Performance Analysis on Transaction Datasets using Different Classification Algorithm

Rashmi Amardeep[1], Sanjana Cholaraju[2], N Chandana[3]
[1]Faculty, Dept. Of ISE Sir.M.VIT, rashmi_is@sirmvit.edu
[2,3]VI Semester, Dept. Of ISE, Sir.M.VIT, sanjana.c1328@gmail.com, chandananadimpalli@gmail.com

Abstract: Market basket analysis is an important analysis in finding out customer purchasing patterns in transactional database. The analysis helps in developing the strategies in sales and services. Customer satisfaction is a key role in business and hence has drawn many researchers in this area. The datasets are real datasets considered from a retail store. Implementation is by Weka tool using different classification algorithm. We have applied different classifiers to the data sets to know most appropriate values. An evaluation metrics indicate Bayes net & in Meta classifier, attribute selected classifier is superior in terms of accuracy.

Key words: Association rules, Classification algorithm

1. INTRODUCTION

Association rules in mining algorithms discover interesting relationships between data items that occur frequently together. Since their introduction in 1993 by Argawal et al. [15], the association rules mining problems have received a great attention. Within the past decade, hundreds of research papers have been published presenting new algorithms or improvements on existing algorithms to solve mining problems more efficiently

Classification can be defined as a function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Using the classifiers Bayes net, Naïve Bayes, Stochastic Gradient Descent (SGD), MLP Classifier, IBK Classifier, Locally Weighted Learning (LWL), Decision table, JRipper (JRIP), Logistic Model Tree (LMT), J48 Classifier, Attribute Selected classifier, Bagging Classifier in finding approximate value. Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

A sale is a transaction between two parties where the buyer receives goods and assets in exchange for money. It is also a document form of the amount of products which are sold in a given amount of time and on the bases of region. It helps the manager to get an idea of what products to be displayed on sales in order to achieve profit. Sales transactions can be of three types: (a) Cash sales: Cash is collected by the employee when the employee delivers the product which is required by the customer. (b) Credit sales: Cash isn't collected by the employee from the customer after the product is delivered and they are given a particular duration within which the customer has to credit the purchase amount to the employee or organisation. (c) Advance payment sales: The customer pays the amount for the product to the employee beforehand he delivers the product.

Many applications took the benefit of association rules mining to improve the decision making process, such as market basket analysis, catalog design, cross marketing, and sales transaction. Market basket analysis is a typical example of association rules analysis that discovers buying behaviour of customers. The discovered association rules can help decision makers develop marketing strategies. The paper presents different classifiers such as Bayes classifier, Functions, Rules, Tress, Meta classifiers and its variants. Analysis of the sales of the product for a particular month is considered in this paper.

2. LITERATURE SURVEY

Yen-Liang *et al* [1], propose a new method of discovering customer purchasing patterns by extracting associations .The existing methods, fail to discover important purchasing patterns in a multi-store environment, because of an implicit assumption that products under consideration are on shelf all the time across all stores. To overcome this weakness they have proposed new method store-chain association rules, an Apriori like algorithm for automatically extracting association rules in a multi-store environment. The proposed method has advantages over the traditional method especially when the numbers of stores and periods are large, stores are diverse in size, and product mix changes rapidly over time. Further, running time is obtained by averaging the running times of all the data sets, and the simulation results show that the algorithm is computationally efficient.

Jochen Hipp *et al* [2],dwelt with association rule mining in the context of a complex, interactive and iterative knowledge discovery process. The proposed method address algorithmic complexity by presenting a rule caching schema that significantly reduces run times of the mining algorithms. Furthermore presented efficient integration with modern database systems key factors in practical mining applications and enhanced the traditional association rule mining framework to select interesting rules.

Rakesh *et al* [3], proposed Apriori and AprioriTid algorithms and AprioriHybrid of discovering association rules for large sales transaction database in discovering association rules. The proposed method is compared with the earlier algorithms AIS and SETM algorithms. The results evaluated showed that the proposed algorithms always outperform AIS and SETM. The performance gap increased with the problem size and ranged from a factor of three for small problems to more than an order of magnitude for large problems. In addition the execution time decreases a little as the number of items in the database increases. As the average transaction size in creases while keeping the database size constant the execution time increases only gradually. These experiments demonstrate the feasibility of using AprioriHybrid in real applications involving very large databases.

