A DETAILED ANALYSIS OF DIFFERENT DATA MINING ALGORITHMS WITH HYPOTHYROID DATA SET

[1]Dr.B.Radha, [2]Mrs. A.Pavithra,

[1]Assistant Professor , [2]Assistant Professor,
[1]Department of Information Technology,
[1]Sri Krishna Arts and Science College, Coimbatore, India

ABSTRACT: ***Data mining is the procedure of mining knowledge from the massive amount of data. The data can be deposited in databases and information warehouses. Data mining work can be separated into two models descriptive and predictive model. In the Predictive model, we can expect the values from a different set of sample data, they are ordered into three types such as classification, regression and time series. The descriptive model permits us to control patterns in a sample data and sub-divided into clustering, summarization and association rules. Data mining is a familiar practice used by health organizations for classification of diseases such as dengue, diabetes, hypothyroid and cancer in bio informatics research. In the proposed approach we have used WEKA with 10 cross validation to evaluate data and compare results.***

In this paper we have primarily ordered the hypothyroid data set and then related the different data mining techniques in weka through Explorer and Experimenter interfaces. Also we can know which attribute has its significance with their ranking values using select attributes and info gain function.

Keywords: Data Mining, Association Rule Mining, Spatial Data Mining, RDBMS, Medical Database, Large Database, Distributed Database.

---

1. INTRODUCTION

Hypothyroidism, also called under active thyroid or low thyroid, is a communal illness of the endocrine system in which the thyroid gland does not yield enough thyroid hormone.

It can cause a numeral of indications, such as poor ability to bear cold, a feeling of drowsiness, constipation, despair, and weight increase. Occasionally there may be bulge of the front part of the neck due to goiter. Untreated hypothyroidism for the period of pregnancy can tip to delays in growth in the baby, which is called cretinism. Here the hypothyroid dataset is ised in arff format.

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. This document descibes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the data mining book written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date attributes, and sparse instances).

Information Gain

Information gain measures the expected reduction in entropy, or uncertainty.

$$Gain(S, A) = Entropy(S) - \sum_{veValues(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

Values(A) is the set of all possible values for attribute A, and Sv the subset of S for which attribute A has value v Sv = {s in S | A(s) = v} the first term in the equation for Gain is just the entropy of the original collection S,the second term is the expected value of the entropy after S is partitioned using attribute A.

Methods

In order to carry out investigations and executions Weka was used as the data mining tool. Weka (Waikato Environment for Knowledge Analysis) is a data mining tool written in java established at Waikato. WEKA is a very worthy data mining tool for the users to categorize the accuracy on the basis of data sets by relating different algorithmic methods and equated in the field of bio informatics. Explorer and Experimenter are the crossing point existing in WEKA that has been used by us. Her in this paper we have used these data mining methods to analyze the hypothyroid disease over classification of different algorithms accuracy. Fig1visualizes the interface of WEKA Data mining tool.

Fig1: Interface of WEKA Data mining tool

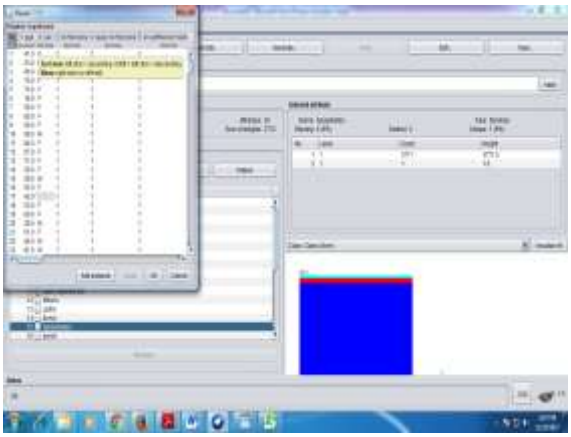

Fig2: Screen shot view of hypothyroid Dataset

2. CLASSIFICATION

In data mining tools classification compacts with classifying the problem by detecting features of diseases among patients and spot or predict which algorithm shows best performance on the basis of WEKA's statistical output. Table 1 shows the WEKA data mining techniques that have been used in this paper beside with other requirements like data set format etc. by using different algorithms.

