**Review**

Enhanced data mining analysis in higher educational system using rough set theory

P. Ramasubramanian¹*, K. Iyakutti² and P. Thangavelu³

¹Department of CSE, Dr.G.U. Pope College of Engineering, Sawyerpuram, India.
²Dept. of Microprocessor and Computer, Madurai Kamaraj University, Madurai, India.
³Department of Mathematics, Aditanar College of Arts and Science, Tiruchendur, India.

Accepted 6 August, 2009

One of the biggest challenges that higher education faces today is predicting the behavior of students. Institutions would like to know, something about the performances of the students group wise. Behrouz–Minaei-Bidgoli (2004) investigated a web based educational system using data mining techniques. He proposed a problem to investigate the performances of the students when the large data base of Students information system (SIS) is given. Generally students' problems will be classified into different patterns based on the level of students like normal, average and below average. In this paper we attempt to analyze SIS database using rough set theory to predict the future of students.

**Key words:** Data mining, knowledge discovery, rough set, lower approximation and upper approximation, characteristic sets, missing attribute values.

**INTRODUCTION**

Rough set theory is a fundamental area that has made important contributions to the Knowledge discovery in data base (KDD). Data mining is the process of autonomously extracting useful information or knowledge from large data stores or sets. It involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining consists of more than collecting and managing data; it also includes analysis and prediction. These tools can include statistical models, mathematical algorithms and machine learning methods such as neural networks or decision trees etc by using rough set algorithms.

Data mining (DM) can be performed on a variety of data stores, including the world wide web (WWW), relational databases, transactional databases, internal legacy systems, personal document format documents (PDF) and data warehouses.

Data mining has been applied into many application domains such as Biomedical and DNA analysis, Retail industry and marketing, telecommunications, web mining, computer auditing, banking and insurance, fraud detection, financial industry, medicine and education. Data mining is popularly known as knowledge discovery databases.

As the educational systems are capable of collecting large amount of students profile data, data mining and rough set techniques can be applied to find interesting relationships between attributes of students.

Figure 1 shows the concept of data mining, which involves three steps:

1. Capturing and storing the data.
2. Converting the raw data into information.
3. Converting the information into knowledge.

Data in this context comprises all the raw material an institution collects via normal operation. Capturing and storing the data is the first phase that is the process of applying mathematical and statistical formulas to “mine” the data warehouse. Mining the collected raw data from the entire institution may provide new information as to how students, parent’s and the institutions own processes really perform. Converting the raw data into information is the second step of data mining.

Our survey on the current works in data mining field shows that one of the application domains that can take advantage of data mining benefits in education. Our focus in this paper is data mining in higher educational...
system. Student Information System (SIS) data is involved with three kinds of large data sets:

1. Educational resources such as student databases, fees collection and individualized problems designed for use on assignments and examinations.
2. Information about users who create, modify, assess, or use these resources.
3. Activity log databases which log actions taken by students in behavioral characteristics and exam results.

The objective of this paper is to predict the future of students who show the weak performance for the academic programmed, by accessing a group of students, likelihood of success, those who are graduating and not graduating and those who need additional training based on the features, which are extracted from their (and others) data. We design, implement and evaluate a series of pattern classifiers with various parameters in order to compare their performances in a real data set from the SIS system. To complete this task we require the knowledge of Rough set theory.

**Rough set logic**

The Rough set theory is a recent mathematical theory employed as a data mining tool with many favorable advantages. Since this theory has been applied to various domains, the majority of these applications are used to solve the classification problems, which exclude the temporal factor in data sets. The rough set analysis is presented as a technique to direct the knowledge discovery process from data.

Pawlak (1982) introduced the Rough Set Theory which was initially developed for a finite universe of discourse in which the knowledge base is a partition, which is obtained by any equivalence relation defined on the universe of discourse.

Let \( U \) be any finite universe of discourse. Let \( R \) be any equivalence relation defined on \( U \). Clearly the equivalence relation partitions \( U \). The pair \((U, R)\) is called the approximation space. This collection of equivalence classes of \( R \) is called as knowledge base. Then for any subset \( A \) of \( U \), the lower and upper approximations are defined respectively as follows:

\[
RA = \bigcup \{W_i : W_i \subseteq A\} \\
\overline{RA} = \bigcup \{W_i : W_i \cap A \neq \emptyset\}
\]

The ordered pair \((RA, \overline{RA})\) is called rough set. In general, \( RA \subseteq A \subseteq \overline{RA} \). If \( RA = \overline{RA} \) then \( A \) is called exact. The lower approximation of \( A \) is called the positive region of \( A \) and is denoted by \( \text{POS}(A) \) and the complement of upper approximation of \( A \) is called the negative region of \( A \) and is denoted by \( \text{NEG}(A) \). Its boundary is defined as \( \text{BND}(A) = \overline{RA} - RA \). Hence, it is trivial that if \( \text{BND}(A) = \emptyset \), then \( A \) is exact.

