IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

A Study On Tuberculosis Analysis Using Data Mining Techniques

N. Suresh[1], K. Arulanandam[2]

Research Scholar, Research Department of Computer Applications, Government Thirumagal Mills College, Gudiyattam. India[1]

Research Supervisor, Research Department of Computer Applications, Government Thirumagal Mills College, Gudiyattam. India[2]

Abstract: Tuberculosis (TB) is an infectious disease usually caused by the bacterium Mycobacterium tuberculosis (MTB). Tuberculosis generally affects the lungs, but can also affect other parts of the body. Most infections do not have symptoms, in which case it is known as latent tuberculosis. About 10% of latent infections progress to active disease which, if left untreated, kills about half of those infected. The classic symptoms of active TB are a chronic cough with blood-containing sputum, fever, night sweats, and weight loss. The historical term "consumption" came about due to the weight loss. Infection of other organs can cause a wide range of symptoms. In previous years TB classification has been done using various algorithms like color segmentation, thresholding, histogram equalization. The main objective of this research Data Mining analysis uses efficient techniques and statistical measures for analyzing the data to predict the possible causes for the health issues and its impact on individual patients. Enormous data mining techniques are available for analysing the outcome accurately. When Classification technique is used in conjunction with the clustering technique , it produces considerable improvement in learning the accuracy particularly in detecting the Outliers. The main objective of using K-Means algorithm is to find the common factors between tuberculosis patients. Clustering is a useful technique of data distribution and finding patterns in the data. In the end, results are being evaluated after classification and testing on the basis of performance parameter such as accuracy, recall, precision, false acceptance ratio, and false rejection ratio.

Keywords: Tuberculosis(TB), Data mining, Neural Network, K-Means Algorithm.

1. INTRODUCTION

1.1 Data Mining

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

1.2 TUBERCULOSIS(TB)

Tuberculosis is a highly 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 estimates 9 million people get sick with TB in a year.
- Women of age 15 to 44 gets affected by TB and is is among the top 3 causes of death.
- TB symptoms may be mild for several months, and the infected people spreads TB up to 10-15 other people through close contact .
- TB is an 'Airborne Pathogen'- spread through air from person to person.

Tuberculosis conditions

If the bacteria that causes TB enters the body the following three things may happen.

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

Tuberculosis in other parts

Tuberculosis infection in bones	--	leads to joint destruction and spinal pain.
Tuberculosis infection in brain	--	leads to meningitis.
Tuberculosis infection in kidney and liver	--	leads to blood in the urine.
Tuberculosis infection in heart	--	leads to cardiac tamponade.

Types

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

Latent TB

The bacteria asleep in the body in an idle state. They have no symptoms and it is not transmissible. But still they have the ability to form as active. To control the disease the disease should be identified and treat properly which is generally carried out for several months.

Risk Factors

Since there is no symptoms for Latent TB the risk factors include:

- HIV infection,
- Recent contact with an infectious people,
- Under treatment of 'Antitumor necrosis factor (TNF)',
- Undergoing dialysis,
- Receiving transplantation,
- Silicosis,
- Being an immigrant from highly affected TB burden countries,
- Being an illicit drug user.

Active TB

This bacteria have symptoms and can be spread to others. If the body resistance is minimum the bacteria leads to cause active tuberculosis. The active bacteria begin to increase in number in the body and cause active tuberculosis. It attacks the body and destroy the tissue. If the lungs get affected by this kind of bacteria then it actually create a hole in the lung. Depend on the body immunity power the people is affected by the bacteria soon or later. Normally Babies and young children often have weak immune systems so they can easily gets affected.

Conditions for weak immune system

- Substance abuse,
- Diabetes mellitus,
- Silicosis,
- Cancer in head or neck,
- Leukemia or Hodgkin's disease,
- Severe kidney disease,
- Low body weight,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

- Certain medical treatments ,
- Specialized treatment for rheumatoid arthritis or Crohn's disease.

Symptoms of TB
The most common symptoms of active TB includes:

- Coughing, sometimes with mucus or 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 depends on a person's age, health, resistance to drugs, type of TB whether latent or active and the location of infection . Patient with latent TB may need one kind of TB antibiotics, whereas patient with active TB will require a prescription of multiple drugs and for relatively long time. The course of TB antibiotics is about 6 months.
Prevention
A few measures can be taken to prevent the spread of active TB mostly by using,
(a) TB vaccination ("BCG"),
(b) Finish the medication completely while the patient in latent TB.

2. METHODOLOGY

2.1 Proposed system
Proposed Work
The proposed system classifies the tuberculosis dataset for high risk and low risk and clusters the patients according to the category of tuberculosis.
Data set
The tuberculosis dataset consists of 1250 data collected from a medical practitioner in the city hospital.
Flow of Work
The Existing system works in three phases.
Phase I – Pre-Processing,
Phase II – Classification
Phase III – Clustering.
Flow Chart

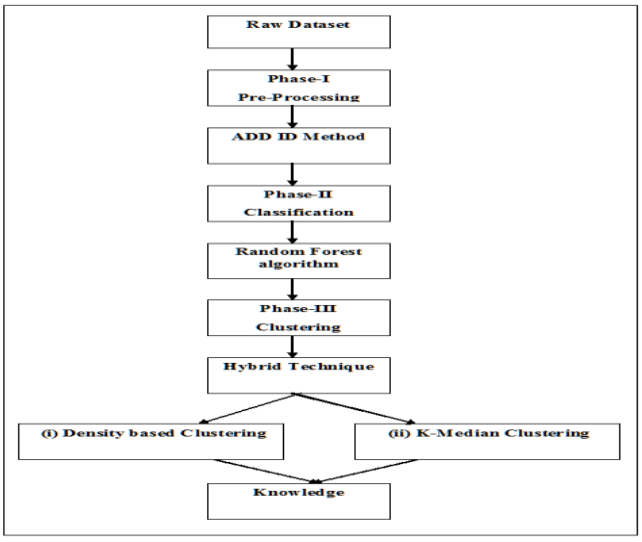

Figure 2.1 Flowchart of Proposed system

Figure 2.1 shows the proposed system flow chart. The system works in three phases. Data Mining Techniques used in the proposed system Phase-I → Pre-Processing

