The Status of Research on Rough Sets for Knowledge Discovery in Databases

Hayri Sever*
The Department of Computer Science & Engineering
Hacettepe University
06532 Beytepe, Ankara, Turkey

Vijay V. Raghavan and Thomas D. Johnsten
The Center for Advanced Computer Studies
University of Southwestern Louisiana
Lafayette, LA 70504, USA

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. Many aspects of KDD have been investigated in several related fields. The emphasis of ongoing research is to extend existing results to handle characteristics of real-world databases. In this article, we outline the fundamental issues of KDD as well as describe the current status of research on applying rough set theory to KDD.

1 Introduction

In the last decade, we have seen an explosive growth in our capabilities to both collect and store data. In fact, it is estimated that the amount of information in the world doubles every 20 months. Our inability to interpret and digest large quantities of data has created a need for a new generation of tools and techniques. Consequently, the discipline of knowledge discovery in databases (KDD), which deals with the study of such tools and techniques, has spurred the interest of researchers from several different disciplines to work on the problems of KDD. Among those, a foundational area that has made important contributions to the knowledge discovery process is rough set theory.

We believe that the extension of rough set theory to solve problems of KDD is a challenging research topic and has the promise of high payoffs in many business and scientific

*Dr. H. Sever is on leave to Department of Computer Science & Engineering, University of Nebraska at Lincoln.

domains. In this article, we asses the current status of and trends in rough set theory from the point of view of KDD.

2 Knowledge Discovery Issues

We discuss knowledge discovery issues under three subsections. First, we outline the steps that comprise the discovery process. We then describe the key characteristics of real-world data that should be accounted for in order to maximize the discovery process. Finally, we present a taxonomy of knowledge discovery tasks (or data mining queries). Although not exhaustive, our taxonomy nevertheless includes those queries commonly cited in the literature.

2.1 Knowledge Discovery Process

In this subsection we briefly outline the KDD process as proposed by Fayyad et al. in [7].

Data Selection: The formulation of a data set that is appropriate for the current discovery task. This step may require joining together several existing data sets in order to obtain an appropriate set of examples.

Data Cleaning and Preprocessing: The elimination or modification of examples from the selected data set that are either noisy or contain missing values. This step improves the overall quality of the discovered information.

Data Reduction: The elimination of non-relevant features (attributes) as well as duplicate examples from the selected data. This step reduces the time required to execute a discovery, or data mining, query.

Data Mining: The selection of a data mining query type (classification, characterization, association, clustering, sequence analysis, etc.) as well as a specific method to execute the selected query type.

Evaluation: The evaluation of the discovered information with respect to validity, novelty, usefulness and simplicity.

2.2 Characteristic Features of Real World Data

Ultra Large Data: The volume of data in real world database systems has already reached the level of giga (or tera) bytes and continues to grow rapidly. Therefore, it is impossible to apply knowledge discovery techniques involving an exhaustive search over a data set.

Noisy Data: There is little support by commercial DBMSs to reduce errors that occur during data-entry and there is virtually nothing that a DBMS can do to catch errors that occur during the collection process. As a result, a knowledge discovery system must be tolerant to the occurrence of noise.

Incomplete Data: The available knowledge in many practical situations is often incomplete and imprecise. This happens either if the examples are not representative all the distinct cases that the system may encounter, or if the decision with respect to certain examples are contradicted by others. Under such circumstances, a knowledge discovery system should have the capability of providing approximate decisions with some confidence level [15].

Redundant Data: A given data set may contain redundant or insignificant attributes with respect to the problem at hand. Fortunately, there exist many near-optimal solutions, or optimal solutions in special cases, with reasonable time complexity that eliminate insignificant (or redundant) attributes.

2.3 Data Mining Queries

The taxonomy provided in [4] is summarized below:

Hypothesis Testing Query: Hypothesis testing queries are fundamentally distinct from other classes of data mining queries since they do not explicitly discover patterns within the data. Instead their purpose is to evaluate a stated hypothesis against a selected database. This form of analysis is particularly useful in refining or expanding already discovered knowledge.

