International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:10, 2013

# An Educational Data Mining System for Advising Higher Education Students

Heba Mohammed Nagy, Walid Mohamed Aly, Osama Fathy Hegazy

***Abstract*—**Educational data mining is a specific data mining field applied to data originating from educational environments, it relies on different approaches to discover hidden knowledge from the available data. Among these approaches are machine learning techniques which are used to build a system that acquires learning from previous data. Machine learning can be applied to solve different regression, classification, clustering and optimization problems.

In our research, we propose a "Student Advisory Framework" that utilizes classification and clustering to build an intelligent system. This system can be used to provide pieces of consultations to a first year university student to pursue a certain education track where he/she will likely succeed in, aiming to decrease the high rate of academic failure among these students. A real case study in Cairo Higher Institute for Engineering, Computer Science and Management is presented using real dataset collected from 2000−2012.The dataset has two main components: pre-higher education dataset and first year courses results dataset. Results have proved the efficiency of the suggested framework.

***Keywords*—**Classification, Clustering, Educational Data Mining (EDM), Machine Learning.

## I. Introduction

WITHIN recent few years, the number of educational institutes that adopted an information system has been growing very quickly; consecutively the amount of data available in each educational institute database has also increased. Educational data mining is intuitively applied to discover hidden information from this data that would improve the quality of the whole educational system. Educational data mining can be applied to discover patterns in untrusted datasets to automate the decision making process of learners, students and administrators.

Educational data mining methods belong to a diversity of literatures. These literatures include data mining, machine learning, information visualization, and computational modeling.

Machine learning approaches include neural network, naive Bayesian, K-nearest neighborhood, decision tree, support vector machine (SVM), linear regression, and rule induction. All these techniques can be used to discover association rules, classification, clusters, and outliers within educational datasets.

This paper uses machine learning techniques to develop an intelligent student advisory framework. This framework improves the student's performance and the quality of the education by reducing the failure rate of first year students. One of the main reasons for this high failure rate is the incorrect selection of the student's department/section.

The framework acquires information from the datasets which stores the academic achievements of students before enrolling to higher education together with their first year grade after enrolling in a certain department. After acquiring all the relevant information, a new student can challenge the intelligent system to receive a recommendation of a certain department in which he/she would likely succeed.

The remaining parts of this paper are organized as follows: Section Two presents the basic information of machine learning with a special concern on the algorithms used in paper. Section Three presents related works in educational data mining. Section Four introduces the proposed intelligent framework for a student advisory system. Section Five presents the case study explained in this research. Section Six shows the implementation of the framework, and then the conclusion follows at the end.

## II. Machine Learning

Machine learning aims at building an intelligent system which will be intelligent enough to determine a decision or calculate output based on new inputs after passing the learning phase and being fed with a set of training data.

According to the definition of Tom Mitchell [1]: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

Learning can be a supervised learning where the correct output in the training set is made available. Supervised learning is used to solve regression or classification problems. Supervised as the learning Application of machine learning includes classification and regression. Examples of classification problems include identifying an email as a spam, face recognition and hand writing recognition while regression problems include building a model for a system that can be used to predict the output value of the system for a given input.

The other type of learning is unsupervised learning where the exact output is unknown. This type of learning is used

Heba Mohammed Nagy is a Staff Assistance in Computer Science Department ,Cairo Higher Institute For Engineering , Computer Science and Management, Cairo, Egypt (phone: 02-01223287734; e-mail: h.nagy@chi.edu.eg).

Walid. Mohammed. Aly is a Prof. Assistance in College of Computing and Information Technology , the Arab Academy for Science, Technology and Maritime Transport, Cairo, Alex (e-mail: walid.ali@aast.edu).

Osama Fathy Hegazy is a Head of Department of Computer Science, Cairo Higher Institute For Engineering, Computer Science and Management, Cairo, Egypt (e-mail: oshegazy eg@yahoo.com).

typically to solve clustering problems. Two of the machine leaning techniques is described below.

### A. C4.5

C4.5 is a supervised learning algorithm for producing decision tree , it was proposed by Ross Quinlan as an extension of the basic ID3 algorithm. C4.5 is considered a statistical classifier as it can deal with both continuous and discrete attributed and data set with missing attributes values. The standard C4.5 algorithm is as follows:

1) Read set S of examples described by continuous or discrete attributes.
2) Identify base case.
3) Find the attribute which has the highest informational gain (Abest).
4) Divide S into S1, S2, S3... according to the values of Abest.
5) Repeat the steps for S1, S2, S3 etc . . .

