International Journal of the Physical Sciences Vol. 5(9), pp. 1428-1437, 18 August, 2010
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 ©2010 Academic Journals

Full Length Research Paper

Data mining in topology education: Rough set data analysis

Serkan Narli[1]* and Z. Ahmet Ozelik[2]

[1]Department of Primary Mathematics Education, Faculty of Education, Dokuz Eylul University, Izmir, Turkey.
[2]Department of Mathematics Education, Faculty of Science, Dokuz Eylul University, Izmir, Turkey.

Accepted 24 August, 2010

This study explores rough sets, which provide mathematical description for uncertain concepts that can not be defined clearly by traditional logic. The study further examines data mining, which helps to discover meaningful information from large data sets by incorporating it with rough set theory. An investigation in topology instruction was undertaken to exemplify the usability of rough set theory in qualitative data analysis. The study provides an example of educational application of data mining in a topology course.

Key words: Artificial Intelligence, rough sets, educational data mining, topology.

INTRODUCTION

Is Serkan a young person? This is an unfortunate question for Serkan who is 33 years old. Looking for the answer to this question in mathematics is meaningless for many mathematicians. Because, mathematics requires the concepts to be certain, or many scientists accept this situation like that. In other words, a concept is mathematical if it possesses a certainty (Frege, 1904).

However, as in the question "Is Serkan a young person?" vague concepts are abundant in daily life. These vague, which we may also call uncertain knowledge, occupied the human mind for centuries. According to Frege (1904), uncertain concepts are those that are related to boundary-line view. That is, an uncertain concept is the one that has some objects not only outside or inside of it, but also on its boundary. Philosophers, psychologists, and currently computing engineers and mathematicians have shown interest in this topic. Now, we are faced with questions such as how can we understand uncertain knowledge? Or how can we formulate uncertain knowledge?

Currently, scientists, particularly those who focus on artificial intelligence are seeking for answers to such questions. It is not an easy task to formulate uncertain concepts that may not involve mathematically definite results. Therefore, there is need for alternative mathematical concepts for mathematical formulation of such concepts. To answer such questions, mathematicians have began to look for different fields of research.

These new mathematical approaches are seen mostly in a fundamental concept of mathematics, sets. For this purpose, several set theories have been developed that are alternative to George Cantor's leading set theory. The mereology theory by Lesniewski (1915), alternative set theory by Vopenka (1970), the penumbral set theory by Apostoli and Kanada (1999), Fuzzy, intuitive, and soft set theories by Zadeh (1965), and rough sets by Pawlak (1982) are some examples of the new set theories (Pawlak, 1997).

The first successful application of uncertainty approaches is the Fuzzy sets, defined by Zadeh in 1965. In this approach, membership of an element in a set is defined via a membership function. In other words, in Fuzzy sets, one can not say that an element certainly belongs to a set or not, one can only say that an element belongs to a set at a certain degree.

Another successful uncertainty approach is the rough sets defined by Pawlak (1982). After being introduced, these sets have been used as a mathematical tool to extract information from incomplete or uncertain data (Pawlak, 1991, 1995). Rough set theory can be used in

*Corresponding author. E-mail: serkan.narli@deu.edu.tr. Tel: +90.232 420 48 82-1365, +90.505 525 00 17(Mobile). Fax: +90.232.420 48 95.

data reduction, detection of dependences, estimation of the importance of data, forming control algorithms from data, approximate classification of data, detection of similarities and differences within data, detection of patterns in data, and detection of cause-effect relationships (Pawlak and Slowinski, 1994; Aydoğan and Gencer 2007). Rough sets are used for these purposes as illustrated in the literature (Kent, 1994; Lin and Cercone, 1997; Nings et al., 1995; Pawlak et al., 1995; Polkowski and Skowron, 1998 a, b; Zhong and Showron, 2000; Yorek and Narli, 2009; Polkowski, 2002; Slowinski, 1992).

Educational data mining

The field of education is a place where one can often face ambiguous situations. There is a growing interest among researchers that use data mining in educational technologies, or educational data mining (EDM). Baker (in press) defines EDM as a field of scientific research that focused on development of methods for investigating a particular type of data obtained from educational settings, and to use those methods to improve students' learning and the context, in which they learn. EDM studies mostly concentrate on detecting patterns in educational data, but there are studies that investigate the ways of using these patterns in student modeling (Amarshi and Conati, 2009).

