Rough Set Approach to KDD

Hung Son Nguyen, Andrzej Skowron

Institute of Mathematics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl and son@mimuw.edu.pl

Abstract. This tutorial is a survey on rough set theory and some of its applications in Knowledge Discovery from Databases (KDD). It will also cover the practice guide to analysis of different real life problems using rough set methods as well as the presentation of Rough Set Exploration System (RSES) what can be treated as a preliminary material for the main conference and associated workshops.

1 Introduction

Rough Set theory was introduced by Zdzislaw Pawlak in the early 80’s and has currently reached a level of high visibility and maturity [33, 32, 36, 35, 34]. Originally, rough sets, whose main philosophy is based simply on indiscernibility and discernibility of objects, were presented as an approach to concept approximation under uncertainty. This brilliantly simple idea has been successively expanded in the last twenty years. Many effective methods for data analysis have been developed on the basis of rough set theory.

In recent years, a growth of interest in rough set theory and its applications can be seen in the number of research papers submitted to international workshops, conferences, journals and edited books, including two main biannual conferences on rough sets and the special sub-line of LNCS series. A large number of efficient applications of rough sets in Knowledge Discovery for various types of databases have been developed. Rough sets are applied in many domains, such as medicine, finance, marketing, telecommunication, conflict resolution, text mining, intelligent agents, image analysis, pattern recognition, bioinformatics (e.g., see, [36, 35, 34] and the bibliography in these papers).

This tutorial is intended to fulfill the needs of many researchers to understand the rough set methodologies for mining of standard and nonstandard data. The methodology based on rough sets can serve as a useful tool to complement capabilities of other data mining methods. The tutorial should help the audience to find out if some of the presented methods may support their own KDD and data mining (DM) research.

The tutorial is intended to occupy four slots of 45 minutes each and to cover the following topics:

– Fundamentals of rough set theory;
– Rough set approach to KDD;
Examples of rough set based methods for data reduction, rule extraction, discretization, decomposition, hierarchical learning;
Practical guide for Rough Set Exploration System (RSES);
Some exemplary applications of RSES (with exercises);
Emergent challenging problems.

The first part of this tutorial is targeted to those participants who want to learn rough set theory by examples at the basic level in the context of KDD problems. The second part of this tutorial is targeted to young researchers who want to apply the rough set approach to real-life problems in KDD.

2 From Rough Sets and Boolean Reasoning to Knowledge Discovery from Databases

The rapidly growing volume and complexity of modern databases make the need for technologies to describe and summarize the information they contain increasingly important. Knowledge Discovery in Databases (KDD) and data mining are new research areas that try to deal with this problem. In [8], KDD was characterized as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data, while data mining is a process of extracting implicit, previously unknown and potentially useful patterns and relationships from data, and it is widely used in industry and business applications. As the main step in KDD, data mining methods are required to be not only accurate but also to deliver understandable and interpretable results for users, e.g., through visualization. The other important issue of data mining methods is their complexity and scalability. Presently, data mining is a collection of methods from various disciplines such as mathematics, statistics, logics, pattern recognition, machine learning, non-conventional models and heuristics for computing (see, e.g., [9, 10, 14, 15]).

Concept approximation is one of the most fundamental issues in machine learning and data mining. The problem considered in this paper is the creation of a general framework for concept approximation. The need for such a general framework arises in machine learning and data mining. Classification, clustering, association analysis or regression are examples of well-known problems in data mining that can be considered in the context of concept approximation. A great effort by many researchers has led to the design of newer, faster and more efficient methods for solving the concept approximation problem (see, e.g., [22, 47, 34]).

Rough set theory has been introduced in [32, 33] as a tool for concept approximation under uncertainty. The idea is to approximate the concept by two descriptive sets called the lower approximation and the upper approximation. The lower and upper approximations must be extracted from available training data. The main philosophy of the rough set approach to concept approximation problem is based on minimizing the difference between upper and lower approximations (also called the boundary region). This simple, but brilliant idea, leads to many efficient applications of rough sets in machine learning and data mining.
like feature selection, rule induction, discretization or classifier construction (see, e.g., [44, 22, 34]).