Classification Query: Classification queries result in the induction of a classification function that partitions a given set of examples into meaningful disjoint subclasses as defined by the values of some "decision" attributes. A classification query discovers patterns that distinguish examples belonging to one concept from those belonging to other concepts.

Clustering Query: Clustering queries partition examples into subgroups or clusters according to certain natural criterion of cohesion among examples. This form of discovery differs from classification queries which require a pre-classified set of examples.

Association Query: Association queries discover data items (values) occuring as a group (called bags) within examples. This type of association, or relationship, among data items are interesting only if they occur frequently enough within a collection of data.

Characterization Query: Characterization queries, unlike classification queries, discover the common features of a concept independently of the characteristics of other concepts. Hence, a characterization query may discover commonalities which are not unique to a given concept.

3 The state of Rough Set Computation

Rough set theory, which is based on the indiscernibility relation of objects, is used to reason about data. In this section, we review the rough set terminology as well as survey existing work in the area of rough sets as related to KDD.

3.1 Terminology of Rough Set Theory

Let the pair $A = (U, R)$ be an approximation space, where U is a finite set of objects in the universe and R is the set of blocks in a partition of U. A member of R is called an elementary set. A *definable* set in A is obtained by applying a finite number of union operations over elements of R. Let the concept of interest X be a subset of U. The least definable set in A containing X, $Cl_A(X)$, is called *closure set* (also known as *upper set*) of X in A. Similarly, the greatest definable set in A that is contained in X, $Int_A(X)$, is called *interior set* (also known as *lower set*) of X in A. We say that the set X is definable in A if $X \in R^*$; otherwise X is said to be a *rough* set or *non-definable*.

For a given $x \in U$, a decision algorithm, denoted by $D_A(X)$, yields one of these three answers: a) x is in X, b) x is not in X, c) *unknown*. We now define the corresponding sets of X in A for each answer. Let $POS_A(X)$ be a set of objects in which each object is considered a member of the concept X by $D_A(X)$. Let $BND_A(X)$ be a set of objects in which $D_A(X)$ gives the answer of *unknown*. Finally, let $NEG_A(X)$ be a set of objects that are not regarded as members of X by $D_A(X)$.

Let $R = \{R_1, R_2, \cdots, R_k\}$. To provide alternative definitions of a positive region of a concept X, rough set theory can utilize elementary sets as shown below: (a) $POS^l_A(X) = Int_A(X) = \bigcup^k_{R_i \subseteq X} R_i$, – interior set approximation; (b) $POS^u_A(X) = Cl_A(X) = \bigcup^k_{R_i \cap X \neq} R_i$, – closure set approximation; and (c) $POS^e_A(X) = \bigcup^k_{|R_i \cap X|/|R_i| \geq \tau} R_i$, where τ is a threshold and$0.5 < \tau \leq 1$, – elementary set approximation. These rules are utilized by the decision algorithm to decide if $x \in X$ for a given $x \in U$. The degree of approximation quality is expressed by $\mu_A(X) = |POS_A(X) \cap X|/(s_1 * |POS_A(X)| + s_2 * |X|)$, where s_1 and s_2 are scaling factors and their sum must equal one. These scaling factors quantify the user's preference as to the increase in accuracy of $D_A(X)$ relative to a certain loss in accuracy of X (or vice versa.)

Let $F = \{X_1, X_2, \ldots, X_n\}$, where $X_i \subseteq U$, be a set of mutually exclusive concepts in A. Positive sets of F in A are defined as the family $POS_A(F) = \{POS_A(X_1), POS_A(X_2), \cdots, POS_A(X_n)\}$. A classification problem is described as generating a decision algorithm, $D_A(F)$, that relates definable sets to concepts. If $D_A(F)$ is a relation then it is called *an inconsistent decision algorithm;* otherwise, it is *a consistent decision algorithm.* Since $POS_A(F) = \bigcup_{X \in F} POS_A(X)$, the classification quality $\varphi_A(F)$ is equal to $\frac{1}{|U|} \sum_{i=1}^n |X_i| \, \mu_A(X_i)$. If $\varphi_A(F)$ is equal to one the classification is *definable* (or *perfect*); otherwise it is *roughly definable* classification.

