ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 3(B), pp. 24757-24762, March, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

BUILDING A MODEL FOR IDENTIFICATION OF ONLINE RETAIL E-COMMERCE POTENTIAL CUSTOMERS BY USING DATA MINING CLASSIFICATION TECHNIQUES

Sridevi D[1]., Pandurangan A[2]., Gunasekran S[3] and Kumaravel A[4]

[1]Department of Computer Applications, Valliammai Engineering College, Chennai, Tamilnadu
[2]Department of Mathematics, Bharath University, Chennai, Tamilnadu
[3]Department of Applied Research, Gandhigram Central University, Dindugal, Tamilnadu
[4]Department of Information &Technology, Bharath University, Chennai, Tamilnadu

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0903.1723

ARTICLE INFO

Article History:
Received 15th December, 2017
Received in revised form 25th January, 2018
Accepted 28th February, 2018
Published online 28th March, 2018

Key Words:

Data Mining; Web Mining; E-Commerce; WEKA; Classification Algorithms

ABSTRACT

Web plays a vital role in our day to day activities. If we are in need of any thing we need not go and buy when we are busy with our works. Online shopping is becoming more and more common in our daily lives. All made simple by this online e-commerce. From their place they click, they are able to receive their material in the hands. This reduces the work tension, minimises time, low cost and Customers can easily select products from different providers without moving around physically. The big industries started rethinking their business. Defined business can be found by carrying E-business. Web mining has become more popular and widely used in various applications. The E-commerce techniques run more efficiently with the application of data mining techniques, especially web mining is considered as the best.

INTRODUCTION

Data mining is a process that involves requirements gathering, information accumulation, database storage, business intelligence, and deployment .In simple the Data mining is said to be large volume of data-value-based, operational and other non conventional types of data is altogether gathered and stored in database warehouses. This information should be preprocessed, cleaned and changed before it can be mined. Most of the commonly used data mining techniques are classification, clustering, association analysis and regression analysis as stated in J. Han and M Kamber, "Data mining concepts and techniques", San Francisco, USA, Morgan Kaufmann, 2001.

Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data. As the name proposes, this is information gathered by mining the web. It makes utilization of automated apparatuses to reveal and extricate data from servers and web2 reports, and it permits organizations to get to both organized and unstructured information from browser activities, server logs, website and link structure, page content and different sources. Web mining can be divided into three different types-Web usage mining, Web content mining and Web structure mining [2].

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web structure mining uses graph theory to analyze the node and connection structure of a web site. Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content as prescribed in https://en.wikipedia.org/wiki/Web_mining

The e-Commerce industry utilizes data mining technology. E-commerce is the utilization of data and communication technology through the Internet platform to share business data, keep business connections, and lead business transactions. The increase in the cost of drawing in new customers on the Internet and the relative trouble in holding them make customer loyalty a fundamental resource for many e-commerce businesses. In the traditional (non-internet) marketplace,

Corresponding author:* **Sridevi D
Department of Computer Applications, Valliammai Engineering College, Chennai, Tamilnadu

customer loyalty is fundamentally the result of excellent service quality and the trust that is established for a particular business organization. Establishing online customer loyalty and maintaining existing customers is the necessity for many e-commerce retailers.

Businesses are moving swiftly to influence the internet to market and sell their products through virtual stores or websites. The success of an enterprise is directly linked to e-commerce quality, and for businesses to remain competitive, evaluation of their e-commerce quality is important. The framework for E-commerce Data Analysis is shown in Fig 1.

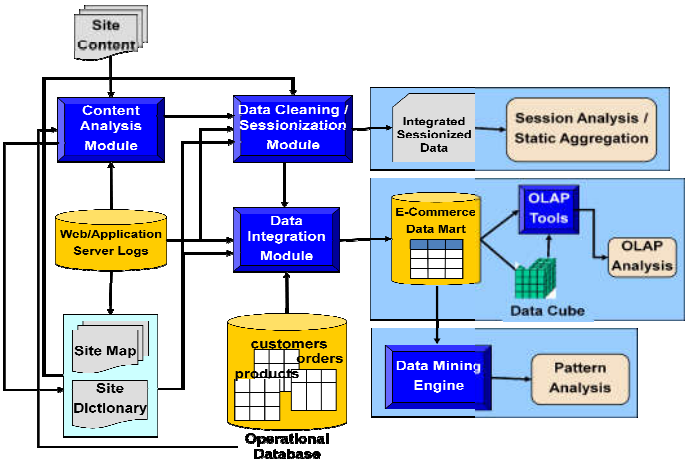

Fig 1 Basic Frame work for E-Commerce Data Analysis

Classification

Data mining techniques can be classified into both unsupervised and supervised learning techniques. Unsupervised learning technique is not guided by variable and does not create a hypothesis before analysis. Based on the results, a model will be built.