Software	Datasets	Weka Data Mining Technique	Classification Algorithms	Operating System	Dataset File Format	Purpose
WEKA	Hypothyroid	Explorer Experimenter	Naïve Bayes J48 SMO REP Random	Windows 7	CSV	Classification
			Info Gain			Select Attributes

Table 1.Weka data mining technique by using different algorithms

Three techniques have been implemented in this paper, the first technique usages explorer interface and rest algorithms like Naïve Bayes, SMO, J48, REP Tree and RANDOM Tree, used in zones to characterize, apply and learn the statistical knowledge and major results have been achieved.

The second technique practices Experimenter interface. This preparation lets one to design experiments for running algorithms such as Naïve Bayes, J48, REP Tree and RANDOM Tree on datasets. These procedures can be run on experimenter and test the results. It organizes the test choice to use cross validation 10 folds. This interface offers ability for running all the algorithms together and thus a comparative result was obtained.

The third technique uses Selecting attributes. In this study we know the significance of different attributes on data sets and compared the results to know which attributes shows best performance. In that Infogain function gives all the details of the attributes in their ranking order.

The algorithms used by us were practical to a hypothyroid data set described in fact. In order to get improved correctness 10 fold cross validation was done. For every classification we selected training and testing sample casually from the base set to sequence the model and then test it in order to guess the classification and accuracy degree for each classifier. \

3. DATA MINING TECHNIQUES

The data mining technique which have been used us by weka data mining tool for classification and accuracy by applying different algorithms approaches. The interfaces of weka used in this paper are the following:

3.1. Explorer Interface

It first preprocesses the data and then cleans the data. Users can then contents the data file in CSV (Comma Separated Value) format and then examine the classification exactness result by selecting the succeeding algorithms using 10 cross validation: Naïve Bayes, J48, SMO, REP Tree, and Random Tree.

3.1.1. Naïve Bayes

Naïve Bayes is one of the algorithms that workings as a probabilistic classifier of all qualities contained in data sample separately and then categorizes data problems. Running the algorithms by Naïve Bayes we examine the classifier output with several statistics founded output by using 10 cross validation to brand a prediction of each instance of the dataset. Afterward running these algorithms we attained a classification accuracy of 95% for 3594 correctly classified instances, error rates achieved i.e. Mean Absolute Error is 0.0357, time taken for building model is 0.03 seconds and ROC area is 0.929 these outputs are got later the algorithms are run. The output obtained by scoring of Naïve Bayes algorithm accuracy of is given in Table 3 on the basis of time, accuracy, error and ROC.

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
Naïve Bayes	0.03	95.281	4.719	0.0357	0.929

Table 2: Naïve Bayes algorithm accuracy

3.1.2. J48 Tree

We have also used J48 Tree on our hypothyroid disease dataset. Later running this algorithm we studied the outputs attained from the classifier, the output contributed numerous statistics based on 10 cross validation to mark a prediction of each instances of dataset. Figure 8 shows the classification accuracy achieved from this algorithm i.e. 99.57% is the correctly classified accuracy for a batch of 3756 instances, mean absolute error obtained is 0.003, time taken to build this model is 0.13 seconds, and ROC area is 0.993.

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
J48	0.13	99.57	0.4242	0.003	0.993

Table 3. J48 algorithm accuracy

3.1.3. SMO

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. SMO is one of the methods used for classification. In this paper we have used this algorithm to divide the data on the base of dataset. Running this algorithm we evaluated the classifier output with diverse statistics founded on output by using 10 cross validation.

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
SMO	3.88	93.6108	6.3892	0.256	0.594

Table 4. SMO algorithm accuracy

3.1.4. REP Tree

REP Tree has been used in this paper to form a decision and decreases errors by arranged values of numeric attribute and splits the instances into bits to classify the accuracy. Running the algorithm we examine the classifier output with statistics based outputs by using 10 cross validation. In figure 10 classification accuracy achieved shows that 99.5758 % are correctly classified accuracy for 3756 instances, 0.4242% incorrectly classified accuracy for 16 instances, error rates that is mean absolute error is 0.0041, time taken to build model is 0.06 seconds and ROC area is 0.993 these are mentioned in output.

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
REP Tree	0.06	99.5758	0.4242	0.0041	0.993

Table 5 REP Tree algorithm accuracy

3.1.5. Random Tree

Random Tree has been used in this paper for accidentally selecting k attributes at each node to let the estimate of class probabilities. Running the algorithm we analyze the classifier output with statistics based output by using 10 cross validation to make of each instances of dataset.