Schonrost *et al* [4], studied a market basket analysis through association rules mining on transactional data for decision making in identifying the common products purchased in single transactions. The datasets considered for four months. Association rules were generated for k-item sets and this helped in find the frequent items purchased during each month. During the analysis it is discovered the total number of products sold for the four month period and in each categories the number of products purchased. Lift was used for evaluation measure for the association rules.

3. METHODOLOGY

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. The datasets are classified on the bases of Bayes, Functions, Lazy, Meta, Rules, and Trees. Table 2.1 gives the description of various classifiers considered in the study.

Table 3.1 Classifiers and its variants

Classifiers	Variants	Description
Bayes classifier	**Naïve Bayes classifier** Chitra *et al,* august 2012,[5]	In this paper they have used naive Bayes algorithm with the help of Bayes theorem as it is very simple to understand, build and it performs advanced classification methods.
	Bayes net Faltin *et al,*2007,[6]	Bayesian network is executed using a graphical model that is called as Directed acyclic graph. In this particular graph it makes sure that every single node has either a parent node or a child node.
Functions	**SGD classifier** Mu Li *et al*,[7]	SGD stands for stochastic Gradient descent which is mostly used in large scale for optimisation problems. If there is an increase in min batch there will be a gradual decrease in the convergence.
	MLP classifier Hassan *et al*,[8]	MLP stands for multi layered perceptron. It works on forward based neural network which consists of three nodes they are input nodes, hidden nodes and output nodes respectively. The hidden nodes and output nodes can be termed as neuron which uses the “nonlinear function”.
Lazy	**IBK** A B M Shawkat *et al*,[9]	The time taken to classify a test instance with a nearest neighbor classifier increases linearly with the number of training instances that are kept in the classifier.
	LWL Christopher *et al*.[10]	LWL stands for locally weighted learning. In this classification the solution for the obtained query is found by using “relevant data in the database.”
Rules	**Decision Table** C. Lakshmi *et al*, December 2011,[11]	In decision table classifier there are two variants which are decision table majority and decision table local. The comparison is made between the decision table and a new instance and if the result is empty then it returns the

		majority of training set and it doesn't contain any training instances becomes decision table majority but when the returns answer from the local neighbourhood then it becomes decision table local.
	JRipper C. Lakshmi *et al*, December 2011,[11]	JRipper is proposed by William W.JRip. This classifier is used for error reduction by implementing propositional rule learner and repeated incremental pruning.
Trees	**LMT** Purva *et al*,[12]	LMT stands for logistic model tree. LMT is purely integrated with "supervised training algorithm". LMT is the product of combination of "logistic prediction and the decision tree learning."
	J48 Purva *et al*,[12]	Based on available data of different attribute values it helps the user to give a "target value of a new sample."
Meta Classifiers	**Attribute selected classifier** Dr Gnanambal *et al*,[13]	The datasets contain many attributes which should not be relevant to make this possible attribute selection is used.
	Bagging Jan N. van Rijn *et al*,[14]	"It increases accuracy and predictive performance on data streams. The performance gains that can be obtained from this are small."

4. EXPERIMENTS & RESULTS:

This section introduces a comparison between different classification algorithms using real datasets: sales transactions database obtained from a retail database. The experiments were run on Pentium *M* computer with a clock rate of 1600 MHz and 256 Mbytes of main memory in Weka software.

The aim of this study was the market basket analysis of purchases by mining association rules on transactional data from a supermarket in order to provide greater insight into the buying behaviour of their customers. Attributes considered in datasets are month, product 1, product 2, product 3, visitor type and weekend .All the values of the attributes are nominal values. The numbers of records are 204 from real time store.

Table 4.1 Classification instances of different classifiers

Classifiers	**Varients**	**Correctly classified instances**	**Incorrectly classified instances**
Bayes	Bayes net	93.1373%	6.8627%
	Naïve bayes	90.1961%	9.8039%
Functions	Multilayer perception	87.7451%	12.2549%
	SGD	91.6667%	8.3333%
Lazy	LWL	92.1569%	7.8431%
	IBK	88.2352%	11.7647%
Rules	Decision Table	93.1373%	6.8627%
	JRIP	94.1176%	5.8824%
Trees	LMT	91.1765%	8.8235%
	J48	90.6863%	9.3137%
Meta	Attribute selected classifier	93.1373%	6.8627%
	Bagging	89.7059%	10.2941%

Table 4.1 displays the number of correct and incorrect classified instances in different algorithms. By considering the above classifiers, it is found that among all the classifiers JRipper has the highest correctly classified instances that is 94.1176% and IBK has found to have the highest incorrectly classified instances as 11.7647%, according to the analysis done, it can be concluded that JRipper has more accurate value.

