Data Mining: Analysis of student database using Classification Techniques

K. Sumathi, PhD
Assistant Professor,
Department of CS and IT,
Kalasalingam University,
Kirishnan Kovil.

S. Kannan, PhD
Associate Professor,
Department of Computer Applications,
MKU, Madurai.

K. Nagarajan
Chief Architect of Business Intelligence,
Tata Consultancy Services,
Chennai

ABSTRACT
Data Analysis can be categorized into two forms. One is used for extracting models describing important classes; another is to predict future trends. Data classification can be used to generate models which are further used to predict the unknown classes. The accuracy of the models can be examined by checking the percentage of correctly classified instance. Lot of classification algorithms is available nowadays. One of the most commonly used algorithms is decision tree because of its simplicity of implementation and easier to understand when compared to other classification algorithms. J48 is the one of the effective classification method. In this paper, J48 algorithm is applied for analyzing student dataset which includes academic year, department, academic grade and job position

Keywords
Data mining, Classification Techniques, Student Database

1. INTRODUCTION
Data mining is a step in Knowledge Discovery in Databases (KDD) and aims to discover useful information from huge amount of data [1]. The major role of data mining is applying various procedures and algorithms in order to retrieve patterns from huge data [2]. Nowadays data can be taken from different kind of large volume of datasets in various formats like flat files, videos, records, texts, images, audios, scientific data and new kind of data formats. The data collected from different sources require proper data analysis for efficient decision making process.

In earlier days data analysis process was manual and tough because domain knowledge was needed and understanding of statistical techniques was also needed. This manual process could not be applicable while facing the rapidly growing sizes data and large dimensions of the data. A community of researchers introduced the term named "data mining" to solve automating data analysis problem and discover the implicit information from the huge amount of data [3] (Giordana and neri,1995). One of data mining techniques is data classification which is used to generate models describing important data classes. The common classification methods used in data mining are 1) k-nearest-neighbor classifiers 2) decision tree classifiers 3) genetic algorithms 4) case-based reasoning, 5) Bayesian classifiers 6) rough sets and 7) fuzzy logic techniques. Decision tree algorithm is the most commonly used data classification algorithm because of its easier understanding and implementation.

Decision tree algorithms can be implemented in either serial or parallel form or in both serial and parallel form. Parallel implementation of decision tree algorithms is very useful in quickly generating results especially with the classification of huge data. When small or medium data sets are involved, serial implementation of decision tree algorithm will be used.

The major objective of this work is to use data mining methodologies to analyze student's job-position based on their academic performance. Data mining provides many tasks that could be used to analyze the performance of the student. In this research, the classification is used to predict the student job-position based on their academic performance. As there are many algorithms are used for data classification, in this paper the decision tree method is used for classification.

2. RELATED WORKS
Data mining has been widely applied in the higher education field as private arts and science colleges, Engineering Colleges, Polytechnic Colleges and universities provide huge amount of data. Some of the application is to study features that affect student retention through monitoring the academic performance and providing powerful methods to intervene as proposed by[4]. Bassil [2] proposed a model for typical university information system that based on transforming an operational database whose data are extracted from an already existing operational database.
The purpose of the model generation is to help decision makers and university principles for efficient decision making. Romero and Ventura in [4] gave a survey in applications of data mining in learning management systems and a case study class with the Moodle system. Fadl Elsid and Mirghani A. Eltahir [6] applied C4.8 algorithm for student database to predict the student academic performance and a case study with faculty of computer science and information technology, Nile Valley University. Ahmed et al. [5] used the classification process to predict the final grade, by presenting an analysis which can help the student's instructors to improve the student's performance.

3. DATA MINING
Data mining is the process of analyzing data in different angles and summarizing results it into useful information. Data mining software is analytical tools which allow the users to analyze data from many different dimensions, categorize the data, and summarize the relationships among data. Technically, data mining is the process of finding correlations among many numbers of fields in huge dataset.

Five major elements of data mining are: 1) Select, Transform and store data onto the data warehouse system. 2) Keep and handle the data in a multidimensional database system. 3) Allow business analysts and information technology professionals to access the data 4) Use application software to analyze the data 5) Display the data in a human understandable format, such as a graph or table.

