DEVELOPING DIAGNOSE TUBERCULOSIS DISEASES MODELS BY USING DATA MINING: CLASSIFICATION AND CLUSTERING ALGORITHMS

[1]HOZIFA MOHAMED, [2]TARIG MOHAMED AHMED

[1]MSc, University of Khartoum, SUDAN
[2]Assoc. Prof., Department of MIS, Prince Sattam Bin Abdul-Aziz University, KSA
Assoc. Prof., Department of Computer Sciences, University of Khartoum, SUDAN
E-mail: [1] hozifa85@gmail.com, [2] Tarig_Harbi71@hotmail.com

ABSTRACT

Nowadays, Tuberculosis is considered as one of the largest cause of death from infectious diseases worldwide. There is increasing evidence that the genetic diversity of Mycobacterium tuberculosis bacteria may have important clinical consequences. So, the combination of clinical genetic data, social and demographic is crucial to understand the epidemiology of these infectious diseases. To help the doctors to predict and diagnosis the tuberculosis disease, this research proposed two models by using Data Mining. The two models could be used as decision support tool in clinics to help in diagnosing Tuberculosis. To develop the models, a tuberculosis dataset was collected from medical field from Tropical area teaching hospital in Khartoum state – Sudan, the size of this data set is 265 patient records. The first model was built by using classification algorithms. After conducting intensive experiments, three classification algorithms had been selected: Naïve Bayes, CN2 and Classification Tree. The model was implemented using Orange application. The model had been validated and according to the result, the classification Tree was selected as best algorthim with accuracy 0.9358, Sensitivity0.9632 and Specificity 0.8933. The second model was built by using clustering algorithm called K-means. Also, this model was implemented by using Orange application. The result of the clustering model had been discussed and evaluated. The two models could be used to cross-check the diagnosis and predict the Tuberculosis diseases

Keywords *Tuberculosis, Data Mining; Classification, Clustering.*

1. INTRODUCTION

Tuberculosis is considered as the second largest cause of death from infectious diseases worldwide [1]. Proportion to the emergence of multiple samples and wide Real-resistant TB threatens to make non-curable. There is increasing evidence that the genetic diversity of Mycobacterium tuberculosis bacteria may have important clinical consequences. Therefore, the combination of clinical and genetic data, social and demographic is critical to understand the epidemiology of these infectious diseases, and how virulence and other phenotypic traits evolve over time.[2] This requires dedicated bioinformatics platforms, and able to integrate and analyze all these heterogeneous data collected from more than one source. To control the evolution of this infectious disease, we need to identify risk factors for transmission. And so that we can achieve this, we need to clinical and social information and detailed demographic. From several sources and must be detailed this data it is necessary to identify the source of infection patient in order to prevent the activation of other injuries. On the other hand, in the communities where rare occurrence of transmission be the main objective is to identify people who have a hidden, as more cases of the disease are the result of latent infection and active thereafter [3].

Nowadays, the world where vast amount of data are collected daily, analyzing such data is an important need, Terabyte or Petabyte of data is pour into computer network. These massive data volumes are a result of computerization of our society and powerful data collection and storage units. So we used databases concepts to organized this data and handling and retrieved it. There is still problem when need to get useful and interesting knowledge or information from big data. [4]

Data mining is a process of knowledge discovery by obtaining useful pattern and knowledge from a large amount of data, the source of data can include flat file, databases, data warehouse or the web, and other information repositories, also data that are entered to system directly [5]. Also, it could be defined as a process of automatically fining a large of amount of data for pattern and recognition of specific characteristic [6].

In medical sector, data mining could be used to help manage health care, and can develop applications to extract knowledge to identify and track chronic conditions and high-risk patients. It could be also used to design appropriate interventions in a timely manner, and reduce the number of patients admitted to hospitals and claims urgent. For example, the developments of better diagnosis and treatment protocols, data network looks to re-admissions and reduce the use of resources and compares its data with the current scientific literature to specify the best treatment choices. Thus the using of evidence to support health care, group health cooperative matching patient by demographic characteristics and medical conditions to specify the groups to use most of the resources. It develops programs to help educate these people and prevent or manage their health group has been cooperative involved in data mining several efforts to give care and better health at lower cost. [7].