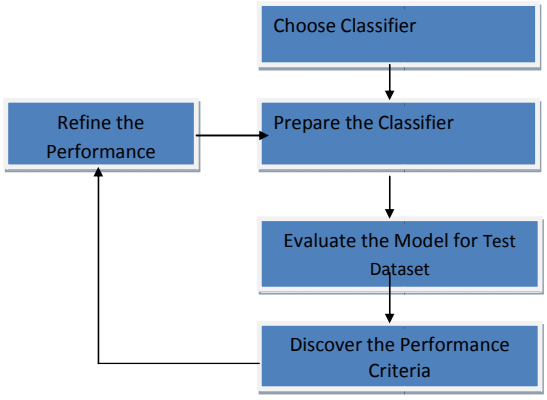

Fig 2 Classification Steps

Classification is the most commonly applied data mining technique, and employs a set of pre-classified examples to develop a model that can classify the population of records at large. The major goal of the classification technique is to predict the target class accurately for each case in the data. There are several classification mechanisms that are used in analyzing online retail data. These include Naive Bayes Algorithm, SMO classifier, Bagging, LogitBoost and J48.

Naïve Bayes

The NaiveBayes Classifier technique is based on Bayesian theorem and is particularly used when the dimensionality of the inputs is high. The Bayesian Classifier is capable of calculating the most possible output based on the input. It is also possible to add new raw data at runtime and have a better probabilistic classifier. A NaiveBayes classifier considers that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given. For example, a fruit may be considered to be an apple if it is red, round. Even if these features depend on each other or upon the existence of other features of a class, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple as prescribed by Bhoomi Trivedi *et al* (2012)

SMO Classifier

SMO Classifier- Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. Bhoomi Trivedi *et al* (2012) states SMO is widely used for training support vector machines and is implemented by the popular libsvm tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers. 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. Because of the linear equality constraint involving the Lagrange multiplier, the smallest possible problem involves two such multipliers. The algorithm proceeds as follows: Find a Lagrange multiplier that violates the

- Karush–Kuhn–Tucker (KKT) conditions for the optimization problem. Pick a second multiplier and optimize the pair.
- Repeat steps 1 and 2 until convergence.
- When all the Lagrange multipliers assure the KKT conditions, the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair off of multipliers so that it can accelerate the rate of convergence.

Bagging

Given a set, D, of tuples, bagging works as follows. For iteration i (i = 1,2…k), a training set, Di, of d tuples is sampled with replacement from the original set of tuples, D. Note that the term bagging stands for bootstrap aggregation. Each training set is a bootstrap sample. Because sampling with replacement is used, some of the original tuples of D may not be included in Di, whereas others may occur more than once. A classifier model, Mi, is learned for each training set, Di. To classify an unknown tuple, X, each classifier, Mi, returns its class prediction, which counts as one vote. The bagging can be applied to the prediction of continuous values by taking the average value of each prediction for a give test tuple. The bagged classifier often has significantly greater accuracy than a single classifier derived from D, the original training data. It

will not be considerably worse and is more robust to the effects of noisy data. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers. For prediction, it was theoretically proven that a bagged predictor will always have improved accuracy over a single predictor derived from it as prescribed by BhoomiTrivedi *et al* (2012).

LogitBoost

Boosting needs weight. But can adapt learning algorithm or can apply boosting without weights and resample with probability determined by weights. Training error decreases exponentially and works if base classifiers are not too complex. Maximizes probability, if base learner minimizes squared error. LogitBoost optimizes probability/likelihood instead of exponential loss .It can be adapted to multiclass problems Shrinking and cross validation based selection apply.

