Journal of Engineering and Applied Sciences 12 (22): 6069-6075, 2017
ISSN: 1816-949X

Adaptive Model for Campus Placement Prediction using Improved Decision Tree

[1]Subitha Sivakumar and [2]Rajalakshmi Selvaraj
[1]Department of Computer Science and Information Systems, Faculty of Computing,
Botho University, Gaborone, Botswana
[2]Department of Computer Science and Information Systems, BIUST, Gaborone, Botswana

Abstract: Student's academic achievements and their placement in campus selection becomes as challenging issue in the educational system. Monitoring the student's progress for their campus placement helps in monitoring the student's progression in the academic environment. Recently, educational data mining provides a deep motivation to students for taking an effective decision as academic planners. This also helps the educational institutions to have good intake of students based on student's academic achievements and appointments through the campus interview. In academic units, implementing this method will help in evaluating and analyzing students and help the educators and institutions to make important decision that will assist the students. This study demonstrates a novel method named "Improved Decision Tree (IDT)" for segregating the eligible students for the campus selection based on the academic performance measures. This model, based on the evaluated result obtained, it provides the suggestions in student's placement predicting. By using the proposed method, the relationship among student's academic performance and their campus placement is analyzed. Under this study information related to student's performance measures is analyzed in different perspectives to learn the achievements of the students through their activities.

Key words: Improved Decision Tree (IDT), Educational Data Mining (EDM), dataset, predicting, measures, achievements

INTRODUCTION

Education Data Mining (EDM) mines the data under the education environment (Baker, 2010). On applying Data Mining (DM) techniques under the educational environment helps in discovering useful information to assist the educationalist for the formative evaluation and to establish the teaching basis for decision making. Also, the data mining methods are applied to analyze relevant information about the students (Bresfelean, 2007) and produce data about the dissimilar perceptions where to analysis in detail with student's involvements. EDM is anxious about the growth of novel approaches in determine information through informative databases and to make fruitful decisions in educational system (Bharadwaj and Pal, 2011a).

In real world, student's performance prediction is a challenging task. Currently, the education system (Bray, 2007) is considered as strength of progress for student's carrier. Professional education is one of the supports of student selection in campus placement (Kabra and Bichkar, 2011). Data mining (Ayesha *et al*., 2010; Written and Frank, 2000) methods are implemented to retrieve data and knowledge and emerging area with numerous applications. Data mining as an emerging discipline in education environment, named EDM (Pal and Pal, 2013). This provides educational environment with lot of promising solutions. EDM methods is useful to interpret the student information and to explore the hidden knowledge from their academic and non-academic information.

With the aid of educational data mining, the answers for the various questions related to the student's issues are solved like:

- Determining the students with the chances of placement in the campus
- Determining the student's quality vs. institutional quality
- Identifying the courses that the institution can attract and enroll large number of students?
- Identifying the student's activities and their performance to their placement in the campus selection and so on

Among the above listed issues in this study we are going to concentrate more on the factor of identifying the student's activities and their performance to their placement in the campus selection.

Corresponding Author: Subitha Sivakumar, Department of Computer Science and Information Systems, Faculty of Computing, Botho University, Gaborone, Botswana

In this study, the student's earlier academic information and present academic information is collected. The collected data are analyzed by using the information gain feature selector for selecting the most important features that helps the students for their placements. The different features selected by the feature selector are evaluated by traditional classification algorithms and compared with improved decision tree algorithm. Based on the comparative results related to the accuracy, precision, recall and other measures, the student's academic performance for training and placement is analyzed (Moucary, 2011) and recommended for campus placement.

On analyzing the classifier's results, the improved decision tree classifier gives the best result on comparing with the other classifiers. This study presents a proposed model named "Improved Decision Tree (IDT)" based on normal decision tree to improve the decision tree results with the property of improved technique to predict student's placement. On measuring the performance measures of improved decision tree and its comparison with the traditional classifiers proves our proposed IDT is best of all in analyzing Student's academic achievements and their placement in campus selection.

Literature review: In past survey, ID3 and J48 decision tree algorithms are used on student's data for predicting the students prospective in continuing the postgraduate degrees, these models were applied on student group pursuing two dissimilar academic areas, the accuracies with 88.68%, 71.74% attained using C4.5 (Quinlan, 1993; Wu and Kumar, 2009).

The other conventional classifiers such as neural networks and SVM are implemented in secondary student dataset (Kovacic, 2010) in predict the students grade relation with education system by Cortez and Silva (2008). The result obtained using differentclassification techniques was found inferior on comparing with the tree based algorithms like decision tree.