To help the doctors to predict and diagnosis the tuberculosis disease, this paper proposed a predictive model by using data mining. The model could be used as decision support tool in clinics. Data mining is very rich researched areas in computer science and information technology owing to the wide influence exhibited. To develop the model, a tuberculosis dataset was collected from medical field from Tropical area teaching hospital in Khartoum state – Sudan, the size of this data set is 265 patient records. After conducting intensive experiments, three classification algorithms had been selected: Naïve Bayes, CN2 and Classification Tree. The model was implemented using Orange application. The model had been validated and discussed.

The reset of the paper was organized as following: section 2 presented some related works conducted by researchers. The proposed model was explained in section 3 including the model components, experiments and the results with full discussion. This work was concluded in section 4 with recommends as future work.

2. RELATED WORK

According to statistics from the World Health Organization nearly a third of the world's population has been infected with M. Tuberculosis, and new infections accurate rate one per second. However, not all infections with M. Tuberculosis cause disease, tuberculosis, and many of the injuries area symptomatic. In 2007 there were an estimated 13.7 million chronic active cases, and in 2010 there were8.8million new cases and 1.45 million deaths, mostly in developing countries. 0.35 Million of these deaths occur in those co-infected with HIV. [8]

There are several reasons and motives of data mining in the health sector, for example, using data mining tools in health insurance companies and health units (hospitals – clinic – center's) in the attempt to reduce public spending on health. Also of the factors that led to use of data mining and the presence of large amounts of data generated by health information systems. The influencing factors in the use of data mining in health care is that data mining is always generates information of interest to very large in all areas of the health sector. It helps technicians' laboratories to analyze samples and helps pharmacists in determining the amount and the impact of the property. In addition it assists health services providers in determining treatment effective dosage and quantity.

Data mining applications takes significantly advantage in the health care industry; however, they are not without limitations. It can be limited to extract data from health care through access to the data, because the raw input data extraction are often found in the settings and different systems, such as management, clinics, laboratories and others. Thus, the data must be collected and integrated before it can be extracted data. While many writers suggested that researchers are building a data warehouse before attempting to extract the data, which can be a costly project and take a long time. On the positive side, it has been a data warehouse built successfully by health care from five different sources of data warehouse. It has been suggested distribution network topology rather than the data warehouse to extract data more efficiently, has documented a case study from Maccabi health care services by using existing databases to guide subsequent data mining [9,10].

Secondly, the problems that arise from the same data these include missing or damaged, inconsistent, or non-uniform data, such as a piece of information recorded in different forms in different data sources. In particular, the absence of a standard clinical vocabulary constitutes a serious obstacle to extract the data. The data are also problems in the field of health care are the result of the size, complexity and heterogeneity of medical data and poor characterization in mathematics and non - canonical form . Moreover, there may be issues of ethical, legal and social , such as data ownership and privacy issues , data-related health care. Quality of data mining results and applications based on the quality of the data all these problems limit the use of data mining model [11].

Thirdly, a sufficiently exhaustive mining of data will certainly yield patterns of some kind that are a product of random. This is especially true for large data sets with many variables. Hence, many interesting or significant patterns and relationships found in data mining may not be useful.

Finally, healthcare organizations developing data mining applications must make a substantial investment of resources, particularly time, effort, and money. Data mining projects can fail for a variety of reasons, such as lack of management support, unrealistic user expectations, poor project management, inadequate data mining expertise, and more. Data mining requires intensive planning and technological preparation work. In addition, physicians and executives have to be convinced of the usefulness of data mining and be willing to change work processes. Further, all parties involved in the data mining effort have to collaborate and cooperate [12].

.

There are some researches discuss how to use data mining in health industry for example we can use data mining in diagnosis many diseases and effectiveness of treatment, in hospitals to management, customer relation management and fraud and abuse in health insurance companies[13].

Data mining application can put criteria for the effectiveness of medical drugs, through comparing and contrasting the causes and symptoms of the disease of schedule, and the amount of treatments, and the data can be extracted by analyzing courses of action that have proven effective, For example, handle the results of patient groups with different drug for the same disease or condition can be compared to determine which treatments work best and most cost-effective.[14]

Data mining could be used to help manage health care, and can develop applications to extract data to identify and track chronic conditions and high-risk patients, and design appropriate interventions in a timely manner, and reduce the number of patients admitted to hospitals. [15].

3. MATERIAL AND METHOD

This section explained the process of constructing data mining healthcare model to diagnosis tuberculosis disease by using data mining techniques and algorithms.

