Journal of Software Engineering and Applications, 2015, 8, 470-477
Published Online September 2015 in SciRes. http://www.scirp.org/journal/jsea
http://dx.doi.org/10.4236/jsea.2015.89045

Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study

S. Olalekan Akinola, O. Jephthar Oyabugbe

Computer Science Department, University of Ibadan, Ibadan, Nigeria
Email: solom202@yahoo.co.uk, gentrukky@gmail.com

Received 21 June 2015; accepted 13 September 2015; published 16 September 2015

Abstract

Two important performance indicators for data mining algorithms are accuracy of classification/prediction and time taken for training. These indicators are useful for selecting best algorithms for classification/prediction tasks in data mining. Empirical studies on these performance indicators in data mining are few. Therefore, this study was designed to determine how data mining classification algorithm perform with increase in input data sizes. Three data mining classification algorithms—Decision Tree, Multi-Layer Perceptron (MLP) Neural Network and Naïve Bayes—were subjected to varying simulated data sizes. The time taken by the algorithms for trainings and accuracies of their classifications were analyzed for the different data sizes. Results show that Naïve Bayes takes least time to train data but with least accuracy as compared to MLP and Decision Tree algorithms.

Keywords

Artificial Neural Network, Classification, Data Mining, Decision Tree, Naïve Bayesian, Performance Evaluation

1. Introduction

A large volume of data is poured into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume emanates as a result of the computerization of our society and the fast development of powerful data collection and storage tools [1].

Data mining is used for the extraction of information (patterns, relationships, or significant statistical connec-

How to cite this paper: Akinola, S.O. and Oyabugbe, O.J. (2015) Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study. *Journal of Software Engineering and Applications*, **8**, 470-477.
http://dx.doi.org/10.4236/jsea.2015.89045

tions) from very large databases or data warehouses [2]. Data mining is a powerful technology that converts raw data into an understandable and actionable form, which can then be used to predict future trends or provide meaning to historical events [3]. The extraction of hidden knowledge, exceptional patterns and new findings from huge databases is considered as the key step of a detailed process called Knowledge Discovery in Databases (KDD) which in other words is defined as the non-trivial process of identifying valid, novel, and ultimately understandable patterns in large databases [4]. Data mining consists of more than collection and managing data; it also includes analysis and prediction.

Efficiency and scalability are always considered when comparing data mining algorithms. Data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams. In other words, the running time of a data mining algorithm must be predictable, short, and acceptable by applications. Efficiency, scalability, performance, optimization, and the ability to execute in real time are key criteria that drive the development of many new data mining algorithms [1].

This study was set out to empirically study the performance of three classification algorithms in terms of the times taken for training and accuracies of their predictions. The algorithms in question are Decision Tree (DT), Multi-Layer Perceptron (MLP) Neural Network and Naïve Bayes.

In the rest of this paper, related works are highlighted in Section 2 while methodology adopted is discussed in Section 3. The results obtained from the experiment are discussed in Section 4 while conclusion is drawn in Section 5.

2. Related Works

Classification has been identified as an important problem in the emerging field of data mining. Over the years, there has been quite a number of tremendous studies on classification algorithms [2] [4], analysis of classification techniques [5] [6], performance evaluation [7]-[10], comparisons and evaluations of different data mining classification algorithms [11]-[14] alongside their applications in solving real world problems such as in the areas of medicine, engineering, business etc.

While classification is a well-studied problem, in recent times there has been focus on algorithms that can handle large databases. Applications of classification arise in diverse fields, such as retail target marketing, customer retention, fraud detection and medical diagnosis. Several classification models have been proposed over the years, such as Artificial Neural Networks (ANNs), statistical models, decision trees and genetic models [15]. Classification is a classic data mining technique based on machine learning [12]. It is a model finding process that is used for portioning the data into different classes according to some constraints. In other words we can say that classification is the process of generalizing the data according to different instances [2]. Basically, classification is used to classify each item in a set of data into one of predefined set of classes or groups [12]. The conventional models used for classification are decision trees, neural network, statistical and clustering techniques [4].