The Notion of Information System

An information system (also known as information table) can be viewed as an application of rough set theory in which each object is described by a set of attributes. Formally, such a system is defined as a quadruple $S = (U, Q, V, \rho)$; where, U is the finite set of objects; Q is the set of attributes; V is the union of domains of attributes in Q; and $\rho : UXQ \Rightarrow V$ is a total description function. The set of attributes in Q is divided into condition attributes, denoted by CON, and decision attributes, denoted by DEC, if one's interest is in the classification of objects. In the context of classification, the information system is called *decision table*. Let $U/\widetilde{P}$ denote the set of blocks in the partition defined by the values of P

on U. A decision algorithm, induced from S, relates the elements of $U/\widetilde{CON}$ to those of $U/\widetilde{DEC}$.

Let $S(P)$ denote a substructure of S such that $S(P) = (U, Q' = P \cup DEC, \bigcup_{a \in P} V_a, \rho')$, where $P \subseteq CON$, ρ' is a restriction of ρ to the set UXQ'. It is said that $CON - P$ is *θ-superfluous* in S iff

$$\varphi_{S(P)}(U/\widetilde{DEC}) = \varphi_S(U/\widetilde{DEC})(1 - \theta),$$

where $0 \leq \theta \leq 1$. Similarly, P is a *θ-reduct* of CON iff $CON - P$ is a *θ-superfluous* in S and no $P' \subset P$ is *θ-superfluous* in $S(P)$.

3.2 Rough Sets for KDD

This subsection is organized around research work, based on rough set theory, addressing nature of real world data, handling data mining queries, and the computational aspects of rough set theory.

Nature of Real World Data

Rough set theory, as originally proposed, approximates given concept(s) using lower and upper sets of the concept(s). Given that the uncertainty in a data set is caused by *noisy* or *incomplete* data, this approach is not always desirable because it does not exercise opportunities to discover/generalize a valuable pattern that is distorted by noise or that is almost certain. This drawback has been addressed however by numerous works. The alternative is to generalize rough approximation methods by adopting alternative definitions of positive (and boundary) regions[5, 9]. For example, in the *elementary set approximation* of an unknown concept[5], an elementary set is mapped to the positive region of an unknown concept if its degree of membership is greater than a user defined threshold value. Another approach is to shift the problem definition to where a probabilistic approximation space, instead of an algebraic approximation space, is adopted [14].

In rough set based classification, the terms `inconsistent' and `nondeterministic' decision algorithms (or rules) have been used interchangeably, though they are different concepts. Recently, it has been argued that `inconsistency' is attributed to the result of a classification method while `nondeterminism' is attributed to the interpretation of that result. It is shown in [5], that inconsistent decision algorithms, under an appropriate representation structure, can be interpreted either deterministically or nondeterministically. This is an important result, particularly when the background knowledge is *incomplete and dynamic*.

Redundant data can be eliminated by pruning insignificant attributes with respect to the problem at hand. In the context of rough set theory, the emphasis, however, is on a more restricted version of the redundancy problem, called reduction of an information system. It is the process of reducing an information system such that the set of attributes of the reduced information system preserves the discrimination power of the original table and no further attribute elimination is possible without loss of some information from the system. The resulting information system is called a *reduct*. Given that an exhaustive search over all possible attribute combinations will require time that is exponential in the number of attributes, it may not be computationally feasible to find a reduct. Furthermore, finding just a single reduct may be too restrictive for some data analysis problems. One plausible approach is to utilize the idea of *θ-reduct* defined in the previous subsection[6].

