It is as follows,

a. Initialize proximity to zero.
b. For any tree, extend it to all cases,
c. If set 'i & j' terminate in the same node, increase the proximity level by one.
d. Cumulate all in RF and make it normal by twicing the number of trees.

The above steps create a matrix. An instance that is sameness will have proximity close to 1.

The matrix is used in the following cases,

- Missing value imputation,
- Outlier detection.

Variable importance

For every tree grown put the out-of-bag cases and add the votes. And permute the values in the out-of-bag cases and put down in the grown tree. Subtract the votes for the correct class in the permuted one from un permuted out- of-bag. The calculated average vote is the score for N.

If this average is independent form everyone then by using standard computation error is calculated.

If the variable is very high in number, forests can run once, then run once more with the most important variables.

Gini importance

The gini impurity criteria for the two descent nodes must be small than the ascent node. Adding the gini value gives a fast variable importance. It measures the imbalance among values of a frequency. A coefficient of zero represent perfect equality and one represent maximal inequality.

Interactions

The variable N and K is considered as interacting if a split on one variable makes a split on another. Gini is computed for each and it is denoted as "rank". The differentiations of the ranks are median over all trees. This rank is also computed with the axiom that the two attributes are independent of each other.

1. Missed content restoration for the training set

It can be done in two ways:

First method

(a) Computing Median for all the values and restore it.

(b) Restore with the most often non-missing value.

The first method is the fastest and cheapest way.

Second method

It replaces missing values only in the tutorial set. Initially it does a rough filling. Further it computes proximities to fill.

If a continuous value - Fill a mean over the existing values.

If a categorical value - Fill a most often existing value.

2. Missed content restoration for the test set

It can be done in two ways based on whether the label exists or not.

If label exists – The content are abstracted from the training set.

If label do not exists – Each set is replicated by the number of classes exists and the first replicated set is taken as class 1 and used to restore the place. The class 2 use the class 1.

Mislabelled cases

The sets are usually having the class labels which are assigned manually. So in some cases it will promote high level of mislabelling. These cases can be solved by using outlier analytics.

Outliers

Outliers are taken as an unfit to the available data. It can be defined as, the "cases whose proximities are mini than other cases in the data".

The average can be calculated as,

$$\bar{P}(n) = \sum_{cl(k)=j} \text{prox}^2(n,k)$$

----- **(1)**

The outlier can be defined as

$$\text{nsample}/\bar{P}(n)$$

----- **(2)**

The cost value is max if the proximity is small.

The outlier is calculated by,

"Difference in the median and the absolute deviation from the median".

Predicting error

In some cases, the error between classes is not balanced to handle. Some have a low error while others have a high. This normally appears when one is much higher than the another. The 'random forest algorithm' minimize the overall error rate' by keeping the error rate less on the large and high on smaller classes.

Unsupervised learning

Random forest predictors measure dissimilarity between the observations. So it can be defined a random forest works in non-labelled data. The "observed" instances must be distinguishes from the "synthetic data". The unlabeled data are taken as the observed case and the other is obtained from a reference distribution. This handles mixed types and is robust to outlying instances.

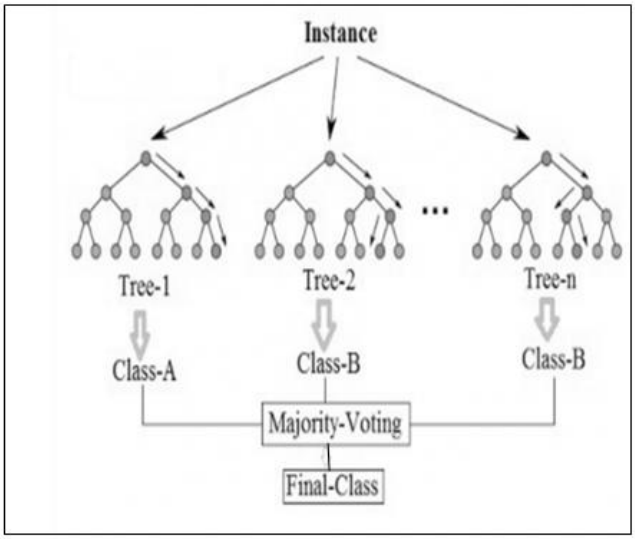


Figure 3.1 Random Forest Tree

Figure 3.1 Shows the Random Forest tree growing process.