There are several fields from which EDM methods are derived such as data mining and machine learning, psychometrics and other areas of statistics, information visualization, and computational modeling. Romero and Ventura (2007) categorized EDM studies as the follows:

1. Statistics and visualization
2. Web mining:

i) Clustering, classification, and outlier detection
ii) Association rule mining and sequential pattern mining
iii) Text mining

A second viewpoint on educational data mining is given by Baker (in press), which classifies work in educational data mining as follows:

1. Prediction:

i) Classification
ii) Regression
iii) Density estimation

2. Clustering
3. Relationship mining:

i) Association rule mining
ii) Correlation mining
iii) Sequential pattern mining
iv) Causal data mining:

4. Distillation of data for human judgment
5. Discovery with models

EDM studies have focused mostly on tutoring systems where structured problem solving (e.g., (Sison et al., 2000; Zaiane, 2002; Baker et al., 2008)) or drill and practice activities (e.g. (Beck, 2005)) are supported. EDM methods may differ from the broader data mining literature in explicitly utilizing the levels of meaningful hierarchy in educational data.

As a data mining method, rough set data analysis is used in EDM. For instance, Narli (2010) discusses the usability of rough sets in the analysis of attitude data obtained in educational studies. In addition, there are biology literature that involve the modeling of the construction of life concept with the help of Fuzzy-rough sets (Yorek and Narli, 2009) and the use of rough sets in the classification of attitudes toward nature (Narli at al., 2010).

How students' incorrect or partially correct ideas affect their learning is an important area of research. It is not an easy task to develop a theory to determine students' misunderstandings. Questions such as 'How student responses represent the level of their understanding? Are their responses similar to each other?' These are important question in working toward explaining student ideas. This study seeks the usability of rough set data analysis to answer such questions. Therefore, in this study, rough sets and data analysis using rough sets are described and a sample rough set data obtained in a topology course are analyzed.

PRELIMINARIES

In the sense of traditional set concept, the set is a well defined collection of objects. In other words, an element either belongs to a set or not. For instance, the set of odd numbers is that kind of a set because a number is either odd or not. However, in daily life not everything can be seen or defined in this clarity.

Let us think about a set of young people. Unlike the set of odd numbers, the set of young people can not be defined with precise boundaries. In fact, the majority of concepts often used in daily life are uncertain concepts, which have no definite boundaries. This situation forced researchers to investigate and look for alternative set theories. In the following section, rough sets, a successful example of the alternative set theories, will be explained.

Rough sets

Rough set theory is an extension of traditional set theory. In this set, a subset of a universal set is defined by two sets called lower and upper approximations. The basic tool in Pawlak's rough sets is an equivalence relation. The lower and upper approximations are built through equivalence classes (Aktas and Cagman, 2005). After Pawlak's definition, other rough set theories are suggested using different algebraic structures instead of an equivalence relation (Bonikowaski, 1995; Jiashang, Congxin and Degang, 2005; Kumar, 1993; Kuroki, 1997; Pomykala and Pomykala, 1998; Narli and Ozcelik, 2008).

Scientists all over the world showed interest in Pawlak's rough set theory. Pawlak defined rough sets as the follows:

Let U be a finite universal set; $R \subset UxU$ is an equivalence relation, and $A \subset U$:

i. The union of all equivalence classes, included in the set A, formed in U according to relation R is called the lower approximation of the set A according to relation R, ($R_*(A)$);
ii. The union of equivalence classes which formed in U according to relation R and which have non-empty intersection with the set A is called the upper approximation of the set A according to relation R, ($R^*(A)$);
iii. The difference of upper approximation set from the lower approximation set is called the boundary region of the set A according to relation R, ($B_R(A)$).

Now, let R(a) represent the equivalence class of the element $a \in A$ and these can be defined with the following relations:

$R_*(A) = \cup_{a \in U} \{ R(a) : R(a) \subset A\}$
$R^*(A) = \cup_{a \in U} \{ R(a) : R(a) \cap A \neq \varnothing\}$
$B_R(A) = R^*(A) - R_*(A)$.

According to rough set theory, the set $R_*(A)$, depending on the property defined by the relation R, is formed with elements, which definitely belong to set A. The elements of the set $R^*(A)$, depending on the property defined by the relation R, are the elements that potentially belong to set A.

In view of these definitions, set A is called a crisp set if its boundary region is empty, and a rough set if its boundary region is non-empty. These definitions are schematically shown in Figure 1.

Lower and upper approximation sets posses the following properties (Pawlak, 1997):

1. $R_*(X) \subseteq X \subseteq R^*(X)$
2. $R_*(\varnothing) = R^*(\varnothing) = \varnothing; R_*(U) = R^*(U) = U$
3. $R^*(X \cup Y) = R^*(X) \cup R^*(Y)$
4. $R_*(X \cap Y) = R_*(X) \cap R_*(Y)$
5. $R_*(X \cup Y) \supseteq R_*(X) \cup R_*(Y)$
6. $R^*(X \cap Y) \subseteq R^*(X) \cap R^*(Y)$
7. $X \subseteq Y \rightarrow R_*(X) \subseteq R_*(Y); \mathrm{R}^*(X) \subseteq R^*(Y)$
8. $R_*(-X) = -R^*(X)$
9. $R^*(-X) = -R_*(X)$
10. $R_*R_*(X) = R^*R_*(X) = R_*(X)$
11. $R^*R^*(X) = R_*R^*(X) = R^*(X)$