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Pre processing is a technique that involves changing raw data into an clear format that can be understandable. The data which available in real may be incomplete, inconsistent or lacking some important certain characteristics and may contain many errors. This technique has many methods for resolving such issues. It prepares or transforms the raw data for further processing.

Major steps involved in data pre processing are :

(a) Data cleaning,
(b) Data integration,
(c) Data reduction and
(d) Data transformation.

(a) Data Cleaning

The raw data may be noisy, incomplete and inconsistent. Data cleaning procedures make a attempt to fill the missing values, smooth the noisy data while finding outliers and rectify the inconsistencies in the data.

Methods used in Data Cleaning

(i) Missing Values

The data can have more missing entries and it leads to misclassification.

1. Ignoring the tuple: This is carried out when the class label is missing. While ignoring the tuple, the remaining attributes is also not used in the tuple. But this not effective, if the tuple does not contains several attributes with missing values. And it is considered poor when the percentage of missing values per attribute varies.
2. Filling the missing value: The missed values are entered manually but this is time consuming.
3. Replace the missing value: Replace all missing values by a unique constant. Though this method is simple, it is not foolproof.
4. Use a mean or median to fill the missing value: which pick the "middle" value of a data to fill the missed value. The symmetric distributions require the mean value, while skewed distribution requires median value to fill or replace.
5. Use the most possible value to fill the missing value: Using a Bayesian method or decision tree induction this may be determined .

(ii) Noisy Data

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

The following methods are used to recover from noise.

Binning: Binning methods smooth a sorted data value based on the values around it. Then the sorted values are allotted into a number of "buckets " or "bins".

Smoothing by bin is a variation in which each value is replaced by the mean value of the bin.

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

Outlier analysis: It may be detected by the method clustering. The similar values are formed as a group in which the most unfit one forms a outlier.

(b) Data Integration

It refers to the combining of data from multiple data sources. Care should be taken to avoid redundancies and inconsistencies after merging the data. If it is avoided it can help to improve the accuracy and speed of the data mining process. The well known implementation of data integration is building a data warehouse. The data warehouse enables to perform analyses based on the data in the warehouse.

Data Integration Techniques

- Manual Integration – User operate with the relevant information.
- Application Based Integration – A particular application do all the integration process.
- Middleware Data Integration – The integration process is transferred to the middleware.
- Virtual Integration – Data resides in the source system and defines a set of unified view for accessing.
- Physical Data Integration – Creates a new system and have a copy of the data from the source system .

(c)Data Reduction

It is a reduced representation of the data set that is much smaller from the original size but maintains the integrity of the original data. Mining on the reduced data set should reflects the same analytical results as the original data set. Data reduction increase storage efficiency and reduce costs.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Data Reduction Strategies
It includes,

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

Dimensionality reduction
It is the process of reducing the number of variables or attributes based on some consideration.
And it includes,

- Wavelet transforms - Transforms the original data onto a smaller space.
- Principal components analysis - Selects the important attributes.
- Attribute subset selection - redundant attributes are detected and removed.

Numerosity reduction
This technique replace the original data volume with the other forms of data representation. It may be,

- Parametric - Data parameters are stored instead of actual data.
  Ex: Regression and Log-linear models.
- Non-parametric - Stores reduced representations of the data.
  Ex: Histograms, Clustering, Sampling and Data cube aggregation.

Data Compression
The transformations are applied to obtain a "reduced or compressed" form of the original data.
It may be,

- Lossless – Original data can be reconstructed without loss .
- Lossy – Cannot reconstruct the actual 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 techniques is also considered one form of data compression.
(d) Data Transformation and Data Discretization
In data transformation the data is changed into a form which is appropriate for the mining process.

Data Transformation Strategies
(i) Smoothing - It removes noise from the data. 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 - Attribute data are scaled within a smaller range .
(v) Discretization - Normal values are replaced by interval labels .
(vi) Hierarchy generation - Attributes can be generalized from higher to lower level .
Data Transformation by Normalization
Normalizing the data means to assign an equal weight to all the attributes. It is particularly used for classification algorithms which includes 'neural networks , nearest-neighbour classification and clustering'.

Phase –II Classification
Classification aims to assign data item from a collection to the specified categories or classes. The goal is to accurately predict the specified class for each case in the data. The task begins with a data set in which the class labels are known. Classifications are discrete. Continuous, floating-point values are denoted by numerical rather than categorical value. A numerical target used for predictive model use a regression algorithm instead of classification algorithm. The simplest form of classification is binary classification. In this classification, the target attribute has only two possible outcomes either high or low. There are also multiclass targets which have more than two values low, medium and high. The classification algorithm finds relationships between the values and the target.

Process
The data is divided into two sets one for building the model and the other for testing the model. The training set is used to build the model. The balance is used for testing.

Classification process includes two steps:

- Building the Classifier or Model

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

This is referred as the "learning step or the learning phase". It is built from the training set from the database tuples with their associated class labels. Each tuple in the training set is referred to as a "category or class". Also is referred as sample, object or data points.

