Volume 5, Issue 11, November 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Evaluation and Performance of Classification Methods for Medical Data Sets

Vajenti Mala*, D. K. Lobiyal
SC&SS, Jawaharlal Nehru University,
New Delhi, India

***Abstract*— In this paper, we have evaluated the performance of J48, BayesNet, NaiveBayesUpdatable, MultilayerPerceptron classifier for the medical data sets from the UCI repository. These medical data from the repository include Diabetes, Breast Cancer, and Lung Cancer data sets. Experiments have been performed by using "WEKA" (Waikato Environment for Knowledge Analysis) tool. The result of the experiment shows that Naïve Bayes Classifier gives the most efficient results. Further, results analysis also shows that BayesNet algorithm performs the best in term of mean absolute error since it give the least classification error in comparison to other algorithms considered in this study.**

***Keywords*— Data Mining, Data Mining Techniques, Classification, Medical Data Sets**

I. INTRODUCTION

Medical data mining is an active interdisciplinary area of research that applies artificial intelligence and data mining techniques to health care data and patient records. The data generated by health department its vast and complex and makes the analysis difficult to make important decision regarding patient health. This data contains the details of hospitals patients, medical claims treatment cost *etc*. Therefore, it is very important to explore a data mining tools that can analyse and extract meaningful information from the data. In this work is, we have used different classification techniques by exploiting WEKA for the data Diabetes, Breast cancer and Lung Cancer data sets.

The analysis of the health data by applying data mining techniques can be used to improve the performance of patient's data management. The classification technique selected in ths study can also be used to identify disease by analysing various factors that are responsible for diseases.

The main problem in medical domain is the lack of analysis of correct and important information in medical science. For the diagnosis, most of the tests normally involve in techniques of clustering and classification for the huge amount of datasets. However, on the other side other side, a complicated test for the main diagnosis makes it difficult to obtain the final results. Therefore, such problem can be dealt with by using machine learning algorithms. These algorithms could be used to directly find the final result by exploiting various data mining techniques of classification.

II. RELATED WORK

In the literature, use of data mining techniques has been reported to analyse the medical data for diagnosis of various diseases. In this section, some of the work, pertaining to classification techniques used in medical domain has been summarized.

In [20], Gupta et al, have worked on the summarization of various articles on the breast cancer diagnosis. They presented and carried out some enhanced techniques for breast cancer diagnosis.

In [21], Othman et al, evaluated performance of different classification methods applied on classifiers such as naïve Bayes, K- Nearest Neighbour, Decision Tree to determine the most important classification methods.

In [22] Phyu has worked out on the basic classification i.e Bayesian network decision tree induction K- Nearest neighbour classifier and fuzzy logic techniques.

In [23] Sokolova, evaluated the performance of a classifying model. They have performed evaluation based on the accuracy of the models.

III. CLASSIFICATION METHODS

There are various classifiers that play significant role in the evaluation of the data sets. In this section, some of classifiers used in our study are discussed.

A. *Bayes Network Based Classification Method*

Bayesian network or naïve Bayes is a simple and powerful probabilistic network used for classification. In this network, all attributes are independent according to the given value of the class variable. The network is based on statistical properties which determine the textual document. In other words, we may say that the term is assigned by a probability which belongs to particular category.

The Bayesain classifiers can be calculated by the probabilities of every class C_k given by a document D_j as:

$$P\left(C_k \mid D_j\right) = \frac{P(C_k)P(\overrightarrow{D_j} \mid C_k)}{P\left(\overrightarrow{D_j}\right)} ......(1)$$

Where,

$P\left(\overrightarrow{D_j}\right)$ is the probability with randomly picked document has a vector form $\overrightarrow{D_j}$.

