International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | Apr-2016 www.irjet.net p-ISSN: 2395-0072

Review on Tuberculosis Detection Using Various Data Mining Techniques

Rupali Zakhmi[1], Jyoti Arora[2]

[1]*Research Scholar, Dept. Computer Science and Engineering, Desh Bhagat University, Punjab, India*
[2]*Assistant Professor, Dept. Computer Science and Engineering, Desh Bhagat University, Punjab, India*

Abstract - *Data Mining is one of the most important and inspiring area of research with the goal of discovering purposeful information from massive data sets. Data mining plays an important role in healthcare field to detect causation of various diseases, their treatment methods .Tuberculosis is one of the well-known disorders among all the persons in the nation including India. Tuberculosis is a virus which strikes the immune system of an individual, usually transmits through air. It is primarily occurs in lungs. It is the typical cause of necrosis. This paper discussed about various data mining techniques to detect tuberculosis such as Classification, Clustering and Association. There are some parameters that are also useful to detect Tuberculosis like Age, Cough, Fever, Chest pain and Weight Loss etc.*

Key Words: **Tuberculosis, Classification, Clustering, Association, Data Mining**

1. INTRODUCTION

Tuberculosis is a serious problem and transmits through bacteria known as Mycobacterium Tuberculosis. It is prime mover of demise. If right treatment is not given to the patient at proper time then it is very difficult to cure from TB. This disease is found on cattle, birds as well as in human being. The organs such as lungs are badly influenced by tuberculosis in most of the tuberculosis cases. TB attacks both grown-ups and kids. In early days, many different techniques were used such as sputum smear microscopy, chest radiography etc. These methods have several drawbacks, such as these methods require expertise to operate citified tools. These methods are suitable to obtain better results on time. Sometimes some symptoms of tuberculosis are same with other diseases, it leads to death. Incomplete information given by the patient or patient's family can stand in the way to find right treatment.

To control over these problems, some researcher use images, sounds or variables as inputs parameters. In this research, we will take some variables as input parameters to detect and identify tuberculosis. Most commonly data mining techniques have discussed in this research [2].

The aim of using data mining is to find significant information from vast data sets. It is also useful in the field of healthcare where unforeseen and relevant information are identified. Medical sector uses data mining techniques to know about various diseases, their causes and treatments. Techniques of data mining are very effective at the time of decision regarding patient heath. Medical data contains all about number of patients, cost of treatment, medical facilities etc. Analyzing this data, healthcare introduced powerful tool which extract important information that is necessary for patient's recovery. It also verifies how much time is taken by patients for diagnosis. Identification of tuberculosis at right time is very important. To enhance the performance of patient's treatment - Classification, Clustering and Association approaches have been introduced. Results of using data mining approaches provide benefits to healthcare domain by grouping the patients having same types of health issues [4].

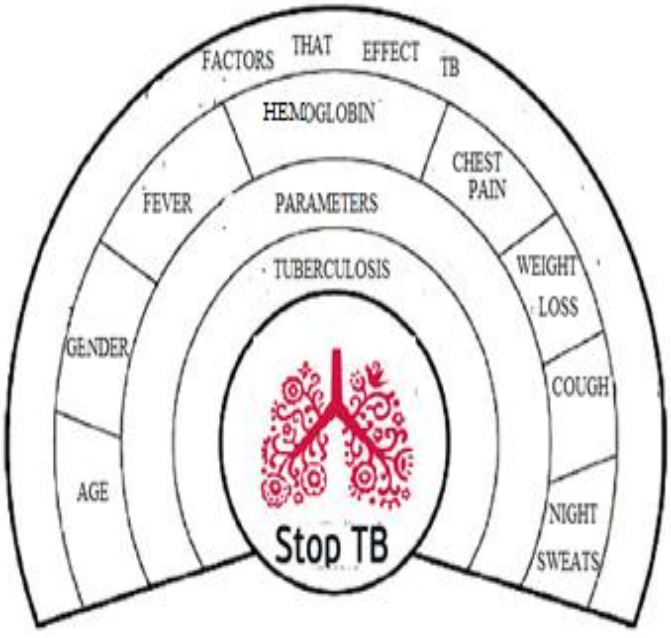

Fig -1: Variables taken as input parameter to detect TB

In Fig-1,Eight variables are used to identify and detect tuberculosis that are Age, Gender, Fever, Chest Pain, Weight Loss, Cough, Night sweats and Hemoglobin. Brain H. Tracey [8] uses cough detection algorithm to recover patient from pulmonary tuberculosis. Author proposed classification technique to decrease the cough count of patient.