- Using Classifier for Classification

The classifier is used for classification. The test data is used to estimate the accuracy of classification rules instead of the training set.

Training and Test data set

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

Classification algorithms includes,

- Linear classifiers
- Fisher's linear discriminant
- Logistic regression
- Naive Bayes classifier
- Perception
- Support vector machines
- Least squares support vector machines
- Quadratic classifiers
- Kernel estimation
- k-nearest neighbour
- Boosting
- Decision trees
- Random forests
- Neural networks
- FMM Neural Networks
- Learning vector quantization

Method Used

Random Forest Algorithm

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

The algorithm was first created by "Tin Kam Ho" by using the random subspace method where his formulation is used. "Stochastic Discrimination" approach 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 Breiman's "bagging" method with "Random selection of features" .

In the random forest classifier, the higher the number of trees gives the high accuracy results. It averages the multiple decision trees, trained on completely variant components to minimize the variance.

Features

- The algorithm can be used 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 grow larger.

Terminologies used in random forest algorithm

Bagging

Each training data set picks a sample of instances with replacement and it is referred as "Bootstrap Sample" from data set. By "Sampling with Replacement", the instances can be replaced in each training data set. So, N models are created using the N bootstrap samples and finally it is combined by averaging the output or votes .

For Training data - 2/3rd of the total data (63.2%) is used.

For Testing data - Balance 1/3 of the total data (36.8%) is used.

Trees are grown only for the training data.

Out of Bag Error

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

It is equal to validation data or test data. There is no separation for validating the result.
It is estimated during the process as follows,
The test data is not used in building that tree and it is called as "Out of bag error estimation".

Bootstrap Sample
The sample is chosen by random with replacement sampling method. The sample which is selected for training set is again put back into the set. So it can be picked up again and it is referred as "Random with Replacement".
Proximity
Random Forest calculates proximity or similarity between two observations.
It is as follows,
a. Initialize proximity to zero.
b. For any given tree, apply the tree to all cases,
c. If case i and j both terminate in the same node, increase proximity between i and j by one .
d. Cumulate all trees in RF and normalize it by twicing the number of trees.
The above steps create a proximity matrix. Instances that are "alike" will have proximity close to 1.

Proximity matrix is used in the following cases,
- Missing value imputation,
- Outlier detection ,

Variable importance
For every tree grown in the forest put the out of bag cases and count the number of votes for the class. And randomly permute the values of variable N in the out of bag cases and put these cases down in the tree. Subtract the number of votes for the correct class in the permuted out of bag data from the number of votes for the correct class in the un permuted out of bag data. The calculated average votes in the forest are the raw importance score for variable N.
If this average is independent form tree to tree then by using standard computation error is calculated.
If the number of variables is very large in number, forests can run once with all the variables, then run again by using only the most important variables.

Gini importance
Every time a split of a node is made on variable N the gini impurity criteria for the two descendent nodes must be less than the parent node. Adding the gini value gives a fast variable importance . It measures the inequality among values of a frequency. A Gini coefficient of zero represent perfect equality if all values are the same. A Gini coefficient of one represent maximal inequality.

Interactions
The variable N and K is said to be interact with each other if a split on one variable N in a tree makes a split on the another variable K. Gini index is computed for each tree and it is called as "rank". The absolute difference of the ranks are averaged over all trees. This rank is also computed under the hypothesis that the two variables are independent of each other.

1. Missing value replacement for the training set
The missing value replacement can be done in two ways:
First method
(a) Computing Median value of all the values replace all missing values of the N th variable in class j if the variable to is not categorical.
(b) The replacement will be the most frequent non-missing value in class j, if the variable is categorical.
The first method is considered as a fastest and cheapest way to replace the value.
Second method
It replaces missing values only in the training set. Initially it begins by doing a rough filling of the missing values. Then it computes proximities to fill.
If it is a continuous value - Fill an average over the non-missing values of the N^{th} variables weighted by the proximities.
If it is a categorical value - Fill it by the most frequent non-missing value where frequency is weighted by proximity.

2. Missing value replacement for the test set
The missing value replacement can be done in two ways based on whether the class label exists or not.
If class label exists – The values are abstracted from the training set.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

If label do not exists – Each test set is replicated number of classes exists the first replicated set is taken as class 1 and it is used to replace missing values. The class 2 use the class 1.

Mislabelled cases

The training sets are usually have the class labels which is assigned manually. So in some areas this leads to a high level of mislabelling. These mislabelled cases can be detected by using the outlier measure.

Outliers

Outliers are considered as unfit to the available data. It can also be defined as, the cases whose proximities are small than other cases in the data.

The average proximity for case n in class j to the rest of the training data class j can be calculated as,

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

The raw outlier measure for case n can be defined as

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

The value will be large if the average proximity is small. The outlier is measured by the difference in the median and the absolute deviation from the median.

Balancing prediction error

In some cases, the prediction error between classes is highly unbalanced to handle. Some classes have a low prediction error while others have a high. This is normally appears when one class is much higher than the another. The random forest algorithm minimize the overall error rate in this case by keeping the error rate low on the large class and high on smaller classes.

Unsupervised learning in random forests

Random forest predictors lead to a dissimilarity measure between the observations. So it can also be defined a random forest dissimilarity measure between unlabeled data. The "observed" data must be distinguishes from the "synthetic data". The unlabeled data are taken as the observed data and the synthetic data are obtained from a reference distribution. This dissimilarity

handles mixed variable types and is robust to outlying observations.