Tree Growing

Each grown as follows,

- If the count in the set is M, it is drawn at "random with replacement", and this is taken as the train set .

- If there are N inputs, a number 'n<<N' is specified, and n variables are selected j from N and the best split is used for further splitting. The value is kept remind while the growing of the forest .
- Each is grown to the deepest length possible. There is not at all pruning.

Error Rate

It relies on two things:

Correlation: Increasing this will increases the error rate.

Strength: A tree with lessen error rate is a strong classifier.

Increasing this will decrease the error rate.

Working Principle

1. **Random File Selecting**: It is trained on 2/3rd of the trained data, drawn at random with replacement from the data.
2. **Random Variable Selecting:** Variables are get at random and the best on these is used to split.
3. By using the leftover data, calculate the out-of-bag error rate. Summate error from all to find the overall error rate for the classification.
4. Each tree ends in a "votes" for particular class. The forest selects which have the most votes over all the trees. For a binary variable, the vote is "1 or 0" and this is taken as the score .

Procedure for tree growing

Input: **Raw Data**
Output**: Random Forest Tree**
Method
Step 1**:** Randomly select "k" features from total "m" features. Where k << m; ***Step 2*****:** Among the "k" features, calculate the node "d" using the best split point. ***Step 3*****:** Split the node into daughter nodes using the best split. ***Step 4*****:** Repeat 1 to 3 steps until "l" number of nodes has been reached. ***Step 5*****:** Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

Procedure for Classification

Input**: Random Forest Tree**
Output: **Classified Data**
Method
Step 1**:** Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome . ***Step 2*****:** Calculate the votes for each predicted target. ***Step 3*****:** Consider the high voted predicted target as the final prediction from the random forest algorithm.

Advantages

- No over fitting problem even many branch are grown.
- Identify the features in the available features.
- It is excelled in accuracy among current algorithms.
- Run efficiently on high dataset.
- Without deleting any variable it can handle numerous inputs.
- Generates an unbiased estimate of the generalization error.
- Estimates missed content and maintains certainty even a large proportion of the series are missed.
- Grown forests can be conserved for further application.
- Models are created on the given information about variables and the classifiers.
- It computes proximities that provide interesting views to the data.
- Handle unlabeled data.
- Detects attribute interactions.

4. RESULTS AND DISCUSSION

Microsoft Excel

Excel could be a program developed by Microsoft for 'Windows, mac OS, golem and iOS'. Its options are unit calculation, graph tools, tables, and a macro artificial language known as Visual Basic for applications. It's been a really wide applied program for these platforms , particularly since version five in 1993, and it's replaced Lotus 1-2-3 because the business customary for spreadsheets. Excel forms a part of workplace.

WEKA Tool

"Waikato Environment for edge Analysis" could be a in style suite of machine learning software package coded

in Java, built at the 'University of Waikato, New Zealand'. It's free software package authorized underneath the wildebeest General Public License. It's associated in the open source environment within which we are able to apply varied techniques.

WEKA supports many customary data processing task and a lot of specifically information in 'Pre-Processing, Clustering, Classification, Regression,' and thus have choice. It supports solely 'ARFF' files. The file are often simply reborn to ARFF format if it's 'CSV' file. All the routines square measure predicated on the idea that the info is accessible jointly file or relation, wherever every information is represented by a hard and fast variety of attributes.

WEKA provides access to SQL infos victimization Java info property and may method the result came back by a database question. It is powerless of multi-relational data processing, however there's separate software package for changing a group of coupled info tables into one table that's appropriate for process victimization.