As boolean algebra has a fundamental role in computer science, the boolean reasoning approach is also an ideological method in Artificial Intelligence [7]. SAT-planing (see [12], [13]) can be mentioned here as a representative example. In recent years, boolean reasoning approach has proven to be a powerful tool for designing effective and accurate solutions for many problems in rough set theory [42, 41, 22]. Boolean reasoning is an algebraic methodology for problem solving. The idea is to encode the studied problem by a Boolean equation, i.e., an equation of form \( f = 0 \) or \( f = 1 \) for some Boolean expression \( f \); and to resolve the corresponding problem over this equation (using (Boolean) algebraic methods) [7]. It has been shown that most problems over Boolean equations are related to the problem of searching for prime implicants of a Boolean formula.

Boolean reasoning methodology plays an important role in intelligent systems as it delivers a universal tool for designing automatic problem solvers. Selman et al [38] have presented an excellent summary of the state of the art in Boolean reasoning, and sketch challenges for the next 10 years. Our research was motivated by the last three which are related to encoding techniques and their influence on computational properties.

This tutorial presents a general framework, called Rough Sets and Approximate Boolean Reasoning Approach, to the concept approximation problem. This is a combination of rough set theory [33, 32, 36, 35, 34], Boolean reasoning methodology [7] and data mining [8–10, 14, 15]. We present the theoretical foundation of the proposed method, including some extensions of the classical rough set theory, as well as its applications to many problems in DM and KDD.

3 Rough Set Based Methods for KDD

In the supervised machine learning paradigm [9, 14–16], a learning algorithm is given a training data set, usually in the form of a decision system \( \mathcal{A} = (U, A, d) \), prepared by an expert. Each such a decision system classifies elements from \( U \) into decision classes. The purpose of the algorithm is to return a set of decision rules together with matching procedure and conflict resolution strategy, called a classifier, which allow to classify unseen objects, i.e., objects that are not described in the original decision table.

In this section, we provide a number of rough set methods that can be used in constructing classifiers. Most of the techniques discussed below are based on searching for prime implicants for computing different kinds of reducts. They are computationally hard. However, many heuristics have been developed which turned out to be very promising. The results of experiments on many data sets, reported in the literature show a very good quality of classification of unseen objects using these heuristics. Moreover, a successful methodology, based on different types of reducts, has been developed for to solve of many problems

\footnote{For simplicity we consider decision systems with one decision.}
like attribute selection, decision rule generation, association rule generation, discretization of real-valued attributes and symbolic value grouping. For further readings the reader is referred to [2, 40, 48] (attribute selection); [25, 17, 22] (discretization); [18, 19] (discretization of data stored in relational databases); [24] (reduct approximation and association rules) as well as to [28, 34].

3.1 Reducts in Information and Decision Systems

In general, reducts are minimal subsets (with respect to the set inclusion relation) of attributes which contain a necessary portion of information about the set of all attributes. The notion of information is as abstractive as the notion of energy in physics, and we will not be able to define it exactly. Instead of explicit information, we have to define some objective properties for all subsets of attributes. Such properties can be expressed in different ways, e.g., by logical formulas or, as in this section, by a monotone evaluation function which is described as follows.

For a given information system $S = (U, A)$, the function

$$
\mu_S : \mathcal{P}(A) \rightarrow \mathbb{R}^+,
$$

where $\mathcal{P}(A)$ is the power set of $A$ and $\mathbb{R}^+$ is the set of non-negative reals, is called the monotone evaluation function if the following conditions hold:

1. the value of $\mu_S(B)$ can be computed using the information set $INF(B) = \{(a, a(x)) : a \in B \text{ and } x \in U\}$ for any $B \subseteq A$;
2. for any $B, C \subseteq A$, if $B \subseteq C$, then $\mu_S(B) \leq \mu_S(C)$.