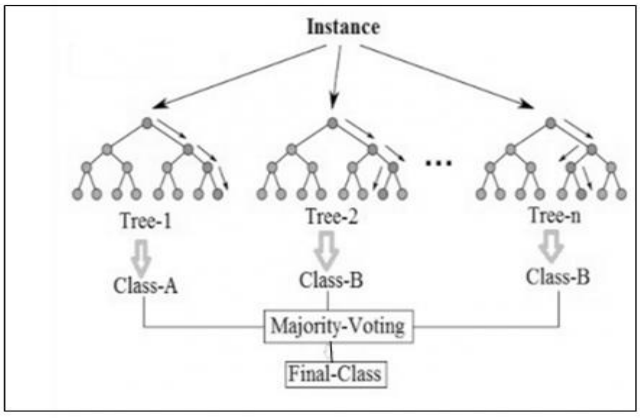

Figure 2.2 Random Forest Tree

Figure 2.2 Shows the Random Forest tree growing process.

Tree Growing

Each tree is grown as follows,

- If the number of cases in the training set is M, it is drawn at "random with replacement", from the data and this is taken as the training set .
- If there are N input variables, a number 'n<<N' is specified, at each node, n variables are selected at random from N and the best split on these n is used to split the node. The value of n is kept constant during the growing of the forest.
- Each tree is grown to the largest length possible. There is no tree pruning.

Error Rate

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

It depends on two things:
Correlation: Increasing the correlation increases the forest error rate.
Strength: A tree which has a low error rate is a considered as a strong classifier.
Increasing the strength will decrease the forest error rate.
Working Principle

1. Random Record Selection: Each tree is trained on roughly 2/3rd of the training data , drawn at random with replacement from the data.
2. Random Variable Selection : Predictor variables are selected at random out of all the variables and the best split on these is used to split the node.

For each tree, by using the leftover data, calculate the out of bag error rate. Aggregate error from all trees to calculate the overall error rate for the classification. Each tree ends in a classification, and is referred as "votes" for that particular class. The forest selects the classification which have the most votes over all the trees. For a binary dependent variable, the vote will be "1 or 0" and this is taken as the Random Forest 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

- There is no over fitting problem even many trees are grown.
- Used to identify the most important features out of the available features.
- It is excelled in accuracy among current algorithms.
- The algorithm can run efficiently on large data bases.
- Without deleting any variable it can handle large number of input variables.
- The algorithm generates an internal unbiased estimate of the generalization error while the forest grows.
- Estimates missing data and maintains accuracy even a large proportion of the data are missing.
- Grown forests can be saved for further use to apply on other data.
- Prototypes are computed based on the given information about the relation between the variables and the classification.
- It computes proximities which can be used in clustering, locating outliers or give interesting views to the data.
- Can also handle unlabeled data.
- It offers a method for detecting variable interactions.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Phase III
Clustering: is the process of binding a group of various objects into a similar classes by applying distance metric.
Features

- The data objects in a single cluster can be treated as one group.
- Cluster analysis is based on data similarity metric.
- It is adaptable to changes and helps to differentiate important features that is used to analyse different groups.

Requirements of Clustering

- Scalability − Highly scalable clustering algorithm is needed to deal with large databases.
- Handling different kinds of attributes – Algorithm has the ability to deal with different kinds of attributes.
- Handling arbitrary shape cluster – Capable of detecting clusters with arbitrary shape.
- High dimensionality − Able to handle heterogeneous data set even with high Spatial environment.
- Handling noisy data – Algorithms should not be sensitive to noisy data.
- Interpretability − The results should be explainable, comprehensible.

Method Used

(i) Density based clustering
This method is based on the idea of density. The basic theme is to "continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold".

3. IMPLEMENTATION

Microsoft Excel

Microsoft Excel could be a program developed by Microsoft for Windows, mac OS, golem and iOS. Its options area unit calculation, graphing tools, pivot 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 Microsoft workplace.

Basic Operation
Microsoft Excel has the fundamental options of all unfold sheets employing a grid of cells organized in numbered rows and letter-named columns to arrange knowledge manipulations like arithmetic operations. it's A battery of equipped functions to answer applied math, engineering and money desires. Additionally, it will show knowledge as line graphs, histograms and charts, and with a really restricted three-dimensional graphical show.
Number of rows and columns
Version 7.0 - 16K (2^{14} = 16384) rows.
Version 8.0 - 64K (2^{16} = 65536) rows and 256 columns.
Version 12.0 - 1M (2^{20} = 1048576) rows and 16384 columns.

File Formats

Microsoft Excel 2007 version used a proprietary computer file format referred to as Excel computer file Format (.XLS) as its primary format. Excel 2007 primary file format is Open XML, associate XML-based format that followed once a previous XML-based format referred to as "XML Spreadsheet" (".XMLSS"), initial introduced in Excel 2002. Though supporting and inspiring the utilization of latest XML-based formats as replacements, Excel 2007 is compatible with the normal binary formats. And most versions of Microsoft Excel will browse 'CSV, DBF, SYLK, DIF', different formats. Support for a few older file formats was removed in Excel 2007. The file formats were in the main from DOS-based programs.