J48

J48 are the improved versions of C4.5 algorithms or can be called as optimized implementation of the C4.5. The output of J48 is the Decision tree. A Decision tree is similar to the tree structure having root node, intermediate nodes and leaf node. Each node in the tree consist a decision and that decision leads to our result. Decision tree divide the input space of a data set into mutually exclusive areas, each area having a label, a value or an action to describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node given by Bhoomi Trivedi *et al* (2012).

Tool

Weka is Java based open source data mining tool. It is easy to use for beginners and has the ability of running several learning algorithms and comparing.

Features

- It is platform independent.
- It performs various data mining tasks including
- Data pre-processing, Classification rules, regression, Clustering, association rules, visualization, feature selection and improving the knowledge discovery.

Weka has 49 Data pre-processing tools, 76 Classification/regression algorithms, 8 Clustering algorithms, 3 algorithm for finding association rules, 15 attribute/subset evaluator plus 10 search algorithms for feature selection as stated by D.Sridevi *et al* (2017).

- There are various built in features.
- There is no programming and coding language required.

Advantages

- Easy to manipulate the data.
- Provide access to SQL databases.
- It provides two options for the user to interact through Explorer and Command line as stated by D.Sridevi *et al* (2017).
- It provides various machine learning algorithms for data mining tasks.
- It supports various standard Data mining tasks that include: Data pre-processing, Clustering and Classification, Regression, Visualization and Feature selection as stated by D.Sridevi *et al* (2017).

Weka Explorer

Weka Explorer has six (6) tabs, which can be used to perform tasks such as pre-process, classify, associate etc. as shown in fig 3

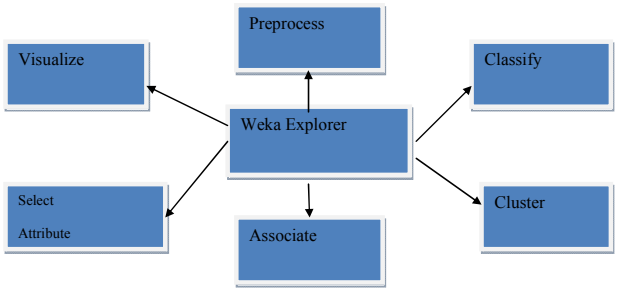

Fig 3 Six tabs of Weka Explorer

Preprocess

Pre-processing is one of the important and prerequisite step in data mining. Feature selection (FS) is a process to select features which are more informative but some features may be redundant, and others may be irrelevant and noisy as stated by C. Velayutham and K. Thangavel,(2011). When the data set consists of meaningless data that is incomplete (missing), noisy (outliers) and inconsistent data, pre-processing of the dataset is required.

For the first 3 ways of pre-processing we have option of "filter" in WEKA. In filter option itself there are two types of filters: Supervised and unsupervised. In both the categories we have filters for attributes and instances separately. After data cleaning, integration and transformation the data reduction is performed to get the task relevant data. For data reduction we have "Attribute Selection" option. It consists of various types of feature selection programs for wrapper approach, filter approach and embedded approach.

Classify

Classify tools can be used to perform further analysis on preprocessed data. If the data demands a classification or regression problem, it can be processed using Classify tab. A classification model produced on the full trained data. Weka consists of all major learning techniques for classification and regression: Bayesian classifiers, decision trees, rule sets, support vector machines, logistic and multi-layer perceptrons, linear regression, and nearest-neighbor methods. It also contains "meta-learners" like bagging, stacking, boosting, and schemes that perform automatic parameter tuning using cross-validation, cost-sensitive classification, etc. Learning algorithms can be evaluated using cross-validation or a hold-out set, and Weka provides standard numeric performance measures (e.g. accuracy, root mean squared error), as well as graphical means for visualizing classifier performance (e.g. ROC curves and precision-recall curves). It is possible to visualize the predictions of a classification or regression model, enabling the identification of outliers, and to load and save models that have been generated as prescribed by Bouckaert RR *et al* (2013).

Cluster

Weka contains "clusters" for finding groups of instances in datasets. Cluster tools give access to Weka's clustering algorithms, such as k-means, a heuristic incremental hierarchical clustering scheme. Cluster assignments can be visualized and compared to actual clusters, defined by one of the attributes in the data as prescribed by Bouckaert RR *et al* (2013).

Associate

Associate tools have generating association rules algorithms. It can be used to identify relationships between groups of attributes in the data as prescribed by Bouckaert RR *et al*(2013).

