Classification and prediction of Pulmonary Tuberculosis by Evaluating its Influences Using Data Mining Techniques

Faisal Arafat
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
faisalarafat016@gmail.com

Faisal Kabir
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
faisalkabir1993@gmail.com

***Abstract*— Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining has become a main strategy in many industries to improve outputs and decrease costs. Now days in healthcare management this field will become very useful. Data mining techniques has became great potential for the healthcare industry to predict health diseases by using systematic data. In this paper we are applying the various classification techniques over tuberculosis disease dataset for the prediction of disease and non disease person. The tuberculosis database is preprocessed to make the mining process more efficient. We are analyzing the various classification techniques like, Naive Bayes, ZeroR, J48 and showing their accuracy.**

***Keywords*— Tuberculosis, Pulmonary Tuberculosis data, Naive Bayes, ZeroR Disease prediction and Decision Table.**

I. Introduction

All Tuberculosis (TB) was believed to be almost under control but it has once again become a serious world-wide problem. Tuberculosis disease is caused by a bacterium which is called as mycobacterium tuberculosis. This disease can spread among humans and the patients who suffer from tuberculosis might die unless they get the right treatment. This microorganism widely exists on humans, cattle, sheep and birds. All of the organs in the body can be affected by tuberculosis. But most of the tuberculosis cases are occur in lungs. Nowadays disease prediction plays an important role in data mining. There are different types of disease predicting in data mining nearly diabetic heart disease, breast cancer, lung cancer, brain tumor etc. This paper analyzes the bacteriologically confirmed PTB disease prediction using classification algorithms. We collected the dataset from National Institute of Diseases of the Chest and Hospital. Introducing data mining into medical analysis are to increase diagnostic accuracy to reduce costs and to save human resources. In this context various classification techniques are used to predict the onset of bacteriologically confirmed PTB in Pulmonary TB dataset and showed that which technique is better approach by using their accuracy. In this paper we obtained more than 96 percent accuracy. In this paper, various techniques are implemented for the forecasting of bacteriologically confirmed PTB and concluded with best forecasting techniques which has a maximum accuracy. Implemented techniques are listed below:

1. Naive Bayes
2. Decision Table
3. ZeroR
4. J48

II. Background Study

There have been many studies regarding the Tuberculosis disease. But we mainly focusing on Pulmonary Tuberculosis also known as PTB. We use several data mining classification algorithms like Naïve Bayes, J48, ZeroR and Decision Table to classify the dataset and find out the most accurate predictive classification algorithm based on accuracy of these classifier algorithms. Before creating the set of question we visit some of the Doctor's for generating an effective and relevant question sets. After completing feasibility study about TB then we make the question sets and collect the data from patient using the form. We face some difficulties during collecting the data from the hospital as well.

III. The Dataset

The dataset which we used for our work is Pulmonary TB Bangladeshi Dataset. This dataset is collected from National Institute of Disease of the Chest and Hospital. We create set of question for collecting the data.

Number of Instances: 100

Number of Attributes: 17

1. Age
2. Gender
3. Smoking
4. Coughing with mucus
5. Haemoptysis (Coughing with blood)
6. Evening rise of temperature
7. Chest pain
8. Shortness of breath
9. Excessive sweating, especially at night

10. Weight loss
11. Wheezing
12. Diabetes
13. Previously diagnosed as TB
14. Are you living around people who have TB
15. Live in crowed or unclean living condition
16. Do you have any drug addiction
17. Bacteriologically positive PTB (Class variable)

Class Distribution: (Class value 'Yes' is interpreted as "Bacteriologically positive for PTB" and class value 'No' is interpreted as "Bacteriologically negative for PTB")

TABLE I. CLASS DISTRIBUTION

Class Value	Number of Instances
No	62
Yes	38

IV. MATERIALS AND METHODS

A. Related Work

1. Jayalalsmi et al. designed a system that is applied to PIMA Dataset for classification aim. The systems made us of Artificial Neural Network for classification.
2. Patilet et al. produced association rules for PIMA Dataset.

B. Weka

Weka is a powerful platform which is used in the purpose of Data Mining related work. We use this platform for classify the dataset.

V. DATA PREPROCESSING

Data preprocessing is the first step in data mining. In data preprocessing the misclassified data is removed. In this process Cleaning and filtering of the data is carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns. In preprocessing first it selects an attribute for selecting a subset of attributes with good predicting capability. If an attribute has more than 5% missing values then the records should not be detected and it is advisable to impute values where data is missing, using a suitable method.

VI. TRAINING AND TEST DATA SELECTION

In this paper, we used percentile split for dividing dataset into training and testing data. It is an inherent part of machine learning. We partition the data into training set and test set. The training set will be used to train the model parameters. Then the trained model is used to make prediction on test set. Predicted values will be compared with actual data to compute the confusion matrix. In this paper we use 66% of the data set is for training the model and rest of them is test data.

VII. NAÏVE BAYES

Naïve Bayes classifier is mainly suitable when the dimensionality of the inputs is high. Due to its simplicity, Bayes can offer outclass more refined classification methods. This model recognizes the characteristics of patients with heart decease. It is the foundation for many machine-learning and data mining methods. Naive Bayes algorithm considers each of the feature to contribute independently to the probability that the person has a heart disease. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature. This algorithm is used to create models with predictive capabilities.