Scalability implies that as a system gets larger, its performance improves correspondingly. Data mining scalability connotes taking advantage of parallel database management systems and additional CPUs as one can solve a wide range of problems without needing to change the underlying data mining environment [16]. Scalability is concerned with how to efficiently handle large databases which may contain millions of data items. For an algorithm to be scalable, Huidong [17] opined that its running time should grow linearly in proportion to the size of the database, given the available system resources such as main memory and disk space.

A Neural Network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples.

Decision trees are powerful and popular for both classification and prediction. They are also useful for exploring data to gain insight into the relationships of a large number of candidate input variables to a target variable [19]. It is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset [20]. Each branch represents an outcome of the test and the leaf nodes represent classes or class distributions. Unknown samples can be classified by testing attributes against the tree. The path traced from root to leaf holds the class prediction for that sample [21].

Naive Bayesian is a simple but important probabilistic model, because the Naïve Bayesian classifiers are statistical classifiers. In simple terms, a naive Bayesian classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The Naïve Bayesian classifier is one of the most popular data mining techniques for classifying large dataset. The classification task is to map the set of attributes of sample data onto a set of class labels, and naïve Bayesian classifier particularly suitable as proven universal approximates [22].

Daniela, Christopher and Roger [11] compared Neural Networks (NN), Naïve Bayes (NB) and Decision Tree (DT) classifiers for the automatic analysis and classification of attribute data from training course web pages. In this study Naïve Bayesian classifier was shown to be the best choice for the training courses domain, achieving an impressive F-Measure value of over 97%, despite it being trained with fewer samples than any of the classification systems.

Abirami, Kamalakannan and Muthukumaravel [6] analyzed several data mining classification techniques (including Naïve Bayesian, ID3 and C4.5 algorithms) using WEKA machine learning tools over the healthcare datasets. Different data mining classification techniques were tested on two heart disease datasets. The standards used were percentage of accuracy and error rate of every applied classification technique. They recommend that the technique, which is suitable for a particular dataset is chosen based on highest classification accuracy rate and least error rate.

Gopala, Bharath, Nagaraju and Suresh [9] made a comprehensive comparative analysis of 14 different classification algorithms for their performance using 3 different cancer data sets. Their results indicate that none of the classifiers outperformed all others in terms of accuracy. Most of the algorithms performed better as the size of the data set is increased.

Anshul and Rajni [12] carried out a performance evaluation of Naïve Bayesian and J48 classification algorithm for a financial institute dataset to maximize true positive rate and minimize false positive rate of defaulters using WEKA tool. The results on the dataset showed that the efficiency and accuracy of J48 and Naive Bayesian are good.

Performance evaluation is a multi-purpose tool used to measure actual performance against expected performance. Evaluating the performance of a data mining technique is a fundamental aspect of machine learning. Evaluation method is the yardstick to examine the efficiency and performance of any model.

3. Methodology

3.1. Hardware and Operating System (OS) Platform Used

The versions of OS used in this evaluation study was Windows 8.1. This was the latest version of Windows OS as at the time of this study. **Table 1** gives the specifications of the hardware and Operating System platforms used.

3.2. Software Tool

Waikato Environment for Knowledge Analysis (WEKA) data mining tool (version 3.6.11) was used for the experiments. Different characteristics of the application using classifiers to measure accuracy, performance metrics and time taken to build models considering different data sizes of the dataset were explored.

3.3. Data Source

The source of data for this study was from a simulated data. An application program using Java Programming Language was developed to simulate Ebola disease data. The simulated data were stored in a MySQL database. The Ebola dataset has its own properties like the number of instances, the number of attributes and number of classes.

3.4. Data Set

The Ebola disease dataset used for the tests was from an anonymous simulated data. The dataset consists of 250 to 10,000 instances (records) with nine attributes (representing symptoms). Each of the attributes being reclassified as 0 for "No" and 1 for "Yes". The target variable (that is, "Remark") consists of two classes: "Yes" for positive to Ebola and "No" for negative to Ebola. The sample structure of the Ebola Disease data set is shown in **Table 2**.

Table 1. Hardware platform with corresponding windows operating system version.

Computer System	Specifications
HP 250	Processor: Intel (R) Pentium (R) CPU N3510 @ 1.99GHz Memory (RAM): 1.9GB System Type: 64-bit Operating System, x64-based processor HDD: 320Gb Windows: Windows 8.1 Single Language

Table 2. The sample structure of the Ebola disease dataset with possible test result.