Any set $B \subseteq A$ is called the reduct relative to a monotone evaluation function $\mu$, or briefly $\mu$-reduct, if $B$ is the smallest subset of attributes that $\mu(B) = \mu(A)$, i.e., $\mu(B') \not\leq \mu(B)$ for any proper subset $B' \subsetneq B$.

This general definition can be used for many different types of reducts. The details of description of some well-known types of reducts in rough set theory, including the notion of approximation reducts, is presented in [22]. A set of attributes is called an approximate reduct if it preserves almost all necessary information relevant for construction of classifiers. Many experimental results are showing that approximate reducts, which are shorter than the exact ones, can construct more accurate classifiers.

The tutorial is intended to present an overview of reduct calculation techniques including efficient implementations of greedy strategy, evolutionary strategy as well as apriori-like strategy. These methods are especially useful in the context of the feature selection problem.

3.2 Minimal Decision Rules

In this section, techniques for constructing minimal rules for decision systems will be considered.
Given a decision table $\mathcal{A}$, a **minimal decision rule** (wrt $\mathcal{A}$) is a decision rule which is true in $\mathcal{A}$ and which becomes not true in $\mathcal{A}$ if any elementary descriptor from the left-hand side of the rule is removed.\(^2\)

The minimal number of elementary descriptors on the left-hand side of a minimal decision rule defines the largest subset of a decision class. Accordingly, information included in the conditional part of any minimal decision rule is sufficient for predicting the decision value of all objects satisfying this part of the rule. The conditional parts of minimal decision rules define the largest object sets relevant for approximating decision classes. The conditional parts of minimal decision rules can be computed using prime implicants.

The main challenge in inducing rules from decision systems lies in determining which attributes should be included in the conditional parts of the rules. Using the strategy outlined above, the minimal rules are computed first. Their conditional parts describe the largest object sets with the same generalized decision value in a given decision system. Although such minimal decision rules can be computed, this approach can result in a set of rules of unsatisfactory classification quality. Such rules might appear too general or too specific for classifying new objects. This depends on the data analyzed. Numerous techniques have been developed for the further tuning of minimal rules.

### 3.3 Value Set Reduction

Let us consider a decision system with a large number of attribute values. There is a very low probability that any new object will be properly recognized by matching its attribute value vector with any of the rows in the decision table associated with the decision system. Hence, in order to construct classifiers with the high classification quality, it is often necessary to reduce the cardinality of the value sets of specific attributes in a training decision table. The task of reducing the cardinality of value sets is referred to as the **value set reduction problem**.

In this section, two methods of value set reduction are considered:

1. discretization, used for real value attributes, and
2. symbolic attribute value grouping, used for symbolic attributes.

### 3.4 MD Decision Tree

The main philosophy of RS & ABR methodology in solving classification problems is based on managing the discernible objects. Thus discernibility becomes an interesting measure for many applications of rough sets in data mining.

Classifications based on decision trees belong to the most popular classification methods. The decision tree construction method based on discernibility measure has been proposed. This method, also known as MD-decision tree\(^3\), creates binary decision tree using cuts on continuous attributes, binary partition

\(^2\) A decision rule $\varphi \Rightarrow \psi$ is true in $\mathcal{A}$ if and only if $\|\varphi\|_\mathcal{A} \subseteq \|\psi\|_\mathcal{A}$.

\(^3\) MD = maximal discernibility
of values for symbolic values. Properties and detailed comparison analysis with other techniques were presented in [17, 29, 22].

Crisp partitions defined by cuts in standard discretization and decision tree methods may cause a misclassification of objects that are very close to those cuts. Soft cuts were proposed as a novel concept for mining data with numeric attributes. Unlike traditional cuts, each soft cut is defined as an interval of possible cuts and represents a family of possible partitions. Modified classification methods based on soft cuts and rough set theory were presented in [21, 22].

