Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 15 (2011) 4028–4032

Procedia
Social and Behavioral Sciences

WCES-2011

A sample study on applying data mining research techniques in educational science: developing a more meaning of data

Kenan Zengin[a], Necmi Esgi[a], Ergin Erginer[b,*], Mehmet Emin Aksoy[a]

[a]Gaziosmanpasa University, Computer and Instructional Technology Education, Tokat 60250 Turkey
[b]Gaziosmanpasa University, Curriculum and Instruction Department, Tokat 60250 Turkey

Abstract

The purpose of this research is to present a sample study analyzing data gathered from an educational study using data mining techniques appropriate for processing these data. In order to achieve this aim, a "Computer Self-efficiency Scale" used in educational sciences was selected and this scale was applied in a study group. Data was analyzed using descriptive statistics (t-test and analysis of variance), and the data mining techniques of decision tree, dependency networks and clustering. The descriptive statistics used were calculated not using common statistical software packages, but by running a program written in Delphi 2009 programming language on Microsoft SQL Server 2008. Microsoft SQL Server 2008 was directly used for the data mining techniques of dependency networks and clustering. Some of the findings of the research, which cannot be obtained by common statistical techniques but can be obtained by data mining methods, were as follows: "those who think they are competent with computer terms and concepts believe they have a special talent in using computers" ; "those who believe they have a special talent in using computers feel as if the computer is part of their body", and "students who have been using computers for more than six years believe they have a special talent in using computers".

Keywords: Educational sciences, data mining.

Introduction

The increase in the volume of the scientific knowledge and its transformation into a means to an end, from being an end in itself, necessitates radical changes in the methods used to gather, analyze, and evaluate data. The quantitative capacity of research data improves reliability of the research and representativeness of the samples, but the increase in quantitative capacity renders analysis of data more difficult. In such cases, technology is here to help the researcher, by offering methods to analyze data faster.

The functional steps of the scientific method used in research, from a macro perspective, are "perception of a difficult situation, discovery and definition of the problem concerning the situation, examination of possible solutions and formulation of hypotheses, and then, based upon the results of hypothesis testing, suspension, retention, rejection, or amendment of the hypotheses" (Erturk, 1986:106), based on Dewey. In this process, the transformation of knowledge into scientific knowledge requires diversification of research methods. It can be observed that in the cycle of scientific method, the process of scientific thinking, from the definition of the problem to testing and evaluation, is processed by human labor, technology and research and the products of this process are then used for the benefit of the human beings (Erginer, 1995: 116).

[*] Ergin Erginer. Tel.: +90-3562521616; fax: +90-3562521546.
E-mail address: erginer@gop.edu.tr

doi:10.1016/j.sbspro.2011.04.408

The necessity to diversify research methods in the process of the transformation of knowledge into scientific knowledge necessitates, in turn, diversification in the methods used to evaluate data. It is obvious that information that is not processed scientifically by empirical processes and not organized cannot be used as reliable data. Because incremental increases in information observed in very short intervals of time makes it difficult to obtain reliable data, more attention needs to be paid to the study of data analysis methods. Data mining has gained a prominent place among these methods in recent years, due to its reliability and conveniences it offers to researchers.

"In the preface to the proceedings book of the conference on Knowledge Discovery in Databases, held for the first time in 1995, the mountains of data created by information technologies are emphasized as follows: It is estimated that the amount of information on earth is doubled every 20 months. What are we supposed to do with this flood of raw information? To the naked eye, only a small fraction of that information is visible. Computers promise to be fountains of wisdom, but they cause floods of data" (Akpinar, 2000:1).

Data mining, which can be described as "the generation of previously unknown, valid and applicable information from large databases and their use in decision making (Silahtaroglu, 2008:10)" provided the researchers with an in-depth flexibility in data analysis, and made it possible to make more sense of the data.

It is seen as a modeling that measures the process of data analysis by the process used in data mining and as an approach that starts with the definition of the problem, proceeds with the examination of data sources and ends with the steps of data preparation, modeling and evaluation, and offers the researcher the opportunity to make very comprehensive and flexible analyses (Akpinar (2000, 1).

Data mining is not a solution in itself, but a tool that supports the decision making process in finding a solution and that provides the information required to solve a problem. Data mining helps the analyst to discover patterns and relationships contained within the data (Baykal, 2006, 96).

"Piatetsky-Shapiro, one of the leading experts in the field, defines data mining as the extraction of previously unknown, tacit and possibly useful information in a non-monotonous process. This process includes many different technical approaches like clustering, data summarization, learning about classification rules, finding out dependency networks, analyzing changes, and detecting anomaly (Akpinar, 2000).