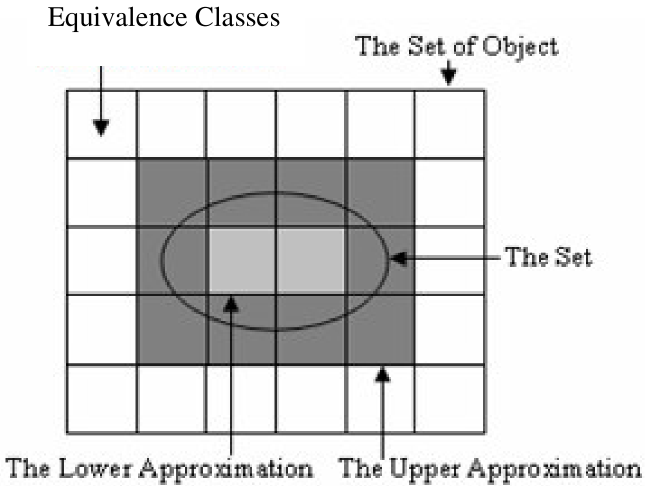

Figure 1. Schematic representation of a rough set.

Rough sets can also be defined via approach membership functions instead of lower and upper approximations. Accordingly, $|A|$ being the cardinality of the set A, approach membership function of the set A according to equivalence relation R, μ_A^R is defined as follows:

$$\mu_A^R : U \rightarrow [0,1]$$

$$x \rightarrow \mu_A^R(x) = \frac{|A \cap R(x)|}{|R(x)|}$$

Approach membership function, μ_A^R determines the possibility of the element x being a member of the set A according to relation R. It is obvious that this possibility will be within closed interval [0,1] and using the membership function, the lower and upper approximation sets and boundary region can be represented as follows:

$$R_*(A) = \{x \in U : \mu_A^R(x) = 1\}$$
$$R^*(A) = \{x \in U : \mu_A^R(x) > 0\}$$
$$B_R(A) = \{x \in U : 0 < \mu_A^R(x) < 1\}$$

Table 1. Patients' flu condition and the symptoms (Pawlak, 1997).

Universal set	Condition attributes			Decision attribute
Patient	**Headache**	**Muscle ache**	**Fever**	**Flu**
H_1	No	Yes	High	Yes
H_2	Yes	No	High	Yes
H_3	Yes	Yes	Very high	Yes
H_4	No	Yes	Normal	No
H_5	Yes	No	High	No
H_6	No	Yes	Very high	Yes

In addition, rough sets can be characterized by a constant that belong to closed interval [0, 1]. The constant that will determine the clarity of the approach, which is defined as:

$$\alpha_R(A) = \frac{|B_*(A)|}{|B^*(A)|}.$$

It should be readily understandable that if $\alpha_R(A) = 1$ then the set A is an exact set; otherwise it is a rough set.

Rough set data analysis

Data in rough set analysis are presented as an attribute-value table such that every row in the table represents an object (or sample) and every column shows a property that characterizes the object. This table is called an information table or decision table. A simple example of an information table is shown in Table 1.

It can be seen that there are three columns in the information table named as 'universal set, condition attributes and decision attribute'. The universal set includes six patients. In rough set data analysis, there may be many condition attributes, but in this case there are three condition attributes namely headache, muscle ache and fever. Similarly, there may be a number of decision attributes. In the above example, only one decision attribute was determined, whether it is having flu or not. The rows of the information table contain individuals, objects, or samples. These are marked by the elements of the universal set. For instance, in Table 1, patients (H1, H2, H3, H4, H5, H6) constitute the rows and the condition, and decision attribute values for each patient are presented in the patient's row (Munakata, 1998). Condition attributes altogether define equivalence relation R, which can also be named as indiscernibility relation. The lower and upper approximation sets are determined by equivalence classes formed according to this relation.

Dependence of attributes

Another important topic in rough set analysis is to determine the dependences among attributes. Intuitively, if a decision attributes set Q is determined by a condition attributes set P, then Q and P can be said to be dependent. In the above example, condition attributes and decision attribute can be defined as P = { Headache, Muscle ache, Fever} and Q = {Flu} respectively. In this case, whether a patient is flu or not may depend on to what degree he/she shows the P attributes. In cases such as this, the relationship between P and Q attribute sets can be determined functionally using rough set theory:

Let P and Q be the condition and decision attributes, respectively. If k is calculated as shown, Q is said to be dependent to P at the level of k *(0≤ k≤ 1*) and is represented as $P \Rightarrow_k Q$:

$$k = \gamma(P,Q) = \frac{|POS_P(Q)|}{|U|}$$

where $POS_P(Q)$, is called positive region of the division U/D according to P and is defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} P_{alt}\ (X).$$

If $k = 1$, *Q* is totally dependent on *P*, if $k < 1$, *Q* is partially dependent on *P*. $\gamma(P,Q)$ defines the closeness of the division *U/Q* and estimation of it according to conditions in *P*. The coefficient *k* represents the degree of dependence (Pawlak, 1997).