3.1 Software package

The main software package used in this research is Orange data mining tool 2.7. It is open-source C++-based data mining software. This software can be free downloaded and installed from Bioinformatics home page http://Orange.biolab.si. Orange is a component-based data mining and machine learning software suite

3.2 Data Set

Dataset was collected from medical field from Tropical area teaching hospital in Khartoum state – Omdurman, the size of this data set is 265 records. Below table explain and describe data set that use in this research.

Table 1: Dataset Description

Attribute Name	Data Type	Meaning (Description)
Patient _ ID	Continues	Patient number
Age	Continues	-
Gender	Discrete	-
State	Discrete	State that patient come from
Locality/ Address	String	Locality or address of patient
Site	Discrete	Type of tuberculosis(pulmonary , Extra Pulmonary)
Sputum/smear	Discrete	Test of sputum
HIV	Discrete	Test of AIDs (HIV virus)

3.3 Instrumentation

To implement this research the following instruments were used:

- For this research using laptop pc, TOSHIBA Satellite C850. Processor : Intel Core i3 – 2.30 GHZ (4 CPUs) Memory 4096 MB RAM.OS Windows 8.1 Pro 32-bit
- Orange data mining tool (version 2.7)

3.4 Research Model

3.4.1 Model Implementation Using Classification

To build the research model, the classification algorisms were used. Classification algorithms are one of the most popular and common technique applied in data mining operations, it employee part of data set to develops model that can use to classify new data. Classification practically suited in fraud detection and critical risk application. Data classification process include two stages or phases, the first one is learning or construct the model that can classify the data, in this phase we can use the sub set of data called training data to build the model and we call this model by "classifier" and this classifier analyzed by classification algorithms. In the second phase, classification test data or model used to estimate Accuracy of classification rules, if the accuracy is considered acceptable, the classifier or this rule can be applied to the classification of new data tuples.

Many experiments were conducted by applying classification algorithms on the data set to finding and detects effeteness of tuberculosis and AIDS on patients, we have to class here “High” and “Normal” , to achieve this task can using three algorithms Naïve Bayes, CN2 and classification tree .

To analyze the database file in order to answer the first query, the following attributes were selected for processing:

- Gender attribute
- Site attribute
- State attribute
- Sputum and HIV attribute.
- Data pre-processing

In this step, the only pre-processing will be done for selecting attributed that participate in this module.

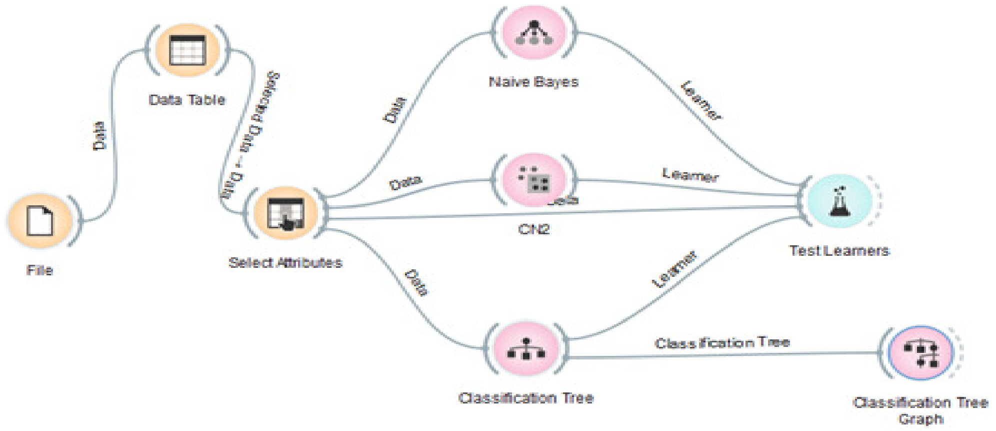

Figure 1: Model Implementation

As we show in figure 1 classification task it achieve by use three algorithms (Naïve bays, CN2 and classification tree) any algorithms icons connect to evaluate the classifier and that done by test learner this evaluate done based on three measurements CA classification accuracy, Sensitivity and Specificity.

3.4.1.1 Model Evaluation Criteria

Table 2: Confucian Matrix

Measure /algorithms	CA	Sensitivity	Specificity
Naïve bayes	0.9094	0.9684	0.7733
Classification	0.9358	0.9632	0.8933
CN2 rules	0.9283	0.9737	0.8267

- Classification Accuracy: The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.
- Sensitivity: is the proportion of positive tuples that are correctly identified.
- Specificity: is the proportion of negative tuples that are correctly identified.