Considering that computer interface is only a tool in data mining (Silahtaroglu, 2008, 11-15), it has many fields of application from banking to marketing, from insurance to health and genetics and is used in the gathering and analysis of data in many sensitive tasks from detection of fraud to gene mapping and to the detection of tendencies on any subject.

In the studies so far, use of data mining in the field of education is very recent (Peña, Domínguez ve Medel, 2009) and in Turkey the focus is directly on data applications rather than methodological studies (Ataseven, 2008; Ozcinar, 2008; Sezer, 2008; Sadic, 2008; Tosun, 2007; Dogan, 2006; Kaygulu, 1999). Most of these studies are conducted within the disciplines of computer science and business management. Studies that are expected to contribute to educational sciences because they increase the predictability of exam results abound, although they are not prepared by educational sciences departments.

Method

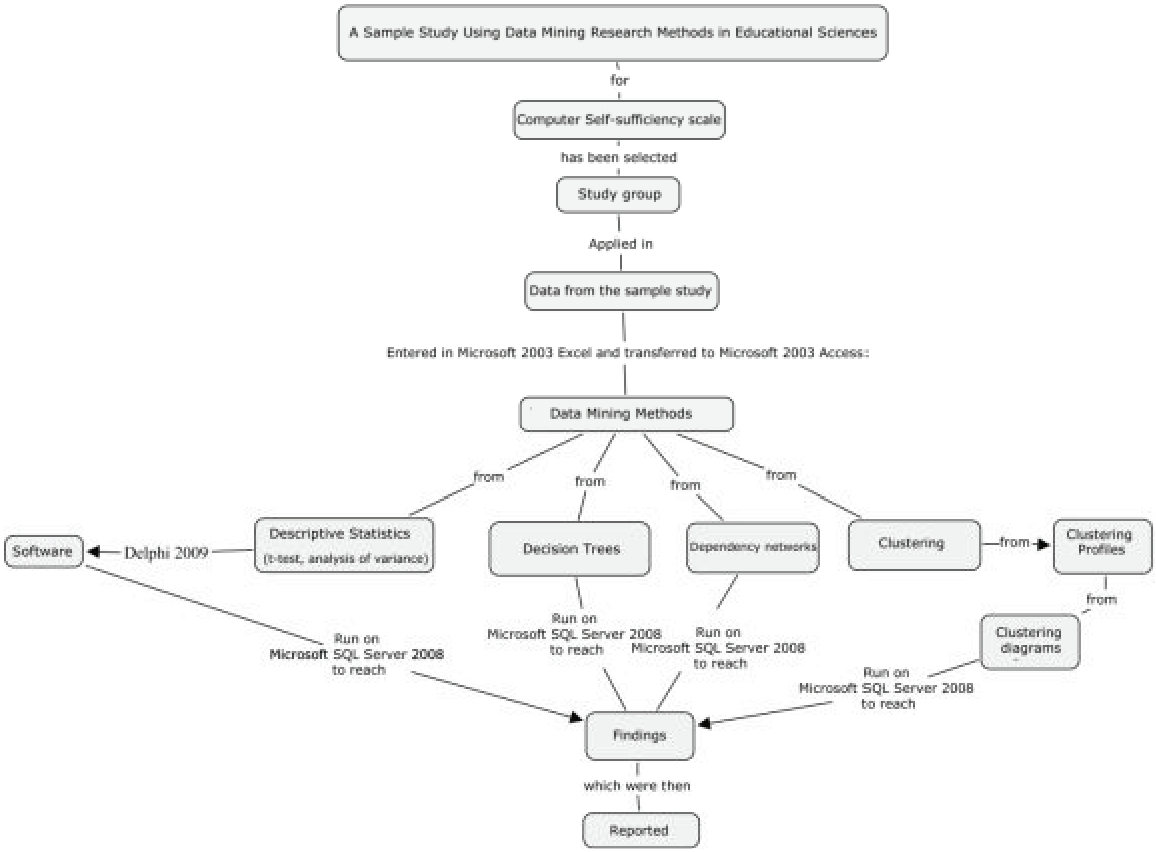
Research Aim: To present a sample study analyzing data gathered from an educational study using data mining techniques appropriate for processing these data.

Significance of Research: In educational science studies, most of the time descriptive statistics (t-test, analysis of variance, etc.) are used. In other social science branches, data mining methods started to be frequently used in recent years. Besides descriptive statistics, data mining allows use of methods like clustering, dependency networks, and decision tree and enriches the evaluation of studies conducted in many ways. In data mining, significant and even seemingly insignificant results obtained using descriptive statistics can be processed in other ways. It is sometimes possible to reach an interpretation not provided by a method via another method. In other words, data obtained from the databases are "mined" to unearth different relationships and definitions. Data mining provides the researcher with flexibility in the selection of methods to be used in research, statistics to be calculated and the interpretation to be made of the results. In addition, the researcher can interpret the data obtained from the research without being dependent on the use of statistical software packages (like SPSS, MYSTAT and SSTAT). He or she can select multiple methods appropriate for the data and analyze the data in depth, using different dimensions.