$P(C_k)$ is also a probability with randomly picked document which belongs to $C_{k.}$

To determined $P\left(D_j \mid C_k\right)$, Naïve network find the probability of a word or term which is independent to the other term that is appear in the same document. By using this simplification, it is necessary to evaluate $P\left(\overrightarrow{D_j} \mid C_k\right)$ is the product of the probabilities of each and every term which is appear in the document, therefore $P\left(\overrightarrow{D_j} \mid C_k\right)$ may calculated as:

$$P\left(\overrightarrow{D_j} \mid C_k\right) = \prod_{i=1}^{|t|} P\left(W_{ij} \mid C_k\right)......(2)$$

Where, $\overrightarrow{D_j} = \left(W_{1j,...,}W_{|t|j}\right)$.

A.1 NaiveBayesUpdatable:-

NaiveBayesUpdatable is an improved version of Naive Bayes. This classifier gives higher precision. In Naïve bayes classifier, where a class has no parents, each attribute has the class as its sole parent. When we build the classifier with zero training instances of 0.1 for numeric data attributes, it also called as incremental update.

A.2 BayesNet:-

This classifier is based on the Bayes theorem and conditional probability. Bayesian network gets formed; each node is obtained first time. BayesNet is an acyclic graph. The classifier's attributes are nominal however, there is no missing value. These values can be replaced globally. In this classifier, the output can be found in terms of graph.

B. Decision Tree Based Method

A Decision Tree Based technique is a hierarchical structure, which contains nodes and the directed edges and organizes a series of questions regarding the (attributes) predictors. This is the way to make possible answer in an effective way.

B.1 J48:-

J48 Classifier is also known as (C4.5 DT). This classifier is used for optimized experimentation. The J48 classifier is a decision tree which has the same tree structure having different nodes, such as root, leaf node and intermediate nodes.

C. Neural Network Based Method

An ANN comprises of interconnected collection of nodes and the directed links. This technique is widely used. The simplest of all ANNs is the Perceptron that consists of two kinds of neurons or units. They are input nodes which represent input attributes, and the output nodes. Each input node is connected to the output node by the weighted link.

C.1Multilayerperceptron:

This classifier is more complex structure than Perceptron model which is the multilayer. The classifier is defined as a Neural Network and an Artificial Intelligence. A MultilayerPerceptron is a neural network which has one or more layers. The classifier contains three layers: input layer, hidden layer and output layer. The hidden layer consists of more than one layer and each and every neuron (node) in a layer is interconnected to every neuron in an adjacent layers. The input layer is connected with the training and testing vectors which may further be processed by the hidden and output layers. The details are in given [15, 16].

IV. EXPERIMENTAL DESIGN

A. Data Sets

In this paper, we used data set in the form of arff file [14], which is WEKA extension. We have three different medical data sets i.e. Diabetes, Breast Cancer, Lung Cancer. For an experiment, the data is divided into two parts, training and testing data. These data are used to build and validate the classifiers. In the Diabetes we have 768 instances and 9 attributes, in Breast Cancer we have 286 instance and 10 attributes and in lung cancer we have 32 instance and 57 attributes. We have applied 10 cross validation experiments to process on all data sets. At the last, the results are recorded in terms of accuracy and mean absolute error rates.

A.1 Algorithm Implementation:-

Classifying the documents we may pre-processed the data by the various techniques:

See in algorithm

1. Stop word Removal
2. TF-IDF
3. Case Folding
4. Normalization

After then we applied classification methods as shown in algorithms to classify the whole documents into the training set. Furthermore we applied classifier and calculate accuracy by comparing different classifiers according to the mentioned algorithm. We may determine the accuracy which defined as the correct classified documents is given below:

$$\text{Accuracy} = \frac{\text{Correct Classified Documents}}{\text{Total Documents}}$$