5. RESULTS

Classification Techniques

Algorithm Used

: Random Forest

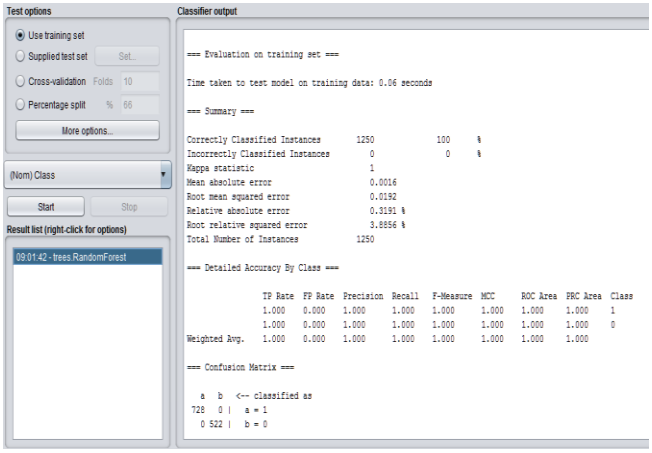


Figure 5.1.1 Classification

Figure 5.1.1 shows the classification result.
Class 0: 522;
Class 1: 728;

Class 0

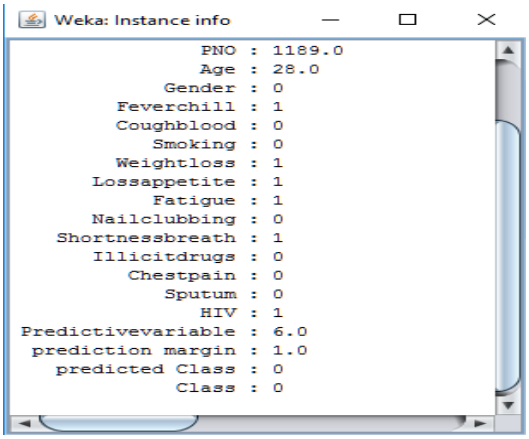


Figure 5.1.2 Classification for Class 0

Figure 5.1.2 shows the clustered output for Class 0. It has few sign of TB but the test result Sputum is negative. Since the patient is suffered from HIV they may be affected by TB. So they are considered as Latent TB.
Entry 1 shows the positive and high value.
Entry 0 shows the negative and low value.

Class 1

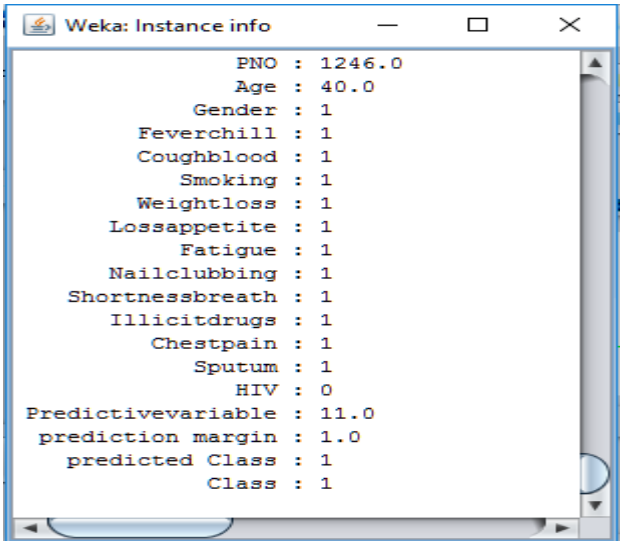
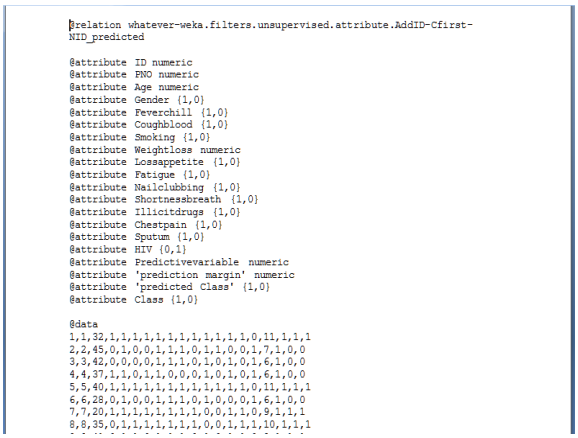


