ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Accuracy Estimation of Classification Algorithms with DEMP Model

Dr. Mahendra Tiwari[1],Manmohan Mishra[2]

Assistant Professor,UCER Allahabad.[1]

Research Scholar,MNIET,Allahabad,U.P.[2]

Abstract: Data Mining is the extraction of hidden information from large databases. It is a technology with great potential to help organizations focus on the most important information in their data warehouses or database. Decision making with data mining is huge and complex task. In our work, we collected data from UCI repository and then apply five different classification methods for classifying different types of data based on their size. We compare these algorithms of classification and check which algorithm is optimal. By this task we extract knowledge that describes algorithms' performance in various dataset. This work helps in making decision about the accuracy of classification algorithms.

Keywords: Data Mining, accuracy, classification algorithms, datasets.

I. INTRODUCTION

Rapid growth in technology has brought vast exposure in the volume of the data available on the various information repositories available, but it is difficult to manually organize, analyze and retrieve information from this data. This generates an essential and eminent need of methods that can help users to effectively navigate, summarize, and organize the data so that it can further be used for decision making in various application areas such as market analysis, fraud detection and customer retention etc.

Therefore, the techniques which perform data analysis and investigate data patterns are called data mining.

The paper is divided into seven parts. In introduction, Data Mining is briefly introduced and classification process is discussed. In section 2, related work performed by various authors are elaborated which allows performance evaluation of different classification techniques on different types of datasets. In section 3 proposed model for accuracy measurement of data mining algorithm is illustrated. In section 4,several steps of DEMP model are discussed and analyzed. Experimental setup and strategy evaluation for accuracy estimation are described in section 5. In section 6,analysis of classification algorithm are presented and some decisive outcomes are underlined. Finally, a conclusion and some future perspectives are highlighted in section.

Classification is a data mining technique used to classify each item in a set of data into one of predefined set of classes or groups.

Data classification is a two step process. In the first step, a model is built by analyzing the data tuples from training data having a set of attributes. For each tuple in the training data, the value of class label attribute is known. Classification algorithm is applied on training data to create the model. In the second step of classification, test data is used to check the accuracy of the model. If the accuracy of the model is acceptable then the model can be used to classify the unknown data tuples..

II. RELATED WORK

A lot of work have been performed by various authors on performance evaluation of Data Mining algorithms on various different tools, accuracy estimation, and enhancement are done on the basis of comparing various classifiers with different types of data set ,we presented their result as well as tool and data set which are used in performing evaluation

Ying Liu,wei-keng Liao *et al.* [15] in his article "performance evaluation and characterization of scalable data mining algorithms" investigated data mining applications to identify their characteristics in a sequential as well as parallel execution environment .They first establish Mine bench, a benchmarking suite containing data mining applications. **Osama Abu Abbas** [1] in his article "comparison between data clustering algorithms " compared four different clustering algorithms (K-means, hierarchical, SOM, EM) according to the size of the dataset, number of the clusters ,type of S/W.**Abdullah *et al.*** [2] in his article "A comparison study between data mining tools over some classification methods" conducted a comparison study between a number of open source data mining S/W and tools depending on their ability for classifying data correctly and accurately.**T. velmurgun** [6] in his research paper "performance evaluation of K-means & Fuzzy C-means clustering algorithm for statistical distribution of input data points" studied the performance of K-means & Fuzzy C-means algorithms. These two algorithm are implemented and the performance is analyzed based on their clustering result quality.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

As literature review suggests that most of work is based on classification by using specific approach. The objective is not only the problem solution but identifying the alternative approach for accuracy estimation of classifiers. Here in the present work some classifiers are selected and applied to the dataset to choose best classifier. With comparative analysis the optimum approach is established.

III. PROPOSED MODEL

A typical data mining application includes data collection from different data sources, preprocessing on data and finally some data mining algorithms are applied on data for various purposes such as prediction,
Accuracy estimation, analysis etc.

There are two most common data mining methodologies are used to build a classification model.

CRISP-DM(Cross Industry Standard Process for Data Mining) which consists five steps: Business understanding, data understanding, data preparation, modeling, evaluation , and deployment while other methodology proposed by SAS institute is SEMMA[8](Sample Explore Modify model Assess) which involves five step process: sampling, exploration, modify, model, and assess.

The proposed model is established for decision making from evaluation of classifier models. With decision making researchers and analysts can create a new way of thinking about classification algorithms.