Table I. Algorithm for classification of data sets

```
Algorithm:  Input: - Unstructured Medical data set
                   Output: - Accuracy and Mean of different classifiers
Method
Begin:   Step-1:- Take the input of unstructured Medical data sets
                   (Diabetics, Breast Cancer, Lung Cancer)
         1.1 Convert unstructured data into structured form for applying classification
             techniques
         Step-2:- Pre-processing the whole data sets. Such as stop-word removal, tokenization,
                  stemming etc.
         Step-3: - Apply the particular model of classification on all medical datasets.
                  Such as KNN, NB, NN, DT, SVM etc.
         Step-4: - Repeat step 3 for each model you have applied on the whole data sets.
                   4.1. Separate particular algorithms suitable for the data sets.
         Step-5:- Choose the best model for further experiment
         Step-6:- Evaluate and Compare the results of accuracy rates as well minimum
                  absolute error of data sets with different classification techniques.
End
```

Table II Classification methods with executed time on the data sets

Classifiers/ Data Set	J48		**Byes Net**		Naïve byes Updatable		Multilayers Perceptron	
	Accuracy	Time Taken	**Accuracy**	Time Taken	Accuracy	Time Taken	Accuracy	Time Taken
Diabetes	73.8281	0.03sec	**74.349**	0.03sec	76.3021	0 sec	75.3906	0.65sec
Breast Cancer	75.5245	0 sec	**72.028**	0.2 sec	71.6783	0.01 sec	64.6853	2.98 sec
Lung Cancer	50	0.04sec	**53.125**	0.01 sec	50	0 sec	37.5	6.17 sec

TABLE III COMPARISON OF ACCURACY OF CLASSIFIERS FOR DATA SETS

Classifiers used	Diabetes	Breast Cancer	Lung cancer
J48	73.8281	75.5245	50
BayesNet	**74.349**	**72.028**	**53.125**
NaivbayesUpdatable	76.3021	71.6783	50
MultilayerPerceptron	75.3906	64.6853	37.5

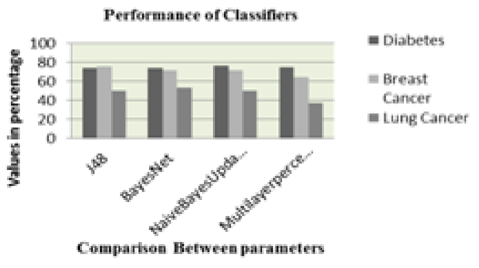

Fig: 1 Performance of different classifiers on data sets

Table IV Comparison of accuracy of classifiers for data sets

Classification Method	**Mean Absolute Error**		
	Diabetes	Breast Cancer	Lung Cancer
J48	0.3158	0.3676	0.3461
BayesNet	**0.2987**	**0.3297**	**0.3297**
NaiveBayesUpdatable	0.2841	0.3272	0.317
Multilayerperceptron	0.2955	0.3552	0.3918

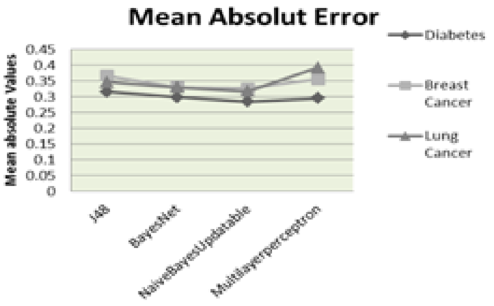

Fig: 2. Less Mean Absolute Error of the classifier

V. ANALYSIS

In the WEKA, all the data could be considered being as instances and features in the data called as attributes. The experimental results can be discussed into various sub items for easier analysis and performance evaluation. We can see an accuracy of the classifier and compared with the methods of Bayes network, Decision tree and neural network based classifier on data set. According to the Table 1 we processed an algorithm to experiment and found that the Naive Bayes classifier shows the best result with accuracy as well as less absolute mean error of Naïve Bayes classifier as shown in Table 2 and Table.3.

VI. CONCLUSION

In the paper, we have experimented and compared the performance of classification methods by using the different classifiers on the medical Data sets from the UCI repository. We found that the performance of various classification techniques in terms of mean absolute error depends upon the data sets. There are different factor that have affected the performance of the classifier on the data sets. These factors are the number of attributes, type of attributes, and the configuration. Further, we observed that Naive Bayes classifier is the best it has the highest accuracy among all the classifier and it has minimum classification errors. However this evaluation may not be same for all the datasets. Therefore, a normal Classifier that should be an adaptable to the various types of datasets may be designed. In the future research work, we will focus on an improvement of classification techniques in terms of the efficiency. Further, we may explore combining of classification techniques to improve the performance.