The data mining tool KEEL, also applied to classify the students based on their performance using C4.5 algorithms. Their results also compared with other statistical classifier such as rule induction and neural networks. Ramesh *et al*. (2011) show the performance of the J48 decision tree algorithm and its performance is compared with the multilayer perception classifier. From the past survey, many of the earlier works deals with the different issues of the student's performance but no work was related to the predicting the student eligibility for the placement of the students for their jobs. With the help of our proposed method, we could clearly predict the factors falls on the placement and helps institutions to educate several programs for analyzing the academic information to predict the existing association among the features that affecting the placement activity.

The tremendous growth in educational data mining remains due to its contributions in the educational systems with main objective of improving the performance of the students for their carrier. Much survey was focused on the student's issues related to their marks, drop outs and so on. Only few work concentrates on the student's performance monitoring and their placements.

The following survey provides the details of the work related to the performance of educational data mining and lists the factors to be considered for the further growth.

Analyzing the data in different dimensions using the data mining algorithms is summarized by Han and Kamber (2000) with the relationships among the several features. According to Han and Kamber (2000) defines DM used in analyzing the features from various sizes, then used to classify the features and to recognition the relation among the features. With the aid of bayesian classification method, 17 attributes related to the student's performance of about 300 academic records from the five various colleges running the BSC in Computer Applications under Dr. R.M.L. Awadh University India is analysed (Yadav *et al*., 2012a, b). It is observed that the student's academic performance is influenced by their performance in senior and higher secondary examinations, student staying place, learning and teaching medium, parent's education, further habitation, household salary and status of the family. Another interesting aspect of DM is to help the institutions to achieve their enrollment goals was developed (Chang and Ed, 2008). This helps the institutions to manage the students enrollment more effectively in the college admissions (Khan, 2005).

Another research by using 600 students in their performance are taken from dissimilar colleges of Dr.R.M.L. Awadh University, India was done (Pandey and Pal, 2011) with the help of Bayes classifier on considering the factors like background qualification, category, language and so on. Based on this study, the performance of the new students could be decided.

Simple linear regression analysis identified that student's academic performance is correlated with parent's academic information and their household income. This conclusion was drawn (Hijazi and Naqvi, 2006) by conducting a research on the performance on a sample of 300 students (225 males, 75 females) from colleges affiliated to Punjab University of Pakistan. According to Ben-Zadok *et al*. (2007) developed a research work to identify the students on risk who need extra support by analyzing the student learning behavior

from the student data and predict their result before their exam, this will help the institution to improve those students. Student retention is the main challenge met by institution now, Yadav *et al*. (2012ab) have identifying the attributes that influence on student retention using the classification techniques on the sample of 398 academic records from Master of Computer Application from VBS Purvanchal University, Jaunpur, India and found that student's graduation stream and grade in graduation play important role in retention.

According to AI-Radaideh *et al*. (2006) proposed that DT method namely ID3, C4.5 and the Naive Bayes had better prediction than other models while predicting the final grade of students doing C++ course in Yarmouk University, Jordan in the year 2005. Elayidom *et al*. (2011) demonstrated that data mining technology can be very effective in predicting the employment opportunities this will help students to choose programme which has future in job market. The general outline for same issues has projected. Bharadwaj and Pal (2011b) predicted student's performance in the final exam with the attributes likes attendance, performance in internal, class presentation, discussion and test marks this will help educators to provide required support and in monitor.

The research objective is to apply different DM algorithms in predicting performance, to place the graduates based on earlier and current academic record. As the novel work the performance of the conventional classifiers are compared with the novel algorithm "Improved Decision Tree (IDT)" and its results proves its prediction accuracy is superior on comparison with the other traditional classifiers. And proves that it is best suited for the educational data mining related data such as based on the student's data set. In this research, comparative analysis of various existing classification algorithms is provided and which is best for the student data set is clearly shown.