Figure 5.1.3 Classification for Class 1

Figure 5.1.3 shows the clustered output for Class 1. It has most of the sign of TB and the result Sputum is also positive. So they are considered as Active TB.
Entry 1 shows the positive and high value.
Entry 0 shows the negative and low value.

Printed out form for Classification



```
@relation whatever-weka.filters.unsupervised.attribute.AddID-Cfirst-
NID_predicted

@attribute ID numeric
@attribute PNO numeric
@attribute Age numeric
@attribute Gender {1,0}
@attribute Feverchill {1,0}
@attribute Coughblood {1,0}
@attribute Smoking {1,0}
@attribute Weightloss numeric
@attribute Lossappetite {1,0}
@attribute Fatigue {1,0}
@attribute Nailclubbing {1,0}
@attribute Shortnessbreath {1,0}
@attribute Illicitdrugs {1,0}
@attribute Chestpain {1,0}
@attribute Sputum {1,0}
@attribute HIV {0,1}
@attribute Predictivevariable numeric
@attribute 'prediction margin' numeric
@attribute 'predicted Class' {1,0}
@attribute Class {1,0}

@data
1,1,32,1,1,1,1,1,1,1,1,1,1,1,0,11,1,1,1
2,2,45,0,1,0,0,1,1,1,1,0,1,1,0,0,1,7,1,0,0
3,3,42,0,0,0,0,1,1,1,0,1,0,1,0,1,6,1,0,0
4,4,37,1,1,0,1,1,0,0,0,1,0,1,0,1,6,1,0,0
5,5,40,1,1,1,1,1,1,1,1,1,1,1,1,0,11,1,1,1
6,6,28,0,1,0,0,1,1,1,0,1,0,0,0,1,6,1,0,0
7,7,20,1,1,1,1,1,1,1,1,0,0,1,1,0,9,1,1,1
8,8,35,0,1,1,1,1,1,1,1,0,0,1,1,1,10,1,1,1
```

Figure 5.1.4 Overall Classification Result

Figure 5.1.4 show the printed output form for classification.

Chart for Classification

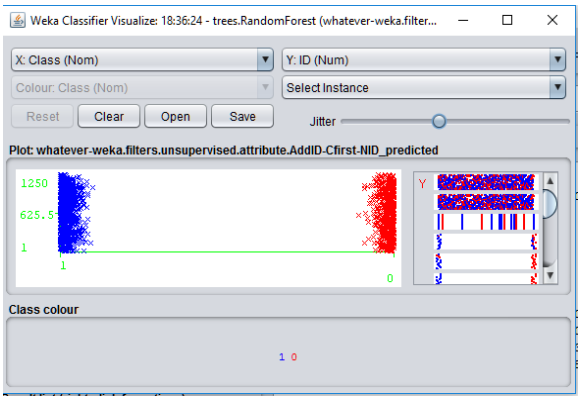


Chart 5.1 Classification

Chart 5.1 shows the classification. It is classified as,

Class 0: Latent TB;

Class 1: Active TB;

