Expert Systems with Applications 41 (2014) 6400–6407

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Student data mining solution–knowledge management system related to higher education institutions

Srečko Natek [a,*], Moti Zwilling [b]

[a] International School for Social and Business Studies, Slovenia
[b] Netanya Academic College, Israel

ARTICLE INFO

Keywords:
Data mining
Knowledge management system
Student's success rate
Data mining for small data set
Higher education institution
Educational data mining

ABSTRACT

Higher education institutions (HEIs) are often curious whether students will be successful or not during their study. Before or during their courses the academic institutions try to estimate the percentage of successful students. But is it possible to predict the success rate of students enrolled in their courses? Are there any specific student characteristics, which can be associated with the student success rate? Is there any relevant student data available to HEIs on the basis of which they could predict the student success rate? The answers to the above research questions can generally be obtained using data mining tools. Unfortunately, data mining algorithms work best with large data sets, while student data, available to HEIs, related to courses are limited and falls into the category of small data sets. Thus, the study focuses on data mining for small student data sets and aims to answer the above research questions by comparing two different data mining tools. The conclusions of this study are very promising and will encourage HEIs to incorporate data mining tools as an important part of their higher education knowledge management systems.

1. Introduction

In knowledge management process, data mining technique can be used to extract and discover the valuable and meaningful knowledge from a large amount of data. Nowadays, data mining has given a great deal of concern and attention in the information industry and in society as a whole. This technique is an approach that is currently receiving a great attention in data analysis and it has been recognized as a newly emerging analysis tool (Osei-Bryson, 2010; Sinha & Zhao, 2008; Tso & Yau, 2007; Wan & Lei, 2009; Zanakis & Fernandez, 2005; Zhuang, Churilov, Burstein, & Sikaris, 2009; Becerra-Fernandez, Gonzales, & Sabherwal, 2004).

In the literature, one may notice that they are many areas which adapted this approach to solve their problems such as in finance, medical, marketing, the stock market, telecommunication, manufacturing, health care and customer relationship. However, the data mining application has not attracted much attention from people in relation to educational systems, despite the majority of HEIs admit, that course's students success rates are very important for their prestige and senior management of the academic institution. Thus the possibilities to deeply understand the reasons and accurately predict the student success have become very valuable (Dermol & Čater, 2013; Marjetič & Lesjak, 2012; Natek & Lesjak, 2013; Rojko, Lesjak, & Vehovar, 2011; Trunk Širca, Babnik, & Breznik, 2013).

OLAP solutions and statistical methods are well established tool to analyze data but data mining enable fresh approach to understand hidden patterns and data prediction. In particular, data mining has turned to be a very popular among researches because many "approachable" standalone or desktop data mining tools are available in the market. From the observation of different tools, one can notice the following as an example: Microsoft Excel, SPSS, Weka, Protégé as Knowledge Acquisition System and Rapid Miner. Some of them (e.g. MS Excel Mining tool) are normally available to HEI's professor and they can benefit from existing knowledge of the Excel. That is the reason to include it in the research. Weka was choosed among other desktop DM tools because of supporting the data mining analysis in very different way, compared with MS Excel. Moreover, Weka is considered as a very legitimate tool analysis where education management data is involved (Cristóbal, Sebastián, & Enrique, 2008; Romero & Ventura, 2006).

HEIs recently promote knowledge management as encouraging data mining environment for their professors and researchers.

* Corresponding author. Tel.: +386 41630053.
E-mail addresses: srecko.natek@mfdps.si (S. Natek), Moti.Zviling@gmail.com (M. Zwilling).

http://dx.doi.org/10.1016/j.eswa.2014.04.024
0957-4174/

In the recent years, data mining has been recognized as a contemporary tool for building knowledge management systems (Jashapara, 2011). A knowledge management approach to data mining process, otherwise associated with business intelligence technology, brings some new synergy and added value to data mining solutions (Wang & Wang, 2008). This study was focused on HEI's knowledge management systems as part of their educational and research work (Baker & Yacef, 2009; Delavari, Somnuk, & Beikzadeh, 2008).

Unfortunately most data mining techniques work best with very large samples. Following that, Andonie (2010), reported that some data mining techniques, such as neural networks may not be able to accomplish the learning task (during the training phase) since small datasets cannot provide enough data to fill the gaps between too small samples. As a consequence, several authors concluded that small datasets limit the scope of data mining technique (Yuan & Fine, 1998).

In the real-life there are many academic situations, where relatively small data sets are normal, at least from the HEIs point of view. Therefore, the data which is collected on courses that students take is a good example of such situation. Even if the course is attended by a relatively large group of students the relevant data is considered to be a small dataset.

