INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Mining: Student Database

Mahesh Vilas Jadhav
Student, Computer Engineering Department, Pune University, India
Swati Chandurkar
Professor, Computer Engineering Department, Pune University, India

Abstract:
Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Educational data mining is an emerging discipline concern with developing methods for exploring the unique type of data that come from educational setting and using those methods to better understand the student's performance. Mining in educational environment is called educational data mining. For mining educational DB the concepts named clustering, classification and association are used. It is concerned with developing new methods to discover knowledge from educational database. Educational data mining provides a set of techniques, which can help the educational system to overcome the performance issues. This paper presents the concepts needed to reduced drop-out ratio to a significant level and to improve the performance of students by discovering knowledge for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on.

Key words*: Educational Data Mining (EDM), Classification, Knowledge Discovery in Database (KDD), k-means Algorithm*

1. Introduction

The data gathered from various applications require proper method of extracting knowledge from large repositories for better decision making. The goal of data mining also referred as Knowledge discovery in databases (KDD), is to discover and extract patterns of stored data by applying various methods and algorithm. The data mining can be use in education. The field, called Educational Data Mining, deals with developing methods that discover knowledge from data originating from educational environments. Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- means, and many others. Knowledge can be discovered using these techniques such as association rules, classifications and clustering. The discovered knowledge can be used to predict the performance of a student, marks of students in a particular subject etc. The main objective of this paper is to study the performance of a student by data mining techniques. This paper presents various clustering and classification mechanisms to evaluate students performance.

1.1. Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes: operational or transactional data such as, sales, cost, inventory, payroll, and accounting nonoperational data, such as industry sales, forecast data, and macro economic data meta data - data about the data itself, such as logical database design or data dictionary definitions

1.2. Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when. Knowledge Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

2. Data Mining

Data mining also known as Knowledge Discovery in Database refers to mining or extracting knowledge from huge amounts of data. The techniques of Data mining are used to discover hidden patterns and relationships from the large volumes of data which will be helpful in decision making. Data mining provides the link between the transaction and analytical system. Data mining software analyzes the relationships and patterns in stored transaction data. Several types of analytical software are available: statistical, machine learning, and neural networks.

Various techniques like clustering, classification, Association rules are used for knowledge discovery from databases.

2.1. Classification

Classification is the data mining technique, which employs a set of pre-classified examples to develop a model that can classify the records. Classification function assigns items in a collection to target categories or classes. Stored data is used to locate data in predetermined groups.The goal of classification is to accurately predict the target class for each case in the data. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials. Another example of classification model could be used to identify loan applicants as low, medium, or high credit risks. The classification of data process includes learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Classification can be used to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, Cars can be classified into different types, by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you might apply it into a particular class by comparing the attributes with our known definition.

Classification technique in data mining uses variety of algorithms and the specific algorithm used can affect the way records are classified. A common approach for classifiers is to use decision trees to partition and segment the records. New records can be classified by traversing the tree from the root through branches and nodes, to a leaf representing a class. The path a record takes through a decision tree can then be represented as a rule. For example, "Income<$30,000 and age<25, and debt=High, then Default Class=Yes).

2.2. Clustering

Clustering technique finds clusters of data objects that are similar in some sense to one another. The members of a cluster are more like each other than they are like members of other clusters.Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high. Like classification Clustering is also used to segment the data but unlike classification, clustering models segment data into groups that were not previously defined. Classification models segment data by assigning it to previously-defined classes, which are specified in a target. Clustering models do not use a target. Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data pre-processing step to identify homogeneous groups on which to build supervised models. Clustering can also be used for anomaly detection. Once the data has been segmented into Clusters, it is find that some cases do not fit well into any clusters. These cases are anomalies or outliers.

Clustering can be said as identification of similar classes of objects. Clustering techniques helps to discover the overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object, but it becomes costly so clustering can be used as a preprocessing approach for attribute subset selection and classification.

2.3. Association

Data can be mined to identify associations.Association discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyze sales transactions. For example, it might be noted that customers who buy cereal at the grocery store often buy milk at the same time. In fact, association analysis might find that 85% of the checkout sessions that include cereal also include milk. This application of association modelling is called market-basket analysis. It is valuable for direct marketing, sales promotions, and for discovering business trends. Association modelling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for Web page personalization. An association model might find that a user who visits pages A and B is 70% likely to also visit page C in the same session. Based on this rule, a dynamic link could be created for users who are likely to be interested in page C.