Reduced attribute sets

Let the set of condition attributes in an information table be P, the set of decision attributes Q, and the indiscernability (equivalence) relation defined by the set P be *IND(P)*. If the attribute subset $B \subset P$ encloses indiscernability relation *IND(P)*, then *P-B* attributes can be disregarded. These attributes are redundant and disregarding them does not deteriorate the classification. The subsets that do not include attributes that could be disregarded are called attribute sets. Reduced attribute set of an information table can be presented as follows:

If

$IND(B) = IND(P)$

and

$IND(B-\{a\}) \neq IND(P)$

then

$B \subseteq P$ is the smallest attribute set.

The intersection of all reduced attribute sets is called core. The core can be an empty set. The set of all reduced sets of an attribute set P is denoted as RED(P).

METHODS (APPLICATION IN A TOPOLOGY COURSE)

The sample

The sample of this study consists of pre-service mathematics teachers in a state university in Turkey. In total, 70 students who enrolled in a topology course were administered a written test and the data were analyzed using rough set theory.

Materials and procedure

All the students completed a written test, which was developed by the researcher (SN). The content validity of the test was provided by two mathematicians who have taught topology courses before. The expert opinions were positive about the test's validity in measuring the intended concepts. The written test is provided as follows:

The questionnaire

Let R be the set of real numbers; *Q* the set of rational numbers; Z the set of whole numbers*; N the set of* natural numbers *and (R,U) t*he usual topological space:

Q1. According to usual topology, what is the interior of the set of natural numbers N ($\overset{o}{N}$)?
Explain your answer: --

Q2. According to usual topology, what will be the boundary of the set of rational numbers *Q* (∂Q)?
Explain your answer: --

Q3. According to usual topology, what are the isolation points of the set of whole numbers *Z*?
Explain your answer: --

Q4. According to usual topology, determine the closure points of the set of whole numbers *Z.*
Explain your answer: --

Q5. Is the family of $\tau = \{Z \cap T : T \in U\}$ a topological structure on *Z*?
Explain your answer: --

Students were given 90 minutes to complete the test. A rubric was developed by the researchers to score students' papers. The data obtained were analyzed using rough sets. The intention of this analysis was to investigate to what degree the first four questions can explain the last (fifth) question.

Rough sets analysis of the questions

As explained earlier, in rough set analysis, data were tabulated in an attribute-value table such that rows in the table include an object or sample, and columns include attributes that characterize the object. The attribute values are obtained either by measurement or human experience. In this study, each question in the test is set as an attribute and responses to the questions are regarded as measurement of the attributes. Students' scores are determined using a rubric and they were grouped as 2, 1, 0 with respect to their answers such that students who provided completely correct answer were put into group 2, those who had partially correct answer into group 1, and those who provided completely wrong answer into group 0.

Provided by additional information that come from an expert or in most situations resulting from classification, in other words, condition attribute, which are concepts family to be estimated is chosen as the fifth question in the study. According to previous descriptions, the information table for the present study is presented in Table 2.

It can be seen in Table 2 that the students who have same scores in the first four questions, might have different scores from the decision class that is the fifth question. For instance, students x_1, $x_{6,}$ and x_7 provided completely correct answers to all four questions. However, student x_1 could partially answer the last question, student x_6 who answered the last question was completely correct, and student x_7 could not provide a correct answer to the fifth question. How can one argue that student x_6 have the complete correct solution to the fifth question, or that student x_7 do not know the solution, and that x_1 partially knows the solution? How precise this argument can be? For instance, student x_2 got no points from the second question, and he/she obtained partial points from the questions one, three, and four. However, he/she provided completely the correct answer to the fifth question. Can one argue that student x_2 is less successful than students x_1, $x_{6,}$ or x_7? To what degree is the success in the first four questions related to success in the fifth question? The answers to such questions are sought via rough set data analysis.

Indiscernibility relation

The universal set in this study is the set of students, U = { x_1, x_2,..., x_{70}}. Considering all "condition attributes" in Table 2, the equivalence relation (indiscernibility relation) R determined by these attributes will divide the set of students U into the following equivalence classes:

Table 2. Information system for topology dataset.