MATERIALS AND METHODS

Proposed system architecture

System architecture and data set description: Recently the campus placement is the strength of all educational system and it is the only factor that attracts the society and brings the popularity to the educational institutions. The present-day situation is that companies are spending a huge capital for recruitments. If one can make filtering process of resumes easier and simpler then companies may not spend the huge capital upon recruitments. The proposed model for the above problem is to classify the resume of the applicants based on some decision parameters related to the student's data at the institution

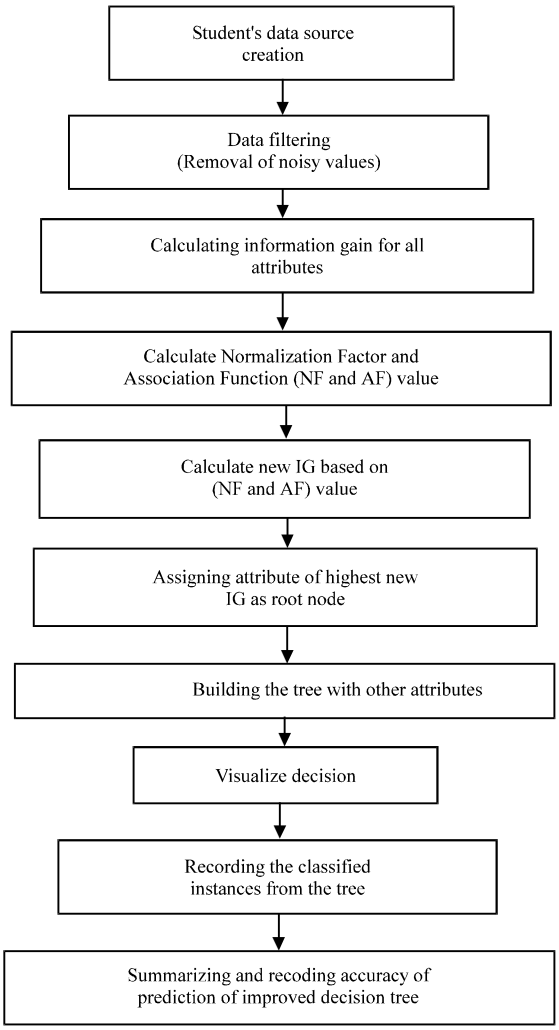

Fig. 1: System flow of enhanced DT Model

itself. The improved decision tree algorithm in the popular domain of data mining provides a significant road map for the classification. The proposed architecture of improved decision tree for the prediction of campus placement is shown in Fig. 1.

The flow diagram reveals the work flow of the proposed model. Various steps involved in working with Improved Decision Tree algorithm can clearly understand through Fig. 1.

Under the student's data source creation phase, the resumes of about 600 students from IT and Computing are collected. Based on vital attributes as shown in Table 1, the student's eligible for the campus selection is considered for future examination.

The clear description about the student related data for the campus selection is provided in Table 1. Table 1 provides all the details about the attributes and their possible values for the campus selection.

Table 1: Student's dataset details

Student's attributes	Description	Possible values
SS	Student's Sex	{Male, Female}
CGPA	Gives information about academic excellence	{Excellent (ce), good (cg), average (ca)}
AL	Aptitude Level	{Poor, Average, Good}
TS	Technical Skills	{Poor, Average, Good}
CS	Communication skills	{Poor, Average, Good}
BS	Beyond the syllabus skills	{Poor, Average, Good}
AS	Achievements	{Poor, Average, Good}
LS	Lab skills	{Poor, Average, Good}
BL (Backlogs)	Existence of backlogs	{Low (bl), Average (ba), High (bh)}
A (Articulate)	Checks how far one can present with clarity and effectiveness	{Good (g), Average (a), Poor (p)}
CSL (Core Skills Level)	Checks the level of core (branch) skills	{Familiar (f), unskilled (u)}
As (Achievements)	Gives information about extra circular activates certifications	{Optimum (o), Nil (n)}
CS (Campus Selection)	Eligibility for the campus selection	{Eligible (E), Not Eligible (NE)}

Algorithm for the improved decision tree: The algorithm given below depicts the steps for the construction of Improved Decision Tree Model for the campus selection.

Algorithm for Improved Decision Tree (IDT) Model:

Step 1: Calculating the Renyi entropy, let D be partitioned by split point into DY and DN:

$$H(D) = -\sum_{i=1}^{k} P\left(\frac{c_i}{D}\right) \log_2 P\left(\frac{C_i}{D}\right)$$

$$H(D_y, D_N) = \frac{n_y}{n} H(D_y) + \frac{n_y}{n} H(D_N)$$

Where $n = |D|$ are number of points in Dataset $n_y = |D|$ and $n_N = |D_N|$ are number of points in DY and DN

Entropy is zero if the point is from same class. If it is mixed class then the equal probability $P(c_i/D) = 1/k$, then the entropy has the highest value, $H(D) = \log_2$.