Fever	Nausea	Headache	Tiredness	Vomiting	Diarrhea	Coughing	Bleeding	Remark
0	0	1	1	0	1	1	1	Yes
0	1	0	1	1	1	0	0	No
1	0	1	0	0	0	0	0	No
0	0	1	0	0	1	0	0	No

3.5. The Experimental Classification Algorithms

The Ebola disease dataset was experimented with three classification algorithms: Decision Tree (J48), Naïve Bayesian (Naïve Bayes) and Artificial Neural Network (ANN, Multilayered Perceptron). Each algorithm was trained with the Ebola Disease data using 66% split and Cross-Validated with 10 Fold option. The training was carried out with respect to different data sets: 250, 500, 1000, 2000, 3000, 3500, 4500, 5000 and 10,000.

3.6. Performance Metrics

Two performance metrics: time to build model (Training Time) and percentage accuracy (correct classifications) were obtained for each of the data sets using the three classification algorithms. The performances were then compared statistically using *Analysis of Variance*, (ANOVA) and simple correlations.

4. Results

4.1. Time Taken for Trainings by the Algorithms

Figure 1(a) and Figure 1(b) show the performance of the three classification algorithms used in the experiments: Decision Tree (J48), Naïve Bayes and Multi-Layer Perceptron (MLP), with respect to their time taken for the different data sizes.

Figure 1(b) was drawn for J48 and Naïve Bayes to show their performances distinctly since they are somehow overlapped Figure 1(a).

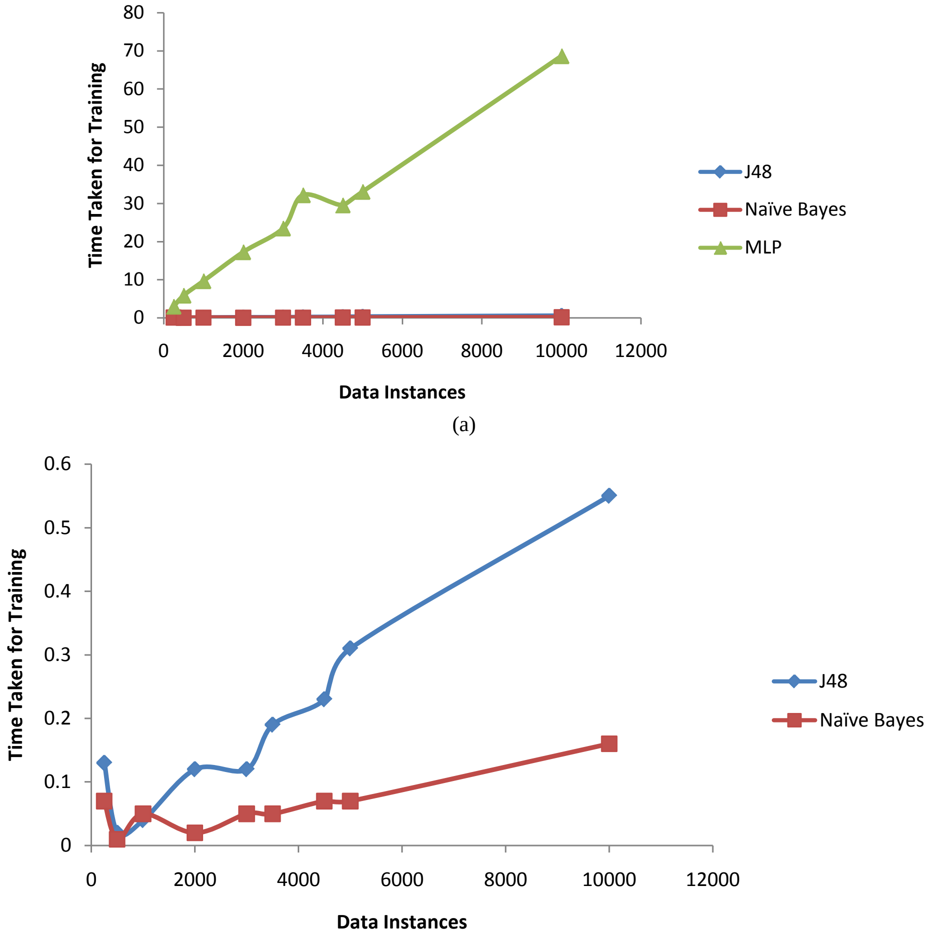
From Figure 1 (a) and Figure 1(b), it could be inferred that as data sizes were increasing, Naïve Bayes classification algorithm's time complexity was the least, followed by J48 (Decision Tree) and ANN (Multi-Layer Perceptron Neural Network) in that order. This means that MLP takes highest times for each of the data instances than the J48 Decision Tree and Naïve Bayes Classifiers.