From above result in table 2, we found the classification tree is the best one to achieve classification task because the accuracy is 0.9358 and Naïve Bayes 0.9094 and CN2 0.9283.

3.4.2 Model Implementation Using Clustering

The cluster simplicity organizes all objects into groups and each group contains similar objects as possible. Unlike classification in the clustering techniques, the class label of each tuple is unknown. These require discovering groups. The common definition of cluster is a process of grouping a set of data object into multiple groups or clusters so those objects with the cluster have high similarity but high dissimilarity to objects in other clusters. Clustering can applied successfully in many field and application such as business intelligence, image recognition, web search, biology and security.

By using K-means algorithm, the research model was implemented using Orange application figure 2 presented all steps.

In figure 2, the file was loaded the data for analysis and wired with the Selected Attributes control. The output of the process was supplied to the data table control for viewing and selecting. The selected data from the data table is passed to K-mean control to be clustered. The last step was connection the clustering control with the visualization control which was the responsibility of the Linear Projection control. In addition to K-mean output, Linear Project was required to be connected to the original data in order to work.

The clustering result is shown in figure 3, below. Reading the cluster layout in the figure 4 is easy. Colors applied to differentiate between the patients and it obvious that 10 clusters are there.

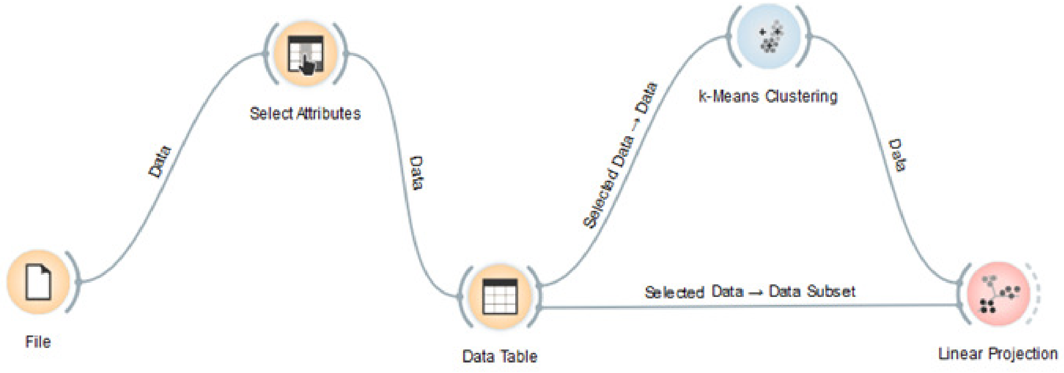

Figure 2: K-means clustering Schema

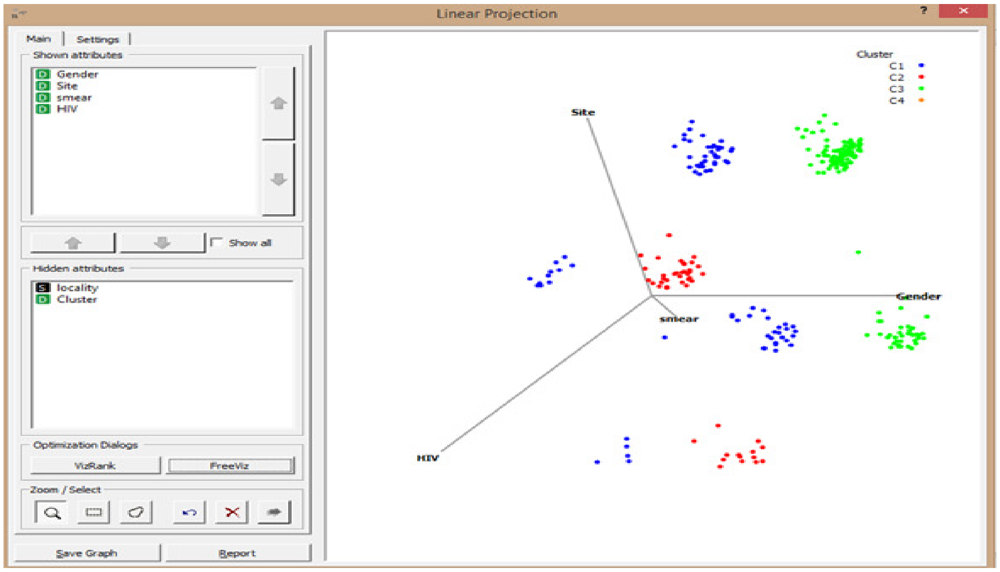

Figure 3: Linear Projections For Clustering Process

3.4.2.1 Evaluation of Knowledge

There three criteria that determine evaluating cluster algorithm output. It includes assessing clustering tendency which tells whether cluster points in space would lead to meaningful clustering. Measuring the cluster quality is about how much the cluster is good. There are several methods can just do that, including extrinsic and intrinsic. Scientifically calculating the expected number of clusters and compare those with the ones produced by the algorithm and it should match.