Association and correlation is usually find frequent item set among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little value.

2.4. Regression

Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors.

A regression task begins with a data set in which the target values are known. In addition to the value, the data might track the age of the house, square footage, number of rooms, taxes, school district, proximity to shopping centres, and so on. House value would be the target, the other attributes would be the predictors, and the data for each house would constitute a case. In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown. Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values. The historical data for a regression project is typically divided into two data sets: one for building the model, the other for testing the model.

Regression modelling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modelling, and environmental modelling.

3 Comparison of Classification

- Algorithms

Algorithm	Function	Description
Decision tree (DT)	Classification	Decision trees extract predictive information in the form of human-understandable rules. The rules are if-then-else expressions; they explain the decisions that lead to the prediction.
Naive Bayes (NB)	Classification	Naive Bayes makes predictions using Bayes Theorem, which derives the probability of a prediction from the underlying evidence, as observed in the data.

4. Comparison of Clustering

Algorithms

Algorithm	Function	Description
k-Means (KM)	Clustering	K-Means is a distance-based clustering algorithm that partitions the data into a pre determined number of clusters. Each cluster has a centroid (centre of Gravity). Cases (individuals within the population) that are in a cluster are close to the centroid
Orthogonal Partitioning Clustering (O-Cluster or OC)	Clustering	O-Cluster creates a hierarchical, grid-based clustering model. The algorithm creates clusters that define dense areas in the attribute space. A sensitivity parameter defines the baseline density level.

5. Comparison of Association

- Algorithms

Algorithm	Function	Description
Apriori (AP)	Association	Apriori performs market basket analysis by discovering co-occurring items Within a set. Apriori finds rules with support greater than a specified minimum support and confidence greater than a specified minimum Confidence...
FP-growth algorithm(Frequent Pattern)	Association	In the first pass, the algorithm counts occurrence of items in the dataset, and stores them to 'header table'.In the second pass, it builds the FP-tree structure by inserting instances.

6. Conclusion

In this study data mining process in the student's database using k-means clustering algorithm and decision tree technique to predict student's learning activities is used. The information generated after the implementation of data mining and data clustering technique may be helpful for instructor as well as for students. This work may improve student's performance; reduce failing ratio by taking

appropriate steps at the right time to improve the quality of education. For future work, we hope to refine our technique in order to get more valuable and accurate outputs, useful for instructors to improve the students learning outcomes.

7. Acknowledgment

We express our sincere thanks to our Guide Prof. Swati Chandurkar, for her constant encouragement and support throughout our project, especially for the useful suggestions given during the course of the project and having laid down the foundation for the success of this work. We would also like to thank our Project Coordinator Mrs. Deepa Abin, for her assistance, genuine support and guidance from early stages of the project. We would like to thank Prof. Dr. J. S. Umale, Head of Computer Department for his unwavering support during the entire course of this project work. We are very grateful to our Principal Dr.A.M.Fulambarkar for providing us with an environment to complete our project successfully. We also thank all the staff members of our college and technicians for their help in making this project a success. We also thank all the web committees for enriching us with their immense knowledge. Finally, we take this opportunity to extend our deep appreciation to our family and friends, for all that they meant to us during the crucial times of the completion of our project.

8. References

1. B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.
2. Hedayetul Islam Shovon, Mahfuza Haque "An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree.",(IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3, No. 8, 2012
3. Oyelade, Oladipupo, Obagbuwa,"Application of K-means clustering algorithm for prediction of student's academic performance.",IJCSIS2010,vol.7,No.1,pp-292
4. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar & M.Inayat Khan, ""Data Mining Model for Higher Education System ", European Journal of Scientific 2010), pp.24 Research, ISSN 1450-216X Vol.43 No.1
5. U. K. Pandey, and S. Pal," "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8,Issue 2,pp. 277-282,ISSN:1694-0814,2011.
6. Vashishta, S. (2011).Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. International Journal of Advanced Computer Science and Applications, 2(4), 77-80