The sample study conducted is significant both because it exemplifies use of data mining in educational sciences, not commonly used in the field and because regarding the field of computer programming, this is the first time a software to calculate descriptive statistics was written using Delphi 2009 programming language and run on Microsoft 2008 SQL Server.

Research Design: Steps of the research design and implementation are given in Figure 1.

Figure 1. Steps of the research design and implementation


A Sample Study Using Data Mining Research Methods in Educational Sciences
for
Computer Self-sufficiency scale
has been selected
Study group
Applied in
Data from the sample study
Entered in Microsoft 2003 Excel and transferred to Microsoft 2003 Access:
Data Mining Methods
from
from
from
from
Software
Delphi 2009
Descriptive Statistics (t-test, analysis of variance)
Decision Trees
Dependency networks
Clustering
from
Clustering Profiles
from
Clustering diagrams
Run on Microsoft SQL Server 2008 to reach
Run on Microsoft SQL Server 2008 to reach
Run on Microsoft SQL Server 2008 to reach
Run on Microsoft SQL Server 2008 to reach
Findings
which were then
Reported

Study Group of the Research: The scale was applied by Aksoy to a total of 531 senior university students, majoring in seven different departments (Computer and Instructional Technology Education, Psychological Counseling and Guidance Education, Elementary Education, Social Sciences Education, Science Education, Music Education and Painting Education) of the Faculty of Education of Gaziosmanpaşa University in the 2009-2010 fall semester.

Data Processing Tool Used in the Study: A five-point likert type educational scale was applied to a group of students. The data gathered from the research were analyzed using appropriate data mining techniques. In the sample study, a five point likert type computer self-sufficiency scale, developed by Aşkar and Umay (2001) and consisting of 18 items was used.

Data Processing/Analyzing Tools Used in the Study: Of the data mining methods, those appropriate for processing the data gathered were used; namely, descriptive statistics (t-test, analysis of variance), decision tree, dependency networks, and clustering. The data gathered using the scale were first entered in a Microsoft Office 2003 Excel file and then turned into a Microsoft SQL Server 2008 database. Then, a software to make the calculations for the study was developed using the Delphi 2009 programming language. The study uses Microsoft SQL Server 2008 Database Management System. Using the Business Intelligence Development Studio contained within the SQL Server 2008, an Analysis Services Project was generated, and Data Mining Applications were run. The algorithms used were, besides the Dependency Networks algorithm, Microsoft Decision Trees and Microsoft Clustering algorithms.

Limitations: The sample study was limited by the views of the 531 senior university students, the 2003 Office Excel, 2003 Office Access, Delphi 2009, Microsoft SQL Server 2008 programs and the results of the computer self-sufficiency scale. In addition, the sample study used descriptive statistics (t-test, analysis of variance), the data mining methods of decision tree, dependency networks and clustering, which were appropriate for the data at hand. For descriptive statistics, one item for each was presented as an example and explained.

Findings

The study exemplifies application of descriptive statistics, decision tree, dependency networks and clustering to educational science data.

1- Descriptive Statistics Examples (t-test and analysis of variance)

Calculation of descriptive statistics is one of the methods of evaluation used in data mining. This method is used in all fields dealing with quantitative data. Also a data mining method, descriptive statistics include statistics like t-test, one way analysis of variance, etc. Besides data mining, the calculations of these statistics are usually done using commercial software packages. Some of these are statistical software packages like SPSS, MYSTAT, and SSTAT. Descriptive statistics can also be calculated without using these software packages.

Calculation of descriptive statistics via data mining techniques provides a lot of flexibility to the researcher, differently from the commercial statistical software packages. Thus, the researcher comes to have a natural control mechanism over the research data. In the sample study, a new software was written using the Delphi 2009 programming language to calculate descriptive statistics from the data gathered. The software, generated using codes, cycles and algorithms, was used to make t-test and analysis of variance calculations.

The sample analysis of variance application contained analysis of variance calculations on whether there are significant differences between the groups (1st group: Computer Education and Educational Technologies; 2nd group: Psychological Counseling and Guidance Education; 3rd group: Primary Education; 4th group: Social Sciences Education; 5th group: Science Education; 6th group: Music Education; and 7th group: Painting Education) with regards to the mean scores received from the first item of the scale (I believe I have a special talent in using computers); and if there are, what causes these differences.