Students	Condition attributes				Decision class
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
x_1	2	2	2	2	1
x_2	1	0	1	1	2
x_3	2	0	2	0	1
x_4	0	0	0	0	0
x_5	2	2	1	2	2
x_6	2	2	2	2	2
x_7	2	2	2	2	0
x_8	2	2	2	2	0
x_9	0	0	0	0	0
x_{10}	2	2	2	2	1
x_{11}	2	2	2	2	0
x_{12}	2	0	2	2	1
x_{13}	2	2	2	2	1
x_{14}	2	0	2	2	0
x_{15}	2	0	2	2	1
x_{16}	2	1	0	0	0
x_{17}	1	0	0	0	0
x_{18}	1	0	0	0	1
x_{19}	0	0	0	0	0
x_{20}	1	0	2	0	1
x_{21}	0	0	0	0	0
x_{22}	0	0	0	1	0
x_{23}	1	0	0	1	0
x_{24}	0	1	0	0	0
x_{25}	0	0	1	1	0
x_{26}	1	0	0	0	1
x_{27}	0	0	1	0	1
x_{28}	2	2	2	2	0
x_{29}	2	0	2	2	2
x_{30}	0	0	0	2	0
x_{31}	0	0	0	0	0
x_{32}	1	0	1	1	0
x_{33}	1	0	2	0	1
x_{34}	2	0	2	1	0
x_{35}	0	0	0	0	0
x_{36}	2	2	2	2	0
x_{37}	2	2	1	1	0
x_{38}	0	0	0	1	1
x_{39}	0	1	0	1	0
x_{40}	0	0	0	0	0
x_{41}	1	0	0	1	1
x_{42}	0	0	2	2	1
x_{43}	0	0	1	1	0
x_{44}	0	0	0	1	1
x_{45}	2	0	1	0	0
x_{46}	1	0	0	0	0
x_{47}	1	0	0	1	1
x_{48}	2	0	2	2	0
x_{49}	2	0	2	0	0
x_{50}	2	2	1	1	1

Table 2. Cont.

x_{51}	1	0	0	2	0
x_{52}	2	0	0	2	0
x_{53}	1	0	2	0	1
x_{54}	2	0	2	0	0
x_{55}	0	0	0	1	0
x_{56}	2	0	2	2	1
x_{57}	0	0	0	1	1
x_{58}	1	0	0	1	0
x_{59}	0	0	1	0	0
x_{60}	0	0	0	0	0
x_{61}	2	2	2	2	0
x_{62}	2	0	2	0	1
x_{63}	1	0	0	2	0
x_{64}	0	0	0	0	0
x_{65}	2	0	0	2	0
x_{66}	0	0	1	1	0
x_{67}	2	0	2	2	0
x_{68}	1	0	0	2	0
x_{69}	2	0	2	0	0
x_{70}	1	0	0	1	1

$U/R = \{\{x_1, x_6, x_7, x_8, x_{10}, x_{11}, x_{13}, x_{28}, x_{36}, x_{61}\}, \{x_5\}, \{x_{16}\}, \{x_{24}\}, \{x_{30}\}, \{x_{34}\}, \{x_{39}\}, \{x_{42}\}, \{x_{45}\}, \{x_{37}, x_{50}\}, \{x_{12}, x_{14}, x_{15}, x_{29}, x_{48}, x_{56}, x_{67}\}, \{x_3, x_{49}, x_{54}, x_{62}, x_{69}\}, \{x_{52}, x_{65}\}, \{x_{20}, x_{33}, x_{53}\}, \{x_2, x_{32}\}, \{x_{51}, x_{63}, x_{68}\}, \{x_{23}, x_{41}, x_{47}, x_{58}, x_{70}\}, \{x_{17}, x_{18}, x_{26}, x_{46}\}, \{x_{25}, x_{43}, x_{66}\}, \{x_{27}, x_{59}\}, \{x_{22}, x_{38}, x_{44}, x_{55}, x_{57}\}, \{x_4, x_9, x_{19}, x_{21}, x_{31}, x_{35}, x_{40}, x_{60}, x_{64}\}\}$.

Lower and upper approximation sets are important concepts that are defined with the help of the equivalence relation of rough set theory. Approximation sets of this study are defined next.

Lower and upper approximation sets

The universal set U, which includes students, can be divided into three subsets with respect to students' responses to the last question. Let us represent the set of students who were group 2 with the set T (true), those who were group 1 with the set M (medium), and the students who coded as 0 with the set F (false). The following section will describe the lower and upper approximation sets of the sets T, M, and F.

Lower and upper approximation sets of set T: Lower and upper approximation sets of set $T = \{x_2, x_5, x_6, x_{29}\}$ are:

$R_*(T) = \cup_{a \in U} \{ R(a) : R(a) \subset T\} = \{x_5\}$

$R^*(T) = \cup_{a \in U} \{ R(a) : R(a) \cap T \neq \varnothing\} = \{x_1, x_2, x_5, x_6, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{28}, x_{29}, x_{32}, x_{36}, x_{48}, x_{56}, x_{61}, x_{67}\}$ respectively.

It can be seen in the lower and upper approximation sets that even though the elements of the set $\{x_1, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{28}, x_{32}, x_{36}, x_{48}, x_{56}, x_{61}, x_{67}\}$ do not belong to set T, they are members of the upper approximation set. Therefore, these elements are potential members of set T.