Select attributes

More interesting in the context of bioinformatics is the fifth tab, which offers methods for identifying subsets of attributes that are predictive of target attribute in the data. Weka contains several methods for searching through the space of attribute subsets, evaluation measures for attributes and attribute subsets. Search methods such as a best-first search, genetic algorithms, forward selection, and attributes ranking. Different search methods and evaluation methods both may be combined, making the system very flexible as prescribed by Bouckaert RR *et al*(2013).

Visualize

Visualization tools show a matrix of scatter plots. Practically visualization is very much useful which helps to determine learning problem difficulties. Weka visualizes single dimension (1D) for single attributes and double dimension (2D) for pairs of attributes. It is to visualize the current relation in 2D plots. Any matrix element can be selected and enlarged in a separate window, where one can zoom in on subsets of the data and retrieve information about individual data points. A "Jitter" option to deal with nominal attributes for exposing obscured data points is also provided as prescribed by Bouckaert RR *et al*(2013).

Preprocess of Data using Weka

Data set used in Weka is in Attribute-Relation File Format (ARFF) file format that consist of special tags to indicate different things in the dataset such as attribute names, attribute types, attribute values and the data. This paper includes the two data sets such as onlineretail.csv and supermarket.arff. Onlineretail.csv taken from UCI repository while supermarket.arff is taken from weka tool website as referred from UCI repository. Onlineretail.csv data set is in the form of text file. Firstly it converts into the .xls format; .xls format to .csv

In this paper, five different classifiers has used for the classification of data. These techniques are applied on two dataset in which one of data set has 730 instances and one third attribute as compare to another data set contains 4627 instances. The fundamental concept to take two datasets is to analyze the performance of the discussed classifiers for small as well as large dataset. To analyze the performance of discussed classifiers, six different parameters are used i.e. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Time Taken, Correctly Classified, Incorrectly Classified instance and Kappa Statistic.

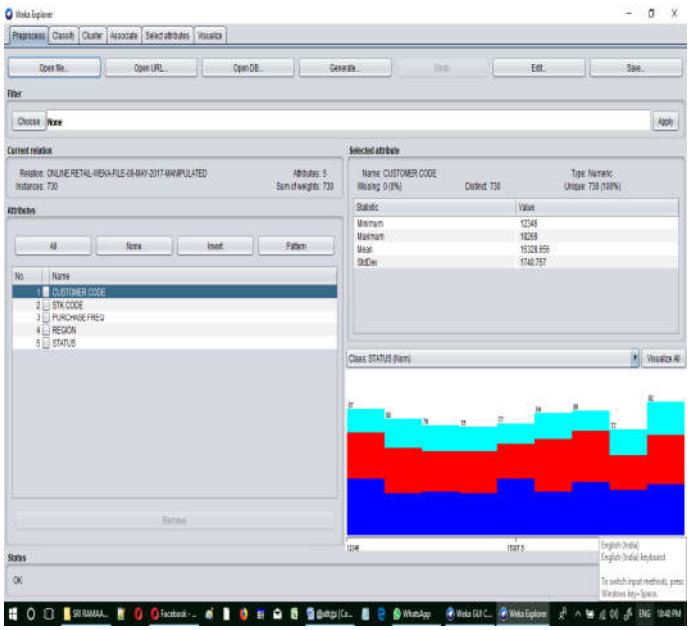

Fig 4 Preprocess of Data using Weka

Classification is a data mining (machine learning) technique used to predict group membership for data instances as stated by Thair Nu Phyu (2009). It is the problem of finding the model for class attribute as a function of the values of other attributes and predicting accurate class assignment for test data. It can be divided in two types: supervised and unsupervised. Supervised is further divided in probabilistic and geometric. Probabilistic is further divided in parametric and nonparametric type. Classification is a two step process: first is model construction i.e. describing a set of predetermined classes and second is using that model for prediction i.e. classifying future or unknown objects.

For the Classification in Weka, we have supervised and unsupervised categories of classifiers. All the classifiers like lazy, tree, rules and naïve comes under these categories only. Meta classifiers are also there to enhance the accuracy of classifiers using various ensemblers.

