ISSN No. 0976-5697

Volume 5, No. 6, July-August 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Assessment of various Supervised Learning Techniques by means of open source API for Qualitative Bankruptcy

K.Thanweer Basha
Reg.No: 124M1D5818, M.Tech.Scholar,
Dept. of CSE, Vemu Institute of Technology,
P.Kottakota, Chittoor Dist, A.P, India

B.Rama Ganesh
B.Tech, M.Tech, (Ph.D), Assoc.Prof & HOD,
Dept. of CSE, Vemu Institute of Technology,
P.Kottakota, Chittoor Dist, A.P, India

Abstract: The improvement of machine learning household tasks such as classification, clustering and association has exposed the call for machine learning algorithms to be applied on huge amount of data. In this paper we present the evaluation of diverse classification techniques and find the suitable finest classification algorithm for taken dataset, using Waikato Environment for Knowledge Analysis API or in short, WEKA. WEKA is an open source which consists of a group of machine learning algorithms for performing data mining tasks. The aim of this paper is to investigate the performance of different classification methods and come across the finest classification algorithm for given set of large data and we also propose the implication rules for Bankruptcy using Apriori. The actual evaluation of learning by example is done with help of confusion matrix and Receiver Operating Characteristics curve. At this time classification algorithms experienced are Bayesnet, Naivebayesclassifier, ConjuctiveRules, DecisionTable/Naivebayes, DecisionTable, Nearestneighbors, and OneR. In this paper I had chosen Bankruptcy dataset for performing data mining tasks.

Keywords: Data Mining, Supervised learning, Time complexity, Confusion Matrix, open source, API, Qualitative Bankruptcy.

I. INTRODUCTION

This paper is an implementation of various data mining supervised algorithms which are applied over a selected data set. This paper provides us a typical idea about how the implementation of various algorithms classification algorithms are applied over the dataset and the properties of each algorithm are deliberate in this paper in different sections. All the section briefly discusses about the data mining supervised learning algorithms evaluation over the Qualitative Bankruptcy dataset. Different classification and association algorithms are selected over the qualitative Bankruptcy dataset and applied over it. This paper describes about the various classification and association algorithms concert over the Qualitative Bankruptcy [11] dataset by weka [6] , and confusion matrix. The performance various Classification and association algorithms are evaluated by their time complexities.

II. BACK GROUND

Data Mining is involved in the process of extracting hidden facts from data. One of the most effective methods for analyzing the data and it is often referred as Business Intelligence. Data mining can also be referred as KDD-Knowledge Data Discovery in Data bases. Data mining algorithms are machine learning algorithms that are classified as Supervised and unsupervised learning algorithms. Classification and association rules are two popular and regularly performed data mining activities which are of supervised and unsupervised learning algorithms. Data mining is performed over a large amount of historic and current data that is preprocessed and stored in a data warehouse. A data warehouse is a subject oriented, non-volatile, integrated and time variant [12] by its features a data warehouse is unique property for every organization in enterprise decision making. Data warehouse is also referred as ETL tool [2]. A data mining activity is performed over a data warehouse for decision making activity. The data warehouse can also be labeled as Online Analytical process (OLAP) [12]. Machine learning is a branch of artificial intelligence which deals with the study of learning and developing a system which can learn from data. Machine learning activity involves representation and generalization. In machine learning activity learning process is performed by the algorithm to learn which can be either representation or generalization. The learning capability of an algorithm lies on the key object of study in the subfield of computational learning.

Classification is a most familiar and most popular data mining technique. Various real world activities such as security and e-commerce applications fraud and intrusion detection, supply chain management etc. are basic classification based data mining applications [3].Classification algorithms are used to classify the data in predefined class intervals. It also pertains with some issues like missing data, measuring performance etc. In our paper we have chosen some popular classification data mining algorithms over the data set,are BAYES NET, NAÏVE BAYES CLASSIFIER,CONJUCTIVE RULE, DTNB, DECISION TABLE, OneR , JRip , and NNge[13,14] over qualitative bank ruptcydataset. Each and every classification algorithm will classify the data set into various classes.The resulted data sets classes are used to analyze the better classification algorithm by taking some measures like confusion matrix, roc curve etc. In our paper we also calculate the time complexity of these algorithms.