2 - Decision Tree Example

All of the data were processed using the decision tree model and as an example, the decision tree differentiation of 12th item on 1st item was identified. Those who “sometimes think” they are competent with computer terms and concepts (12th item, level 2) were differentiated into two groups as those who “often believe” they have a special talent in using computers (1st item, level 4), and those who “rarely believe” so.

All other items had a homogenous distribution.

3-Dependency Network Example

All of the data were processed using the decision tree model and as an example, a dependency relationship was found between the following items.

1) It was found that there was a (two-way, strong) connection between 1st item ⇄ 12th item. In other words, those who think they are competent with computer terms and concepts (12th item) also believe that they have a special talent in using computers (1st item). Similarly, those who believe that they have a special talent in using computers (1st item) also think that they are competent with computer terms and concepts (12th item). A strong and visible two-way connection was found between these two items.

3- Clustering Example

In the clustering algorithm, department and the duration of computer use were entered as inputs. As an example, the 2nd item (I am talented with computers) and the 7th item (I fear doing something wrong or hitting a wrong button when I use a computer) were entered as prediction items. The clustering algorithm run returned 10 separate clusters. The cluster diagrams of the 10 separate clusters were displayed.

Conclusions

The researcher, if he/she has basic statistical and computer skills, can make desired statistical calculations via data mining methods without having to use statistical software packages (S.P.S.S, SSTAT, MSTAT, etc.).

In studies that use scales as data gathering tools, because limitless inquiry opportunities are offered to the researcher when examining the views, conducting an in-depth tendency analysis is made possible.

In quantitative studies, it facilitates categorization and prediction of qualitative data when needed.

Because it enables making in-depth inquiries into the research data, it offers an opportunity to make more sense of the data.

It enables a more visual presentation of the research data via figures and diagrams and facilitates comprehension of research results.

It saves time in cases where in-depth data analysis methods are to be used.

References

Akpinar, H. (2000). Veri tabanlarinda bilgi kesfi ve veri madenciligi, *I.U. Isletme Fakultesi Dergisi*, 29(1), 1-22.

Ataseven, S. (2008). Universitelerin adaylar tarafindan tercih edilme desenlerini veri madenciligi yontemleri ile belirleyen bir model onerisi, *Yayimlanmamis Yuksek Lisans Tezi,* Istanbul: Istanbul Kultur Universitesi, Fen Bilimleri Enstitusu.

Baykal, A. (2006). Veri madenciligi uygulama alanlari, *D.U. Ziya Gokalp Egitim Fakultesi Dergisi,* **7**. 95-107.

Dogan, B. (2006). Zeki ogretim sistemlerinde veri madenciligi, *Yayimlanmamis Doktora Tezi,* Istanbul: Marmara Universitesi, Fen Bilimleri Enstitusu.

Erginer, E. (1995). Bilimsel arastirma ve teknoloji uretiminin yaygınlastirilmasinda neden-nasil sorularinin cozumlenmesini egitimsel bir bakis acisi ile desenleme, *Verimlilik Dergisi*, 1995(2), 109-120.

Erturk, S. (1986). *Diktaci tutum ve demokrasi,* Ankara: Yelkentepe Yayinlari.

Kaygulu, M.S. (1999). Supervised and unsupervised learning techniques in data mining, *Unpublished Master Thesis,* Izmir: Graduate School of Natural and Applied Sciences of Dokuz Eylul University.

Ozcinar, H. (2008). KPSS sonuclarinin veri madenciligi yontemleriyle tahmin edilmesi, *Yayimlanmamis Yuksek Lisans Tezi,* Denizli: Pamukkale Universitesi, Fen Bilimleri Enstitusu.

Peña, A., Domínguez , R., Medel, J. de J. (2009). Educational data mining: A sample of review and study case, *World Journal on Educational Technolog,* 2, 118-139.

Sadic, S. (2008). Data mining including application of cognitive maps and decision tree algorithm, *Unpublished Master Thesis,* Istanbul: Istanbul Technical University, Institute of Science and Technology.

Sezer, U. (2008). Karar agaclarinin birliktelik kurallari ile iyilestirilmesi, *Yayimlanmamis Yuksek Lisans Tezi,* Kocaeli: Kocaeli Universitesi, Fen Bilimleri Enstitusu.

Silahtaroglu, G. (2008). *Veri madenciligi,* Istanbul: Papatya Yayincilik.

Tosun, S. (2007). Siniflandirmada yapay sinir aglari ve karar agaclari: Ogrenci basarilari uzerine bir uygulama, *Yayimlanmamis Yuksek Lisans Tezi,* Istanbul: Istanbul Teknik Universitesi, Fen Bilimleri Enstitusu.