Data Mining on University Database

Jasmin Ramadani[1],Sime Arsenovski[2], Ruben Nuredini[3], Zoran Gacovski[4]

Abstract – **The paper is focusing on using Data Mining technologieson University Database of student data for extracting knowledge to improve the management of enrolment of new students. Using the data of current students placed in Oracle Database, the goal is to use Oracle Data Mining to classify the students in order to predictof distribution of the students based on their characteristics that will help to recognize better the target groups of future students and improve the enrolment process.**

Keywords –**Data Mining, Students, Oracle, ODM, Algorithm.**

I. Introduction

The enormous use of electronic databases on the higher education institutions has brought to large amount of data that are stored and not converted to valuable knowledge. The data is hiding the knowledge which means that we have to use some techniques to extract and use that information. The best suited technique for this purpose is the Data Mining. It can be defined as an analysis of large amounts of data in order to find undiscovered relations and to present them on a new way that they will be understandable and useful [1].Some authors are using the term Data Mining as a synonym for Knowledge Discoverywhile others consider the Data Mining as a part from that process. The process of Data Mining can be used to find new information hidden in the databases which can be used to support many issues of the educational process. In the paper we will describe the process of Data mining on Oracle database of student records in order to extract valuable knowledge.

II. Knowledge Discovery Process

As central part of the process of Knowledge Discovery are the methods of data mining but there are also other parts that are important. The process (Fig.1) can be simply represented as a composition of three main elements [2]:

1. Preparation of data
2. Algorithm of Data Mining
3. Analysis of the data

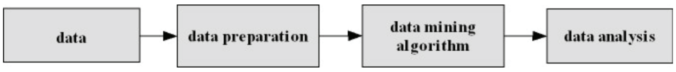


Fig. 1 The process of Knowledge Discovery

On the beginning of the process we must deal with the data, its form and condition. Based on that, we will prepare the data for the process of Data Mining. This is eminent if there are some irregularities in the database. After the Data Mining is done the data should be analyzed so it can be used and properly presented. The process of Knowledge Discovery is an interactive process where every step has it own importance although the step of Data Mining is the most important.

III. Techniques

There are various techniques of Data Mining that can be used to find different patterns in the data. The number of techniques can vary, but we can identify five main techniques of Data Mining [3]:

- Classification
- Prediction
- Clustering
- Association
- Summarization

In the paper we will use the classification technique to get the probabilities for the distribution of the students.The classification is a very frequent technique of Data Mining and it checks the characteristics of a database object after which it sets the object in a class which was defined previous. The classification organizes the data in classes on basis of certain attributes. The goal is to make a model that could be used on unclassified datawhich takes the attributes as input and it gives the class as output. The number of classes can be two, n but also we can have many classes.

IV. Algorithms

There are many algorithms that can be used in the process of Data mining. What algorithm we will use it depends from the technique that we use, the conditions and the technology we use. For the technique of classification most used algorithms are Decision Trees and Naïve Bayes.

For creation of the model of Data Mining in this paper we use the technique of classification which is using the “Naïve Bayes” algorithm.For the Naïve Bayes algorithm we are making classification using the Bayesian classification which is statistic classification of data [5]. We take the calculation of the possibility that one object belongs to one of the classes [6]. To calculate the possibility we use the Bayesian theorem:

[1]Jasmin Ramadani is with the Faculty of Information and Communication Technologies at FONUniversity b.b, 1000 Skopje, Macedonia,E-mail: jasmin.ramadani@fon.edu.mk

[2]Prof. Dr. Sime Arsenovski is with the Faculty of Information and Communication Technologies at FONUniversity b.b, 1000 Skopje, Macedonia, E-mail: sime.arsenovski@fon.edu.mk

[3]Ruben Nuredini is with the Faculty of Information and Communication Technologies at FONUniversity b.b, 1000 Skopje, Macedonia, E-mail: ruben.nuredini@fon.edu.mk

[4]Zoran Gacovski is with the Faculty of Information and Communication Technologies at FON University b.b, 1000 Skopje, Macedonia, E-mail: zgacovski@yahoo.com

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \qquad (1)$$

where *P(X)* is the probability that the event will happen, *P(Y)* is the probability for the event Y while *P(X|Y)* is the prior probability for the event Xunder condition that the event Y has happened and *P(Y|X)* is the posterior probability of X conditioned by Y.