To handle missing values, Grzymala-Busse [8] has transformed a given decision table with unknown values to a new and possibly inconsistent decision table by replacing the unknown attribute value with all possible values of that attribute. In other words, he reduced the missing value problem to that of learning from inconsistent examples. He then used rough set theory to induce certain (and/or possible) rules. Alternatively, instead of using an equivalence relation, we can use a partial order relation on object subsets, which is called the generalized rough set model in [11]. In this case, an information table becomes a set-valued information table; that is, a field in a tuple may assume more than one value from its domain. Assuming that a missing value of a field is equal to all possible values of its domain, the set-valued information table is used in the induction of decision rules. It is worth stating that the notion of a set-valued information table can be conceived as the natural counterpart of a set-valued relation, which has been proposed as a mechanism to handle uncertain patterns.

Data Mining Queries

Rough set based approach for knowledge discovery employs a greedy algorithm technique for reducing search space in order to extract a reduct of given a decision table. These algorithms can be distinguished by the way they explore the search space. More specifically, the examination of various subsets of attributes is performed either by stepwise forward or by stepwise backward selection techniques. Another question that arises, in the design of algorithms to extract knowledge, is how to quantify the objective function. There are a number of metrics proposed for the evaluation of a rule, namely accuracy [15], significance [1], quality [2], and penalty factor [13].

In [1], the knowledge discovery algorithm exploits lower approximations of given concepts. The reduction of search space is based on the stepwise forward selection technique, where initially no dependency between attributes is considered and subsequently, as further iterations are performed, greater and greater degrees (pairs, triplets, etc.) of attribute dependence is recognized. The output of the algorithm is a set of rules, such that each rule covers only one elementary set. The significance (or quality) of a rule is quantified by the proportion of the elementary set over the set of objects in a given decision table. This algorithm assumes a consistent decision table since it is based on the lower approximation. A similar algorithm, called ILA, which employs the notion of an almost elementary set of a given concept, has been proposed in [13]. The heuristic exploited by ILA is based on a well-known metric called “penalty factor.”

When we inspect *the data mining queries* with respect to the rough set methodology, we see that *attribute dependency analysis* and *classification* are well investigated subjects, among others. *Hypothesis testing* and *association* queries can easily be handled by the rough set methodology. A recent theoretical paper by Kent [10] extends the notions of approximation and rough equality to formal concept analysis. An immediate result of this study, in the data mining context, is to use the rough set methodology for the *characterization of a concept* (or more generally for concept exploration). As a final note, rough classifiers face a problem when a new object (coming from outside of the data set) is introduced and the description of the object does not match any of the rules included in a classifier. In such cases, a mechanism to find the closeness of the given object to known concepts at hand is needed. The usual remedy for this problem is to map non-quantitative values into a numerical scale and use a distance function for the evaluation [12].

Computational Aspects of Rough Set Theory

In the literature, there has long been a lack of time complexity analysis of algorithms for frequently used rough set operations. Recently two independent studies have addressed this issue, which is the subject of this section [1, 3].

Time complexity of constructing an equivalence relation is shown to be $O(lm^2)$, where l and m are number of attributes and objects, respectively [3]. This result correponds to the anlysis of an algorithm, reported in [1], where the goal is to obtain the equivalence relation according to the values of a single attribute.

A single concept is defined by a pair of its interior and closure sets. The computation effort for finding either interior (or lower) or closure (or upper) sets is $O(lm^2)$, where l and m are number of attributes and objects, respectively [1, 3].

The intersection of two equivalence relations is mainly used for reducing the search space (e.g., stepwise backward/forward feature selection [3] or knowledge discovery based on forward selection of significant features [1]). This computation is bounded by m^2 for two equivalence relations on the same set of objects.

For a given functional dependency $X \Longrightarrow Y$ that holds in an information table S, we say that $x \in X$ is superfluous (or nonsignificant) attribute for Y in S if and only if, $X - \{x\} \Longrightarrow Y$ 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 $0 \leq \theta \leq 1$ [6]. It is shown in [1] 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.