2. DATA MINING APPROACHES

2.1 CLASSIFICATION

Classification is the one of data mining technique which is used to classify categorical data items. A classification technique includes neural network, Decision tree & Support vector machine classifier. Classification techniques are useful for predicting diseases such as TB, Liver, Lung, Heart disease etc. Some classification techniques are as follows:

2.1.1 Neural Network

Neural Networks are commonly categorized in different layers which are made up of nodes that are interdependent. There are three different layers-input layer, hidden layer and output layer. To show the patterns, input layer is used. Input layer interacts with hidden layers which links the output layer to get derive solution. Weights are assigned to hidden layers .A neural network adjusts its weights value by executing various functions. Shakshi Garg [1] proposed neural network and genetic algorithm to diagnosis tuberculosis. This disease is major issue in the area where treatment chances are less. To prevent TB, different data mining techniques was introduced.

Merits:

- NN needs training at once. There is no need to reprogram.
- Good for handling noisy data.
- Easily identify complex relationships between dependent and independent variables.

Demerits:

- It requires training to perform well.
- It is not suitable for large networks; hence it takes high processing time.
- It does not allows to add new nodes, cannot modified once created.

2.1.2 Decision Tree

It is tree structure method which breaks data items to create decision tree. In decision tree different attributes are used like Age, Gender etc to make decision. Numerous types of decision algorithms are employed –ID3 is one algorithm of the decision Tree.ID3 is used for forming short tree. Decision Tree minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. Decision tree is widely used by many researchers in healthcare field. Sofia Benbelkacem [4] introduced decision support tree for the treatment of Tuberculosis. This research used data mining approach to generate classification rules.

To build decision tree IGSS tool and jCOLIBRI as a platform was used. Indexing method was introduced to decrease the time of verification and adaptive rules defined by upgrading adaptation phase.

Merits:

- It can easily process data with high dimension.
- It is easy to interpret.
- It can handle both numerical and categorical data.

Demerits:

- It is not suitable for non numeric data.
- It is time consuming.

2.1.3 Support Vector Machine

The SVM algorithm was originally introduced by Vladimir Vapnik. In a high or infinite dimensional space, SVM constructs a hyperplane that was used for classification. Hyperplane maximizes the margin and minimizes the error. This plane achieves largest distance to the nearest training data points to good separation results. Tuberculosis is an infectious disease in which SVM classifier used by many researchers. Asha.T [9] proposed SVM classifier along with k-mean clustering. For better treatment planning methods, SVM classifier was used in the study. This achieved highest accuracy 98.7% as compared to other technique.

Merits:

- SVM give better accuracy as compare to other classifier.
- Easily handle complex nonlinear data points.
- No over fitting problem.

Demerits:

- SVM is computationally costly.
- For every dataset different kernel function shows different results. To choose right kernel from right dataset is the main problem of the SVM classifier.
- It takes lots of time during training process, as compare to other method.

2.2 CLUSTERING

Clustering is a technique which is also known as unsupervised learning method, has no predefined classes. The large data sets are divided into small data sets and later same data sets are grouped into single cluster. Objects having high similarity that is grouped into same cluster. The objects which are dissimilar exist in another cluster.

2.2.1 K-means Clustering

It is also known partitioning method. In this method objects are classified that belongs to k-groups. K-mean clustering consists of K-clusters, each object having one cluster. In every cluster there is a centroid where we take real value data. Arithmetic mean are used to find the variables within classes that represent appropriate solutions [6]. Centroid may be needed in other aspects. K-Means Algorithm Properties

- At least one item is present in a cluster
- Cluster are non hierarchical in nature.
- Clusters do not overlap with one another.

- K clusters are always present.
- Cluster's members are closer to its cluster instead of any other cluster.

Merits:

- It is simple clustering approach.
- Easy to understand.
- Complexity is less in K-mean clustering.
- It is very fast, robust.
- No overlapping character is allowed.