Step 2: Calculate the Information Gain (IG) for split points as follows:

$$Gain(D, D_y, D_N) = H(D) - H(D_y, D_N)$$

If information gain is high, the reduction of entropy is high, this lead to better split point so we choose the one that gives the highest information gain.

Step 3: Calculating the Association Function (AF). let A be the attribute and C be category attribute in D as follows:

$$AF(D) = \frac{1}{n} \times \sum_{i=1}^{n} |X_{i1} - X_{i2}|$$

Where, X_{ij} is the ith values of A in D, n is total attributes that A hold.

Step 4: Calculating the normalization factor, let m as attributes and the relation degree function of each attributes value are AF(1), AF (2), …,AF (m), respectively:

$$V(k) = \frac{A(k)}{A(1) + A(2) + A(3) + ..., A(m)}$$

for which $1 < k \leq m$.

Step 5: Calculate the new Information Gain (IG) as follow:

$$New\,gain(D, D_y, D_N) = Gain(D, D_y, D_N) \times V(k)$$

The new gain are the new fragmenting condition for feature selected in constructing the DT which is small, very generalized for minimal datasets, to avoids over-fitting.

Step 6: DT with root node can be built with the feature having high IG value. The splitting criterion tells us which attribute to test at node N by determining the "best" way to separate or partition the tuples in dataset into individual classes.

Step 7: The breaking state in root node is that when all class labels values. falls to single class, otherwise the node is extended with next higher IG value.

Step 8: Repeat the overhead procedure until the condition satisfied.

Various strategies are available to built decision tree and it is one of the simplest method so it can be used by wide range of application. Attribute selection is the main process in decision tree and that may hold many values. Improved decision tree used to get more effective accuracy results from these attribute values. Because the improved decision tree can significantly use Association Function (AF) and Normalized Factor (NF) to further improve the model generation process.

By using IDT algorithm, the information gain for each attribute, normalization factor and association function is calculated. Based on these values the new information gain is computed which will help in constructing the DT through root node being the feature with highest new IG among other attributes. The remaining attributes are made nodes of the decision tree using the new information gain when a new level of abstraction in needed. This process continues till a leaf node is obtained. The last attribute 'CS' is the campus selection is completely based on the features relevant to the student's performance. The relevant attributes extracted by the information gain is evaluated by the improved decision tree for its accuracy.

RESULTS AND DISCUSSON

Implementation results: The raw data is likely to be vulnerable to noise, holding missing values and being inconsistent so, we need to improve the data quality

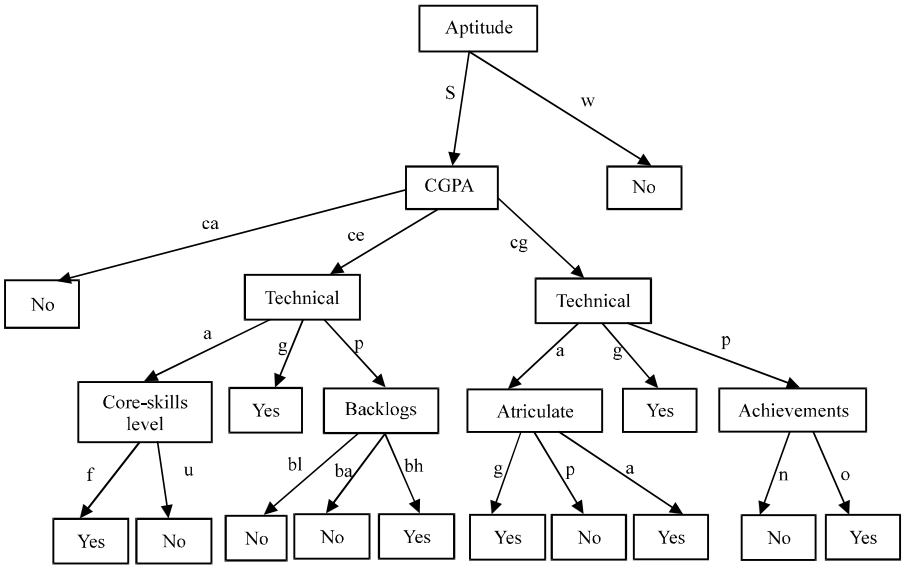

Fig. 2: The conceptual view of improved decision tree

before analyzing it through the data preprocessing techniques. Data preprocessing does the data preparation and filtering processes like data cleaning integration, transformation, normalization and aggregation. Doing this results in improving the accuracy of the classifier and is very must for all the other data mining tasks to obtain the good accuracy however this process takes considerable amount of time.