4 Future Directions

As mentioned in the previous section, some aspects of the nature of data (i.e., incomplete, redundant, and uncertain data) have already been investigated in the rough set methodology, but the resulting algorithms need to be tested using large databases. Slowinski & Stefonowski's study on determining the nearest rule [12], in the case that the description of a new object does not match those of known concepts, is a preliminary contribution in enhancing the performance of a rough classifier when the training data set is poorly designed or sampled from a large data set. Even though it is not stated in the paper, such a measure can be used for *clustering queries*. Although data dependency analysis within the rough set methodology can be applied to the characterization of concepts, the current efforts need to be extended to explicitly include the process of representing relationships between concepts, when a knowledge model contains a set/hierarchy of persistent concepts [10].

References

[1] BELL, D., AND GUAN, J. Computational methods for rough classification and discovery. *Journal of ASIS 49*, 5 (1998), 403–414.

[2] CHOUBEY, S. K., DEOGUN, J. S., RAGHAVAN, V. V., AND SEVER, H. A comparison of feature selection algorithms in the context of rough classifiers. In *Proceedings of Fifth IEEE International Conference on Fuzzy Systems* (New Orleans, LA, 1996), vol. 2, pp. 1122–1128.

[3] DEOGUN, J., CHOUBEY, S., RAGHAVAN, V., AND SEVER, H. Feature selection and effective classifiers. *Journal of ASIS 49*, 5 (1998), 423–434.

[4] DEOGUN, J. S., RAGHAVAN, V. V., SARKAR, A., AND SEVER, H. Data mining: Research trends, challenges, and applications. In *Roughs Sets and Data Mining: Analysis of Imprecise Data* (Boston, MA, 1997), T. Y. Lin and N. Cercone, Eds., Kluwer Academic Publishers, pp. 9–45.

[5] DEOGUN, J. S., RAGHAVAN, V. V., AND SEVER, H. Rough set based classification methods and extended decision tables. In *Proceedings of the International Workshop on Rough Sets and Soft Computing* (San Jose, California, 1994), pp. 302–309.

[6] 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.

[7] FAYYAD, U., PIATETSKY-SHAPIRO, G., AND SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of ACM 39*, 11 (1996), 27–34.

[8] GRZYMALA-BUSSE, J. W. On the unknown attribute values in learning from examples. In *Proceedings of Methodologies for Intelligent Systems*, Z. W. Ras and M. Zemankowa, Eds., Lecture Notes in AI, 542. Springer-Verlag, New York, 1991, pp. 368–377.

[9] HASHEMI, R. R., PEARCE, B. A., HINSON, W. G., PAULE, M. G., AND YOUNG, J. F. IQ estimation of monkeys based on human data using rough sets. In *Proceedings of the International Workshop on Rough Sets and Soft Computing* (San Jose, California, 1994), pp. 400–407.

[10] 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.

[11] LINGRAS, P., AND YAO, Y. Data mining using extensions of rough set model. *Journal of ASIS 49*, 5 (1998), 415–422.

[12] SLOWINSKI, R., AND STEFANOWISKI, J. Rough classification with valued closeness relation. In *Proceedings of the International Workshop on Rough Sets and Knowledge Discovery* (San Jose, CA, 1995).

[13] TOLUN, M. R., SEVER, H., AND ULUDAG, M. Improved rule discovery performance on uncertainty. In *Research and Development in Knowledge Discovery and Data Mining*, X. Wu, R. Kotagiri, and K. Korb, Eds., vol. 1394 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 1998, pp. 310–321. Proc. Second Pacific-Asia Conf. PAKDD-98, Melborne, Australia, April 1998.

[14] YAO, Y. Y., AND WONG, K. M. A decision theoretic framework for approximating concepts. *International Journal Man-Machine Studies 37* (1992), 793–809.

[15] ZIARKO, W. The discovery, analysis, and representation of data dependencies in databases. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. AAAI/MIT, Cambridge, MA, 1991.