Export and Migration of Spread Sheets

Programmers have made genus 'API's to open Excel spreadsheets in a very form of applications and environments aside from Microsoft Excel. These embody gap Excel documents on the net victimisation either ActiveX controls, or plug ins just like the Adobe Flash Player. The Apache dish open supply project provides Java libraries for reading and writing Excel program files. Excel Package is another ASCII text file project that gives server-side generation of Microsoft Excel 2007 spreadsheets. PHP Excel may be a PHP library that converts Excel 5, 2003, and 2007 formats into objects for reading and writing inside an internet application. Excel Services may be a current .NET developer tool that may enhance Excel's capabilities.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

4. RESULT AND DISCUSSION

Dataset in Excel

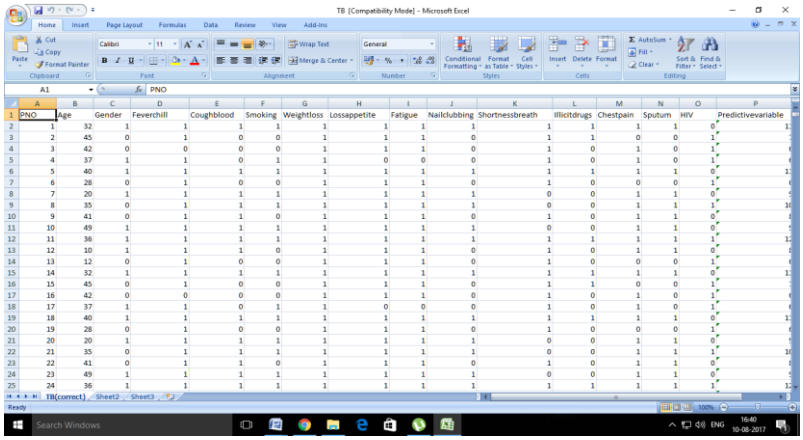

Figure 4.1 Dataset with attributes

Figure 4.1 shows the data set with 14 attributes in Microsoft Excel.

The attributes used in this work:
Age, Gender, Cough with Blood, Smoking, Weight Loss, Loss of Appetite, Fatigue, Nail Clubbing, Shortness of breath, Illicit Drugs, Chest pain, Sputum and HIV.
Phase – I Pre Processing

The raw dataset is collected from a heterogeneous source. So in the pre processing it is cleaned Integrated and Transformed. And also the dataset given by the medical Practitioner did not have the Patient Name or Patient Identification Number for security reason. So in Pre-Processing ADD-ID method is used to create Identification Number to the Patients.

Dataset in WEKA before pre-Processing

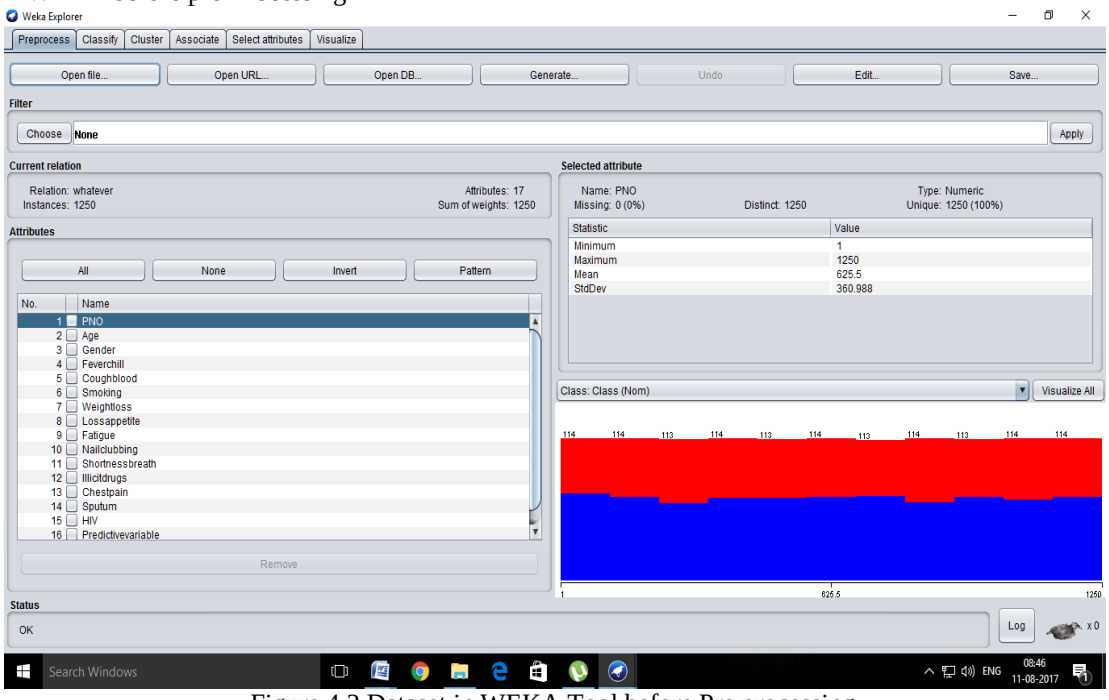

Figure 4.2 Dataset in WEKA Tool before Pre processing

Figure 4.2 shows the dataset opened in WEKA tool before pre -processing.

Phase – II Classification
Algorithm Used: Random Forest Algorithm

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

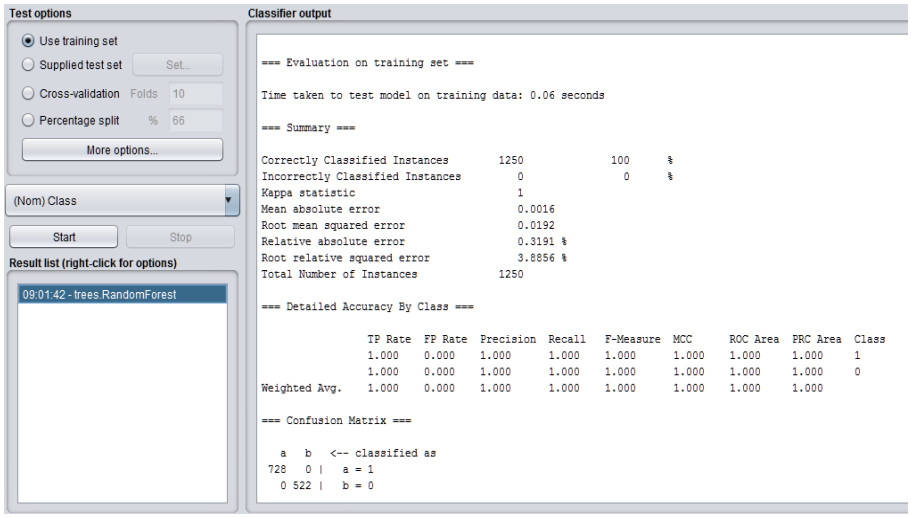

Figure 4.3 Classification in WEKA Tool

Figure 4.3 shows the classification result for the Random Forest algorithm.