Correlations of the Algorithms' Training Times with Increasing Data Size

The rank correlation coefficients of the three algorithms between data sizes and time used for trainings are shown in Table 3.

Table 3 shows that there were strong positive correlations between the algorithms training times and data sizes. As data size increases, so also the training times. MLP and J48 however showed higher positive correlations

ANOVA Result showed a high significant difference in the time complexities among the three algorithms (F = 13.669 and p = 0.0, where p is the level of significance). Further Tukey HSD test indicates that there were significant differences in the time complexities between MLP and J48 (p = 0.0), MLP and Naïve Bayes (p = 0.0) while J48 and Naïve Bayes had no significant differences in their time complexities (p = 1.0). The mean difference was significant at the 0.05 level.



(a)

(b)

Figure 1. (a) Time taken for training against the data instances for all Algorithms; (b) Time taken for training against the data instances for J48 and Naïve Bayes.

Table 3. Rank correlation coefficients (training time versus data size).

Algorithm	Correlation
J48	0.96
Naive Bayes	0.85
MLP	0.99

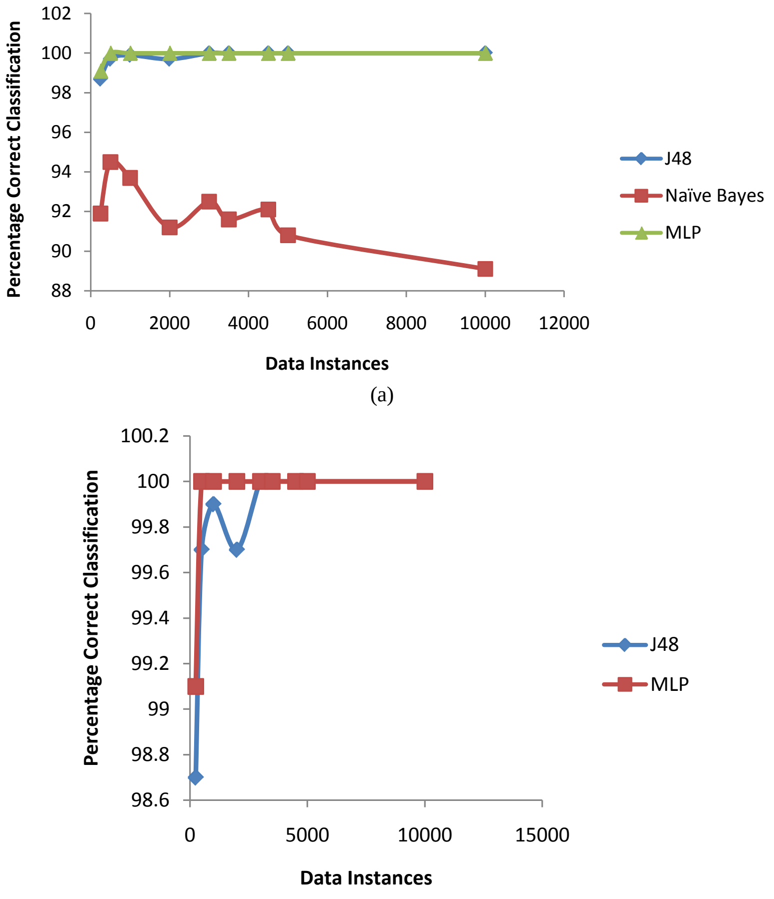
4.2. Percentage Accuracies of the Algorithms with Increase in Data Size

Figure 2(a) and Figure 2(b) show the performance of the three classification algorithms with respect to their correct classifications (accuracies) for the different data sizes.