Classification

Classification is a data mining technique that assigns items in a collection to target categories. The objective of classification is to accurately predict the category which is unknown for each case in the data. For example, a classification model could be used to identify student results as pass, good, very good or excellent.A classification task begins with a data set with known class labels. For example, a classification model which predicts student results might be developed based on observed data for students academic performance over a period of time.

In addition to the data might track previous performance, attendance percentage, general and technical attitude, and so on. Classification algorithm can be applied for categorical data. When the target is numerical the predictive model uses a regression algorithm.

Comparing the values of the predictors and the values of the target gives the accuracy of the classification model. The methods for obtaining relationships can be differed from one classification algorithm to another. If the accuracy percentage is acceptable, the model can then be applied to a different data set in which the class assignments are unknown.The dataset for a classification algorithm is divided into two data sets: 1. Training set is the one for building the model 2. Test set is the one for testing the model. Classification algorithms can be applied to many applications such as biomedical and drug response modeling, customer segmentation, business modeling, marketing and credit analysis.

Accuracy of the model refers to the percentage of correctly classified instance made by the model when compared with the actual classifications in the test data.

Clustering

Clustering analysis forms group of data objects that are similar in some sense to one another.

Members having identical attributes are grouped together to form a cluster. The objective of clustering analysis is to obtain high-quality clusters that minimize the inter-cluster similarity and maximize the intra-cluster similarity.

Clustering is used to segment the data. Clustering models categorize data into groups that were not previously defined where as classification models categorize data by assigning it to previously-defined classes, which are specified in a target.

Clustering algorithms can be used to find natural groupings when there are many cases and no obvious groupings. Clustering can also act as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models.

Clustering is mainly used for anomaly detection. Once clusters are formed, some data in some cases do not fit well into any clusters. These cases are outliers.

Association

Association is a data mining function that discovers correlations among the items in a huge dataset. Association rules are generated based on the relationships between co-occurring items. Association rules are often used to analyze business transactions. This application of association modeling is called market-basket analysis. Association modeling is used in other domains as well.

Data Mining Process

Data mining process discovers patterns from huge datasets involving techniques which combine artificial intelligence techniques, machine learning, statistics, Predictive analytics, and database systems.

Data mining is the process of discovering useful, hidden information from huge dataset. Data mining applies mathematical analysis to derive patterns and trends that exist in huge data. Data mining process is a series of steps and the important steps can be summarized in the following:

Data collection

Data collection is the process of gathering information usually with software. The first step of data mining process is collecting data from different kind of sources.

Data Preprocessing

Data preprocessing is a data mining technique that transforms original data into an understandable format. The real-world data is often incomplete, inconsistent, and is likely to contain many errors. Data preprocessing removes such problems. Data preprocessing prepares collected data for further processing.

Pattern Discovery

Pattern is discovered by classify students according to their graduation degrees, and academic year and performance. The goal of classification is to identify the distinguish characteristics of predefined classes, based on a set of instances, e.g. students performance, academic year and degrees etc., Pattern discovery requires mining and selection of attributes which describes its properties of a given class or category [7].

4. CLASSIFICATION MODEL IMPLEMENTATION

Students' information is collected over a period of time from the Faculty of Science & Technology around Madurai through questionery. The data set is divided in to test set and training set. Model is built using the training set using J48 algorithm. The objective of the work is to apply the model for test dataset and to find the performance of the model by classifying the new instance. This model is generated for obtaining the student job position based on their academic performance. If the performance of the model is acceptable this model may be used for the forth coming graduates to analyze the placement status.

The model to classify the instances of the sample training set which is provided in table 1. The actual data format is ARFF (studentdata.arff), WEKA source file which include academic year, department, final grade of graduated students and job-position, is given in Figure 1. The predicted instances (tested set), the file ARFF (datanew.arff) is given in Figure 2. The attribute section is identical to the training data. The file includes the tested instances of the attribute ("position"). The values of "position" attribute is left as "?", thus WEKA has no actual values to which it can compare the predicted values of new instances.

Table 1: Sample of students' academic year, department, academic results and job- position

Academic Year	Department	Grade	Position
2010	CS	Excellent	ITJobs
2010	IT	VG	IT Jobs
2011	CS	Good	ITJobs
2011	IT	Good	CallCenters
2012	CS	Good	ITJobs
2012	IT	Excellent	ITJobs
2013	CS	Good	Others
2013	IT	Good	ITJobs

Once the model is generated, it can be used to classify new instance. Here classification is used to classify new instance using J48 algorithm which is implemented in WEKA by the classifier.