V. Oracle Data Mining

The Oracle Data Mining (ODM) is very capable tool for the process of Data Mining. The software is part of the Oracle Database Enterprise Edition. The ODM package is constructed of two parts [8]:

- Data Mining API
- Data Mining Server

The first part contains java classes and methods for creatingmodels of Data Mining. The second part is server side component which is making the Data Mining.

ODM supports the most used techniques of Data Mining. It is chosen to be used for the process of Data Mining in this casebecause it has many advantages. The complete process is done inside the Oracle Database which means that there is no need to export the data in to other environments. This is a very important advantage because in this way the process is more secure, more stable and the number of errors is smaller. It is very important to keep the database in the system which will remove the possibility of error and inconsistence which can appear if the database is moved. Also ODM is avoiding the possibility of using old data which is nit update because it uses the current database condition.The complete preparation of the data like cleaning and transforming is done with the ODM tools.Another great advantage is that if we use ODM, the models are staying in the Oracle database.

The Oracle Data Miner is another part of ODM which is the graphic interface that enables the visual creation of the process of Data Miningand it gives access to all the activities and functions.

VI. Data Mining Integration in Higher Education

The higher education institutions are storing large quantities of data about the students.This means that the institutions are storing data that are strategic resourcethat can be used for improving of the quality of different processes in the institution [9].The Data Mining as a process is well used in other areas like banking, trade, insurance where the data amount is very large. This means that we can try to implement the Data Mining in the educational system. In order to extract the hidden knowledge for the database we will use the techniques of Data Mining. The knowledge found can be used to support decisions and resolve problems in the management of the institution like the increasing of the efficiency, marketing decisions, enrolment of new students etc. In our case we want to find information that will help the University to get some information from its own database about the characteristics of the students on several campuses that will lead to improvements.

VII. Goal of the Research

The universities and faculties are defining different strategies they will use in the process of enrolment of new students. The higher education institutions could use the data collected when the student enroll to find some relations and knowledge that can predict some directions in the next year of enrolment. The data collected contains personal and demographic data like gender, place of living, department of studying, faculty etc. This kind ofinformation can help the institution to concentrate the efforts of marketing using the demographic factors. The classification of the students based on the above mentioned characteristics can group and predict the future student behavior. In this paper we have chosen to define the grouping of the students on the faculties of Detectives and Security and the Faculty of Economics based on the campus location. This could help the University to predict what percent of the students are going to enroll on different campuses on the mentioned faculties. The faculties are chosen based on their greater popularity among all the faculties on the University. We will try to build the models of data mining using the following students attributes in the database:

- Campus
- Faculty

Where the attribute campus contains the city where the campus is located and the attribute faculty is showing the name of the faculties. We will try to show any tends about that how the students are divided on base of the location of the campus on particular faculty. We will use the Data Mining to find the trends and to predict the distribution of the students. The benefits of the research will be the information we will get that will show the management of the institution about the possible interests of the students to enroll on particular students on the different campuses. According to that information the management can decide either to maintain or close some faculties on different campuses and to find out which faculties need more marketing on the campuses where there is lack of interest,

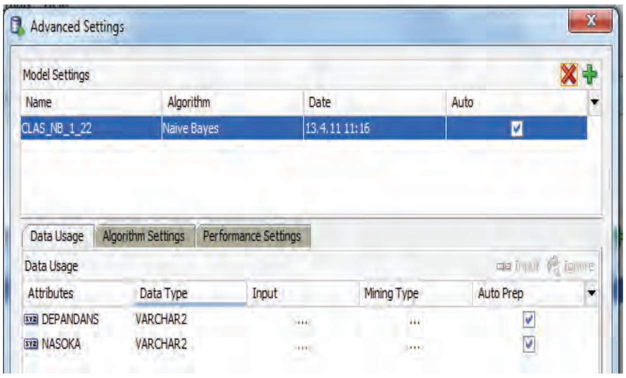
VIII. Modelling

After resolving the problem we will choose this case is concerning the possible management of the enrolment strategybased on the data enrolled students in 2009 on the FON University in Skopje. The data is stored in Oracle Database where with ODM we will try to show the information which was extracted.For the model we will use the attributes faculty and campus which are stored in the database of the students.

The data is prepared where we are treating the missing values in the tables of the database and other irregularities such the duplicate entries in some attributes like the name of the Faculties where it was found that there are different names for the same faculty in the student database. This needs correction of the duplicates in order to avoid the errors in the results.

After this step we choose the model of Data Mining that will be used. ODM gives support for several techniques and algorithms.After the analysis of the data and the possible outcomes it is chosen to use the model of classification using the Naïve Bayes algorithm.