The performance criteria for evaluating the classifiers are: classification accuracy, specificity, sensitivity/recall, precision, AUROC curve, kappa statistics, mean absolute error, root mean squared error, Relative absolute error, root relative squared error, Time.

RESULTS AND DISCUSSION

In this paper, the following parameters are used to evaluate the performance of above mentioned classification techniques

Classification accuracy: It is the ability to predict categorical class labels. This is the simplest scoring measure. It calculates the proportion of correctly classified instances.

Specificity*:* Specificity relates to the classifier's ability to identify negative results. It is also called true negative rate.

Sensitivity/Recall: Sensitivity is the proportion of actual positives which are correctly identified as positives by the classifier. It is also called true positive rate.

Precision: This is a measure of retrieved instances that are relevant.

AUROC curve: It is the graph between false positive and true positive rate. The area measures discrimination, that is, the ability of the classifier to correctly classify the test data.

Kappa Statistics: The kappa measure of agreement is the ratio

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times the k raters agree i.e. percentage agreement between classifier and ground truth, and P(E) is the proportion of times the k raters are expected to agree by chance alone i.e. the chance agreement. K=1 indicates perfect agreement and K=0 indicates chance agreement. The value greater than 0 means classifier is doing better. Higher the kappa statistic value betters the classifier result as given in Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

Mean absolute error (MAE): Mean absolute error is a calculation of how close predictions are to reality as given in Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/.

Root mean squared error: Small value of RMSE means better accuracy of model. So, minimum of RMSE & MAE better is prediction and accuracy as given in Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

Fig.5 shows the comparison of mean absolute error parameter for large dataset i.e. 4627 having 30 attributes as well as small dataset i.e. 730 instance having 7 attributes. The analysis of MAE parameter according to Fig. 5 shows that LogitBoost provide better result in case of small dataset and in large set. This parameter states that minimum of MAE tends to better performance of the classifiers because this parameter measure the difference between the predicted value and actual value.

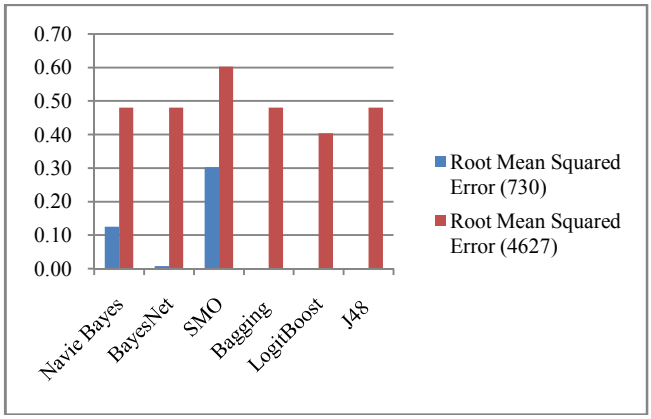

Fig 6 Comparison of Root Mean Squared Error Parameter

The analysis of RMSE parameter which has shown in Fig. 6 describes that the model formed by Bagging, LogitBoost and J48 classifier is better i.e. 0.00 than others. Because the minimum value of RMSE, better is prediction.

Table 1 Tabular Comparison of the different Classifiers

Algorithm (Total Instance : 730/4627)	Correctly Classified Instances % (Value)		Incorrectly Classified Instances % (Value)		Time Taken (Sec)		Kappa Statistic		Mean Absolute Error		Root Mean Squared Error	
NaiveBayes	96.44	63.71	3.56	36.29	0.01	0.08	0.94	0.00	0.05	0.46	0.13	0.48
BayesNet	100.00	63.71	0.00	36.29	0.02	0.29	1.00	0.00	0.00	0.46	0.01	0.48
SMO	92.19	63.71	7.81	36.29	0.15	0.38	0.88	0.00	0.24	0.36	0.30	0.60
Bagging	100.00	63.71	0.00	36.29	0.06	0.91	1.00	0.00	0.00	0.46	0.00	0.48
LogitBoost	100.00	76.03	0.00	23.97	0.17	0.58	1.00	0.46	0.00	0.31	0.00	0.40
J48	100.00	63.71	0.00	36.29	0.03	0.09	1.00	0.00	0.00	0.46	0.00	0.48