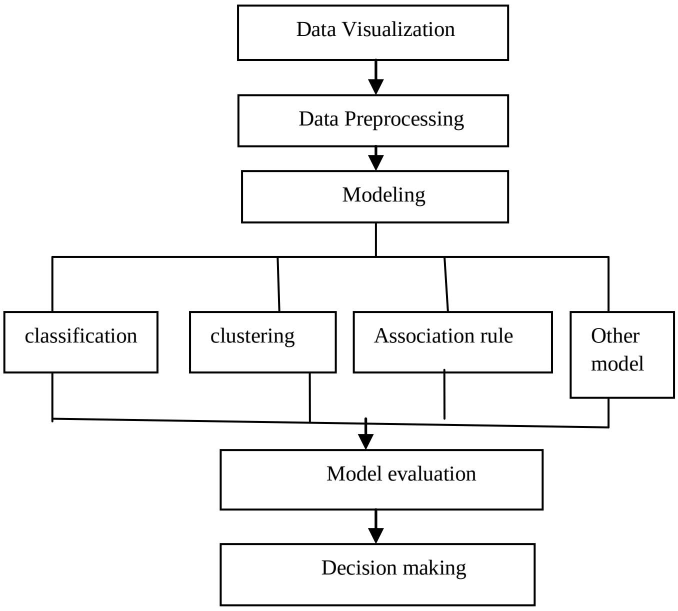

Figure 1:DEMP model

The proposed model focus on decision making based on accuracy of classifier models, the model does not only cover classification algorithms but it also covers clustering, Association rule and other stat based models. The strength of model is decision after evaluation so we call it name DEMP stands for Decision with Evaluation and modeling process.

IV. DEMP MODEL

The proposed DEMP model includes five steps: data visualization, data preprocessing, modeling, model evaluation, and decision making. Data visualization elaborates source of data ,collection of data, description of data, and quality of data. The data used in my experiment is obtained from UCI data repository[7] and from widely accepted dataset available in Weka[9] toolkit. We used 2 data set for evaluation with classifier one of them from UCI Data repository that are Zoo data set and supermarket data set is inbuilt in WEKA 3-6-6 Zoo data set are in csv file format and supermarket data set are in arff file format.

Table 1: Dataset

Name of dataset	No, of attributes	No. of instances	Missing values
Zoo	17	101	No
Supermarket	217	4627	No

The data preprocessing step includes data construction and transformation. We used weka tool for experiment and weka provides working with attributes section such as, use all the attributes of data set or select some specific attributes along with statistic show about the values stored in the attributes like numeric or nominal.

Modeling step contributes in applying the classification techniques to build classifier models. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. while classification predicts categorical labels(classes), prediction models continuous valued functions. Model evaluation step demonstrate the evaluation result of classifiers in graphical form . Finally decision making provides comparative result of various classifier to better judgment of classifier's accuracy.

V. EXPERIMENTAL SETUP

The performance of classifier can most simply be measured by counting the proportion of correctly predicted examples in training and test dataset. This value is the accuracy. The accuracy of a classifier on a test set is the percentage of test set tuples that are correctly classified by the classifier. It is also referred as recognition rate of classifier.

H/W tool: I conduct my evaluation on Pentium 4 Processor platform which consist of 512 MB memory, Linux enterprise server operating system, a 40GB memory, & 1024kbL1 cache.

S/W tool: In all the experiments, I used Weka 3-6-6, I looked at different characteristics of the applications- using classifiers to measure the accuracy in different data sets.

There are five classification techniques used in experiment, and the test design for evaluation is cross validation, the name of techniques and their corresponding classifiers are listed below:

Classification Techniques	Classifier
Decision Tree	Decision stump, REP,CART
Lazy learner	IBK, Kstar
Rules based	OneR, ZeroR
Naïve bayes	Naive
Regression	Linear regression

Figure 2 Classification techniques

The result of evaluation are based on predictive accuracy and error rate of classifier. Predictive accuracy measures in correctly classified instances while error rate illustrates incorrectly classified instances.

Classifier	instances	predictive accuracy	Error rate
SMO	108	64	44
Ibk	108	38	70
Naïve	108	72	36
Zeror	108	73	35
oner	108	70	38
Cart	108	72	36
decision	108	74	34
kstar	108	70	38

Figure 3 :Evaluation of classifier on Zoo data set with cross validation test mode

Classifier	instances	predictive accuracy	Error rate
SMO	108	63	44
Ibk	108	37	70
Naïve	108	63	36
Zeror	108	67	35
oner	108	63	38
Cart	108	63	36
decision	108	64	34

Figure 4: Evaluation of classifier on Zoo data set with cross validation test mode

VI. ANALYSIS OF CLASSIFICATION ALGORTITHM

Data mining algorithms are analyzed on performance basis and it involves performance analysis of accuracy of data mining algorithms.