### B. k-Means Clustering

K-means [2] is an unsupervised learning algorithm. It is one of the partitioning clustering procedures. It is dependent on distance-based that split ”n” dataset into the specific predetermined number of clusters in which each cluster is associated with the centroid and then each point in the dataset follows the cluster in the nearest centroid. The basic k-means algorithm which is standard and simple as given below:

1) Select K points as the initial Centroidsrepeat
2) Form K clusters by assigning data points to nearest Centroid
3) Recalculate the Centroid of each cluster
4) Until the Centroids do not change.

## III. Related Work

Many researchers have contributed to the field of data mining in higher education. In this section, the researchers will give an overview on a few representative works.

Abu Tair and El-Halees [3], gave a case study from the higher education stage. The main purpose of their study is to show how useful data mining can be in the educational domain in order to discover many kinds of knowledge by applying the graduate student data set on the different educational data mining techniques by using Rapid Miner software to discovered classification, clusters, association rule, and outlier detection then gave description to their importance in the education domain. M. Sukanya, S. Biruntha, S. Karthik and T. Kalaikumaran [4] applied the Bayesian classification technique on the existing higher education student. The main goal of their study is to predict the number of upcoming students in the next year based on the valid number of enrolled students in the previous years. This study helps decision makers to manage the number of resources and staffs they need to administer the outcomes of a student. This study helps also the teachers to know at early stage the students that need more attention to facilitate taking the correct action at the suitable time to reduce the failure in the academic approach and improve the student’s academic performance. Md. Hedayetul Islam Shovon and Mahfuza Haque [5] implemented a k-means cluster algorithm. The main goal of their study is to help both the instructors and the students to improve the quality of the education by dividing the students into groups according to their characteristics using the application which have been implemented. Er. Rimmy Chuchra, M. tech [6] gave a case study from the higher education university. Their study was based on applying the educational data mining techniques on the existing student data set from the database university to discovered cluster, decision tree and neural network to show how to evaluate the performance of students with the usage of these techniques. Brijesh Kumar Bhardwaj and Saurabh Pal [7] applied Baysian classification on the student database from the higher education stage. This study aimed at identifying those students which needed more attention to reduce the drop out ratio and take action at a right time which helped to improve the performance of the students and the instructors. Md. Hedayetul Islam Shovon and Mahfuza Hague [8] applied a hybrid procedure that was based on Decision Tree and Data clustering from Data Mining methods. The main goal of their study is to predict the GPA which helped the teachers to reduce the drop out ratio to improve the performance of the students and the academics.

## IV. Intelligent Framework for a Student Advisory System

### A. Framework Description

The proposed framework uses both classification and clustering techniques to suggest recommendations for a certain department for a student or an educational dataset that is required. This framework shall include attributes representing:-

- Student academic level before entering college
- Department chosen by student
- Student grade in the first year

### B. Classification Phase

In this phase, a classification algorithm is applied on the educational dataset to find an efficient classifier. The role of the classifier is to output the department recommended for the student. The steps in this phase are as follows:-

1) Remove all the records for the student who failed in his/her first year
2) Use this training dataset, and apply different classification algorithm with the Department attribute as the class.
3) Record the set of rules for the classification algorithm with highest F-Measure

### C. Clustering Phase

In this phase, a clustering algorithm is applied on the educational dataset to divide student records into a number of clusters based on marks’ similarity. The steps in this phase are as follows:-

1) Remove all attributes regarding the Department chosen by student.
2) Remove all attributes regarding first year grade.

3) Choose the number of clusters.
4) Use K-means algorithm to identify the clusters.
5) Identify the distribution percentage of each department along all clusters
6) Record the set of rules.

### D. Request an Output from the System

A user can ask the system to acquire a recommendation for a certain educational department. The steps of this phase can be summarized as follows:

1) The new student will enter his/her data
2) The purposed system will read the data and validate its soundness
3) Predict the cluster (Xcluster) according to rules declared by clustering phase
4) Output the department with the highest percentage rate in Xcluster
5) Predict the department according to rules declared by classification phase. If both predict the same department, the output will be one choice, otherwise the output will be two choices where the first choice will be the one with the highest accuracy and the other will be the second choice.

## V. Case Study

The Student Data used in this case study is obtained from "Cairo Higher Institute for Engineering, Computer Science, and Management" (CHI which is located in Cairo, Egypt. The institute has four departments:

1) Management Information System (MIS)
2) Computer Science (CS)
3) Architecture Engineering (AE)
4) Computer Engineering (CE)

### A. Data Set

The student data is collected from CHI during the period from2000to2012 to form a dataset known as (CHISDS). CHISDSincludes1866records, each record has 21 attributes.