The classification algorithms are added to the data mining tool weka and the targeted data set is selected and converted the data set into CSV file format is called as Comma Separated Value file format. The data set now is named as QB.csv. We apply classification algorithms over the dataset and visualize the results using weka [1]. Association Rules are used in data mining to determine the mode of the dataset. Association rules generate the most common association in the data. The association rules are

also used for affinity analysis. Here we use apriori algorithm over the Bankruptcy dataset to perform affinity analysis [4]. WEKA is formally called as Waikato Environment for Knowledge Learning. It is a computer program which was developed by university of Waikato, New Zealand. The tool supports various standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection [12]. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Weka uses ARRF- Attribute File Format for performing data mining tasks. Weka has various GUI's for performing various types of data mining tasks they are Explorer, Experimenter, Knowledge workflow, Simple CLI. It can also be used to develop new machine learning schemes. The bankruptcy data set was chosen from UCI Machine learning repository [11]. The dataset was available in the UCI Machine Learning Repository website.

III.METHODOLOGY

Datamining supervised learning algorithms are used for performing various operations, we use the classification and association algorithms for analyzing given dataset and generate rules .here thedataset is source, association is used to generate rules that are relevant to dataset. Here our approach is to identify the best algorithm which can perform the analysisfaster, and better analysis. The data set here we choose is a banking related dataset named as qualitative Bankruptcy .This data set contains various fields which provide the relevant information. Various steps that are involved in process. The steps involved are:

a. ***Load dataset:*** The data set is generally is an excel sheet. This sheet is converted in to an attribute relation file format (arff) which is the standard file.

b. ***Apply Classification and association algorithms:*** After the data set is loaded we use the following classification algorithms namely BAYES NET, NAÏVE BAYES CLASSIFIER, CONJUCTIVE RULE, DTNB, DECISION TABLE, OneR, JRip, NNge and association algorithms Apriori [6].

c. ***Algorithm investigation:*** The selected algorithms are applied over the loaded dataset. For each and every algorithm time complexity is calculated, and also calculate correctly classified tuples percentage and receiver operating characteristic (ROC) is a graphical design which point up the performance of a supervised algorithm with help of sensivity or recall. In ROC plotting is done in case of sensivity, by plotting fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate).

d. ***Recognition of Optimized Algorithm:*** After analyzing the algorithms we prepare a Confusion matrix for each supervised algorithm; to compare them among percentages of correctly classified instances we can choose a better algorithm for analysis over the given data set. In regard to time complexity also we can choose suitable an algorithm for fast computation. Better rules proposed for given dataset, using association algorithm such Apirori over the dataset we derive some association rules which can be interpreted for decision making over a qualitative Bankruptcy.

IV.DATASET

The attributes or parameters which we used for collecting the dataset qualitative bankruptcy data are as followsattribute information furnished below. Dataset contains 255 around recordswhich are to be analyzed [11].The data set have various attributes like

a. ***Industrial Risk (IR)****:* It is a Nominal attribute with possible values positive, negative, average. This attribute provides us considerations about Government policies and International agreements Cyclicality, Degree of competition. The price and stability of market supply. The size and growth of market demand. The sensitivity to changes in macroeconomic factors. Domestic and international competitive power and Product Life cycle.

b. ***Management Risk (MR)****:* It is a Nominal attribute with possible values positive, negative, average. This attribute provides us considerations about Ability and competence of managementStability of management. The relationship betweenmanagement / owner. Human resources management. Growth process / business performance. Short and long term business planning. Achievement and feasibility.

c. ***Financial Flexibility (FF)****:* It is a Nominal attribute with possible values positive, negative, average. This attribute provides us considerations about Government policies and International agreements. Cyclicality, Degree of competition. The price and stability of market supply. The size and growth of market demand. The sensitivity to changes in macroeconomic factors. Domestic and international competitive power and product life cycle.