Soft decision trees, i.e., decision trees using soft cuts, have some advantages in comparison with the traditional ones. Firstly, this approach can overcome the overfitting problem without pruning. Secondly, it is possible to efficiently construct soft decision trees from large data bases [19, 21]. Two techniques called rough decision tree and fuzzy decision tree were proposed in [20].

3.5 Association Rules

In this section [24, 26], we show how rough set techniques can be used to extract association rules [1] from information systems. Association rules, which play an important role in the field of data mining, provide associations among attributes. A real number from the interval [0,1] is assigned to each rule and provides a measure of the confidence of the rule.

There are two basic steps used in methods used for generating association rules. (Below s and c stand for support and confidence thresholds wrt a given information system A).

1. Generate as many templates $T = D_1 \land \ldots \land D_k$ as possible, such that $support_A(T) \geq s$ and $support_A(T \land D_i) < s$, for any descriptor $D_i$ different from all descriptors $D_1, \ldots, D_k$.

2. Search for a partition $\{P, Q\}$ of $T$, for each $T$ generated in the previous step, satisfying
   (a) $support_A(P) < support_A(T)$
   (b) $P$ has the shortest length among templates satisfying (a).

Each such partition defines an association rule of the form $P \Rightarrow Q$ with the confidence greater than $c$.

The second step, which is crucial to the process of extracting association rules, can be solved using rough set methods (see, e.g., [24, 26, 22, 34]).

3.6 Approximate Boolean Reasoning: Discretization of Large Data Sets Stored in Relational Databases

In this section (see [18, 19, 22, 34]), we discuss an application of approximate Boolean reasoning in efficient searching for cuts in large data sets stored in relational databases. Searching for relevant implicants is based on simple statistics that can be efficiently extracted from relational databases.

\footnote{called as frequent items (see, e.g., http://fimi.cs.helsinki.fi/)}
Searching algorithms for optimal partitions of real-valued attributes, defined by cuts, have been intensively studied. The main goal of such algorithms is to discover cuts which can be used to synthesize decision trees or decision rules of high quality wrt some quality measures (e.g., quality of classification of new unseen objects, quality defined by the decision tree height, support and confidence of decision rules).

In general, all those problems are hard from computational point of view (e.g., one can show that the searching problem for minimal and consistent set of cuts is NP-hard). In consequence, numerous heuristics have been developed for approximate solutions of these problems. These heuristics are based on approximate measures estimating the quality of extracted cuts. Among such measures discernibility measures are relevant for the rough set approach.

We outline an approach for solution of a searching problem for optimal partition of real-valued attributes by cuts, assuming that the large data table is represented in a relational database. In such a case, even the linear time complexity wrt the number of cuts is not acceptable because of the time needed for one step. The critical factor for time complexity of algorithms solving that problem is the number of SQL queries of the form

SELECT COUNT FROM a_Table WHERE (an_Attribute BETWEEN value1 AND value2) AND (additional condition)

necessary to construct partitions of real-valued attribute sets. We assume the answer time for such queries does not depend on the interval length\textsuperscript{5}. Using a straightforward approach to the optimal partition selection (wrt a given measure), the number of necessary queries is of order $O(N)$, where $N$ is the number of preassumed cuts. By introducing some optimization measures, it is possible to reduce the size of the searching space. Moreover, using only $O(\log N)$ simple queries, suffices to construct a partition very close to optimal.

4 RSES - Rough Set Exploration System

RSES 2.2 - Rough Set Exploration System 2.2 is a software tool that provides the means for analysis of tabular data sets with use of various methods, in particular those based on Rough Set Theory (see [6]).