The training data set was retrieved from student's admission records. If the tuple found having missing values in any attribute then it will be ignored from training set as it couldn't be predicted or not should be set with default value. As there will be low chances of getting missing values, ignoring them will not affect the accuracy adversely. The basic step in preprocessing of the presented data is to eliminate the irrelevant parameters like registration numbers, names, etc., on which the prediction is not based. To evaluate the performance of different classifier, various test options are used. Few of them are listed.

Use training test: It takes the student dataset for training Supplied test set; After taking the dataset for training set, the new dataset which is to be predicted is fed to supplied test and then the classification is done.

Cross validation: The testing in cross-validation is based on FOLDS. The total database is to be divided into the numerous folds as mentioned in fold's column. Out of them one-fold is used for testing while others for training.

Percentage-split: The testing here depends on percentage-split. The total database is split into the two groups namely training and testing as mentioned in this test option. The percentage mentioned in the test option indicates the training dataset while other are used as test data. The conceptual view of the enhanced DT obtained is in Fig. 2.

Once the classification process initiates, the classifier output column would give the necessary information about the accuracy and various parameters related to the process. The various fields that present the classifier output are.

Run information: It gives the basic information such as scheme, relation, attributes and number of instances.

Classifier model: It gives the pruned tree along with size and instances of the tree.

Stratified cross validation: It is the field which accounts for accuracy, percentage of classified instances, mean absolute error, etc.

Confusion matrix: It is a 2×2 matrix whose elements give different possibilities of prediction based on true positive, false positive, true negative and false negative. With these predictions, precision, recall, accuracy and other performance measures are computed.

Table 2-4 show the performance for IDT algorithm when compared with performance of Decision Tree (DT), NB (Naive Bayes), ANN (Artificial Neural Network) algorithms.

Table 2: Comparison of evaluation measures

Classifier	TP	FP	Precision	Recall	Class
IDT	0.818	0.094	0.900	0.818	Yes
	0.906	0.182	0.829	0.906	No
DT	0.768	0.081	0.800	0.801	Yes
	0.815	0.078	0.725	0.878	No
ANN	0.788	0.188	0.813	0.788	Yes
	0.813	0.212	0.788	0.813	No
NB	0.758	0.250	0.758	0.758	Yes
	0.750	0.242	0.750	0.750	No

Table 3: Details of accuracy and error measures

Evaluation criteria	Classifiers			
	IDT	DT	ANN	NB
Accuracy	98.5%	95.16%	92.16%	90.48%
Kappa statistic	0.7234	0.6513	0.6001	0.5076
Mean Absolute Error (MAE)	0.2338	0.2201	0.2212	0.3156
RMSE	0.3427	0.3107	0.4234	0.453
Relative Absolute Error (RAE)	46.7085%	44.1023	44.2036%	63.0499%

Table 4: Confusion matrix

Classifier	Yes	No	Class
IDT	295	6	Yes
	3	296	No
DT	285	16	Yes
	13	286	No
ANN	275	24	Yes
	23	278	No
NB	273	26	Yes
	25	278	No

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives+Fasle positives}} \quad (1)$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives+Fasle negatives}} \quad (2)$$

$$\text{Accuracy} = \frac{(\text{True positives+True negatives})}{\text{True positives+True negatives+ False positives+False negatives}} \quad (3)$$

CONCLUSION

The proposed Improved Decision Tree Model helps in reducing the cost of conducting interviews by corporate to a greater extent by means of saving the human resources. As conclusion, to evaluate and investigate student's campus placement opportunity after completion of their graduation based on their performance is evaluated by the selected classification algorithms such as DT, NB, ANN and compared with the proposed Improved Decision Tree (IDT). The best result for the student data is obtained by the proposed IDT Model in a least computation time with improved accuracy. The result of IDT is superior than the other classifiers such as DT, NB and ANN with the yielded accuracy of 98.5%.

SUGGESTION

The other measures suggest that among the various machine learning algorithm examined, improved decision tree classifier has the potential to significantly improve the conventional classification methods for use in placement.

REFERENCES

Al-Radaideh, Q.A., E.M. Al-Shawakfa and M.I. Al-Najjar, 2006. Mining student data using decision trees. Proceedings of the 2006 International Arab Conference on Information Technology (ACIT'06), December 19-21, 2006, Yarmouk University, Irbid, Jordan, pp: 1-5.