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
Random tree	0.06	97.1103	2.8897	0.0148	0.901

Table 6. Random Tree algorithm accuracy

3.2. Experimenter Interface

Experimenter Interface has been used in this paper to analyze data by experimenting through algorithms such as Naïve Bayes, J48, REP Tree and Random Tree to classify the data using train and test sets.

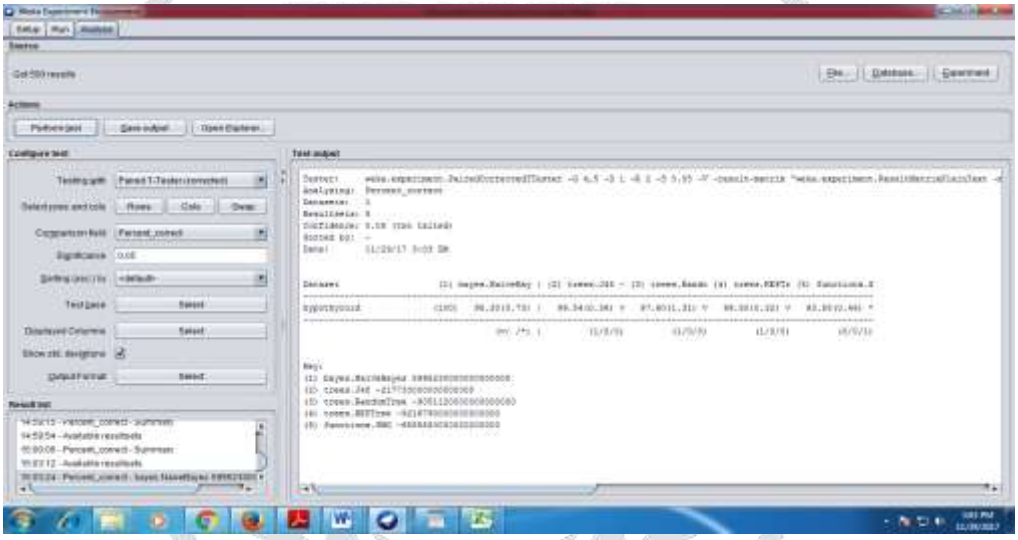

Fig.3. Screenshot view of Experimenter Algorithm Accuracy

Scoring accuracy of all algorithm is given in Table 7 Experimenter algorithms accuracy

Algorithm	Best Accuracy(v)	Worse Accuracy(*)
J48	99.54	-
SMO	-	93.58
REP	99.50	-
RANDOM	97.60	-

3.3 Selection Attribute

Here in this paper we can know which attribute has its significance with their ranking values using *select attributes and info gain function*. The Infogain function gives the importance of attributes which makes the result positive or even negative by discarding some attributes.

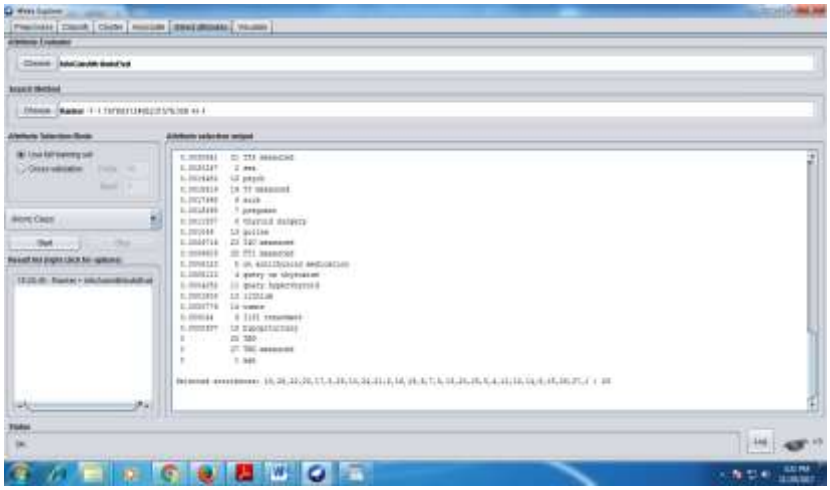

Fig. 4. Screenshot view of select attribute

In figure 4 it shows that there are nearly three attributes which are of 0 significance in dataset.