Random Forest Algorithm

A. Most accurate learning algorithm.
B. Runs efficiently on large databases.
C. Can handle thousands of input variables without variable deletion.

Class 0: 522;
Class 1: 728;
Class 0

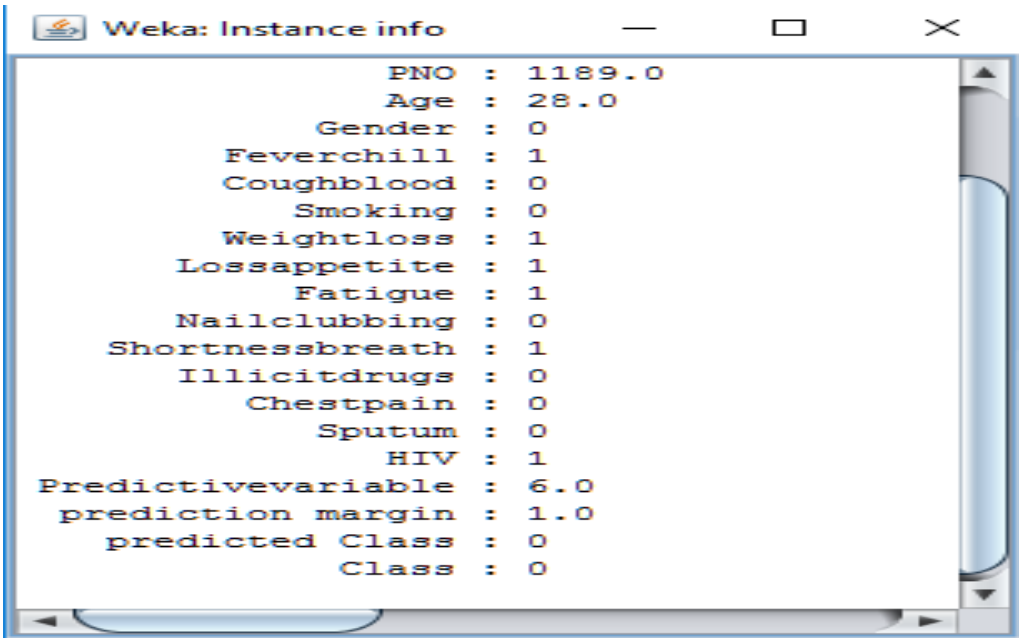

Figure 4.4 Classification for Latent TB

Figure 4.4 shows the clustered output for Latent TB.

It has some of the symptoms of TB but the test result Sputum is negative. Since the patient is suffered from HIV they may be affected by TB.
Entry 1 shows the positive and high value.
Entry 0 shows the negative and low value.
Class 1

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

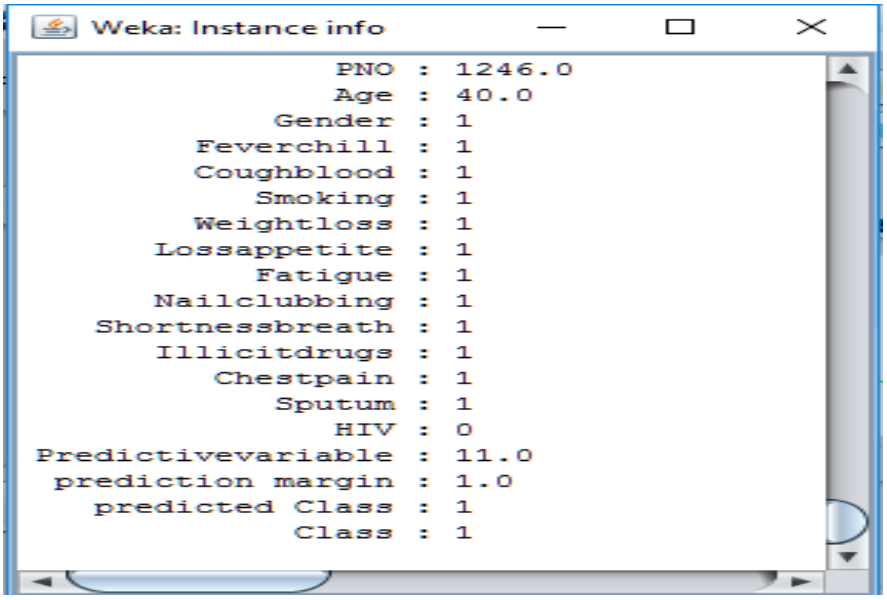

Figure 4.5 Classification for Active TB

Figure 4.5 shows the clustered output for Active TB. It has most of the symptoms of TB and the test result Sputum is also positive.

Entry 1 shows the positive and high value.
Entry 0 shows the negative and low value.
Phase – III Clustering
Algorithm Used: Hybrid Technique
(i) Density Based Clustering,
(ii) K-Median Clustering.