Comparison Chart for Existing and Proposed System

Chart for True Positive

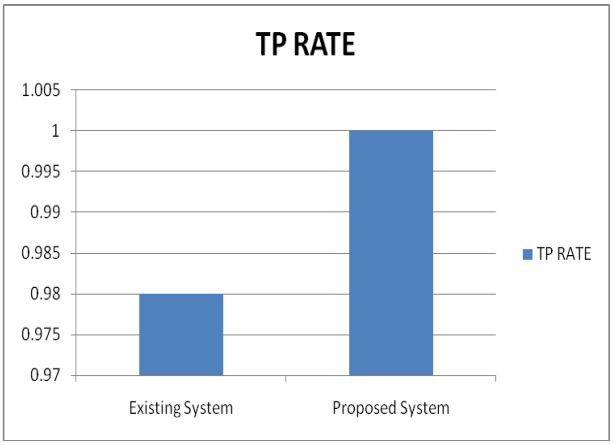


Chart 5.2 TP Rate

Chart 5.2 shows the assessment of the Existing and the Proposed work on the True positive rate. The existing has the excessive rate.

Chart for False Positive Rate

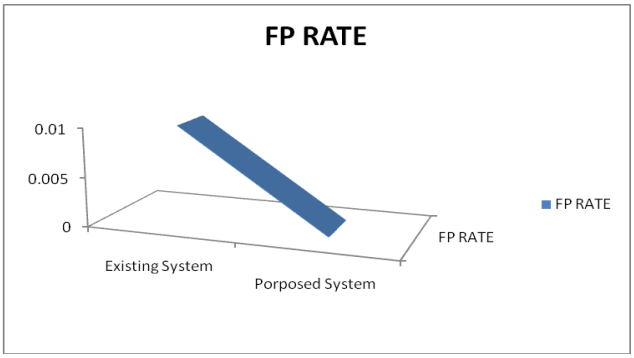


Chart 5.3 FP Rate

Chart 5.3 shows the assessment of the Existing and the Proposed work on the false positive rate. The existing system has the minimum error .

Chart for Precision

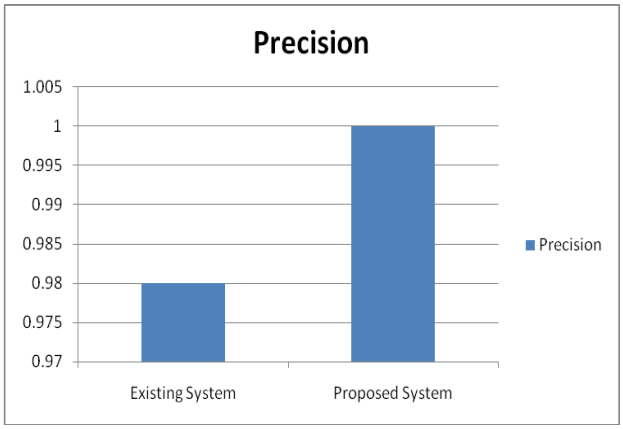


Chart 5.4 Precision

Chart 5.4 shows the resemblance of the Existing and the Proposed work on the Precision. The existing system has the maximum rate.

Chart for Recall

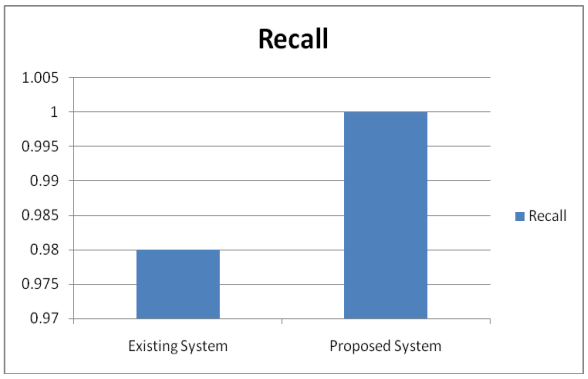


Chart 5.5 Recall

Chart 5.5 shows the resemblance of the Existing and the proposed work on the Recall. The existing system has the maximum rate.