SIGNIFICANCE	ATTRIBUTES
0	TBG
0	TBG Measured
0	age

Table 8. zero significance attributes

4. Conclusion and Future Work

The main goal of this paper is to forecast hypothyroid disease using WEKA data mining tool. It has four interfaces. Out of these four we have used two interfaces: Explorer, Experimenter. Each interface has its individual classifier algorithms. We have used five algorithms i.e. Naïve Bayes, J48, SMO, REP Tree and Random tree for our experimentation. Then these algorithms were executed using WEKA data mining technique to explore algorithm accuracy which was gained after running these algorithms in the output window.

After running these algorithms the outputs were compared on the basis of accuracy achieved. These algorithms compare classifier accuracy to each other on the basis of correctly classified instances, time taken to build model, mean absolute error and ROC Area. Experimenter result showed that scoring accuracy of REP Tree is 99.50% and J48 is 99.54% as compared to Naïve Bayes and Random tree so we can conclude that in Experimenter interface REP Tree and J48 are the best classifier algorithms for accuracy of hypothyroid disease survival on the basis of symptoms given in dataset among patients. The uses of Weka can be extended more to medical field for diagnosis of different diseases like cancer, dengue, etc. It can also support in resolving the harms of clinical research using different bids of Weka. Additional benefit of using Weka for forecast of diseases is that it can simply diagnose a disease even in case when the number of patients for whom the prediction has to be done is enormous or in case of very large data sets spanning lakhs of patients. Even though Weka is a dominant data mining tool to examine the outline of classification, clustering, Association Rule Mining and visualization of result in medical health to predict disease among patient but we can use other tools such as Matlab, R tool, Rapid miner in order to extra classify different data sets .The proposed approach is used with hypothyroid data set but we plan to extend this methodology in upcoming for prediction of other diseases such as cancer, dengue etc.

References

[1]. Dhamodharan S , Liver Disease Prediction Using Bayesian Classification , Special Issues , 4th National Conference on Advance Computing , Application Technologies, May 2014

[2]. SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014.

[3]. Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323, October 2014.

[4]. David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.

[5]. Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.

[6] Kumar M. N., Alternating Decision trees for early diagnosis of dengue fever .arXiv preprint arXiv:1305.7331,2013.

[7]. Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.

[8]. Sugandhi C , Ysodha P , Kannan M , Analysis of a Population of Cataract Patient Database in WEKA Tool , International Journal of Scientific and Engineering Research ,2(10) ,October ,2011.

[9]. Yasodha P, Kannan M, Analysis of Population of Diabetic Patient Database in WEKA Tool, International Journal of Science and Engineering Research, 2 (5), May 2011.

[10]. Bin Othman M. F , Yau, T. M. S., Comparison of different classification techniques using WEKA for breast cancer, In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, Springer Berlin Heidelberg, 520-523,January 2007.

[11] Ian H. Witten, Eibe Frank & Mark A. Hall., "Data Mining Practical Machine Learning Tools and Techniques, Third Edition." Morgan Kaufmann Publishers is an imprint of Elsevier.

[12] Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp- 155-160
[13] Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifier", Int. Journal of Engineering Research and Applications, Vol. 4, Issue 8(Version 5), August 2014, pp- 91-93
[14] Tina R. Patil, and S. S. Sherekar. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications Vol. 6, No.2, (Apr 2013), 256 – 261
[15]A. Anderson, M. Corney, O. de Vel, and G. Mohay."Identifying the Authors of Suspect E-mail". Communications of the ACM, 2001.
[16] Shlomo Hershkop, Ke Wang, Weijen Lee, Olivier Nimeskern, German Creamer, and Ryan Rowe, "Email Mining Toolkit Technical Manual". (June 2006) Department of Computer Science Columbia University.
[17] Bron, C. and J. Kerbosch. "Algorithm 457: Finding all cliques of an undirected graph." (1973).
[18] Ding Zhou et al and Ya Zhang, "Towards Discovering Organizational Structure from Email Corpus". (2005) Fourth International Conference on Machine Learning and Application.
[19] Giuseppe Carenini, Raymond T. Ng and Xiaodong Zhou , "Scalable Discovery of Hidden Emails from Large Folders". Department of Computer Science, University of British Columbia, Canada.
[20] Hung-Ching Chen el al, "Discover The Power of Social and Hidden Curriculum to Decision Making: Experiments with Enron Email and Movie Newsgroups". Sixth International Conference on Machine Learning and Applications.
[21] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz and Anand Swaminathan, "Mining Email Social Networks". (May 22-23, 2006). Dept. of Computer Science, University of California, Davis.