The classification model generation process using studentdata.arff in WEKA is shown in Figure 3. Applying the same model to test dataset is shown in Figure 4. The running information of model generation is given in Figure 5, the running information of test dataset is given in Figure 6. The graphical versions of the decision tree are appeared in Figure 7 and Figure 8.

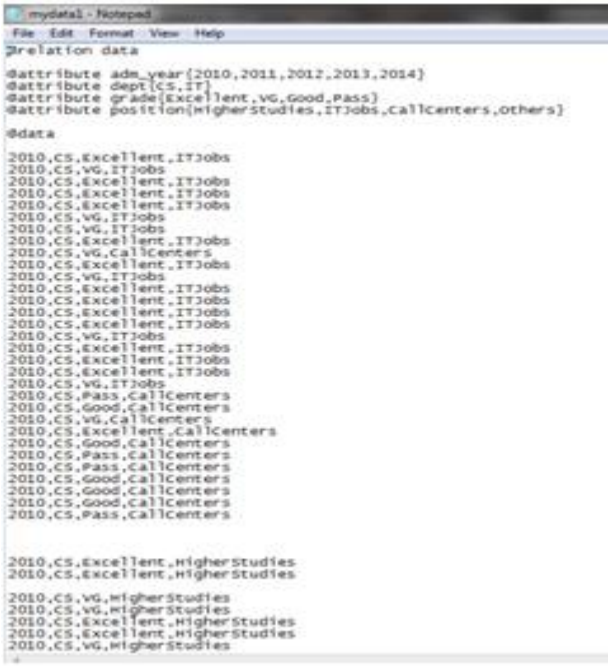

Fig : 1 Training Set Instance (.arff File)

The student database contains 78 attributes such as personal information, physical information, academic information etc., For this analysis the relevant attributes such as academic year, department of the graduate, final academic results and job-position of the graduate only used.

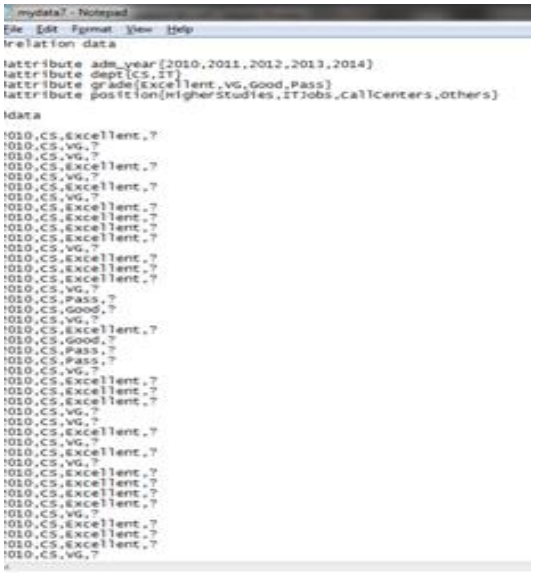

Fig : 2 The Testset Instance (.arff File)

The testset includes only three attributes such as academic year, department of the graduate, final academic results. job-position is the class attribute which will predicted by the classification model.

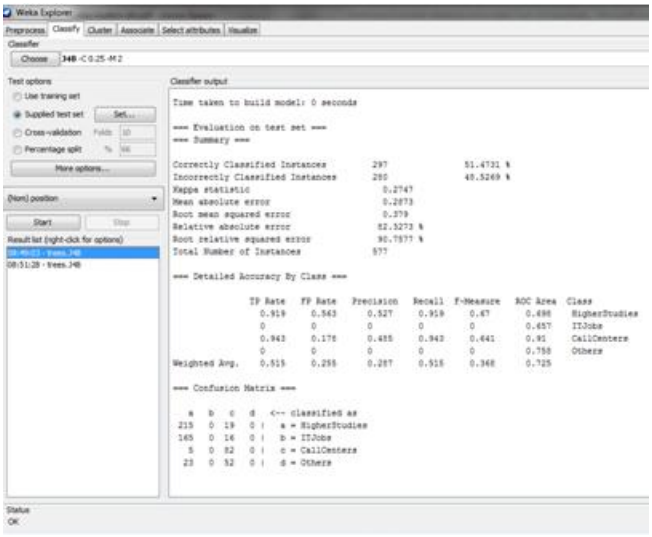

Fig : 3 Applying C4.5 (J48) classifier, WEKA for Training Set – Screenshot

Once the model is generated, the details of the model such as total number of instance in the dataset, the number of correctly classified instance, incorrectly classified instance, and accuracy of the model will be displayed