d. ***Credibility (CR)****:* It is a Nominal attribute with possible values positive, negative, average. This attribute provides us considerations about Ability and competence of management Stability of management. The relationship between management / owner. Human resources management. Growth process / business performance. Short and long term business planning. Achievement and feasibility.

e. ***Competitiveness (CO)****:* It is a Nominal attribute with possible values positive, negative, average. This attribute provides us considerations about direct financing. Indirect financing and other financing.

f. ***Operating Risk (OP)****:* It is a Nominal attribute with possible values positive, negative, average. This attribute provides us considerations about Credit history, reliability of information, the relationship with financial institutes.

g. ***Class:*** It is a Nominal attribute with possible values Bankruptcy, and non Bankruptcy which is used to classify the tuples in the dataset.

V. CONFUSION MATRIX

Error matrix usage to assess the eminence of the productivity of a classifier on the bankruptuacy data set. The diagonal elements signify the number of points for which the predicted label is equal to the true label, while off-

diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions [9].

VI.ROC CURVE

ROC curve in signal theory is a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot which illustrates the performance of the binary classifier system. Here the true positive values (TPR) are plotted out of total actual positives. Recall.Figure below illustrates the ROC curves generated for various classification algorithms [7].The table below illustrates use each and everyalgorithms performanceover the selected dataset.

VII. RESULTS

After applying various classification algorithms over the Bankruptcy dataset. The execution time for various classification algorithms are shown as an output window which is shown below. For better analyzing of the results we are using the graphical representations of graphs for better visualization and understandability. We also choose an algorithm based on execution time and percentage of correctly classified.

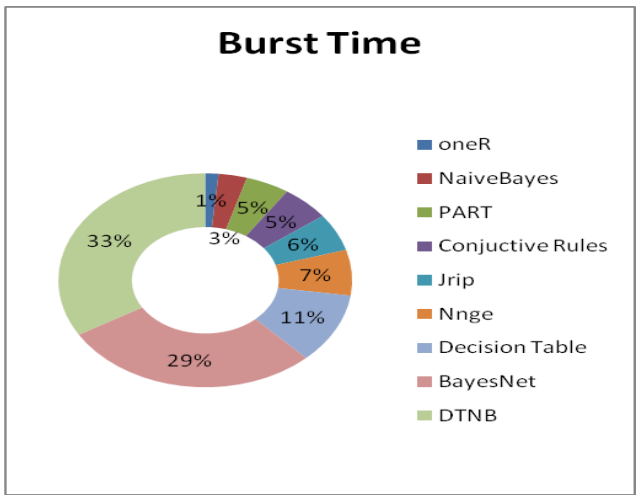

Figure 1: Execution times of various data mining algorithms

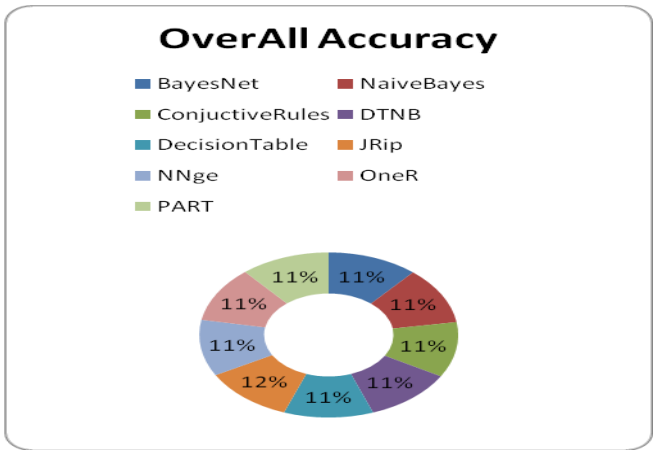

Figure 2: percentages of correctly classified tuples using various classifications

The accuracy of each supervised algorithm that is used over bankrtpancy dataset is shown below table

