International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 3, Issue 2, 2016, PP 10-14
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

Application of Data Mining Techniques to Predict Students Placement in to Departments

Getaneh Berie Tarekegn
MSC, Department of Computer Science
Faculty of Engineering and Technology
Assosa University, Assosa, Ethiopia
getaneh_berie@yahoo.com

Dr.Vuda Sreenivasarao
Associated Professor Department of ISE
School of Computing Adama Science and Technology University, Adama, Ethiopia
vudasrinivasaro@gmail.com

Abstract: *Data mining is one of the techniques to extract useful information from a huge data and support to make decision in various aspects. In academic institutions like universities and colleges the students placement in to different departments is one of the activity that data mining can be applied to predict the departments which the students will be placed based on the order of their preference. Educational Data Mining is concerned with developing new methods to discover knowledge from educational database and can used for decision making in educational system. In this study, we collected the student's data that have different information about their entrance exam result and then apply different classification algorithm using Data Mining tools (WEKA) for analysis of student's placement into departments. The study used three algorithms J48, Naïve Bayes and Random Forest to build a prediction model for placement of students. The analysis result shows that Random Forest algorithms has performed well and can be used as a best predicting model algorithm than the other two algorithms. Finally the output shows that most of the students were placed based on their first choice and only some of them were placed without their first choice.*

Keywords: *Knowledge Discovery in Databases, Data Mining, Cross-Validation, Classification Model, Classification, WEKA.*

1. INTRODUCTION

Data Mining is a process of extracting previously unknown, valid, potential useful and hidden patterns from large data sets. As the amount of data stored in educational databases is increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used [9].

An institute with efficient Data Warehousing and Data Mining approach can find out novel way of improving student's behavior, success rate and course popularity. All these effort may finally improve the quality of education, better student intake, better career counseling and overall practices of education system.

The number of students that join higher education in Ethiopia is growing rapidly from time to time. These students choose a field of study of their interest and universities they want to learn after they took the entrance exam. Based on their field of interest and entrance exam result they placed to different universities. And the universities placing students in to departments based on their field of interest (choice) and the exam result they have scored on the entrance exam of higher education.

Students are placed based on different criteria that are nationally prepared for all universities. The students choose the departments and give numbers according to their order of preference. The students who has higher exam score, females and developing regions are given a priority to be placed based on their first choice up to the department acceptance capacity, the other students also placed by their second, third choices.

The student data is the main source of information that can be processed and interpreted in different ways so that the universities can predict or classify and know about their students.

Classification is the most commonly applied data mining technique, which employs a set of pre-classified attributes to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms [8].

The remaining part of this paper is organized as follows. Section two discusses the present system, in section three we present methodology of the study, in section four discusses related works. Experimental results are reported in Section five. Section six presents conclusion.

2. PRESENT SYSTEM

Placing higher education students into different departments is not easy task. Several problems can be raised related to the department placement of students in different perspectives. From these problems predicting the departments of the students that are not placed based on some training data, clustering the students in to different clusters as students which are placed based on their first, second, third choice and etc., knowing the frequent departments chosen first, second, third choice etc. Such problem is happened in student's placement in University of Gondar (UoG).. This study is mainly concerned on how the department placement is held and how many of the students are placed to the departments based on their choice of preference. On the other hand, which algorithm performs well for classifying students in to different departments and can build a best predicting model is the other question can be raised.

3. METHODOLOGY

To achieve the objective of this study the following research methodologies were employed

3.1. Literature Review

A review of relevant literature has been conducted to assess data mining technology, both concepts and techniques, and researches in this field. Related books, journals, articles, and papers from the Internet pertaining to the subject matter of data mining and Knowledge Discovery Process in Databases (KDD) have been reviewed to understand the potential applicability of data mining in the placement of students into departments.

3.2. Data Source

The source of data used to undertake this study was taken from the assistance registrar office of Gondar University. The data set contains about 1496 instances of placed students. The data source obtained from the assistant registrar office is recorded on MS Excel. The data of faculties selected for this study are Faculty of Natural and Computational Science, Faculty of Agriculture, and Faculty of Engineering.

3.3. Data Preparation

After the data were collected, tasks such as processing and cleansing are imposed in order to make the data more suitable for the particular data mining software, which are used in the study. This comprises attribute selection, defining target classes (attributes for classification), handling noisy data, accounting for missing data fields, and preparing the processed data in a file format acceptable to the Weka software.

The data of the students is includes name, id, total entrance mark, the departments to be selected and gender of each student. From these all data the gender, total mark order of departments of the students' choice and the department in which each student placed is selected for the experiment to discover the department placement process. And to partition the data into training and test dataset, cross-validation and supplied test set were used.

3.4. Data Selection

Some of the attributes in the initial dataset that was not pertinent (relevant) to the data mining experiment goal were ignored. The attributes name, registration number and faculty were not used as having no data mining value to classify the students in to different departments. The main attributes used for this study are sex, total_mark, biology, biotechnology, chemistry, geology, maths, physics, sportscience, statistics, plantscience, animalproduction, veternarypharmacy, waterresource, agriculturaleconomics, naturalresource, rdae, electricaleng, civileng, mechanicaleng, constractioneng, and placement.

3.5. Data Integration and Transformation

The data used for this study were collected from the same source and have the same format and year. Therefore, there is no any data integration and transformation techniques used to integrate and

transform the data. But since it has taken from three different faculties the data of students of these two faculties are merged.

3.6. Tools and Techniques

The tools used in this study are MS Excel and Weka software that are used for data preprocessing and classification algorithms respectively. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato. Weka is freely available software. Therefore, Weka is a very good data mining tool which could be used in the field of education, in that it is going to be used for classification technique. The classification techniques of data mining help to classify the data on the basis of certain rules [3].

For this study classification algorithms such as J48, Naïve Bayesian, and Random Forest were applied to discover the distribution of the students through different departments.

4. RELATED WORKS

Data mining techniques has evolved its research very well in the field of education in a massive amount. This tremendous growth is mainly because it contributes much to the educational systems to analyze and improve the performance of students as well as the pattern of education. Various works had been done by a large number of scientists to explore the best mining technique for performance monitoring and placement. Few of the related works are listed down to have a better understanding of what should be carried on in the past for further growth.

Data mining is a powerful tool for academic intervention. Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success [6].

Jing Luan [6] conducted that by using well defined algorithms from the disciplines of machine learning and artificial intelligence to discern rules, associations, and likelihood of events, data mining has profound application significance. If it were not for the fast, vast, and real-time pattern identification and event prediction for enhanced business purposes, there would not have been such an exponential growth in dissertations, models, and the considerable amount of investment in data mining in the corporate world. Data mining can conducted to predict the likelihood of an applicant's enrollment following their initial application may allow the college to send the right kind of materials to potential students and prepare the right counseling for them.

Erdogan and Timor (2005) used educational data mining to identify and enhance educational process which can improve their decision making process. Finally Henrik (2001) concluded that clustering was effective in finding hidden relationships and associations between different categories of students [3].

Mining in educational environment is called Educational Data Mining. Data mining applications in higher education given in [2] concluded that data mining techniques on student's data base is helpful for executives for training & placement department of engineering colleges, and classified the categories of student's performance in their academic qualifications.

5. EXPERIMENT AND RESULTS

The study attempts to discover the placement of students in to departments and predict the distribution of the students after they have chosen the departments in order of their preference. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large[2].

The algorithms selected to classify the data are J48, Naïve Bayes and Random Forest that are in different categories from the algorithm implemented in Weka software. The size of the data used to predict the department placement of students is 1496 instances. The data has been experimented by cross validation with the total data set using 10 fold and by supplied test data that has 1197 instances for training set and 299 instances for test set which means 80% for training and 20% for test.

To build the predictive model, the arff format of the selected dataset was given to Weka and two major experiments were performed for each selected algorithms. J48, Naïve Bayes and Random Forest algorithms were experimented by using default parameters to build the predictive model.

In the experiments, variable placement was set as dependent variable and the remaining other attributes were set as independent variables. Another four experiments were done with each algorithm by using cross validation and supplied test set with two different numbers of attributes.

5.1. Summary of Experimental Result and Findings of the Study

The results obtained from the various data mining algorithms i.e. J48, Naive Bayes and Random Forest on the data set for the three different faculties of students are given in the following table and the performance of the classifiers is analyzed.

Table1. *The summary of the results found*

Experiment	Algorithm	Precision (%)		Recall (%)		F-measure (%)		Accuracy (%)	
		10 fold CV	20% test set	10 fold CV	20% test set	10 fold CV	20% test set	10 fold CV	20% test set
1	**J48**	**88.4**	**93.8**	**88.4**	**94**	**85.4**	**93.7**	**88.435**	**93.979**
2	Naïve Bayes	**70.1**	**75.3**	**69.1**	**71.9**	**67.7**	**71.7**	**69.050**	**71.906**
3	Random Forest	**86.8**	**99.3**	**86.6**	**99.3**	**86.5**	**99.3**	**86.564**	**99.33**

5.2. Experiment on Decision Tree (J48)

In experiment one, from the above table we can observe that J48 with 22 attributes scored accuracy of 88.435%, 85.4% F-measure ***with 10 fold cross validation*** and accuracy of 93.979% and 93.7% F-measure with 20% ***supplied test set.***

5.3. Experiment on Naïve Bayes

Experiment two on Naïve Bayes algorithm with 22 attributes shows accuracy of 69.050% and 67.7% F-measure with cross validation and accuracy of 71.906% and 71.7% F-measure with 20% supplied test set.

5.4. Experiment on Random Forest

The third experiment on Random Forest using 22 attributes has shown an increase in accuracy of 86.564% and 86.5% F-measure with cross validation, and accuracy of 99.33% and 99.3% F-measure with 20% supplied test set.

The above analysis results shows that Random Forest with accuracy of 99.33% using 22 attributes and 20% supplied test set is more appropriate for the data given **to build the predicting model for department placement with high accuracy than the two algorithms used in experiment one and two.**

On the other hand, the Naïve Bayes algorithm performs less accuracy which has great difference with the Random Forest algorithm. J48 which has performed accuracy of 93.979% with 20% supplied test set using 22 attributes can be selected as a good algorithm next with the Random Forest technique.

The other thing observed from the experiments result above is using 20% supplied test set has performed better accuracy than 10 fold cross validation on all of the three algorithms using 22 attributes. Therefore, using supplied test data is better than cross validation for this data to get better accuracy when using these algorithms.

Therefore, the result gained from the experiments show that the best prediction model using Random Forest algorithm is more acceptable than the remaining two algorithms used in this study.

6. CONCLUSION AND RECOMMENDATION

This work is an attempt to use Data Mining techniques to analyze students' entrance exam result to predict students' placement into departments. In this work I apply three classification methods on students' data i.e. J48, Naïve Bayes, and Random Forest classification methods. From the analysis result of the experiments J48 performs with the accuracy of 91.304%, Naïve Bayes 74.581% and Random Forest 95.652%.We notice that according to experimental result Random Forest Classifier is most suitable method for this type of student dataset.

As a conclusion, results from the study have shown that the problem of predicting departments of students in the University could be supported by the use of data mining technique.

The study is done for academic purpose but it can be implemented for the registrar office of the University and other academic institutions. The data used for the study is only one year data and three faculties; therefore, if other projects and researches are conducted on this area by using different year and all faculties' data it can be helpful for the office to implement the application of data mining. In addition, only three algorithms were used from many classification algorithms available, so other better algorithms can be selected by performing better than these algorithms used in this study. Therefore, it is a good attempt if anyone can apply and get best algorithm to build better predicting model for this type of data.

ACKNOWLEDGMENT

The Authors would like to thank University of Gondar by giving the student data set to do this research and development. And also would like to thank the reviewers for their constructive comments.

REFERENCES

[1] Tadesse Dejenie. (2010, September). Student's Placement by First Choice and Without at MekelleUniversity: The Impact on Academic Performance. Ethiopian Journal of Education & Science.Vol. 6 No 1 September 2010. Available: www.ju.edu.et/ejes/sites/.../STUDENT% 20PLACEMENT.pdf

[2] Samrat Singh, Dr. Vikesh Kumar. (2012, August). Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education. International Journal of Computer Science and Network (IJCSN) Volume 1, Issue 4, August 2012, India.Available: http://ijcsn.org/IJCSN-2012/1-4/IJCSN-2012-1-4-63.pdf

[3] V.Ramesh, P.Parkavi, P.Yasodha. (2011, August). Performance Analysis of Data Mining Techniques for Placement Chance Prediction. International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011.Available: http://www.ijser.org/researchpaper /Performance-Analysis-of-Data-Mining-Techniques-for-Placement-Chance-Prediction.pdf

[4] Micheline Kamber. (2006).Data Mining: Concepts and Techniques.(Second Edition). Jiawei Han University of Illinois at Urbana-Champaign.

[5] Selam Assamnew. (2011, July). Predicting the Occurrence of Measles Outbreak in Ethiopia Using Data Mining Technology. Addis Ababa University, Ethiopia.

[6] Jing Luan. (2004) Data Mining Applications in Higher Education. PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories. Available: http://www.spss.ch/upload/1122641492_Data%20mining%20applications%20in%20higher%20education.pdf

[7] Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, Vol. 2(1), 121-167.Available: http://research.microsoft.com/ pubs/67119 svmtutorial.pdf

[8] Ajay K. and Saurabh P. (2013), Classification Model of Prediction for Placement of Students,.J.Modern Education and Computer Science, 2013, 11, 49-56

[9] Samrat S, Dr. Vikesh K, (2013), Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques, Samrat Singh et al | IJCSET |February 2013 | Vol 3, Issue 2, 31-37

AUTHORS BIBLIOGRAPHY

Getaneh Berie Tarekegn, received his B.Sc. Degree in Computer Science from University of Gondar (UoG). Currently perusing.M.Sc. in Computer Science, School of Computing and Electrical Engineering, IOT, Bahir Dar University, Ethiopia. His main research interest is Neural Network, Data Mining and Big Data.

Dr. Vuda Sreenivasarao, received his M.Tech degree in computer science and engineering from Sathyabama University from 2007.He received PhD degree in computer science and engineering from Singhania University, Rajasthan, India from 2010. Currently working as Associated Professor in School of Computing, Adama Science and Technology University, Adama, Ethiopia.

His main research interests are Data mining, Fuzzy logic, Mobile communication and Network Security. He has got 16 years of teaching experience. He has published more than 35 research papers in various international journals and one Springer international conference paper. He has 115 Editorial Board / Reviewers memberships in various international journals. He is a life member of various professional societies like IEEE, ACM, MAIRCC, MCSI, SMIACSIT, MIAENG, MCSTA, MAPSMS, MSDIWC, SMSCIEI, SNMUACEE and MISTE.