Not all the attributes will be used in the data mining process, some of the attributes in the dataset such as the Student ID, Student Name, Address, or Home Phone Number present personal information of the students' data. These attributes are not useful in the mining process because they do not expand any knowledge for the dataset under processing. Feature selection process is applied to choose the relevant attributes which would affect the success of student in a certain department; this process resulted in selecting only seven attributes. The selected attributes are shown in Table I.

TABLEI
Data Set Meta Data

| Attribute | Data Type | Range |
|---|---|---|
| Secondary Stage Type | Discrete | 9 values(SSA1,SSA2,...SSA9) |
| Total Marks SS | Continues | 0-420 |
| English Mark | Continues | 0-50 |
| Math Marks | Continues | 0-100 |
| Physics Marks | Discrete | 0-50 |
| First Year Grade | Discrete | 8 values (A,B+,B-,C+,C,D+,D,F) |
| Department | Discrete | 4 values (AE,CE,CS,MIS) |

### B. Results for Recommendation Using Classification

Due to the existence of a lot of classification algorithms, we tested a number of algorithms on educational datasets.TheC4.5 proved to be efficient and robust. Fig. 1 shows the average F-measure percentage for different classification algorithms.

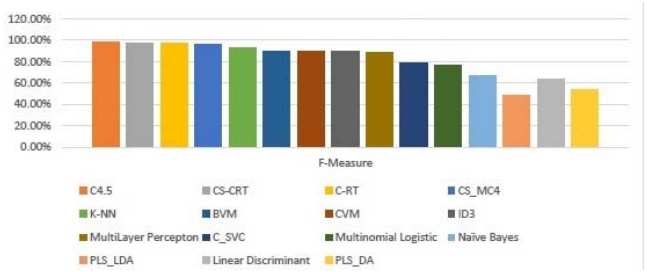


Fig. 1 F-measure for different classifiers

Applying the C4.5 algorithm as classifier resulted on classification of recommended department. The F-measure for classification is as shown in Table II.

TABLE II
F-Measure

| Attribute | F measure |
|---|---|
| MIS | 0.98 |
| CS | 0.99 |
| AE | 0.98 |
| CE | 0.99 |

Table III shows the Confusion Matrix for the C4.5 classifier output.

TABLE III
Confusion Matrix

| | MIS | CS | AE | CE | SUM |
|---|---|---|---|---|---|
| MIS | 319 | 0 | 4 | 0 | 323 |
| CS | 0 | 400 | 0 | 0 | 400 |
| AE | 6 | 7 | 534 | 4 | 551 |
| CE | 0 | 0 | 0 | 409 | 409 |

Fig. 2 shows the decision tree produced by C4.5.

Fig. 3 shows the correctly classified instances and the mistakenly classified instances for each department.

### C. Results for Recommendation Using Clustering

Applying the K-means algorithm on the available data set resulted on having four different clusters with Centroids as shown in Table IV.