In the theory of rough sets, the decision table of any information system is given by \( T = (U, A, C, D) \), where \( U \) is the universe of discourse, \( A \) is a set of primitive features, \( C \) and \( D \) are the subsets of \( A \), called condition and decision features respectively.

For any subset \( P \) of \( A \), a binary relation \( \text{IND}(P) \), called the indiscernibility relation is defined as \( \text{IND}(P) = \{(x, y) \in U \times U : a(x) = a(y) \text{ for all } a \in P\} \).

The Rough set in \( A \) is the family of all subsets of \( U \) having the same lower and upper approximations. Generally Rough set theory is explored using the decision tables that are information tables.

This table describes cases (also called objects or examples) using attribute values and a decision. Attribute values are independent denoted by ‘a’, while decision values are dependent denoted by ‘d’. Rows of the table are labeled as cases denoted by ‘U’. The set of all attributes is denoted by ‘A’. The following notations are used in this paper:

- \( U \): Set of all cases.
- \( A \): Set of all Attributes.
- \( V \): Set of all attribute values.
- \( D \): Decision values.
- \( B \): Non-empty subset of \( A \).

Here \( P \) is a function defined from \( U \times A \) to \( V \).

**Rough set analysis on SIS data**

Consider a completely specified student data as shown Table 1. In rough set theory it is called an information table. The table consists of three attributes of a student namely academic (A), non-academic (NA) and human behavior relationships (HBR). The value of the academic attribute is calculated on the basis of student performances in theory, practical, attendance, participating seminars, paper presentations, interactions, reading books and participating department activities. The value of the non-academic attribute is calculated on the basis of student performances in Sports, NSS, NCC, YRC and Social activities. The value of the Human behavior rela-
Table 1. SIS data set.

<table>
<thead>
<tr>
<th>S.No</th>
<th>Academic</th>
<th>Non-academic</th>
<th>Human behavior relation</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.4</td>
<td>0.9</td>
<td>Normal</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>Average</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
<td>0.3</td>
<td>0.7</td>
<td>Below Average</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.4</td>
<td>0.9</td>
<td>Below Average</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>Below Average</td>
</tr>
<tr>
<td>7</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>0.3</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>Below Average</td>
</tr>
<tr>
<td>11</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Table 2. Characteristic Relation.

<table>
<thead>
<tr>
<th>a</th>
<th>K_x(a)</th>
<th>K_y(a)</th>
<th>K_z(a)</th>
<th>K_{empty}(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{1,4,5,7,8,9}</td>
<td>{1,3,9,10}</td>
<td>{1,7,8,9,11}</td>
<td>{1,9}</td>
</tr>
<tr>
<td>2</td>
<td>{2,4,5,11}</td>
<td>{1,2,3,4,5,9,10}</td>
<td>{2,5,11}</td>
<td>{2,5}</td>
</tr>
<tr>
<td>3</td>
<td>{3,4,5}</td>
<td>{1,3,9,10}</td>
<td>{3,11}</td>
<td>{3}</td>
</tr>
<tr>
<td>4</td>
<td>{4,5}</td>
<td>{1,2,3,4,5,9,10}</td>
<td>{4,11}</td>
<td>{4}</td>
</tr>
<tr>
<td>5</td>
<td>{4,5}</td>
<td>{1,2,3,4,5,9,10}</td>
<td>{2,5,11}</td>
<td>{5}</td>
</tr>
<tr>
<td>6</td>
<td>{4,5,6}</td>
<td>{1,3,6,9,10}</td>
<td>{6,10,11}</td>
<td>{6}</td>
</tr>
<tr>
<td>7</td>
<td>{1,4,5,7,8,9}</td>
<td>{1,3,7,8,9,10}</td>
<td>{1,7,8,9,11}</td>
<td>{1,7,8,9}</td>
</tr>
<tr>
<td>8</td>
<td>{1,4,5,7,8,9}</td>
<td>{1,3,7,8,9,10}</td>
<td>{1,7,8,9,11}</td>
<td>{1,7,8,9}</td>
</tr>
<tr>
<td>9</td>
<td>{1,4,5,7,8,9}</td>
<td>{1,3,9,10}</td>
<td>{1,7,8,9,11}</td>
<td>{1,7,8,9}</td>
</tr>
<tr>
<td>10</td>
<td>{4,5,10}</td>
<td>{1,3,9,10}</td>
<td>{6,10,11}</td>
<td>{10}</td>
</tr>
<tr>
<td>11</td>
<td>{2,4,5,11}</td>
<td>{1,3,9,10,11}</td>
<td>{11}</td>
<td>{11}</td>
</tr>
</tbody>
</table>