Thus, it can be said that students $\{x_7, x_8, x_{11}, x_{14}, x_{28}, x_{32}, x_{36}, x_{61}, x_{48}, x_{67}\}$ who could not solve the last question and those $\{x_1, x_{10}, x_{12}, x_{13}, x_{15}, x_{56}\}$ who could partially solve it, potentially belong to set T and thus it can be said that they could potentially solve the last question. The only student x_5 who were included in the lower approximation set can be said to be definitely successful in solving the last question.

Lower and upper approximation sets of set M: Lower and upper approximation sets of set $M = \{x_1, x_3, x_{10}, x_{12}, x_{13}, x_{15}, x_{18}, x_{20}, x_{26}, x_{27}, x_{33}, x_{38}, x_{41}, x_{42}, x_{44}, x_{47}, x_{50}, x_{56}, x_{57}, x_{62}, x_{70}\}$ are presented as follows:

$R_*(M) = \cup_{a \in U} \{ R(a) : R(a) \subset M\} = \{x_{42}\}$

$R^*(M) = \cup_{a \in U} \{ R(a) : R(a) \cap M \neq \varnothing\} = \{x_1, x_3, x_6, x_7, x_8, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18}, x_{20}, x_{22}, x_{23}, x_{26}, x_{27}, x_{28}, x_{29}, x_{33}, x_{36}, x_{37}, x_{38}, x_{41}, x_{42}, x_{44}, x_{46}, x_{47}, x_{48}, x_{49}, x_{50}, x_{53}, x_{54}, x_{55}, x_{56}, x_{57}, x_{58}, x_{59}, x_{61}, x_{62}, x_{67}, x_{69}, x_{70}\}$

The set of students who were included in the upper approximation of M, but were not a member of M is $\{x_6, x_7, x_8, x_{11}, x_{14}, x_{17}, x_{22}, x_{23}, x_{28}, x_{29}, x_{36}, x_{37}, x_{46}, x_{48}, x_{49}, x_{53}, x_{54}, x_{55}, x_{58}, x_{59}, x_{61}, x_{67}, x_{69}\}$. These students are potentially a member of set M.

Lower and upper approximation sets of set F: The students who could not answer the last question are a member of set F={ x_4, x_7, x_8, x_9, x_{11}, x_{14}, x_{16}, x_{17}, x_{19}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{28}, x_{30}, x_{31}, x_{32}, x_{34}, x_{35}, x_{36}, x_{37}, x_{39}, x_{40}, x_{43}, x_{45}, x_{46}, x_{48}, x_{49}, x_{51}, x_{52}, x_{53}, x_{54}, x_{55}, x_{58}, x_{59}, x_{60}, x_{61}, x_{63}, x_{64}, x_{65}, x_{66}, x_{67}, x_{68}, x_{69}}.

Lower and upper approximation sets of F are given as:

$R_*(F)= \cup_{a\in U}$ { R(a) : R(a) $\subset$ F}={ x_4, x_9, x_{16}, x_{19}, x_{21}, x_{24}, x_{25}, x_{30}, x_{31}, x_{34}, x_{35}, x_{39}, x_{40}, x_{43}, x_{45}, x_{51}, x_{52}, x_{60}, x_{63}, x_{64}, x_{65}, x_{66}, x_{68}}

$R^*(F)= \cup_{a\in U}$ { R(a) : R(a)$\cap$F$\neq\emptyset$}= U-{x_5, x_{42}} respectively.

Since the boundary sets B_R(T), B_R(M), and B_R(F) of the three sets of which lower and upper approximation sets as defined above are different from empty set; T, M, and F are rough sets. These sets can also be characterized by a constant within the closed interval [0.1]. As defined in section two, this constant that will determine the clarity of the approximation is defined as $\alpha_R(A)=\frac{|B_*(A)|}{|B^*(A)|}$. The constants for the three sets are calculated as:

$$\alpha_R(T)=\frac{1}{20}=0{,}050,$$

$$\alpha_R(M)=\frac{1}{44}\cong 0{,}022,$$

$$\alpha_R(F)=\frac{23}{68}\cong 0{,}338.$$

This indicates that answers for the first four questions explain the answers for the last question via a weak relationship.

Interdependence of questions: An important topic in data analysis is to detect dependences among attributes. The primary focus of this study was on to what degree the first four questions determine the last question. Let the first four questions represent attributes set C and the last question represent set D, then to what degree the set C explains the set D, or the degree of dependence of the set D to set C can be found as:

$$k=\gamma(C,D)=\frac{|POS_C(D)|}{|U|}.$$

In this study, the value for this dependence is calculated as:

$$\gamma(C,D)=\frac{|POS_C(D)|}{|U|}=\frac{25}{70}\cong 0{,}357$$

Table 3. Approximation qualities.