Demerits:

- Requires predefined cluster centres.
- Not suitable to handle categorical attributes.

2.2.2. Density Based Clustering

To discover cluster of arbitrary shapes, previous clustering techniques are not appropriate. They are suitable only for spherical shaped cluster. Density clustering methods eradicate this limitation very effectively by handling arbitrary shaped cluster. DBSCAN and OPTICS are two approaches of density based clustering method. In this cluster is discovered by density connectivity analysis. DENCLUE is another approach of density based clustering. Based on distribution value analysis, it assembles various data points of density function. Divya [4] introduced density based method to the region of uniform colour in biomedical images. The density based method separates the injurious skin from healthy skin.It also checks varied shade or dappled part inside the injurious skin. DBSCAN algorithm was used to test wounded skin.

Merits

- No need to specify number of cluster in advance.
- Easily handle cluster with arbitrary shape.
- Worked well in the presence of noise.

Demerits:

- Not handle the data points with varying densities.
- Results depend on the distance measure.

3. ASSOCIATION

Association is the method of data mining in which we can find commonly known patterns and keen relationship among group of data items in the data warehouse. Association has great effect in the field of healthcare to detect the relationship among various diseases and their symptoms. This approach is used in finding relationship among several diseases and drugs. Association rule mining is used to detect fraud in health insurance.

3.1 A Priori Algorithm

In 1993,A priori algorithm was introduced by R.Aggrawal. It is a well known method to discover remarkable relations among databases. The aim of this algorithm is to find logical relationship between various objects having same significance .The support and confidence are the two inputs. To distinguish between frequent and infrequent item sets, these two inputs are used. The research work removed those data items sets from database of transaction which cannot satisfy the given condition such as frequent items minimize the support and confidence constraints. In other words, this algorithm is based on the principle that items do not fulfil the minimum support constraint is removed from the transaction database. To construct the association rule, efficacy evaluation is main factor various methods are used to improve the efficiency of Apriori algorithm such as Hash table, transaction reduction, partitioning etc. Papender Kumar [7] introduced association rule mining to diagnose tuberculosis. In this research, author introduced breath first search and a tree structure to calculate candidate item sets.

3. CONCLUSION

Tuberculosis is an infectious disease which attacks all persons. It is major cause of demise for most developing nations including India. In this research, frequently used data mining techniques are discussed. To identify tuberculosis, several variables age, gender, cough, fever, weight loss, chest pain night sweats are used in this research. From this we conclude SVM is best classifier which gave highest accuracy as compare to neural network, k-means clustering or any other technique.

REFERENCES

[1] Shakshi Garg and Navpreet Rupal,"A Review on Tuberculosis Using Data mining Approaches," In 2015 IJEDR Vol 3, Issue 3,pp. 1-4.

[2] Rusdah, Edi Winarko," Review on Data Mining Methods for Tuberculosis Diagnosis," In 2013 ISICO.

[3] Sofia Benbelkacem, Baghdad Atmani and Mohamed Benamina," Treatment Tuberculosis Retrieval using Decision Tree,"IN 2013 IEEE, pp. 283-288.

[4] Divya Tomar and Sonali Agarwal," A survey on Data Mining approaches for Healthcare,"In 2013, International Journal of Bio-Science and Bio-Technology Vol.5, No.5 , pp. 241-266.

[5] Amandeep Kaur Mann and Navneet Kaur," Survey Paper on Clustering Techniques,"In 2013 International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4,pp. 803-806.

[6] Navjot Kaur, Jaspreet Kaur Sahiwal, and Navneet Kaur," Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining,"In 2012 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3.pp.85-91.

[7] Papender Kumar, Deepak Dangwal And Nidhi Puri," Diagnosis Of Tuberculosis Using Association Rule

Method," In 2012 Journal of Information and Operations Management, Vol 3, Issue 1, pp. 133-135.

[8] Brian H. Tracey, Germán Comina, Sandra Larson, Marjory Bravard, José W. López, and Robert H.Gilman," Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis,"In 2011 IEEE EMBS,pp. 6018-6020.

[9] Asha.T, S. Natarajan, and K.N.B. Murthy," A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification."