International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:11, 2007

An Intelligent Approach of Rough Set in Knowledge Discovery Databases

Hrudaya Ku. Tripathy, B. K. Tripathy, and Pradip K. Das

***Abstract*—**Knowledge Discovery in Databases (KDD) has evolved into an important and active area of research because of theoretical challenges and practical applications associated with the problem of discovering (or extracting) interesting and previously unknown knowledge from very large real-world databases. Rough Set Theory (RST) is a mathematical formalism for representing uncertainty that can be considered an extension of the classical set theory. It has been used in many different research areas, including those related to inductive machine learning and reduction of knowledge in knowledge-based systems. One important concept related to RST is that of a rough relation. In this paper we presented the current status of research on applying rough set theory to KDD, which will be helpful for handle the characteristics of real-world databases. The main aim is to show how rough set and rough set analysis can be effectively used to extract knowledge from large databases.

***Keywords*—**Data mining, Data tables, Knowledge discovery in database (KDD), Rough sets.

I. Introduction

THE current information age is characterized by an extraordinary expansion of data that are being generated and stored about all kinds of human endeavors. An increasing proportion of these data is recorded in the form of computer databases, in order that the computer technology may easily access it. The availability of very large volumes of such data has created a problem of how to extract from them useful, task-oriented knowledge. The Knowledge Discovery from Databases (KDD) is usually a multi-phase process involving numerous steps, like data preparation, preprocessing, search for hypothesis generation, pattern formation, knowledge evaluation, representation, refinement and management. Furthermore, the process may be repeated at different stages when a database is updated. The multi-phase process is an important methodology for the knowledge discovery from real-life data. Although the process-centric view has recently been widely accepted by researchers in the KDD community, few KDD systems provide capabilities that a more complete process should possess [1].

Manuscript received October 9, 2007.

F. Hrudaya Ku. Tripathy is with the Institute of Advanced Computer & Research, Prajukti Bihar, Rayagada, Orissa-765002, India (Phone: 91-6856-236350; fax: 91-6856-235546; e-mail: hrudayakumar@hotmail.com).

S. B.K.Tripathy is with Berhampur University, Bhanja Bihar, Berhampur, Orissa-760007, India (e-mail: tripathybk@rediffmail.com).

T. Pradip K. Das is with the Computer Science & Engineering Department, Indian Institute of Technology Guwahati, North Guwahati, Assam- 781 039, India (e-mail: pkdas@iitg.ernet.in).

A. Knowledge Discovery in Databases (KDD)

Data mining is seen as the key element in the so-called knowledge discovery in databases. KDD is not a new technique. It is an interdisciplinary area in which machine learning, statistics, database technology, expert systems and data visualization come together. The KDD process consists of six stages: data selection, cleaning of data, enrichment of data, coding, data mining and reporting. In data selection the data most relevant to the problem at hand is selected, leaving out redundant data. In the cleaning process, the obvious flaws in the data are corrected. In the coding process, data is coded such that it can be used in data mining algorithms. In the data mining process, the actual inference takes place. In the reporting process the results are of course reported, usually in a visually attractive way [2].

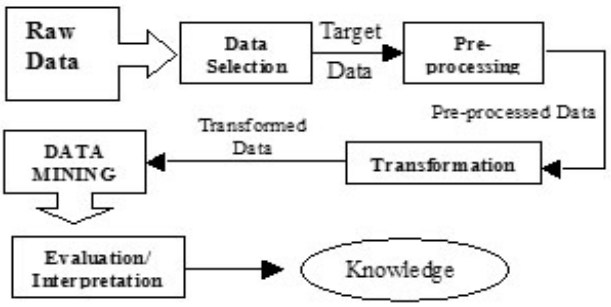

Fig. 1 The KDD Process

B. Rough Sets in Knowledge Discovery

The data contained in the databases of the real world were not collected with learning in mind. The data is uncertain in nature, in deterministic and contains noise due to errors in measurements, data transfer and human errors performed during the data collection process. All of the characteristics presented above contribute to the uncertainty of the data. Uncertainty in the knowledge is inevitable in real life situations, and several different frameworks for reasoning with uncertain knowledge have been introduced.

Rough set theory constitutes a sound basis for KDD. It offers useful tools for discovering patterns hidden in data in many aspects (Lin and Cercone, 1997; Pal and Skowron, 1999; Pawlak, 1982; 1991; Skowron and Rauszer, 1992). It can be used in different phases of the knowledge discovery

process, like attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction (templates, assrociation rules) (Komorowski et al., 1999). Furthermore, recent extensions of rough set theory (rough mereology) have brought new methods of decomposition of large data sets, data mining in distributed and multi-agent based environments and granular computing (Polkowski and Skowron, 1996; Polkowski and Skowron, 1999; Yao and Zhong, 1999; Zhong et al., 1999) [3].

It includes mechanisms for defining partial memberships of sets, but does not introduce additional measures of probabilities or degrees of membership. The basic assumption is that there is some information (data) associated with each object in the universe of discourse. Based on this information, it is possible to tell some of the objects apart, while others are impossible to distinguish. The latter objects are *indiscernible* from each other, and form a set. Each set of indiscernible objects is a *knowledge granule (*atom), and they form the building blocks for knowledge about the universe. The rough set community has been a very active research community since its inception in the eighties, and a large number of rough set methods for knowledge discovery and data mining have been developed. The entire knowledge discovery process has been subject to research, and a wide range of contributions has been made. Data mining technology provides a new thought for organizing and managing tremendous data. Rough set theory is one of the important methods for knowledge discovery. This method can analyze intactly data, obtain uncertain knowledge and offer an effective tool by reasoning. The main feature of rough set data analysis is non-invasive, and the ability to handle qualitative data. This fits into most real life application nicely [4].

II. Process behind KDD

Knowledge discovery in databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [8], [14]. Data is a set of facts F, and a pattern is an expression E in a language L describing the facts in a subset F_E of F. E is called a pattern if it is simpler than the enumeration of all facts in F_E. A measure of certainty, measuring the validity of discovered patterns, is a function C mapping expressions in L to a partially or totally ordered measure space M_C. An expression E in L about a subset $F_E \subset F$ can be assigned a certainty measure c = C(E,F). Novelty of patterns can be measured by a function N(E,F) with respect to changes in data or knowledge. Patterns should potentially lead to some useful actions, as measured by some utility function u = U(E,F) mapping expressions in L to a partially or totally ordered measure space M_U. The goal of KDD is to make patterns understandable to humans. This is measured by a function s = S(E,F) mapping expressions E in L to a partially or totally ordered measure space M_S.

According to the widely accepted description of [5], the (iterative) process of knowledge discovery in databases (KDD) consists of the following steps:

- Developing an understanding of the application domain, the relevant prior knowledge, and the goal(s) of the end user.
- Creating or selecting a target data set.
- Data cleaning and preprocessing: this step includes, among other tasks, removing noise or accounting for noise, and imputation of missing values.
- Data reduction: Finding useful features to represent the data depending on the goal of the task. This may include dimensionality reduction or transformation.
- Matching the goals to a particular data mining method such as classification, regression, clustering etc. Model and hypothesis selection, choosing the data mining algorithm(s) and methods to be used for searching for data patterns.
- Data mining.
- Interpreting mined patterns.
- Acting on discovered knowledge.

III. Rough Set Analysis in KDD

The most common representation of initial knowledge in rough set theory is in a tabular form, similar to a relational table. The column in the table represents attributes, and each row represents an object. There are two different kinds of knowledge representations, namely *information systems* and *decision systems*.

A. Information Systems

An information system is the most basic kind of knowledge. It consists of a set of tuples, where each tuple is a collection of attribute values. Rough Set Analysis in KDD is based on the viewpoint that objects are known up to their description by attribute vectors: An *information system I* consists of a set U of objects, and a set Ω of attributes; the latter are functions $a:U\rightarrow V_a$ which assign to each object x a value $a(x)$ in the set V_a of values which x can take under a .

If $\theta \neq Q \subseteq \Omega$ we denote the feature vector of x with respect to the attributes in Q by $\vec{x}^{Q}$. This operationalisation by Object $\rightarrow$ Attribute data tables assumes the "nominal scale restriction" which postulates that each object has exactly one value of each attribute at a given time, and that the observation of this value is without error. Data reduction is a major feature of rough set analysis. Each $Q \subseteq \Omega$ determines an equivalence relation θ_Q on U by setting,

$$x \equiv \theta_Q\, y \Leftrightarrow (\forall_a \in Q)\; a(x) = a(y)$$

The finest equivalence obtained in this way is θ_Ω. If $Q \subseteq \Omega$ and $\theta_Q = \theta_\Omega$, then the attributes in Q are sufficient to describe the classification induced by Ω, and thus, one can project Ω to Q. Note that only information by the data is used for attribute reduction. A set Q of attributes, which is minimal with respect to above equation, is called *reduct* of I [6].

B. Decision Systems

A decision system is similar to an information system, but a distinction is made between condition and decision attributes.