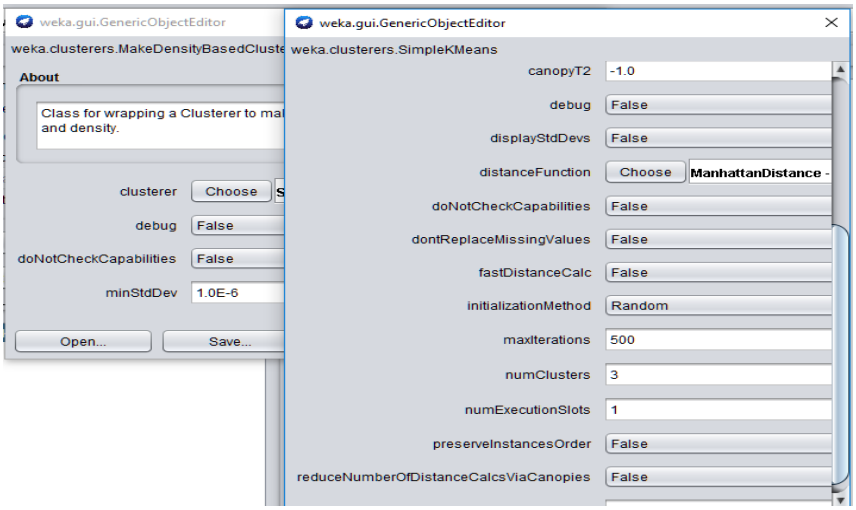

Figure 4.6 Density based clustering with K-Median

Figure 4.6 shows the Density based clustering in association with K-Median.

Density Based Clustering

Fits normal and discrete distributions within each cluster wrapped .

K-Median

K-Means with Manhattan Distance denotes the K-Median algorithm.
Cluster the data based on the Median value.

They are:

Cluster 0→ HIV Patients with Tuberculosis risk.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Cluster 1→Latent Tuberculosis patients without HIV.
Cluster 2→ Active Tuberculosis patients with HIV.
Cluster 0
HIV Patients with Tuberculosis risk.

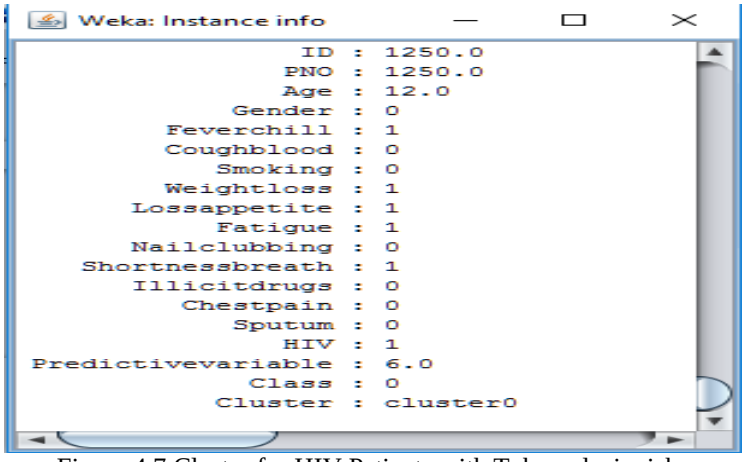

Figure 4.7 Cluster for HIV Patients with Tuberculosis risk.

Figure 4.7 shows the HIV patient with TB risk.

The patient is suffered from HIV and have some of the symptoms of TB.
Cluster 1

Latent Tuberculosis patients without HIV.

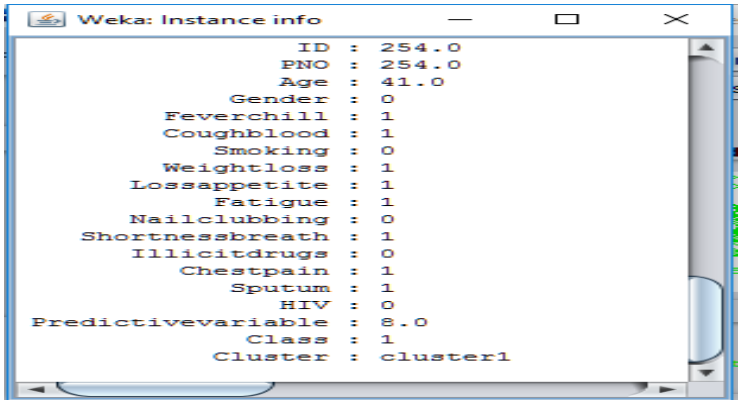

Figure 4.8 Cluster for Latent TB patient

Figure 4.8 shows the Latent TB patient.

Cluster 2
Active Tuberculosis patients with HIV

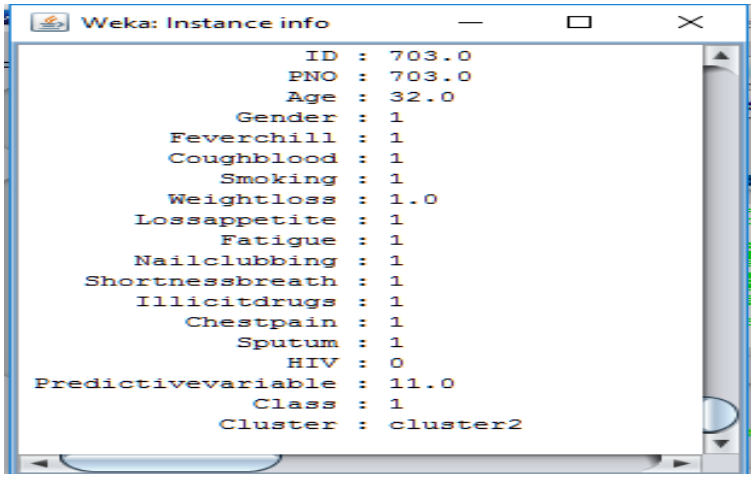

Figure 4.9Cluster for Active TB Patients

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Chart for all the Attributes