Figure 2(a) shows that as the data instances (sizes) increased, the percentage classification correctness (accuracy) of Naïve Bayes reduces. However, J48 and MLP showed high accuracies with low data sizes. At a higher data sizes, J48 and MLP's percentage accuracies became stable at 100%. Figure 2(b) gives an elaborate chart for the J48 and MLP algorithms.

Correlations of the Algorithms' Percentage Correct Classifications with Increasing Data Size

The rank correlation coefficients of the three algorithms between data sizes and percentage correct classifications are shown in Table 4.



(a)

(b)

Figure 2. (a) Percentage correct classifications versus data instances for all algorithms; (b) Percentage correct classifications versus data instances for J48 and MLP.

Table 4. Rank correlation coefficients (accuracy versus data size).

Algorithm	Correlation
J48	0.53
Naive Bayes	−0.82
MLP	0.38

Table 4 shows that Naïve Bayes algorithm demonstrated a very strong negative correlation. As the data size increases, the classification accuracy of Naïve Bayes algorithm decreases. Weak positive correlations were obtained with MLP while J48 gave an average positive correlation.

ANOVA Result showed a high significant difference in the accuracies of the three algorithms (F = 202.96 and p = 0.0, where p is the level of significance). Further Tukey HSD test indicates that there were significant differences in the percentage accuracies between Naïve Bayes and J48 (p = 0.0), Naïve Bayes and MLP (p = 0.0) while J48 and MLP had no significant differences in their accuracies (p = 0.96). The mean difference was significant at the 0.05 level.

4.3. Discussion of Results

Results from this study show that there is a trade-off between accuracy and time complexities of the three algorithms (Multi-layer Perceptron, Naïve Bayes and Decision Tree) used. Low accuracy means low time complexity and vice versa. For instance, Naïve Bayes, having least time complexity for training has low accuracy but

Multi-Layer Perceptron and Decision Tree with higher time complexity had higher accuracy in their classifications. Naïve Bayesian (Naïve Bayes) classification algorithm tends to have more error rate with respect to the growth of the size of data-instances. This result indicates that users have to choose in between accuracy and time needed for training when choosing any of these three algorithms for classification tasks.

Neural networks usually have long training times and are therefore more suitable for applications where this is feasible. They require a number of parameters that are typically best determined empirically such as the network topology or "structure" [1]. Thus, one of the major problems with artificial neural network is that its convergence time is usually very long since the training set must be presented many times to the network. If the learning rate is too low, the network will take longer to converge. On the other hand, if high, the network may never converge. The learning rate has to be selected very carefully [18]. This study confirms that Neural Networks have long training times in consonance with the submission of Jiawei *et al*. [1] but they have good accuracies for classification tasks.

Although decision tree classifiers have good accuracies, as confirmed in this study, however, successful use may depend on the nature and size of data at hand. While decision trees classify quickly, the time for building a tree may be higher than another type of classifier. Decision trees suffer from a problem of errors propagating throughout a tree; a very serious problem as the number of classes increases.

Naïve Bayesian models are popular in machine learning applications, due to their simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This simplicity equates to computational efficiency, which makes Naïve Bayesian techniques attractive and suitable for many domains [11]. Naïve Bayesian algorithm builds and scores models extremely rapidly; it scales linearly in the number of predictors and rows. Bayes' Theorem states that the probability of event A occurring given that event B has occurred (P(A|B)) is proportional to the probability of event B occurring given that event A has occurred multiplied by the probability of event A occurring ((P(B|A)P(A)) [23]. Results obtained in this study confirm that Naïve Bayes classifiers are indeed fast in their trainings. However, their accuracies are low when compared with other classifiers.

5. Conclusion

Performance evaluation of data mining algorithms is very essential as this will help users to choose the best algorithm needed for their classification/prediction tasks. In this study, the performances of Decision Tree, Multi-Layer Perceptron and Naïve Bayes classification algorithms were studied with respect to their times taken for training and accuracy of prediction. The study shows that even though Naïve Bayesian algorithm takes less time for its prediction, its accuracy becomes low as data size increases.