Andonie (2010) even suggested that there is no universal optimal solution to the problem of small datasets. So what possibilities are available to the HEIs if they want to explore the success rate of students in various courses using contemporary but widely available data mining tools? Student's data, available to HEIs is limited and clearly fall into the category of small data set (Table 1). The described approach in this study is very popular and well supported by data mining tools.

The research goal is to focus on small student's data set which is analyzed using data mining approaches, thus trying to answer various HEIs questions, using the student's courses data and the data mining tools for analysis (which are normally available to average HEIs). For HEIs the most important outcome are relatively reliable students' success rate prediction and user friendly experience. Bukowitz and Williams (2000) advocate how technology is blurring the distinction between types of knowledge from unknown knowledge to a known one. Their study supports the conceptual model as data mining on student's data sets which discovered some interesting key influencer and "Final grade" prediction to a chosen course. Moreover, as stated by the authors, small HEIs data sets can be used for analysis with data mining techniques. In these cases specific algorithms can be used to overcome the limit of data which is often not the case since most algorithms are working on a large scale databases (Wu et al., 2008).

2. Methodology

2.1. Building student success rate data mining model using the MS Excel tool

Several authors suggest different steps for building data mining model, focusing mainly on data transformation process. Turban advocate data preparation model with data consolidation, data cleaning, data transformation and data reduction phases (Turban, Sharda, & Delen, 2011).

Berry and Linoff defined comprehensive data mining building model as an iterative process with several steps (Berry & Linoff, 2000). The adopted model was used to establish the student success rate data mining model related to International School for Social and Business Studies in Slovenia:

The first step was data requirements identification where HEIs require predicting the student's success rate related to the specific course.

Data acquisition focused on Informatics courses for Bachelor Study Programs – Economy in Contemporary Society for the academic years 2010/11 (42 students), 2011/12 (32 students) and 2012/13 (32 students). The data was imported from the web application 'Novis' – Higher Education Information System – Professor Module to MS Excel Table Tools with Data Mining add-in and MS SQL Server, which was installed on a laptop computer.

Validate, explore, clean and transpose data were the next activities where irrelevant columns were deleted (e.g. group, remark, individual activities points, registration number), translation from Slovene to English (columns title and content) was implemented and columns order were adjusted accordingly (logical order for success prediction, the last column was set as the predictable attribute: "Final grade"). For more transparent prediction the different statuses were transformed (from several status lists to yes/no value). To avoid personal data problems, the student names were transformed into numbers.

In addition, to increase the informative value of prediction some derived variables were added (additional columns): study year, gender (taken from the student name) and the type of study (by merging full time and part time students).

Creation model data sets is the central step of building data mining model with the following student data sets columns (Table 1): Study year (2010/11, 2011/12, 2012/13), Student (order number 1–106), Gender (female, male), Student year of birth (e.g. 1988), Employment (no, yes), Status, e.g. sport etc. (no, yes), Registration (first, repeat), Type of study (full time, part time), Exam condition (no, yes), Activities points (0–50), Exam points (0–50), Final points (0–50) and predictable Final grade (1–10). The data sets were divided into two groups. The first group of data sets

Table 1
Examples of students data sets (attributes taken are: year of study, Student #, gender, year of birth, employment, status – taken sport course, registration status, Type of study either full-time/part-time, exam condition, points for activity participation, exam points, final points, Final grade out of 10).

Study year	Student	Gender	Year of birth	Employment	Status (sport...)	Registration	Type of study	Exam condition	Activities points (50)	Exam points (50)	Final points (100)	Final grade (10)
2010–2011	1	Female	1988	No	No	First	Full time	Yes	46	46	92	10
2010–2011	2	Male	1990	No	No	First	Full time	Yes	38	33	71	7
2010–2011	3	Female	1990	No	No	First	Full time	Yes	39	30	69	7
2010–2011	4	Female	1990	No	No	First	Full time	Yes	47	35	82	8
2010–2011	5	Female	1989	No	No	First	Full time	Yes	39	36	75	7
2010–2011	6	Male	1990	No	No	First	Full time	Yes	38	30	68	7
2010–2011	7	Female	1990	No	No	First	Full time	Yes	39	36	75	7
2010–2011	8	Female	1990	No	Yes	First	Full time	Yes	39	33	72	7
2010–2011	9	Male	1990	No	No	First	Full time	Yes	39	38	77	8
⋮												
2012–2013	106	Female	1990	No	No	First	Part time	Yes	44	30	74	