In an information system, the information is not interpreted. However, an expert may classify the different objects according to some semantic criteria, thus assigning an expert classification attribute to each object. Adding a decision attribute d to an information system creates a decision system, where the attributes A form the condition attributes. Using a single decision attribute can be done without any loss of generality, as it is possible to represent any k-size attribute set D by a single decision attribute d. Any combination of the values for the decision attributes in D may be represented (coded) by a distinct value for d. Hence, it is sound to assume that D ={*d*}. A *decision system* (DS) $\mathcal{A}$ = (*U*,*A*,{*d*}) is an information system for which the attributes are separated into disjoint sets of condition attributes A and a decision attributes *d* (A ∪ {*d*}=ϕ). Now, it should be apparent that from any given DS $\mathcal{A}$ = (*U*,*A*,{*d*}), it is possible to construct an information system by simply removing the decision attribute d from the system, giving us an information system $\mathcal{A}'$= (*U*, *A*). In the same manner as a decision system is a specialized kind of information system, decision rules are a special kind of pattern. A decision rule represents a probabilistic relationship between a set of conditions and a decision. Given a decision system $\mathcal{A}$ = (*U*,*A*,{*d*}), let α denote a pattern that only involves attributes in *A*. Let *β* denote a descriptor *d* = *v*, where v ∈ V_d. The decision rule is then read as "if *α* then *β*", and is denoted *α*→*β*. *α* is called the rule's *antecedent*, and *β* the rule's *consequent* [4].

In practice however, generating a decision rule from a reduct or a reduct-equivalent means overlaying the attributes in the reduct over an object *x*, and reading off the values of a(*x*) for every a ∈ reduct. This means that the decision rules will always be conjunctions of descriptors (or a single descriptor, in the event that the reduct consists of a single attribute). Rules of this type are said to represent positive knowledge, defined as follows:

Given a DS $\mathcal{A}$ = (*U*,*A*,{*d*}). The decision rule $\alpha \rightarrow \beta$ is said to be a *positive decision rule* if *α* is a conjunction of descriptors that only involve attributes in *A*.

IV. Upper and Lower Approximation

Let $\mathcal{A}$ = (*U*,*A*) be an information system and let $B \subseteq A$ and $X \subseteq U$. We can approximate *X* using only the information contained in *B* by constructing the *B-lower* and *B-upper approximations* of *X*, denoted $\underline{B}X$ and $\overline{B}X$ respectively, where $\underline{B}X$ = {*x* : $[x]_B \subseteq X$} and $\overline{B}X$ = {*x* : $[x]_B \cup X \neq \phi$}.

The lower approximation corresponds to certain rules while the upper approximation to possible rules (rules with confidence greater than 0) for *X*. The *B*-lower approximation of *X* is the set of all objects, which can be with certainty classified to *X* using attributes from *B*. The set *U* – $\overline{B}X$ is called the *B-outside region* of *X* and consists of those objects, which can be with certainty classified as not belonging to *X* using attributes from *B*. The set $BN_B(X)$ = $\overline{B}X$ – $\underline{B}X$ is called the *B-boundary region* of *X* thus consisting of those objects that on the basis of the attributes from *B* cannot be unambiguously classified into *X*. A set is said to be *rough* (respectively crisp) if the boundary region is non-empty (respectively empty). Consequently each rough set has boundary-line cases, i.e., objects, which cannot be with certainty classified neither as members of the set nor of its complement. Obviously crisp sets have no boundary-line elements at all. That means that boundary-line cases cannot be properly classified by employing the available knowledge. The size of the boundary region can be used as a measure of the quality of set approximation (in *U*). It can be easily seen that the lower and upper approximations of a set are, respectively, the interior and the closure of this set in the topology generated by the indiscernibility relation [12].

A. Accuracy of Approximation

A rough set *X* can be characterized numerically by the following coefficient,

$$\alpha_B(X) = \frac{\left|\underline{B}(X)\right|}{\left|\overline{B}(X)\right|},$$

called the *accuracy of approximation*, where $|X|$ denotes the cardinality of $X \neq \phi$ and *B* is a set of attributes. Obviously $0 \leq \alpha_B(X) \leq 1$. If $\alpha_B(X) = 1$, *X* is crisp with respect to *B* (*X* is *exact* with respect to *B*), and otherwise, if $\alpha_B(X) < 1$, *X* is *rough* with respect to *B* (*X* is *vague* with respect to *B*).