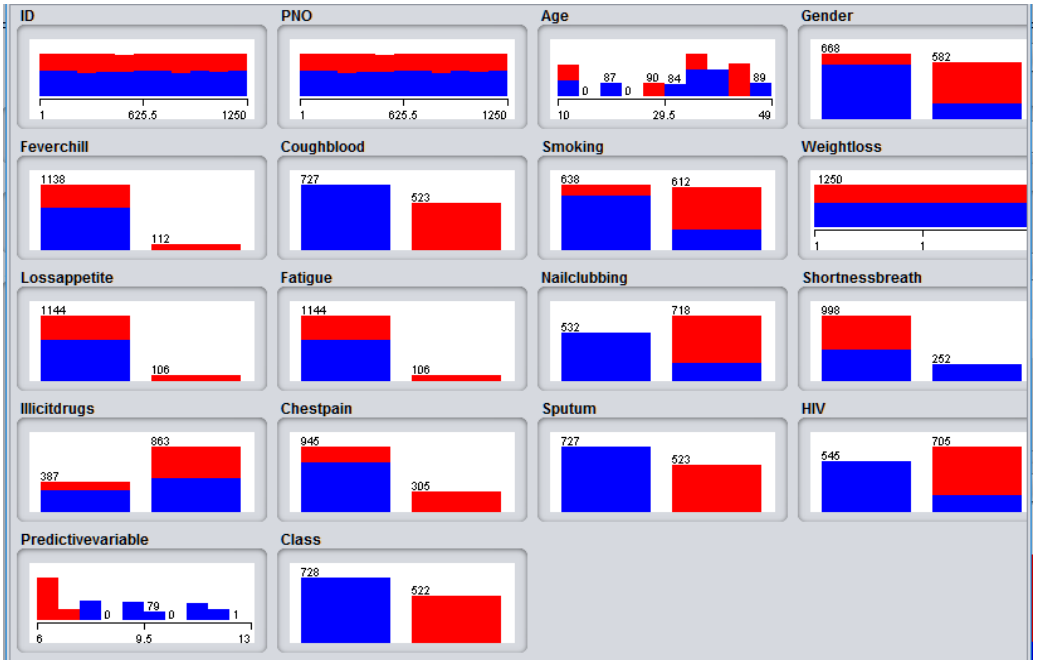

Chart 4.10 Individual attributes

Chart 4.10 shows the distribution of all the attributes in the WEKA Tool.
Class 0: Active TB;
Class 1: Latent TB;

Chart for Clustering

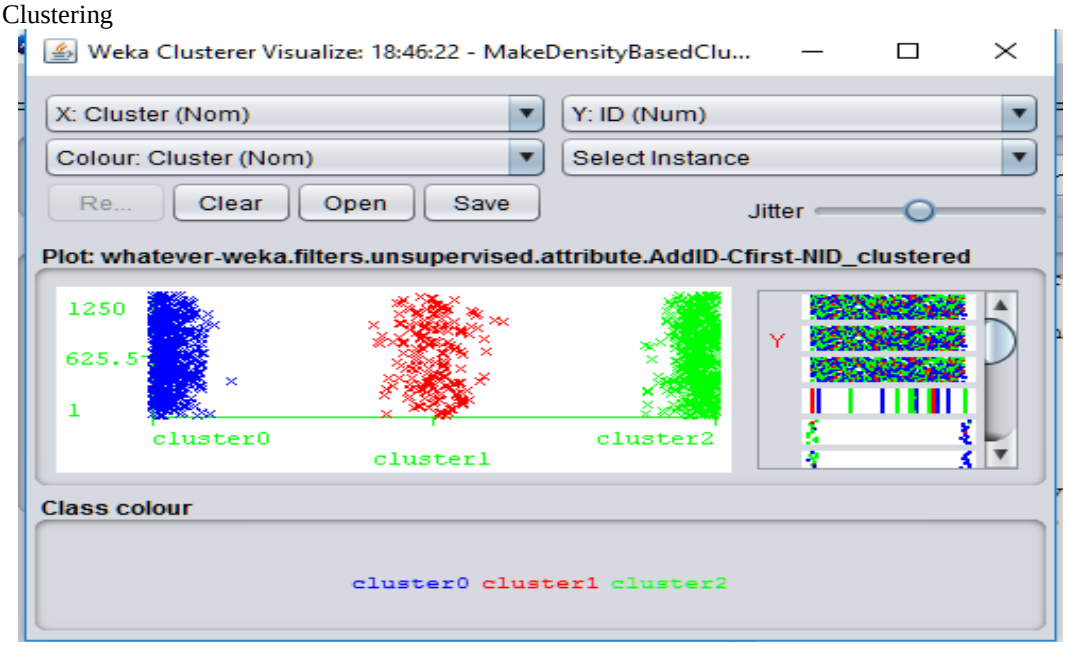

Chart 4.11 Clustering

5. CONCLUSION

Tuberculosis is a typical and frequently savage irresistible disease brought about by mycobacterium; in people it is primarily Mycobacterium tuberculosis. The aim of this research work is to deploy an intelligent system to predict the disease accurately. Data Mining Techniques are used to reveal the hidden patterns from the vast collection of patient's data. Data Mining Techniques such as Classification and Clustering are used for Predictive and Descriptive analysis respectively. When classification is used in conjunction with clustering, it produces high accuracy and also takes part in detecting the Outliers. Nearby is urgent destination to inhibit TB in humans.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

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.