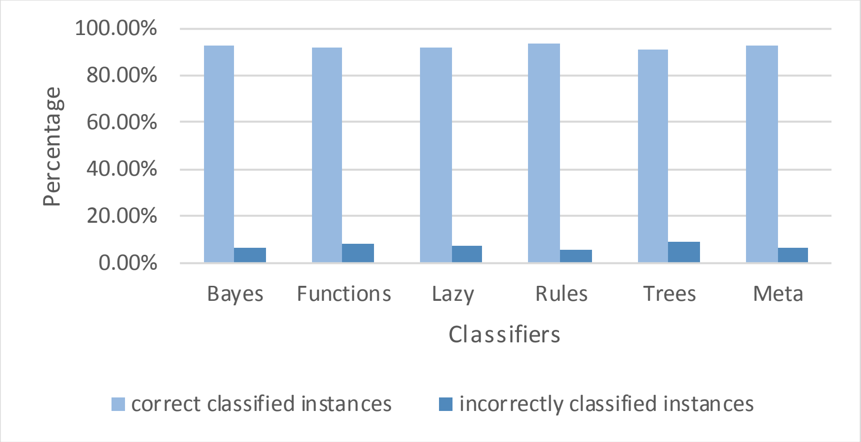

Fig 4.1 Graph of highest variants in a classifier

The evaluation metrics considered are the following:
TP Rate- True positive rate is positives correctly classified/ total positives
FP Rate- false positive rate is negatives in correctly classified/ total negatives
Precision (P) - It is the ration of the predicted positive cases that were correct to the total number of predicted positive cases.

$$P = TP/ (TP+FP)$$

Recall(R) - Recall is determine of completeness. It is the proportion of positive cases that were correctly recognized to the total number of positive cases. It is also known as sensitivity or true positive rate (TPR).

$$R= TP / (TP + FN)$$

F-Measure -The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

$$\text{F-Measure} = (2*\text{recall}*\text{precision})/ (\text{precision} + \text{recall})$$

Receiver Operating Characteristic (ROC) Curve - It is a graphical approach for displaying the trade-off between true positive rate (TPR) and false positive rate (FPR) of a classifier. TPR is plotted along the y axis and FPR is plotted along the x axis. Performance of each classifier represented as a point on the ROC curve.

Table 4.2 gives the details of all the Evaluation metrics on the considered classifiers.

Table 4.2 Performance Analysis

Classifiers	**Variants**	**Performance Metrics**					
		TP Rate	**FP Rate**	**Precision**	**F-Measure**	**ROC Area**	**Class**
Bayes	Bayes Network	0.977	0.939	0.843	0.905	0.569	false
		0.069	0.023	0.333	0.103	0.569	True
		0.828	0.791	0.761	0.775	0.569	
	Naïve Bayes	0.977	0.970	0.839	0.903	0.576	False
		0.030	0.023	0.200	0.053	0.576	True
		0.824	0.817	0.736	0.765	0.576	
Functions	Multilayer Perception	0.936	0.818	0.856	0.894	0.525	False
		0.182	0.064	0.353	0.240	0.525	True
		0.814	0.696	0.774	0.788	0.525	
	Stochastic Gradient Descent	0.965	0.970	0.830	0.897	0.498	False
		0.030	0.035	0.143	0.050	0.498	True
		0.814	0.819	0.725	0.760	0.498	

Lazy	IBK	0.971	0.970	0.838	0.900	0.567	False
		0.030	0.029	0.167	0.051	0.533	True
		0.819	0.818	0.730	0.762	0.562	
	Locally Weighted Learning	0.965	0.970	0.838	0.897	0.537	False
		0.030	0.035	0.143	0.050	0.537	True
		0.814	0.819	0.725	0.760	0.537	
Rules	Decision Tree	1.000	1.000	0.838	0.912	0.472	False
		0.000	0.000	_	_	0.472	True
		0.838	0.838	_	_	0.472	
	JRIP	0.982	1.000	0.836	0.903	0.468	False
		0.000	0.018	0.000	0.000	0.468	True
		0.824	0.841	0.701	0.757	0.468	
Trees	LMT	0.988	1.000	0.837	0.906	0.520	False
		0.000	0.012	0.000	0.000	0.520	True
		0.828	0.840	0.701	0.760	0.520	
	J48	1.000	1.000	0.838	0.912	0.466	False
		0.000	0.000	_	_	_	True
		0.838	0.838	_	_	_	
Meta	Attribute Selected Classifier	1.000	1.000	0.838	0.912	0.466	False
		0.000	0.000	_	_	0.466	True
		0.832	0.832	_	_	0.466	
	Bagging	0.994	0.970	0.842	0.912	0.634	False
		0.030	0.006	0.500	0.057	0.634	True
		0.838	0.814	0.786	0.773	0.634	