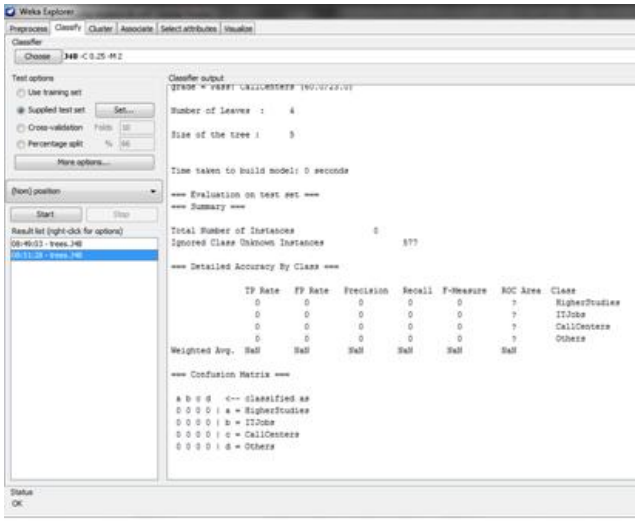

Fig : 4 Applying C4.5 (J48) classifier, WEKA – for Test set - Screenshot

Classifier Model(Training Dataset)

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: data

Instances: 577

Attributes: 4

adm_year

dept

grade

position

Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree

---

grade = Excellent: HigherStudies (288.0/131.0)

grade = VG: HigherStudies (120.0/62.0)

grade = Good: CallCenters (109.0/64.0)

grade = Pass: CallCenters (60.0/23.0)

Number of Leaves : 4

Size of the tree : 5

Time taken to build model: 0 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances 297 51.4731 %

Incorrectly Classified Instances 280 48.5269 %

Kappa statistic 0.2747

Mean absolute error 0.2873

Root mean squared error 0.379

Relative absolute error 82.3273 %

Root relative squared error 90.7577 %

Total Number of Instances 577

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.919	0.563	0.527	0.919	0.67	0.698	HigherStudies
0	0	0	0	0	0.657	ITJobs
0.943	0.178	0.485	0.943	0.641	0.91	CallCenters
0	0	0	0	0	0.758	Others
Weighted Avg. 0.515	0.255	0.287	0.515	0.368	0.725	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
215	0	19	0	a = HigherStudies
165	0	16	0	b = ITJobs
5	0	82	0	c = CallCenters
23	0	52	0	d = Others

Fig: 5 Training Dataset - Running InformationTest mode : user supplied test set: size unknown (reading incrementally)

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: data

Instances: 577

Attributes: 4

adm_year

dept

grade

position

Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree

---

grade = Excellent: HigherStudies (288.0/131.0)

grade = VG: HigherStudies (120.0/62.0)

grade = Good: CallCenters (109.0/64.0)

grade = Pass: CallCenters (60.0/23.0)