The first and second criteria are hard and time consuming to be calculated manually but the third criterion will be calculated to prove that the expected result matches the number of the algorithm output. The rule is as follows

$$m = \sqrt{\frac{n}{2}}$$

Where n is the total number of records

$$\sqrt{\frac{265}{2}} = 11.51$$

As we having 11 clusters in our graph matches the calculated number of cluster of the algorithm output then the comparison is true.

From this task we found three classes it appear the first class consisted of 160 record most of them was Male and Site P was 115 and EP was 44, class two consists 36 record and most of them is Female and Site EP was 13 and P was 23, the last one class three consisted of 69 record all of them if Female and Site P was 42 and Site EP was 27. According to these result, the model was accepted to identify different cluster of patients.

After validating the results of two models, the model could be used by medical centers to help the doctors in taking the right decision when selecting the best treatment plans for patients.

4. CONCLUSION

Tuberculosis is the second largest cause of death from infectious diseases worldwide. There is increasing evidence that the genetic diversity of Mycobacterium tuberculosis bacteria may have important clinical consequences. To help the doctors to predict and diagnosis the tuberculosis disease, this research proposed two models by using Data Mining. The two models could be used as decision support tool in clinics. To develop the models, a tuberculosis dataset was collected from medical field from Tropical area teaching hospital in Khartoum state – Sudan, the size of this data set is 265 patient records. The first model was built by using classification algorithms. After conducting intensive experiments, three classification algorithms had been selected: Naïve Bayes, CN2 and Classification Tree. The model was implemented using Orange application. The model had been validated and according to the result, the classification Tree was selected as best one with accuracy 0.9358, Sensitivity0.9632 and Specificity 0.8933. The second model was built by using clustering algorithm called K-means. Also, this model was implemented by using Orange application. The result of the clustering model had been discussed and evaluated. The two models could be used to cross-check the diagnosis and predict the Tuberculosis diseases.

As limitation of this study, the dataset had been selected from only one medical center. So, future work, we propose to develop data warehouse that can store massive data from several data centers to increase the quality of the models. Also, new algorithms may obtain new knowledge that can help to determine good treatment plans.

References

[1] Dye, Christopher, and Brian G. Williams. "The population dynamics and control of tuberculosis." Science 328.5980 (2010): 856-861.

[2] American Thoracic Society. "Diagnostic standards and classification of tuberculosis in adults and children." Am J Respir Crit Care Med 161 (2000): 1376-1395.

[3] [http://www.who.int/tb/publications/global_report/en/index.html last access 1/9/2013

[4] Zarate Santovena, Alejandro. Big data: evolution, components, challenges and opportunities. Diss. Massachusetts Institute of Technology, 2013.

[5] Willett, Walter. Eat, drink, and be healthy: the care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, July 2013, Pages 127-136

[6] Mehta Neel B, Predictive Data mining and discovering hidden values of Data warehouse, ARPN Journal of Systems and Software Volume 1 No. 1, APRIL 2011.

[7] Kincade, K. (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2),IM2-IM7.

[8] Joshi, Rajnish, et al. "Tuberculosis among health-care workers in low-and middle-income countries: a systematic review." PLoS Med 3.12 (2006): e494.

[9] Oakley, S. (1999). Data mining, distributed networks and the laboratory. Health Management Technology, 20(5), 26-31.

[10] Friedman, N.L. &Pliskin, N. (2002). Demonstrating value-added utilization of existing databases for organizational decision-support. Information Resources Management Journal, 15(4), 1-15.

[11] HianChye& Gerald Tan, Data Mining Applications in Health care , Journal of Healthcare Information Management — Vol. 19, No. 2

[12] Cios, K.J. & Moore, G.W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(1), 1-24.

[13] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." Journal of healthcare information management 19.2 (2011): 65.

[14] Gravenhorst, Franz, et al. "Mobile phones as medical devices in mental disorder treatment: an overview." Personal and Ubiquitous Computing 19.2 (2015): 335-353.

[15] Kincade, K. (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2),IM2-IM7.