Table 1 shows the comparison of the BayesNet, NaiveBayes, SMO, Bagging, LogitBoost, and J48 in tabular form. For the analysis of discussed classifiers the two data sets has been used in which UCI online retail dataset has 730 instance and 7 attributes while the super market data set has 4627 instance and 30 attributes. The aim to take two dataset is to analyze the performance of discussed classifiers with large as well as small data set. Fig. 5 shows that LogitBoost provide better result in case of small dataset and in large set

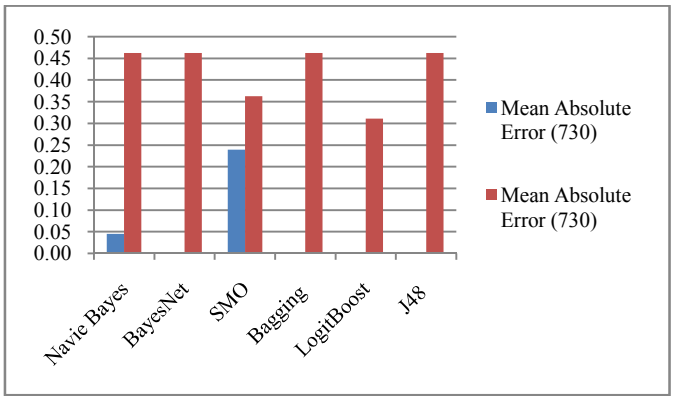

Fig 5 Comparison of Mean Absolute Error Parameter

While BayesNet provides poor performance for small dataset i.e. 0.13 and SMO provides worst result with large dataset i.e. 0.60. The performance of Bayes classifiers with small dataset is almost same. Hence it is conclude that neural network classifiers provide worst result among these classifiers and tree classifiers provide best result with RMSE parameter.

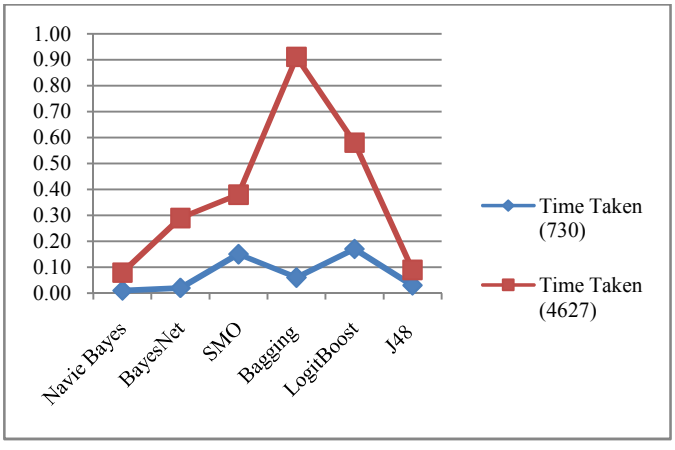

Fig 7 Comparison of Time Taken Parameter

The interpretation of fig 7 shows that Bayes & Tree classifiers have almost same performance whether the dataset is large or small. These models take less time to build the model for both of data sets. But, Bagging, LogitBoost and SMO take largest time to build the model. Hence analysis of time parameter states that Bagging, LogitBoost and SMO classifiers have poor performance with both of dataset.

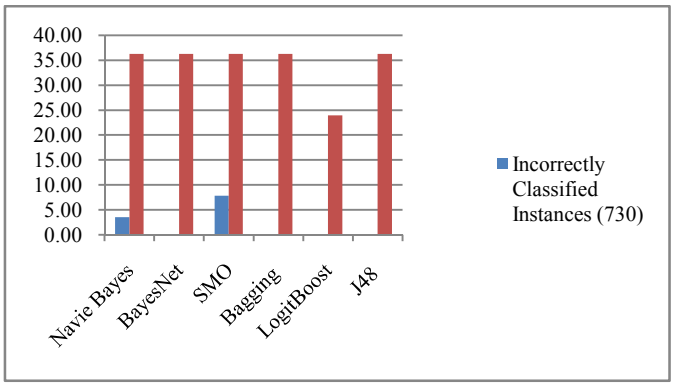

Fig 8 Comparison of Incorrectly Classified Instances Parameters

The comparison of incorrectly classifiers show in Fig. 8 which states that LogitBoost model is provide the best performance with both of dataset i.e. 0.00 & 23.97. While SMO provide poor performance when dataset is small i.e.7.81 and provide poor performance when data set is large i.e. 36.29.