Chart for F-Measure

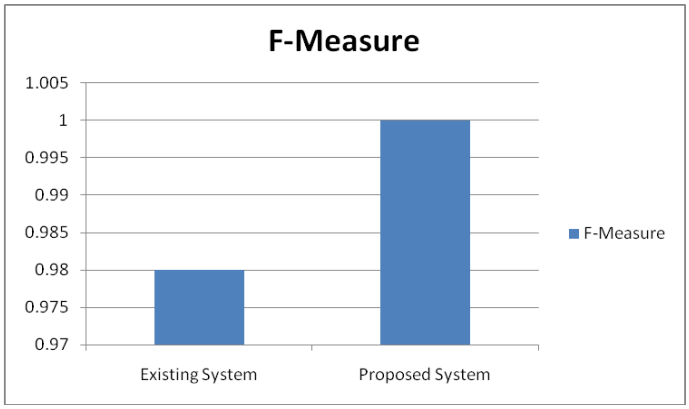


Chart 5.6 F-Measures

Chart 5.6 shows the co-relation of the Existing and the proposed work on the F-Measure. The existing work exceeds the proposed.

Chart for Accuracy

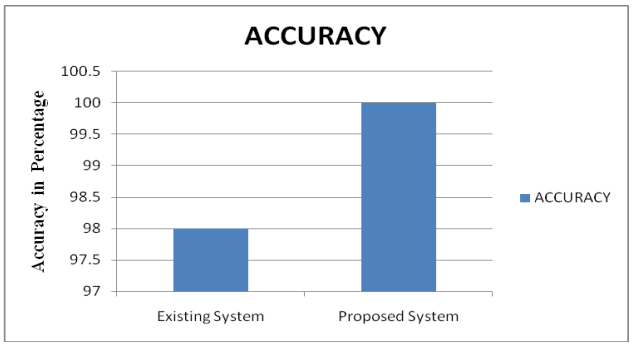


Chart 5.7 Accuracy

Chart 5.7 shows the difference on the Existing and the proposed work on the Accuracy. The existing work is uplifted.

5. CONCLUSION

There will be many prosperity in the aspect of life lead by the patient as the use of analysis on the patient's acquired data increases. The motivation of this work is to deploy an intellectual systematic process to assess the disease flawlessly. Data Mining routines are used to reveal the hidden norms from the vast collection of patient's data. Classification and Clustering are evaluated for Predictive and Descriptive analysis respectively. When the both routines are

allied it produces high precision and also takes part in detecting the Outliers.

For Implementation, this work use WEKA 3.8 tool. It supports several standard task in general and more specifically "Data Pre-Processing, clustering, and classification, Regression, Visualization and Feature Selection". The output is confirmed by analysing the outcome in more forms. And it is clearly viewed by the chart created by WEKA tool.

REFERENCES

[1] Ashfaq Ahmed . K, Sultan Aljahdali and Syed Naimatullah Hussain, "Comparative Prediction performance with support Vector Machine and Random Forest Classification Techniques " , International Journal of Computer Applications, Volume 69-No.11, pp no 12-16. 2013.

[2] Madhuri V. Joseph Data Mining : " A Comparative study in various techniques and methods", IJARCSSE, Volume 3,Issue 2,Feb 2013.

[3] Manish Shukla and Sonali Agarwal, "Hybrid approach for tuberculosis data classification using optimal centroid selection based clustering" DOI: 10.1109/SCES.2014.6880115 Conference: Students Conference on Engineering and Systems (SCES) 2014.

[4] K. R. Lakshmi, M. Veera Krishna, S. Prem Kumar, " Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability" DOI: 10.5815/IJMECS, 02.08.2013 .

[5] Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu, "Tuberculosis disease diagnosis using Artificial Neural networks ", Journal of Medical Systems, Springer, DOI 10.1007/s10916-008-9241, 2008.

[6] Wai Yan Nyein Naing , Zaw Z. Htike, IIUM, Malaysia, "Advances in Automatic Tuberculosis Detection in Chest X-Ray Images " volume 5, number 6, SIPIJ, December 2014.

[7] Collins K. Ahorlu, Frank Bonsu, "Factors affecting TB case detection and treatment in the Sisala East District, Ghana", Journal of Tuberculosis Research, 1 , 29-36. DOI: 10.4236/jtr.2013.13006.