Ayesha, S., T. Mustafa, A. Sattar and I. Khan, 2010. Data mining model for higher education system. Eur. J. Sci. Res., 43: 24-29.

Baker, R.S.J.D., 2010. Data Mining for Education. In: International Encyclopedia of Education, McGaw, B., P. Peterson and E. Baker (Eds.). Elsevier, Oxford, UK., pp: 112-118.

Ben-Zadok, G., A. Hershkovitz, R. Mintz and R. Nachmias, 2007. Examining online learning processes based on log files analysis: A case study. Res. Reflection Innovations Integrating ICT Educ., 2007: 55-59.

Bharadwaj, B.K. and S. Pal, 2011b. Data mining: A prediction for performance improvement using classification. Int. J. Comput. Sci. Info., Security (IJCSIS), 9: 136-140.

Bharadwaj, B.K. and S. Pal, 2011a. Mining educational data to analyze student's performance. Int. J. Adv. Comput. Sci. Applic., 2: 63-69.

Bray, M., 2007. The Shadow Education System: Private Tutoring and its Implications for Planners. 2nd Edn., UNESCO, Paris, France, ISBN:9789280313055, Pages: 101.

Bresfelean, V.P., 2007. Analysis and predictions on students behavior using decision trees in Weka environment. Proceedings of the 29th International Conference on Information Technology Interfaces (ITI'07), June 25-28, 2007, IEEE, Cavtat, Croatia, ISBN:953-7138-09-7, pp: 25-28.

Chang, T. and D. Ed, 2008. Data mining: A magic technology for college recruitment. Master Thesis, Overseas Chinese Association for Institutional Research, Santa Ana, California.

Cortez, P. and A.M.G. Silva, 2008. Using data mining to predict secondary school student performance. Proceedings of the 5th Annual Conference on Future Business Technology (FUBUTEC'08), April 9-11, 2008, EUROSIS, Porto, Portugal, ISBN: 978-9077381-39-7, pp: 5-12.

Elayidom, S., S.M. Idikkula and J. Alexander, 2011. A generalized data mining framework for placement chance prediction problems. Intl. J. Comput. Appl., 31: 40-47.

Han, J. and M. Kamber, 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Burlington, Massachusetts, USA.,.

Hijazi, S. and S. Naqvi, 2006. Factors affecting students performnce: A case of Private Colleges, Bangladesh. J. Sociology, 3: 12-17.

Kabra, R.R. and R.S. Bichkar, 2011. Performance prediction of engineering students using decision trees. Intl. J. Comput. Appl., 36: 8-12.

Khan, Z.N., 2005. Scholastic achievement of higher secondary students in science stream. J. Soc. Sci., 1: 84-87.

Kovacic,., 2010. Early prediction of student success: Mining students enrolment data. Proceedings of the 2010 Conference on Informing Science and IT Education (InSITE), June 19-24 2010, Information Sciences Institute, Cassino, Italy, pp: 1-17.

Moucary, C.E., 2011. Data mining for engineering schools predicting students performance and enrollment in masters programs. Intl. J. Adv. Comput. Sci. Appl., 2: 1-9.

Pal, A.K. and S. Pal, 2013. Analysis and mining of educational data for predicting the performance of students. Intl. J. Electron. Commun. Comput. Eng., 4: 1560-1565.

Pandey, U.K. and S. Pal, 2011. A data mining view on class room teaching language. (IJCSI) Int. J. Comput. Sci., 8: 277-282.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA., USA.

Ramesh, V., P. Parkavi and P. Yasodha, 2011. Performance analysis of data mining techniques for placement chance prediction. Intl. J. Sci. Eng. Res., 2: 1-7.

Witten, I.H. and E. Frank, 2000. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann Publishers, San Francisco, USA., ISBN: 1-55860-552-5, Pages: 373.

Wu, X. and V. Kumar, 2009. The Top Ten Algorithms in Data Mining. CRC Press, Boca Raton, Florida, USA., ISBN-13:978-1-4200-8964-6, Pages: 214.

Yadav, S.K., B. Bharadwaj and S. Pal, 2012b. Mining education data to predict student's retention: A comparative study. Intl. J. Comput. Sci. Inf. Secur., 10: 113-117.

Yadav, S.K., B.K. Bharadwaj and S. Pal, 2012a. Data mining applications: A comparative study for predicting student's performance. Int. J. Innovative Technol. Creative Eng., (IJITCE), 1: 13-19.