Tabel I: Accuracy of classification algorithms

Supervised Algorithm	CCBI	ICCBI	CCNBI	ICCNBI	Accuracy
BayesNet	66	4	103	2	96.19047619
NaiveBayes	66	4	103	2	96.19047619
ConjuctiveRules	62	8	103	2	93.33333333
DTNB	66	4	100	5	94.76190476
DecisionTable	66	4	101	4	95.23809524
JRip	68	2	101	4	96.66666667
NNge	66	4	100	5	94.76190476
OneR	62	8	103	2	93.33333333
PART	66	4	101	4	95.23809524

In the above table I
Correctly classified Bankrtpancy Instances- CCBI
In Correctly classified Bankrtpancy Instances- ICCBI
Correctly classified Non-Bankrtpancy Instances -CCNBI
In Correctly classified Non-Bankrtpancy Instances - ICCNBI
The association results are shown below:

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D
0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     bankrupt_qualitative
Instances:    175
Attributes:   7
              IR
              MR
              FF
              CR
              CO
              OP
              Class
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.25 (44 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15
Generated sets of large item sets:
Size of set of large itemsetsL(1): 17
Size of set of large itemsetsL(2): 12
Size of set of large itemsetsL(3): 5
Size of set of large itemsetsL(4): 1
Best rules found:
a)  FF=N CO=N 59 ==> Class=B 59    conf:(1)
b)  CR=N CO=N 47 ==> Class=B 47    conf:(1)
c)  FF=N CR=N CO=N 45 ==> Class=B 45    conf:(1)
d)  CO=P 75 ==> Class=NB 73    conf:(0.97)
e)  CO=N 64 ==> Class=B 62    conf:(0.97)
f)CR=N CO=N 47 ==> FF=N 45    conf:(0.96)
g)   CR=N CO=N Class=B 47 ==> FF=N 45    conf:(0.96)
h)  CR=N CO=N 47 ==> FF=N Class=B 45    conf:(0.96)
i)CO=N Class=B 62 ==> FF=N 59    conf:(0.95)
j)CR=N Class=B 54 ==> FF=N 51    conf:(0.94)
```

VIII.CONCLUSION

The results provide the evidence of evaluation of the supervised learning algorithms for data mining over bankruptcy dataset. Analyzing the results yield that OneR

classification algorithm is computes at a better pace and conjunctive Rules perform better affinity over the data.

IX. FUTURE WORK

This project can be future enhanced in developing a better classification or association rules generating algorithms that can be performed over a variety of datasets of similar type which in turn helps to develop a robust data mining application based algorithms which can be useful in a variety real world applications.

X. REFERENCES

[1]. DimitriosGiannoulis, EmmanouilBenetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange and Mark D. Plumbley "detection and classification of acoustic scenes and events" IEEE Workshop on Applications of Signal Processing to Audio and Acoustics on 23oct2013.

[2]. P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 37–47, 2009.

[3]. J. Kovacevic, "How to encourage and publish reproducible research," in IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2007, pp. 1273–1276.

[4]. P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quenot, "TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in Proc. of TRECVID 2012. NIST, USA, 2012.

[5]. S. Araki, F. Nesta, E. Vincent, Z. Koldovsk`y, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separationevaluation campaign (SiSEC2011)," in Latent Variable Analysis and Signal Separation. Springer, 2012, pp. 414–422.

[6]. http://www.cs.waikato.ac.nz/ml/weka/

[7]. http://en.wikipedia.org/wiki/Receiver_operating_characteristicFast Computation algorithms

[8]. http://en.wikipedia.org/wiki/Machine_learning

[9]. http://en.wikipedia.org/wiki/Confusion_matrix

[10]. D. Wang and G. J. Brown, Computational auditory scene analysis: Principles, algorithms, and applications. IEEE Press, 2006.

[11]. http://archive.ics.uci.edu/ml/datasets/Qul_Bankruptcy.

[12]. Jiwen & Khamber Datamining : Concepts and techniques 3rd Edition by Elsevier

[13]. F. Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp. 520-523,www.springerlink.com © Springer-Verlag Berlin Heidelberg 2007

[14]. Nikhil N. Salvithal, Dr. R. B. Kulkarn Evaluating Performance of Data Mining Classification Algorithm in Weka, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 10, October 2013, Volume 2, Issue 10, October 2013.