B. Rough Membership Function

In classical set theory either an element belongs to a set or it does not. The corresponding membership function is the characteristic function of the set, i.e., the function takes values 1 and 0, respectively. In the case of rough sets the notion of membership is different. The *rough membership function* quantifies the degree of relative overlap between the set *X* and the equivalence class to which *x* belongs. It is defined as follows:

$$\mu_x^B(x) : U \rightarrow [0,1] \text{ and } \mu_x^B(x) = \frac{\left|[x]_B \cap X\right|}{\left|[x]_B\right|}.$$

The rough membership function can be interpreted as a frequency-based estimate of Pr($y \in X | u$), the conditional probability that object *y* belongs to set *X*, given the information signature $u = Inf_B(x)$ of object *x* with respect to attributes *B*. The value $\mu_X^B(x)$ measures degree of inclusion of $\{y \in U : Inf_B(x) = Inf_B(y)\}$ in *X*.

V. Computational Aspects of Rough Set on KDD

In the literature [7], there has long been a lack of time complexity analysis of algorithms for frequently used rough set operations. Time complexities of constructing an equivalence relation are shown to be $O(lm^2)$, where *l* and *m* are number of attributes and objects, respectively [8]. This result corresponds to the analysis of an algorithm, reported in [9], where the goal is to obtain the equivalence relation according to the values of a single attribute. For a given

functional dependency $X \Rightarrow Y$ that holds in an information table S, we say that $x \in X$ is superfluous (or non-significant) attribute for Y in S if and only if, $X-\{x\} \Rightarrow Y$ still holds in S. A reduct of X for Y in S is a subset P of X such that P does not contain any superfluous attribute. If we have a metric to measure the degree of dependency, then we have a way to explore a reduct of X, with a degree of θ, where $0 \leq \theta \leq 1$ [10]. It is shown that finding a reduct of X for Y in S is computationally bounded by l^2m^2 where l and m is a length of X and the number of objects in S respectively. The time complexity to find all reducts of X is $O(2^lJ)$, where J is the computational cost for finding one reduct, and l is the number of attributes in X [11].

VI. Conclusion

In the paper, basic concepts of data mining/KDD and the rough set theory were discussed. Rough Set Theory has been widely used in KDD since it was put forward. Having important functions in the expression, study, conclusion and etc. of the uncertain knowledge, it is a powerful tool, which sets up the intelligent decision system. The main focus is to show how rough set techniques can be employed as an approach to the problem of data mining and knowledge extraction.

References

[1] Ryszard S. Michalski and Kenneth A. Kaufman, "Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach", *Machine Learning and Data Mining, Methods and Applications,* 1997.

[2] J.N.Kok and W.A.Kosters, "Natural Data Mining Techniques", *European Association for Theoretical Computer Science*, Vol. 71, June 2000, pp.133-142.

[3] Ning ZHONG, Andrzej SKOWRON, "A Rough Set-Based Knowledge Discovery Process", *International Journal of Applied Mathematical Computer Science,* 2001, Vol.11, No.3, pp.603-619.

[4] Terje Løken, "Rough Modeling Extracting Compact Models from Large Databases", Knowledge Systems Group IDI, *A thesis submitted to Norwegian University of Science and Technology*, 1999.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases", *Artificial Intelligence Magazine 17* (1996), pp.37–54.

[6] Ivo Düntsch, Günther Gediga, Hung Son Nguyen, "Rough set data analysis in the KDD process", *published in the Proceedings of IPMU 2000*, pp. 220-226.

[7] Hayri Sever, "The Status of Research on Rough Sets for Knowledge Discovery in Databases", www. cuadra.cr.usgs.gov/pubs/srj98.pdf

[8] Deogun.J,Choubey.S,Raghavan.V and Sever.H, "Feature selection and effective classifiers", *Journal of ASIS 49*, 5 (1998), pp.423–434.

[9] Bell.D, and Guan.J, "Computational methods for rough classification and discovery", J*ournal of ASIS 49,* 5 (1998), pp.403–414.

[10] Deogun.J.S, Raghavan.V.V, and Sever.H, "Exploiting upper approximations in the rough set methodology", *In The First International Conference on Knowledge Discovery and Data Mining (Montreal, Quebec, Canada, aug 1995), U. Fayyad and R. Uthurusamy, Eds*., pp.69–74.

[11] Kent.R. E, "Rough concept analysis", *In Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (Banff, Alberta, Canada, 1993)*, pp.245–253.

[12] Andrzej Skowron, "Rough Sets in KDD", *Institute of Mathematics Warsaw University Banacha* 2, 02{095, Warsaw, Poland.

Hrudaya Ku. Tripatthy is presently working in the Department of Computer Science & Engineering at Institute of Advanced Computer & Research. M. Tech in Computer Science & Engineering from Indian Institute of Technology Guwahati on the year 2006. He is continuing PhD in Computer Science under Berhampur University. He has 10 years of teaching experience in both Graduate & under graduate technical degree course.