Characterizations is calculated on the basis of students’ relationships with their teachers, fellow students, public, family members and their performances in hobbies and entertainment. Now, we wish to analyze the following hypothetical sample information table using rough set theory.

Learning from examples is the most important tool employed in data mining. We compute lower and upper approximations of all decisions in SIS data. By using boundary of the decisions (BND), we extract knowledge of the SIS database. SIS data for data mining are frequently affected by missing attributes values of Academic, Non-academic and Human behaviour relationships. Consider the following incomplete SIS decision table as shown in Table 3, we will assume that all unknown attributes values are denoted by '*'. Such values will be denoted by '?'.

Computing characteristic relation

The characteristic relation R(B) is known if we know characteristic sets K_B(a) for all a ∈ U. Here;

U = {1,2,3,4,5,6,7,8,9,10,11}
B = {x, y, z}.

For each a ∈ U,
K_x(a) = { b ∈ U : p(a, x) = p(b, x) } ∪
{ b ∈ U : p(b, x) = '*' }.
K_y(a) = K_x(a) ∩ K_z(a) ∩ K_{empty}(a). Using this formula we form the Table 2.

Let (U, A) be an information system and let B ⊆ A and X ⊆ U. We can approximate X using the information contained in B by constructing the B-lower and B-upper approximations of X denoted by 

\[ \overline{BX} \text{ and } B^X \text{ respectively, where:} \]

\[ \overline{BX} = \bigcup \{ K_B(a) : K_B(a) \subseteq X \} \]
\[ B^X = \bigcup \{ K_B(a) : K_B(a) \cap X \neq \emptyset \} \]

The set BND(X) = \[ B^X - \overline{BX} \] is called the B-boundary re-
Table 3. SIS data set with missing attribute values.

<table>
<thead>
<tr>
<th>S. No</th>
<th>Academic</th>
<th>Non-academic</th>
<th>Human behavior relation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>*</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.4</td>
<td>0.9</td>
<td>Average</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>0.4</td>
<td>0.4</td>
<td>Average</td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>0.4</td>
<td>0.7</td>
<td>Average</td>
</tr>
<tr>
<td>5</td>
<td>*</td>
<td>0.4</td>
<td>0.9</td>
<td>Average</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
<td>0.5</td>
<td>?</td>
<td>Below Average</td>
</tr>
<tr>
<td>7</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>9</td>
<td>0.8</td>
<td>*</td>
<td>0.8</td>
<td>Normal</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>*</td>
<td>?</td>
<td>Below Average</td>
</tr>
<tr>
<td>11</td>
<td>0.7</td>
<td>0.7</td>
<td>*</td>
<td>Average</td>
</tr>
</tbody>
</table>

region of X, and this consists of those objects that we cannot decisively classify into X on the basis of knowledge in B.

Example 1

Consider the two sets (decisions) $X = \{2, 3, 4, 5, 11\}$ and $Y = \{1, 6, 7, 8, 9, 10\}$. From the above table, the B-lower and B-upper approximations of the two concepts (decisions) are:

\[
\underline{BX} = \{2, 3, 4, 5, 11\}
\]

\[
\bar{B}X = \{2, 3, 4, 5, 11\}
\]

\[
\underline{BY} = \{1, 6, 7, 8, 9, 10\}
\]

\[
\bar{B}Y = \{1, 6, 7, 8, 9, 10\}
\]

Boundary region of X is $\{2, 3, 4, 5, 11\} – \{2, 3, 4, 5, 11\} = \emptyset$

Boundary region of Y is $\{1, 6, 7, 8, 9, 10\} – \{1, 6, 7, 8, 9, 10\} = \emptyset$

Thus we conclude that, the set X and set Y have the same characteristics.