The model of Data Mining is created in Oracle Data Miner 11g Release 2 which can be found in the Oracle Sql Developer 3.0.



Model Settings

Name	Algorithm	Date	Auto
CLAS_NB_1_22	Naive Bayes	13.4.11 11:16	☑

Data Usage

Attributes	Data Type	Input	Mining Type	Auto Prep
DEPANDANS	VARCHAR2	....	...	☑
NASOKA	VARCHAR2	....	....	☑

Fig. 2 Settings of the Data Mining Model

The result will show is it possible to group the students on basis of the city where the campus is placedand they study and the faculty that they will enroll to study. The prediction can be made using the attributes and the data from the current student database. The model should show what faculties are popular on different campuses of the University.

This information can help us to predict the distribution of the students on different faculties. As an example we will show how the students on the Faculty of Detectives and Security (Table1) and the Faculty of Economics (Table2) are dispersed on the different campuses in Skopje, Struga, Strumica and Gostivar.

TABLE 1.PROBABILITIES FOR THE FACULTY OF DETECTIVES AND SECURITY

Attribute	Value	Probability
Campus	Skopje	61,07784431
Campus	Gostivar	20,95808383
Campus	Strumica	12,57485030
Campus	Struga	5,38922156

The results in Table 1 show that there is high probability of about 61 percent that the future students of the Faculty of Detectives and Security will study on the campus in Skopje. That means that more than the half of all students of this faculty will be studying in the campus in Skopje. The smallest percent is found for the campus in Struga which shows that on this campus we cannot expect significant interest.This means that there is place for actions that should be taken by the management to increase the number of enrolled students of this faculty on the other campuses of the University

TABLE 2. PROBABILITIES FOR THE FACULTY OF ECONOMICS

Attribute	Value	Probability
Campus	Skopje	79.33333333
Campus	Struga	10.00000000
Campus	Gostivar	8.000000000
Campus	Strumica	2.666666667

From the results in the Table2 for the Faculty of Economics we can see that the biggest percent of probability found for the campus in Skopje. This shows that according to the technique of classification and the algorithm used that most for the students on the Faculty of Economics, about 79 percent will be studying on the campus un Skopje. In contrary there is very small percent which is about 2.66 percent that there will be students on the campus in Strumica. The information can affect the of the process of enrollment by concentrating the resources on the campuses where there interest about the particular faculty is smaller

IX. CONCLUSION

The process of Data Mining can help the Higher education institutions to use the knowledge hidden in the data they store every year about their students. The data does not show the possible all the relations and information that can help us to find valuables information and recognize future trends. In our case in this paper the Data Mining model shows some information about that how thecharacteristics of the students are connected with the city of the campus where they study and the faculty they are attending. Based on the current data using the model of classification in the Data Mining we can classify the students and predict the future distribution of the students on the campuses divided by the faculty. As an output we have the percent of students on every campus for one faculty.The information can help the University to determine which faculties on different campuses are popular and which are not. This could improve the process of marketing and enrolment of new students showing in which places there is need for more activities in order to increase the number of students. The benefit of the use of Data Mining for the purposes of improvement of the enrollment process will be in the information we will get which shows the trends of the future of the distribution of the students on the different campuses. This will show the justification of the existence of some faculties on different campuses and the need for greater marketing efforts and improvements to increase the popularity of different faculties.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth, “Principles of Data Mining”, MIT Press, 2001.

[2] S. Sumathi, S.N. Sivanandam,”Introducтion to Data Mining and its Applications”, SpringerVerlag,2006

[3] H.H. Hsu,”Technologies in Bioinformatics”, IDEA Group Publishing,2006.

[4] R. Agrawal, T. Imielinski and A. Swami, “Mining as sociation rules between sets of items in large data bases”, in Proceedings of ACM-SIGMOD Conference, Washington, DC, 1993.

[5] R. Hanson, J. Stutz, P. Cheeseman, “Bayesian classification theory”, Technical report, 1991.

[6] M. Pretzer, “Clustering und Klassifikation”, Oldenburg Universitat, 2003.

[7] J. Han, M. Kamber, ”Data Mining Concepts and Techniques”, Morgan Kaufman Publishers, 2001.

[8] K. Hauke, M.L. Owoc, M. Pondel, ”Building Data Mining Models in the Oracle 9i Environment”, Wroclaw University of Economics, Poland, 2003.

[9] N. Delavari,”Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System”, Faculty of Information Technology, Multimedia University (MMU), Cyberjaya, Malaysia.