References

[1] Han, J.W., Kamber, M. and Pei, J. (2012) Data Mining Concepts and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Waltham.

[2] Raj, K. and Rajesh, V. (2012) Classification Algorithms for Data Mining: A Survey. *International Journal of Innovations in Engineering and Technology* (*IJIET*), **1**.

[3] Berkin, O., *et al*. (2006) An Architectural Characterization Study of Data Mining and Bioinformatics Workloads. Evanston.

[4] Pardeep, K., Nitin, V.K. and Sehgal, D.S.C. (2012) A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems. *International Journal of Data Mining & Knowledge Management Process* (*IJDKP*), **2**.

[5] Thirunavukkarasu, K.S. and Sugumaran, S. (2013) Analysis of Classification Techniques in Data Mining. *IJESRT*: *International Journal of Engineering Sciences & Research Technology*, 3640-3646.

[6] Abirami, N., Kamalakannan, T. and Muthukumaravel, A. (2013) A Study on Analysis of Various Data Mining Classification Techniques on Healthcare Data. *International Journal of Emerging Technology and Advanced Engineering*, **3**.

[7] Liu, Y., Pisharath, J., Liao, W.-K., Memik, G., Choudhary, A. and Dubey, P. (2002) Performance Evaluation and Characterization of Scalable Data Mining Algorithms. Intel Corporation, CNS-0406341.

[8] Syeda, F.S., Mirza, M.A.B. and Reena, M.P. (2013) Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis. *IOSR Journal of Computer Engineering* (*IOSR-JCE*), **10**, 1-6.

[9] Gopala, K.M.N., Bharath, K.P., Nagaraju, O. and Suresh, B.M. (2013) Performance Analysis and Evaluation of Dif-

ferent Data Mining Algorithms Used for Cancer Classification. (*IJARAI*) *International Journal of Advanced Research in Artificial Intelligence*, **2**.

[10] Nikhil, N.S. and Kulkarni, R.B. (2013) Evaluating Performance of Data Mining Classification Algorithm in Weka. *International Journal of Application or Innovation in Engineering & Management*, **2**.

[11] Daniela, X., Christopher, J.H. and Roger, G.S. (2009) Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *IJSCI*: *International Journal of Computer Science Issues*, **4**.

[12] Anshul, G. and Rajni, M. (2012) Performance Comparison of Naïve Bayes and J48 Classification Algorithms. *International Journal of Applied Engineering Research*, **7**.

[13] Sampson, A. (2012) Comparing Classification Algorithms in Data Mining. A Thesis, Central Connecticut State University New Britain, Connecticut.

[14] Jyoti, S., Ujma, A., Dipesh, S. and Sunita, S. (2011) Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, **17**.

[15] John, S., Rakeeh, A. and Manish, M. (1997) SPRINT: A Scalable Parallel Classifier for Data Mining. IBM Almaden Research Center, San Jose.

[16] Robert, D.S. and Herbert, A.E. (1997) Scalable Data Mining, Two Crows Corp. http://www.twocrows.com/intro-dm.pdf

[17] Huidong, J. (2002) Scalable Model-Based Clustering Algorithms for Large Databases and Their Applications. Ph.D. Thesis, The Chinese University of Hong Kong, Hong Kong.

[18] Lalitha, S.T. and Suresh, B.C. (2013) Optimum Learning Rate for Classification Problem with MLP in Data Mining. *International Journal of Advances in Engineering & Technology*.

[19] Michael, J.A.B. and Gordon, S.L. (2004) Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. 2nd Edition, Wiley Publishing, Inc., Indianapolis.

[20] Brijesh, K.B. and Saurabh, P. (2011) Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications*, **2**.

[21] Chowdary, B.V., *et al*. (2012) Decision Tree Induction Approach for Data Classification Using Peanut Count Trees. *International Journal of Advanced Research in Computer Science and Software Engineering*, **2**.

[22] Kabir, M.F., *et al*. (2011) Enhanced Classification Accuracy on Naive Bayes Data Mining Models. *International Journal of Computer Applications*, **28**.

[23] Galathiya, A.S., *et al*. (2012) Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. *International Journal of Computer Science and Information Technologies*, **3**, 3427-3431.