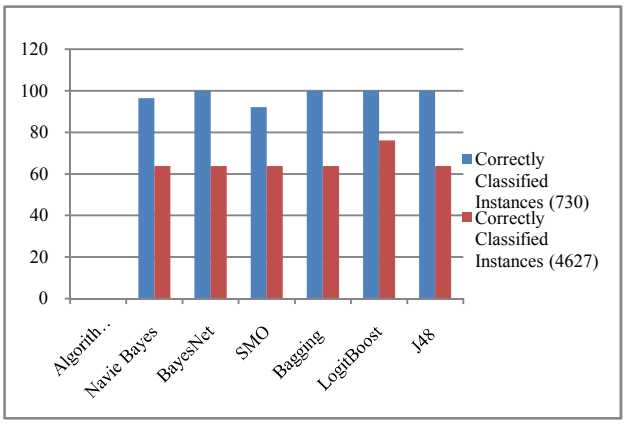

Fig 9 Comparison of Correctly Classified Instances Parameter

Fig. 9 provide the comparison of correctly classified parameter in which LogitBoost classifier classified the instance more correctly whether data set is large as well as small i.e.100 & 76.03,While SMO provide worst result among all these classifiers when data set is small (730 instance) i.e. 92.19 and worst result when dataset is large (4627) i.e. 63.71.

The interpretations of kappa statistic parameter state that value greater than zero means classifiers doing better. Hence, according to kappa statistic parameter that has done in Fig. 10, Fig. 5 shows that LogitBoost provide better result in case of small dataset and in large set So, it is easily concluded that performance of LogitBoost classifier is better in case of kappa statistic parameter.

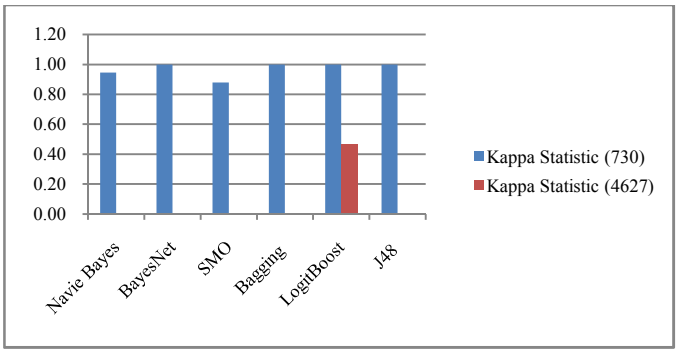

Fig 10 Comparison of Kappa Statistic Parameter

CONCLUSION

In this paper, five different classifiers has used for the classification of data. These techniques are applied on two dataset in which one of data set has one sixth of instance and one third attribute as compare to another data set. After analyze the result of discussed classifiers on the behalf of parameter it concludes that LogitBoost classifier provides better performance (RMSE, MAE, Time Taken, Kappa Statistic, Incorrectly classified instance, correctly classified instance) among all these classifiers for large as well as small dataset. while the SMO classifier provide poor performance in case of MAE, RMSE, time taken, Incorrectly Classified instance, Correctly classified instance and kappa statistic. Hence, it is conclude that Meta classifiers provide better performance among all these classifiers.

References

Bhoomi Trivedi, Ms.Neha Kapadia, INDUS institute of Eng&Tech, TCET, Kandivali(E), Ahmedabad Modified Stacked generalization with Sequential Learning. TCET 2012 on IJCA.

Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, Scuse D. (2013). WEKA Manual for Version 3-7-8.

J. Han and M Kamber, "Data mining concepts and techniques", San Francisco, USA, Morgan Kaufmann, 2001.

D.Sridevi, Dr.A.Pandurangan, Dr.S.Gunasekaran & Dr.A.Kumaravel," Application of Ensemble Design for On-Line Retail E-Commerce for the better Customer Response", *International Journal of Computational Research and Development*, Volume 2, Issue 1,Page Number 102-107,2017

Thair Nu Phyu, "Survey of classification techniques in data mining"; Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I

C. Velayutham and K. Thangavel, "Unsupervised Quick Reduct Algorithm Using Rough Set Theory", *Journal of Electronic Science And Technology,* Vol. 9, NO. 3, September 2011

https://en.wikipedia.org/wiki/Web_mining

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

****