Attributes	γ
Q1	0.000
Q2	0.042
Q3	0.000
Q4	0.000
Q1, Q2	0.042
Q1, Q3	0.057
Q1, Q4	0.042
Q2, Q3	0.042
Q2, Q4	0.042
Q3, Q4	0.114
Q1, Q2, Q3	0.010
Q1, Q2, Q4	0.010
Q1, Q3, Q4	0.342
Q2, Q3, Q4	0.157

This value indicates the degree to which first four questions together explain the last question. In addition, this value can be calculated for individual questions or any two or three-question combination. Table 3 displays these values.

When Table 3 was examined, considering C = {S1, S2, S3, S4}, one can obtain RED(C) = {{S1, S2, S3, S4}} and Core(C) = $\cap$Red(C)= {S1, S2, S3, S4}. One can say that C attributes set is not a reducible set.

Conclusion

Data mining, referred to as knowledge discovery in databases (KDD), is the area of detecting useful information from large data sets. Data mining has application in many fields such as retail sales, bioinformatics and counter-terrorism. Recently, there has been growing interest in using data mining in educational research, which is referred to as educational data mining. The use of alternative means such as Fuzzy sets or rough sets is becoming increasingly common in analyzing vague data. These concepts are also seen in data analysis of educational research in recent years (Yorek and Narli, 2009).

As exemplified in the present study, approach sets, which have a great number of application areas, seems to interpret today's vague information. Loslever and Lepoutre (2004) suggest that humans have intuitively multivariate and complex behavior. According to this context, it may be argued that it will not be easy to evaluate humans within 'exact' and 'definite' categories. The most common data analyses procedures used in educational research are descriptive statistics, ANOVA/MANOVA, correlation, regression, t-test, and psychometric statistics (Hsu, 2005). Rough set data analysis can be either an alternative to these statistics, or at least a supplemental method to these statistics.

The students were placed into three groups according to analysis of the last question. The results of this study indicate that it is possible mathematically to determine, in which other groups the students could be included. For each of these typologies, the degree of vagueness, represent with $\alpha_R(X) = \frac{|\,\mathrm{Rlow(X)}\ \,|}{|\,\mathrm{Rup(X)}\ \,|}$ were calculated. It was determined that there were 25 students who definitely belong to one of the four typologies. It may be said that there can be no clear boundary drawn for the remaining 45 students. The fact that the values of $\alpha_R(T) \cong 0{,}050$, $\alpha_R(M) \cong 0{,}022$, and $\alpha_R(F) \cong 0{,}338$ are close to zero indicate and support that the sets T, M, and F are far from being exact. Similar results were obtained in Narli (2010), where he analyzed fennema-sherman attitude scale using rough set theory. Studies by Yörek and Narli (2009) and Narli et al. (2010) on using rough sets in educational research also revealed more exploratory results. These types of results seem to be impossible to obtain using other statistical means. In this context, rough set data analysis provides great advantage.

In addition, it is determined that first four questions explain the last question at a degree of 0.357. This may be interpreted as there is no significant relation between the first four questions and the last question. Moreover, when data reduction is carried out, this value gets smaller. In other words, no question can be removed from the attribute set C consisting of four questions. When a question is emoved from the set C, the new attribute set will explain the last question even at a weaker level.

As a result, after it has been introduced by Pawlak, rough sets have been used in many fields such as mathematical morphology, genetic algorithm, artificial intelligence, Petri network, decision tables, probability, drug industry, engineering, control systems and social sciences. It is thought that rough sets can also be used to better understand via analyzing behavior, attitude, achievement, beliefs, etc. data obtained from humans.

REFERENCES

Aktas H, Cagman N (2005). Bulanik ve Yaklaşimli Kümeler, Çankaya Üniversitesi Fen-Edebiyat Fakültesi*,* J. Arts Sci. Sayı: 3: 13-*25*

Amarshi S, Conati C (2009). Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments, J. Educational Data Mining, Article 2, 1(1): 1-54

Aydogan EK, Gencer C (2007). Veri Madenciliği Problemlerinde Kaba Küme Yaklaşimi Kullanilarak Sınıflandırma Amaçli Yapilmiş Olan Çalişmalar, Kara Harp Okulu Savunma Bilimleri Dergisi, 6(2): 17-32

Baker RSJD (in press) Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

Baker RSJD, Corbett AT, Roll I, Koedinger KR (2008). Developing a Generalizable Detector of hen Students Game the System. User Modeling User-Adapted Interaction, 18(3): 287-314.

Beck J (2005). Engagement Tracing: Using Response Times to Model Student Disengagement*.* In Proceedings of the International Conference on Artificial Intelligence in Education.

Bonikowaski Z (1995). Algebraic structures of rough sets, Rough sets, Fuzzy sets and Knowledge Discovery, Springer-Verlag, Berlin.

Frege G (1904). Grundgesetze der Arithmentik [Basic principles of arithmetic] In Geach, P. T. & Black, M. (Eds.), Selections from the Philosophical Writings of Gotlob Frege, Oxford: Blackweil.