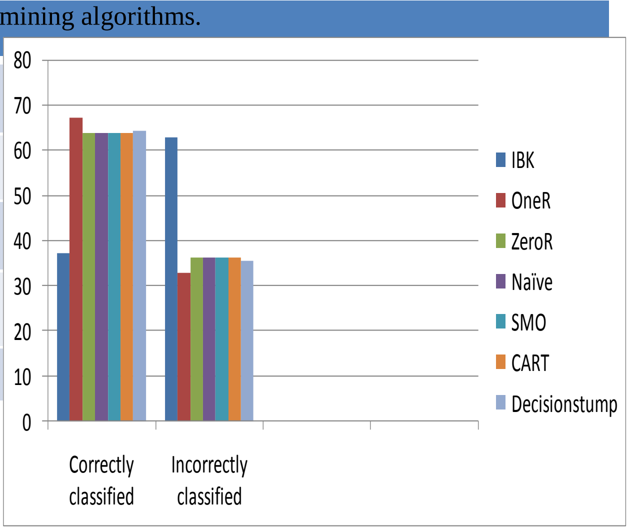

Figure 5: Analysis of accuracy of classification algorithm On supermarket dataset

Analysis has been done on performance of classification algorithms on two types of dataset, varying in size and features. In analysis we found that IBK is much better than others when data size increases, but in small dataset ZeroR performs good,if error rate can be measured then we found that CART is still consistent,it does not change when dataset changed.

VII. CONCLUSION

This paper has concentrated on accuracy estimation of classification algorithms which can be depend on dataset size and test mode ,we used cross validation method for experiment, we tried to cover small dataset range with limited attributes to large dataset with more attributes, working on performance, we found that if we change the test mode , and choose complex dataset with mix attribute types like ordinal, nominal, categorical, and continuous, then result can be unpredictable. in vice versa our selected dataset has no any missing values but if we select un-processed dataset then the decision making on basis of accuracy is difficult to establish.

We introduced DEMP model and all the operations are performed according the steps mentioned in DEMP. Finally we can conclude that accuracy of data mining algorithms always varies on multiple aspects.

REFERENCES

I. Osama Abu Abbas, (2008), "comparison between data clustering algorithms", The international Arab journal of IT, vol 5 no3

II. Abdullah *et al.*, "A comparison study between data mining tools over some classification methods", international journal of advanced computer science & applications(IJACSA),www.ijacsa.thesai.org.

III. Velmurugan T.,T. Santhanam, (2010), "performance evaluation of K-means & fuzzy C- means clustering algorithm for statistical distribution of input data point", Europen Journal of scientific research, vol. 46 No.3.

IV. Jayaprakash *et al.*, "performance characteristics of data mining application using minebench", National science Foundation (NSf) .

V. Pramod S. ,O. Vyas, (2010), "performance evaluation of some online association rule mining algorithms for sorted & unsorted data sets", International Journal of computer Applications, vol. 2 No. 6 .

VI. Kavitha P, T. Sasipraba, (2011), "performance evaluation of algorithms using a distributed data mining framework based on association rule mining", International Journal on computer Science & Engineering (IJCSE).

VII. UCI Data repository (archive-ics.uci.edu/ml/) (kdd.ks.uci.edu/).

VIII. SAS Enterprise mines documentation, what's new in SAS enterprise mines 5.1, SAS Institute Inc.

IX. Weka (w.w.w. cs. waikato.ac.nz/ml/weka/)

X. Thair Nu Phyu, "survey of classification techniques in data mining", proceedings of the international multi conference of engineers & computer scientists 2009, IMECS, vol 1.

XI. Han J., M. Kamber, Data Mining concepts and Techniques, Morgan Kaufmann 2nd edition.

XII. Alex Berson, J. Smith, Data Warehousing, Data Mining & OLAP, TaTa McGraw-Hill edition 2004.

XIII. I.Krishna M., (2010), "Data Mining- Statistics Applications: A Key to Managerial Decision Making", indiastat.com socio – economic voices

XIV. Anderson, J., (2002), "Enhanced Decision Making using Data Mining: Applications for Retailers", Journal of Textile and Apparel, vol 2,issue 3

XV. Ying liu *et al.*, "performance evaluation & characterization of scalable data mining algorithms" (www.users.eecs.nortwestern.edu/~yingliu/papers/pdcs.pdf).

BIOGRAPHY

Dr. Mahendra Tiwari is an Assistant Professor in computer science & engineering department at United College of Engineering & Research. He received MCA, & M.Tech in Information Technology and completed Ph.D. in computer science. He has more than 30 research papers publications in area of Data Mining .