The following table shows the centroids for the clusters

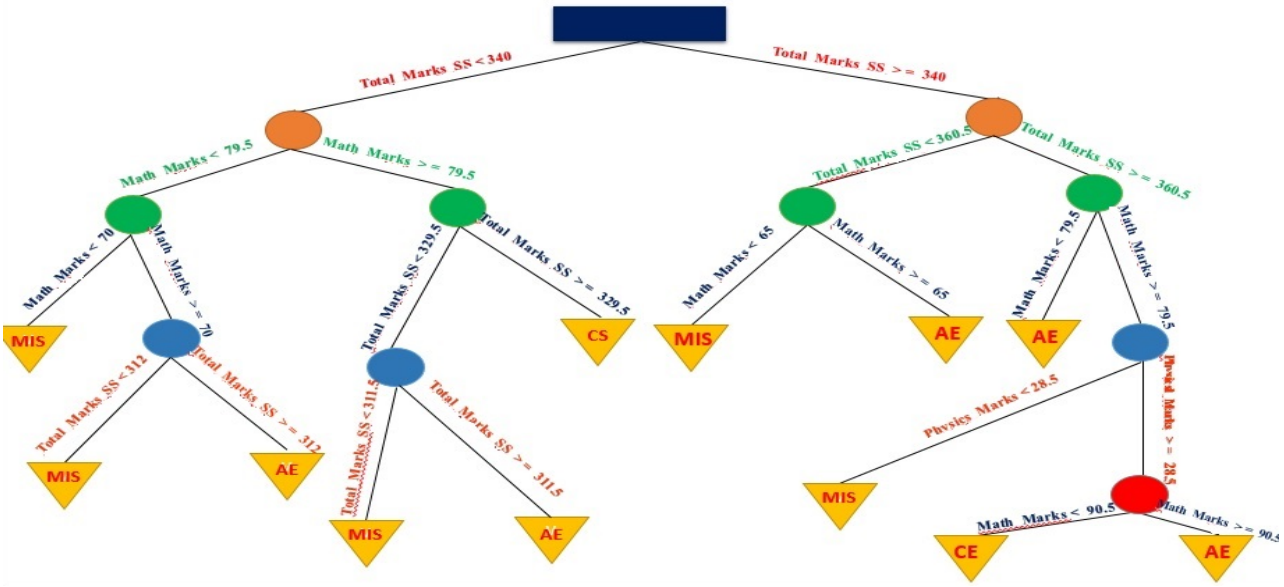


Fig. 2 Decision Tree

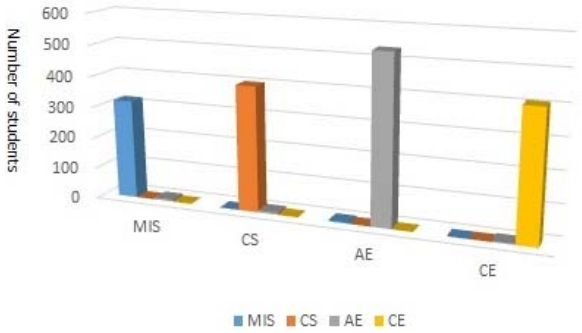


Fig. 3 Classification Output using C4.5

TABLE IV
CENTROIDS OF CLUSTERS

| Attribute | Cluster#1 | Cluster#2 | Cluster#3 | Cluster#4 |
|---|---|---|---|---|
| Total Marks | 269 | 344 | 347 | 351 |
| English Marks | 28 | 40 | 30 | 30 |
| Math Marks | 52 | 81 | 78 | 83 |
| Physics Marks | 28 | 33 | 24 | 37 |

Table V shows departments' distribution over the four clusters.

TABLE V
DEPARTMENT DISTRIBUTION IN CLUSTERS

| | Cluster#1 | Cluster#2 | Cluster#3 | Cluster#4 |
|---|---|---|---|---|
| MIS | 77.71% | 15.48% | 4.95% | 1.86% |
| CS | 0 | 40% | 36.25% | 23.75% |
| AE | 0.36% | 34.48% | 42.65% | 22.5% |
| CE | 0% | 43.03% | 15.16% | 41.81% |

The recommended department for students belonging to a certain cluster is the department with the highest percentage.

## VI. IMPLEMENTATION

Classification and clustering rules are acquired using Tanagra Data Mining software. A system was built using Java SE7 to implement the proposed framework, a graphical user interface was developed as shown in Fig. 4 to enable user to enter the student's data then show him/her are commendation for a department


Fig. 4 Framework GUI

## VII. CONCLUSION

In this paper, we have developed an intelligent Student advisory framework in the educational domain. We classified the students into the suitable department using C4.5 algorithm. Also, we clustered the students into groups as per the suitable education tracks using k-means algorithm. Finally, we have combined the results that came out from classification and clustering operations to predict more accurate results, all of these procedures were applied to improve the level of success of the first year university stage.

A case study was presented to prove the efficiency of the proposed framework. Students data collected from Cairo Higher Institute for Engineering, Computer Science and Management during the period from 2000 to 2012 were used and the results proved the effectiveness of the proposed

intelligent framework.

## REFERENCES

[1] T. M. Mitchell. Machine Learning. McGraw-Hill, New York, 1997.
[2] Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on, pages 63–67, 2010.
[3] A. M. El-Halees M. M. Abu Tair. Mining educational data to improve students' performance: A case study. International Journal of Information and Communication Technology Research, 2(2): 140–146, April 2011.
[4] S. Karthik M. Sukanya, S. Biruntha and T. Kalaikumaran. Data mining: Performance improvement in education sector using classification and clustering algorithm. In Proceedings of the International Conference on Computing and Control Engineering, ICCCE 2012, 2012.
[5] Mahfuza Haque Md. Hedayetul Islam Shovon. Prediction of student academic performance by an application of k-means clustering algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 2(7): 353–355, July 2012.
[6] M. tech Er. Rimmy Chuchra. Use of data mining techniques for the evaluation of student performance: a case study. International Journal of Computer Science and Management Research, 1(3): 425–433, October 2012.
[7] Brijesh Kumar Bhardwaj and Saurabh Pal. Data mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security, 9(4), April 2011.
[8] Md. Hedayetul Islam Shovon and Mahfuza Haque. An approach of improving students academic performance by using k-means clustering algorithm and decision tree. (IJACSA) International Journal of Advanced Computer Science and Applications, 3(8): 146–149, August 2012.