REFERENCES

[1] Ian H. Witten, Eibe Frank and Mark A. Hall. “Data Mining: Practical Machine Learning Tools and Techniques”, Elsevier. ISBN 978-0- 12-374856-0, 3rdEdition, pp. 313-317, 2000.

[2] Yanwei Xing, Jie Wang and Zhihong Zhao “Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease”. *IEEE*. p1-5.

[3] Graduate Institute of Applied Information Sciences **(2009)**. MEDICAL DATA MINING USING BGA AND RGA FOR WEIGHTING OF FEATURES IN FUZZY K-NN CLASSIFICATION. *IEEE*. p1-6.

[4] K. Srinivas, B. Kavihta Rani and Dr. A.Govrdhan, “Applications of Data Mining Techniques in Healthcae and Prediction of Heart Attacks”, International Journal on Computer Science and Engineering, vol. 02, no. 02, **(2010)**, pp. 250-255.

[5] S. H. Ha and S. H. Joo, “A Hybrid Data Mining Method for the Medical Classification of Chest Pain”, International Journal of Computer and Information Engineering, vol. 4, no. 1, **(2010)**, pp. 33-38.

[6] M. Ilayaraja., “Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm”. *IEEE*. **(2013).**

[7] Fayyad, Usama, Gregory Piatetsky-Shapiro and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases",http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf,1996.

[8] Medline Resources http://www.nlm.nih.gov/bsd/pmresources.html.

[9] V.Elango, R.Subramanian and V.Vasudevan, “A Five Step Procedure for Outlier Analysis in Data Mining”,European Journal of Scientific Research, ISSN 1450-216, Vol.75, Issue No.3, pp. 327-339, 2012.

[10] Jau-Huei Lin, M.D. and Peter J. Haug, M.D. “Data Preparation Framework for Pre-processing Clinical Data in Data Mining”, AMIA Annu Symp Proc. 2006; 2006: 489493.

[11] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, Morgan Kauffmann Publishers, 2001

[12] TN. Phyu, “Survey of classification techniques in data mining”. Proc. Int. Eng. Comp. Sci. pp.18-20, **(2009)**

[13] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar;“Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification”, Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA. 1999

[14] https://archive.ics.uci.edu/ml/datasets.html

[15] Zak S.H., (2003), “Systems and Control” NY: Oxford Unininversity Press.

[16] [16] Hassoun M.H, (1999), “Fundamentals of Artificial Neural Networks”, Cambridge, MA: MIT press.

[17] Han, J., Kamber, M., and Pei, J., “Data Mining: Concepts and Techniques”, 3rd edition, Morgan Kaufmann, (2011); (1st ed., 2000-2001) (2nd ed., 2006)

[18] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

[19] F. Firouzi, M. Rashidi, S. Hashemi, M. Kanqavari, A. Bahari, NE. Daryani, MM. Emam, N. Naderi, HM. Shalmani, A. Farnood, M. Zali, “A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software”, Eur. J. Gastroenterol. Hepatol, vol.12, pp.1075-1081,2007. [Pubmed:17998832]

[20] Gupta, S., Kumar, D., and Sharma, A., “Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis”, Indian Journal of Computer Science and Engineering (IJCSE), Vol 2, pp, 188-195, 2011.

[21] Othman & Yau, “Comparison of Different Classification Techniques Using WEKA for Breast Cancer,” *IFMBE Proceedings 15*, pp. 520-523, 2007.

[22] Phyu, Thair Nu, “Survey of Classification Techniques in Data Mining,” *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009*, Hong Kong.

[23] Sokolova, M. and Lapalme, G.2009. A systematic Analysis of performance Measures for classification Task. *Information processing & management.* 427-437.