Confusion matrix and Accuracy:

TP	FP	FN	TN
77.5862	22.4138	32.2751	67.7249

Accuracy:- 100 * (77.8562 +67.7249) / (77.5862 +67.7249+22.4138 + 32.2751) = 64.1155

VIII. DECISION TABLE

Decision tables are a concise visual representation for specifying which actions to perform depending on given conditions. They are algorithms whose output is a set of actions. The information expressed in decision tables could also be represented as decision trees or in a programming language as a series of if-then-else and switch-case statements. Aside from the basic four quadrant structure, decision tables vary widely in the way the condition alternatives and action entries are represented.[8][9] Some decision tables use simple true/false values to represent the alternatives to a condition (similar to if-then-else), other tables may use numbered alternatives (similar to switch-case), and some tables even use fuzzy logic or probabilistic representations for condition alternatives.[10] In a similar way, action entries can simply represent whether an action is to be performed (check the actions to perform), or in more advanced decision tables, the sequencing of actions to perform (number the actions to perform). A decision table is considered balanced.[11] or complete[10] if it includes every possible combination of input variables. In other words, balanced decision tables prescribe an action in every situation where the input variables are provided.[11]

Confusion matrix and Accuracy:

TP	FP	FN	TN
87.0690	22.9310	44.9153	56.0847

Accuracy: 100 * (87.0690 +56.0847) / (87.0690 +56.0847+22.9310 + 44.9153) = 87.1639

IX. ZEROR

Zero Rule or ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for

other classification methods. There is nothing to be said about the predictors contribution to the model because ZeroR does not use any of them. The following confusion matrix shows that ZeroR only predicts the majority class correctly. As mentioned before, ZeroR is only useful for determining a baseline performance for other classification methods.

Confusion Matrix		Pulmonary-TB			
		Yes	No		
ZeroR	Yes	62	37	Positive Predictive Value	62.6263
	No	0	0	Negative Predictive Value	0.00
		Sensitivity	Specificity	Accuracy = 62.6263	

X. J48

J48 is a type of C4.5 decision tree deployed for classification purposes. Please refer to the paper Samer shared for understanding what the C4.5 decision tree is. C4.5 (J48) is an algorithm used to generate a decision tree developed by Ross Quinlan mentioned earlier. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set of already classified samples. Each sample consists of a p-dimensional vector , where the represent attribute values or features of the sample, as well as the class in which falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists. This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Confusion matrix and Accuracy:

TP	FP	FN	TN
84.0690	21.9310	43.9153	56.0847

Accuracy: 100 * (84.0690 +56.0847) / (84.0690 +56.0847+21.9310 + 43.9153) = 82.1641

X. PERFORMANCE AND EVALUATION

Performance of model can be evaluated various performance measures: classification accuracy, sensitivity and specificity. These measures are evaluated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Actual Vs. Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Where TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative These measures showed that The Decision Table can give accuracy of 87.1639%, Which is the best accuracy. Mostly the accuracy result of the Decision Table classifier is placed between 85% to 89% depending upon the number of cross validation applied on the dataset when performing the test.

CONCLUSION AND FUTURE WORKS

Decision Table is more productive than other classifiers. Thus this article introduces a successful Pulmonary Tuberculosis Diagnosing technique which helps to predict the disease. We began with observing the symptoms as it are very difficult to predict Pulmonary Tuberculosis decease finding symptoms. In the second step we preprocess the Pulmonary Tuberculosis database to make the mining process more efficient. Finally, the results are compared with the help of different prediction classifiers Naive Bayes, Decision Table, ZeroR, J48. The results were compared using different performance measures. These measure used true positive (TP), true negative (TN), false positive (FP) and false negative (FN) to calculate results. Performances of our technique were measured by Accuracy: 87.1639%. The proposed approach has demonstrated that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict. Besides these information analysis results can be utilized for further research as a part of upgrading the accuracy of the prediction system in future. Our future goal is to find out the real accuracy using neural network.

REFERENCES

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example: [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in “[3]”—do not use “Ref. [3]” or “reference [3]”. Do not use reference citations as nouns of a sentence (e.g., not: “as the writer explains in [1]”).

Unless there are six authors or more give all authors’ names and do not use “et al.”. Papers that have not been published,

even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Rogers, William T. *"Decision Table Examples: Medical Insurance"*. Saint Xavier University Systems Analysis and Design. Archived from *the original* on March 29, 2007.

[9] "Archived copy". Archived from the original on 2012-05-30. Retrieved 2010-07-07.

[10] Wets, Geert; Witlox, Frank; Timmermans, Harry; Vanthienen, Jan (1996). *"Locational choice modelling using fuzzy decision tables"*. New frontiers in fuzzy logic and computing: 1996 biennial conference of the North American Fuzzy Information Processing Society – NAFIPS. Biennial Conference of the North American Fuzzy Information Processing Society. Berkeley, CA: IEEE. pp. 80–84. *doi*:*10.1109/NAFIPS.1996.534708*. *ISBN 0-7803-3225-3*.

[11] Udo W. Pooch, "Translation of Decision Tables," ACM Computing Surveys, Volume 6, Issue 2 (June 1974) Pages: 125–151 ISSN 0360-0300