Hsu CH (2005). Joint modelling of recurrence and progression of adenomas: a latent variable approach. Statistical Modelling, 5: 210-215.

Jiashang J, Congxin W, Degang C (2005). The product structure of fuzzy rough sets on a group and the rough T- fuzzy group, Information Sci., 175: (1-2) 97-107.

Kent RE (1994). Rough concept analysis, rough sets, fuzzy sets knowledge discovery. In Ziarko W.P. (ed.) & Alta B., Proceeding of the international workshop on rough sets, knowledge, discovery (pp. 248-255), Canada: Springer-Verlag.

Kumar R (1993). Fuzzy Algebra *I,* University of Delhi, Publ. Division.

Kuroki N (1997). Rough ideals in semigroups, Inform. Sci*.,* 100: 139-163.

Le´sniewski St (1915).O Podstawach Matematyki. Przeglad Filozoficzny, 30-34. (1927-1931).

Lin TY, Cercone N (1997). Rough Sets and Data Mining. The Netherlands: Kluwer Academic Publishers.

Loslever P, Lepoutre FX (2004). Analysis of objective and subjective data using fuzzy coding and multiple correspondence analysis: principle and example in a sitting posture study Theoretical Issues Ergonomics Sci., 5(5): 425-443

Munakata Y (1998). Fundamentals of the new artificial intelligence: beyond traditional paradigms. New York, USA: Springer-Verlag.

Narli S (2010). An Alternative Evaluation Method For Likert Type Attitude Scales: Rough Set Data Analysis, Scientific Res. Essays, 5(6): 519-528.

Narli S, Ozcelik A (2008). On Generalizing rough set theory via using a fitler, Int. J. Comput. Mathematical Sci., 2-3: 149-152.

Narli S, Yorek N, Sahin M, Uşak M (2010). Can We Make Definite Categorization of Student Attitudes? A Rough Set Approach to Investigate Students' Implicit Attitudinal Typologies Toward Living Things, J. Sci. Edu. Technol. 19: 456-469.

Nings S, Ziarko WP, Hamilton J, Cercone N (1995). Using Rough sets as tools for knowledge discovery. In Fayyad, U.M. & R. Uthurusamy (eds.), KDD'95 Proceedings first international conference on knowledge discovery data mining *(*pp. 263-268), Montreal, Canada: AAAI.

Pawlak Z (1982). Rough sets, Int. J. Comp. Info. Sci. 11: 341-356.

Pawlak Z (1991). Rough sets-theoretical aspect of reasoning about data, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pawlak Z (1995). Vagueness and uncertainty: A rough set perspective, Comput. Intelligence, 11: 277-232

Pawlak Z (1997). Sets, Fuzzy sets and Rough Sets Available at "http://grammars.grlmc.com/GRLMC/reports/rep29.doc" last accessed 23/01/2009.

Pawlak Z, Grzymala-Busse J, Slowinski R, Ziarko W (1995). Rough sets. Communications ACM, 38(11): 89-95.

Pawlak Z, Slowinski R (1994). Rough Set Approach to Multi-attribute Decision Analysis. Europan J. Operational Res. v 72: s. 443-459.

Polkowski L (2002). Rough Sets – Mathematical Foundations, Advances in Soft Computing, Physica-Verlag, Springer-Verlag Company, 1-534.

Polkowski L, Skowron A (1998a). Rough sets in knowledge discovery (vols. 1-2). In Kacprzyk, J. (series ed.) Studies in Fuzziness and Soft Computing Series. Heidelberg: Physica-Verlag/Springer-Verlag.

Polkowski L, Skowron A (1998b). Rough Sets and Current Trends in Computing. In Goebel, R., Siekmann, J. & Wahlster, W. (series eds.) Lecture Notes in Artificial Intelligence Series. Heidelberg/Berlin: Springer-Verlag.

Pomykala J, Pomykala JA (1988). The stone algebra of rough sets, Bull. Polish Acad. Sci. Math., 36: 495-508.

Romero C, Ventura S (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Syst. Applications, 33: 125-146.

Sison R, Numao M, Shimura M (2000). Multistrategy Discovery and Detection of Novice Programmer Errors. Machine Learning, 38: 157-180.

Slowinski R (1992). (Ed) Intelligent Decision Support-Handbook of

Advances and Applications of Rough Set Theory. Kluwer Academik Publisher.

Yorek N, Narli S (2009). Modeling of Cognitive Structure of Uncertain Scientific Concepts Using Fuzzy-Rough Sets and Intuitionistic Fuzzy Sets: Example of The Life Concept, Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., 17(5): 747-769.

Zadeh L (1965). Fuzzy sets, Information and Control, 8: 338-353.

Zaiane O (2002). Building a Recommender Agent for e-Learning Systems. In Proceedings of the International Conference on Computers in Education.

Zhong N, Showron A (2000). Rough sets in KDD: Tutorial notes. Bull. Int. Rough Set Society, 4(1-2).