Example 2

Consider the two sets (decisions) $X = \{1, 4, 5, 6, 8, 9\}$ and $Y = \{2, 3, 7, 10, 11\}$. From the above table, the B-lower and B-upper approximations of the two concepts (decisions) are:

\[
\underline{BX} = \{1, 4, 5, 6, 9\}
\]

\[
\bar{B}X = \{1, 2, 4, 5, 6, 7, 8, 9\}
\]

\[
\underline{BY} = \{3\}
\]

\[
\bar{B}Y = \{1, 2, 3, 5, 7, 8, 9, 10, 11\}
\]

Boundary region of X is $\{1, 2, 4, 5, 6, 7, 8, 9\} – \{1, 4, 5, 6, 9\} = \{2, 7, 8\}$

Boundary region of Y is $\{1, 2, 4, 5, 6, 7, 8, 9\} – \{1, 2, 4, 5, 6, 7, 8, 9\} = \emptyset$

Hence, we conclude that, any characteristic of X is also the characteristics of Y.

Example 3

Consider the two sets (decisions) $X = \{2, 4, 5, 6, 11\}$ and $Y = \{1, 9, 11\}$. From the above table, the B-lower and B-upper approximations of the two concepts (decisions) are:

\[
\underline{BX} = \{2, 4, 5, 6, 11\}
\]

\[
\bar{B}X = \{2, 4, 5, 6, 11\}
\]

\[
\underline{BY} = \{1, 9, 11\}
\]

\[
\bar{B}Y = \{1, 9, 11\}
\]

Boundary region of X is $\{2, 4, 5, 6, 11\} – \{2, 4, 5, 6, 11\} = \emptyset$

Boundary region of Y is $\{1, 9, 11\} – \{1, 9, 11\} = \emptyset$

From this, we conclude that, the characteristics of X and Y differ significantly.

Algorithm

**Input:** B, the subset of attributes

**Input:** X, the subset of U, the decision space

**Output:** Boundary of X

1. [Initialize]
   
   $G = U$
   
   $a \in U$
   
   Lower(X) = $\emptyset$
   
   Upper(X) = $\emptyset$

2. Repeat step 3 to 8 while $G \neq \emptyset$

3. If $K_B(a) \subseteq X$ then Lower(X) = Lower(X) $\cup K_B(a)$. 

4. If $K_B(a) \cap X \neq \emptyset$ then $\text{Upper}(X) = \text{Upper}(X) \cup K_B(a)$.
5. $G = G - \{a\}$
   [End of the loop]
6. $\text{BND}(X) = \text{Upper}(X) - \text{Lower}(X)$.

**Data mining in higher educational system**

Our main contribution in this paper is addressing the capabilities and strengths of data mining technology in the context of higher educational system. Higher educational system can enhance their educational processes according to our proposed analysis model. We will present and describe an analysis model for this system.

We analyze Student information system (SIS) database with the attributes such as academic, non-academic and human behavior relationships. These attributes can be further divided into corresponding sub attributes whose values will be analyzed by rough set logic. We will use these techniques to mine our SIS.

Data are generally organized as matrix. The matrix contains a set of instances. Each matrix column represents the values associated to an attribute. Attributes can be for the student database for example: register number, name, branch of study, year, percentage of marks, academic, non-academic and human behavior relationships etc. An attribute can be either numerical or nominal. A nominal attribute contains values within a limited set of choices. A special attribute called decision attribute contains information about the decision to take upon the individual’s consideration.

The database consists of student’s data collected from various departments. The advantage of this data is that it includes sufficient number of records of different types of students.

A database consists of a set of tables containing either values of entity attributes or values of attributes from entity relationships.

We use SQL data base. This is the most commonly used query language for relational database in SQL, which allows retrieval and manipulation of the data stored in the tables as well as the calculation of aggregate functions such as average, sum, min, max and count.

Data mining can be benefited from SQL for data selection, transformation and consolidation. SQL provides predicting, correcting, comparing and detecting deviations etc. The various attributes are given in tabular column:

**Conclusion**

Thus we have analyzed the SIS database, using Rough set theory and discussed an algorithm to implement the technique. Using the technique discussed in this paper, one can access the performances of a particular student by accessing a group of students even if we are not aware of certain attribute values of the students.

**REFERENCES**