5. CONCLUSION:

Classification is the most researched topic. This paper has presented a focus on various classifiers. We also performed an extensive experimental study applying all the above classifiers which are considered to the datasets, with the help of Weka tool the results of all the classifiers used is mostly equal but we are at a conclusion where the highest value for correctly classified instances is 94.12% and that is obtained from JRipper which is a variant of Rules classifier. Although there are many other research topics that have been investigated in the literature, we believe that this selected review has covered the most important aspects of in solving classification problems. It is clear that research opportunities are abundant in many aspects of classifiers. In the future we believe that the multidisciplinary nature of the classification research will generate more research activities and bring about more fruitful outcomes in the future.

6. REFERENCES:

[1]Yen-Liang Chena, Kwei Tangb,*, Ren-Jie Shena, Ya-Han Hua, Market basket analysis in a multiple store environment ,Received 1 December 2003; received in revised form 5 April 2004; accepted 5 April 2004 Available online 2 June 2004

[2]Jochen Hipp, Ulrich Gu¨ntzer, and Gholamreza Nakhaeizadeh,DaimlerChrysler AG, Data Mining of Association Rules and the Process of Knowledge Discovery in Databases Research & Technology, 89081 Ulm, Germany jochen.hipp@daimlerchrysler.com rheza.nakhaeizadeh@daimlerchrysler.com 2 Wilhelm Schickard-Institute, University of Tu¨bingen, 72076 Tu¨bingen, Germany guentzer@informatik.uni-tuebingen.de.

[3]Rakesh Agrawal, Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules, IBM Almaden Research Center 650 Harry Road. San Jose. CA 95120.

[4]Schonrost G B1, Paes V C1, Balestrassi P P1*, Paiva A P1, Campos P H S1,Data Mining Association Rules Applied to Supermarket Transactional Data Modeling: a case study in Brazil, International Joint Conference - ICIEOM-ADINGOR-IISE-AIM-ASEM (IJC 2017) Valencia, Spain, July 6-7, 2017

[5] Chitra Nasa, Suman, Evaluation of different classification techniques for web data, International journal of computer applications(0975-8887) volume 52-No9,August 2012.

[6] Faltin F, Kenett R, Bayesian Networks, Encyclopaedia of statistics in quality & reliability, Wiley & sons (2007).

[7] Mu Li, Tong Zhang, Yuqiang Chen, Alexander j.Smola, Efficient mini-batch training for stochastic optimization, Carnegie mellon university baidu, Inc. Rutgers university Google, Inc.

[8] Hassan Ramchoun, Mohammed Amine janati Idrissi, Youssef Ghanou, Mohamed Ettaouil, Multilayer Perceptron: Architecture Optimisation and Training, Modeling and Scientific Computing Laboratory, Faculty of Science and Technology, University Sidi Mohammed Ben Abdellah, Fez, Morocco.

[9] A B M Shawkat Ali, Performance Analysis of Statistical Classifier SMO with other data mining Classifiers, School of Computing and Information Technology, Monash University, Victoria 3842, Australia.

[10] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal, Locally Weighted Learning, ATR Human Information Processing Research Laboratories 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan.

[11] C. Lakshmi Devasena, T. Sumathi, V.V. Gomathi and M. Hemalatha, Effectiveness Evaluation of Rules Based Classifiers for the Classification of iris data Set,Bonfring international Journal of man machine interface, Vol 1,Special Issue, December 2011.

[12] Purva Sewaiwar, Kamal Kant Verma, Comparitive study of various decision tree classification algorithm using weka, International journal of emerging researching management and technology ISSN: 2278-9359

[13] Dr Gnanambal S, Dr thangaraj M, Dr Meenatchi V.T, Dr Gayathri v, Classification algorithms with attribute selection: an evaluation study using weka, Int.J.Advanced networking and applications.

[14] Jan N. van Rijn, Geoffrey Holmes, Bernhard Pfahringeer, Joaquin Vanschoren, Case study on Bagging stable classifiers for data streams.

[15] Agrawal R., Imilienski T., and Swami A., “Mining Association Rules Between Sets of Items in Large Databases”, Proceedings of the ACM SIGMOD Conference, pp. 207-216, 1993.