Number of Leaves : 4

Size of the tree : 5

Time taken to build model: 0 seconds

=== Evaluation on test set ===

=== Summary ===

Total Number of Instances 0

Ignored Class Unknown Instances 577

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0 0 0 0 ? HigherStudies

0 0 0 0 0 ? ITJobs

0 0 0 0 0 ? CallCenters

0 0 0 0 0 ? Others

Weighted Avg. NaN NaN NaN NaN NaN NaN

=== Confusion Matrix ===

a b c d <-- classified as

0 0 0 0 | a = HigherStudies

0 0 0 0 | b = ITJobs

0 0 0 0 | c = CallCenters

0 0 0 0 | d = Others

Fig: 6 Running Information Test Dataset

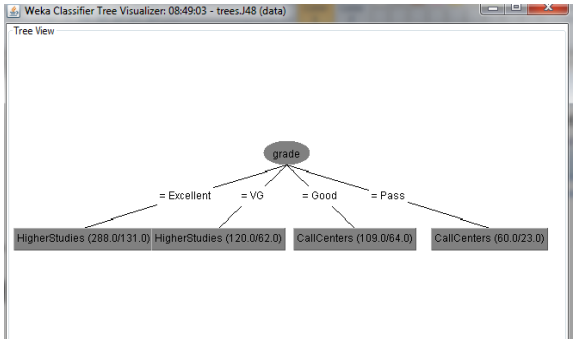

Fig : 7 - Decision Tree - for training dataset

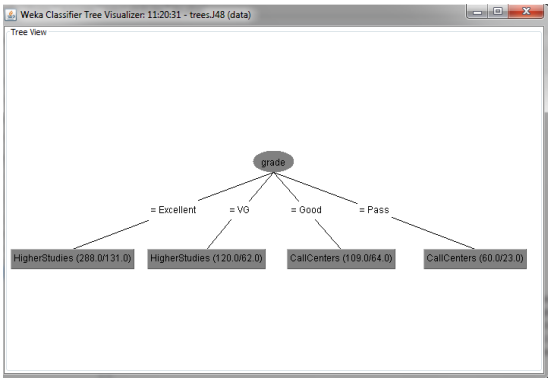

Fig : 8 – predicted decision Tree - for test dataset

Experiment results show that the efficiency and effectiveness of J48 algorithm in predicting the Job-Position based on academic results. The model also can improve the efficiency of the job-position retrieving and evidently promote retrieval precision.

Table- 2- Statistical Accuracy

	TP Rate	FP Rate	Precision	Recall	F - Measure	ROC Area	Class
	0.919	0.563	0.527	0.919	0.67	0.698	HigherStudies
	0	0	0	0	0	0.657	ITJobs
	0.943	0.178	0.485	0.943	0.641	0.91	Call Centers
	0	0	0	0	0	0.758	Others
Weighted Avg.	0.515	0.255	0.287	0.515	0.368	0.725	

In Table- 2 the ROC Area measurement is approximately greater than 0.6 for all classes that means the classification process succeeded in training the set. Thus, the predicted instances is similar to the training set, this proves the suggested classification model. The J48 algorithm can be used extracting and retrieving information to appear unseen information. The extracted information can be used to achieve the quality in many fields.

5. CONCLUSION

In this paper, the student database is analyzed using J48 classification algorithm and placement related information is predicted based on academic results. The various data mining algorithms that can support education system via generating valuable information are discussed. Data mining techniques can be very helpful in areas like prediction of academic results, student's placement and to improve students' academic results in higher institutions. The Data mining techniques also used in analyzing the student's academic results in different aspects as well as predicting the reason. Once data retrieved from the relevant resources, the classification algorithms can be applied to categorize the data. The feature selection of attributes from large data set plays an important role in efficiency of algorithm.

6. REFERENCES

[1] Jasser, Muhammed Basheer, et al. "Mining Students' Characteristics and Effects on University Preference Choice: A Case Study of Applied Marketing in Higher Education." International Journal of Computer Applications 67.21 (2013): 1-5.

[2] Bassil, Youssef. "A Data Warehouse Design for A Typical University Information System." arXiv preprint arXiv:1212.2071 (2012).

[3] Giordana.A and neri.F,(1995) "Search intensive concept induction Evolutionary computations", 3(4) 375-416, Doi10.1162/evco.1995.3.4.375

[4] Yu, Chong Ho, et al. "A data mining approach for identifying predictors of student retention from

sophomore to junior year." Journal of Data Science 8.2 (2010): 307-325.

[5] Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. "Data Mining: A prediction for Student's Performance UsingClassification Method." World Journal of Computer Application and Technology 2.2 (2014): 43-47.

[6] Tariq O. Fadl Elsid, Mirghani. A. Eltahir "Data Mining: Classification Techniques of Students' Database A Case Study of the Nile Valley University, North Sudan", International Journal of Computer Trends and Technology (IJCTT) – volume 16 number 5 – Oct 2014

[7] Jasser, Muhammed Basheer, et al. "Mining Students' Characteristics and Effects on University Preference Choice: A Case Studyof Applied Marketing in Higher Education." International Journal of Computer Applications 67.21 (2013): 1-5.