The RSES system was created by the research team supervised by Professor Andrzej Skowron. Currently, the RSES research and development team consists of: Jan Bazan (University of Rzeszów), Rafał Latkowski (Warsaw University), Michał Mikołajczyk (Warsaw University), Nguyen Hung Son (Warsaw University), Nguyen Sinh Hoa (Polish-Japanese Institute of Information Technology), Andrzej Skowron (Warsaw University), Dominik Sleza (University of Regina and Polish-Japanese Institute of Information Technology), Piotr Synak (Polish-Japanese Institute of Information Technology), Marcin Szczyka (Warsaw University), Arkadiusz Wojna (Warsaw University), Marcin Wojnarski (Warsaw Uni-

\textsuperscript{5} This assumption is satisfied in some existing database management systems.
The RSES system is freely available (for non commercial use) on the Internet. The software and information about it can be downloaded from:

http://logic.mimuw.edu.pl/~rses

The main aim of RSES is to provide a tool for performing experiments on tabular data sets. In general, the RSES system offers the following capabilities:

- import of data from text files,
- visualization and pre-processing of data including, among others, methods for discretization and missing value completion,
- construction and application of classifiers for both smaller and vast data sets, together with methods for classifier evaluation.

The RSES system is a software tool with an easy-to-use interface, at the same time featuring a bunch of method that make it possible to perform compound, non-trivial experiments in data exploration with use of Rough Set methods.

The version 2.2 of RSES system was written in Java with some computational kernel written in GCC. In this way, the portability of the system was achieved and the RSES system is distributed for both Microsoft Windows 9x/NT/2000/Me/XP and Linux/i386.

The tutorial is associated with some practical examples of performing data mining tasks using RSES system. The following scenarios will be presented during the tutorial:

- Reduct calculation;
- Rule based classifier creation and evaluation;
- Induction of rule based classifier with discretization;
- Decomposition tree;
- Rule based classifier improvement;

5 Challenging Problems

In this part of the tutorial, we discuss several challenging problems.

We start from the problem of searching for approximation of relations from data. We show that this problem creates the basic component of many compound tasks. We also present a novel rough set based approach to discovering useful patterns from nonstandard and complex data for which the standard inductive learning methodology fails. The proposed solution is based on a two-layered learning algorithm. The first layer consists of methods that are responsible for searching for (rough) approximation of some relations between objects from the data. At the second layer, the approximated relations induced by the first layer are used to synthesize the solution of the original problem. The critical problem in any layered learning system is how to control the global accuracy by tuning the quality of its components. We present a solution of this problem based on
the changing of the quality of approximate relations. We describe two representative examples related to binary relations to demonstrate the power of the proposed methodology. In the first example, we consider the problem of extracting the optimal similarity relation and present some applications of approximate similarity relations in classification problem. We present the advantages of this method comparing with the standard classification methods [27, 49]. The second example relates to the approximation of the preference relation and its applications in (1) learning of the ranking order on a collection of combinations, (2) predicting the values of continuous decision attribute, (3) optimizing the process of searching for the combination with maximal decision [23]. This method can be applied to mining ill-defined data, i.e., data sets with few objects but a large number of attributes. Results of some initial experiments on medical and biomedical data sets were very promising.

The second example of challenging problems is related to approximation of complex concepts from data and domain hierarchical knowledge. It has been made a substantial progress in developed methods for approximation of complex concepts in numerous real-life projects related to such areas as medicine, control of unmanned helicopter or Sun spots classifications (see, e.g., [4, 5, 28, 31, 30, 39, 46, 45, 3]). Still further progress is needed.

In the data mining community, there is a rapidly growing interest in developing methods for the discovery of structures of temporal processes from data. Works on discovering models for processes from data have recently been undertaken by many renowned centres worldwide (e.g., http://people.cs.vt.edu/~ramakris/kdddm06/cfp.html, www.isle.org/~langley/, soc.web.cse.unsw.edu.au/bibliography/discovery/index.html). This is the next challenging problem for the rough set theory [43].

In [11, 37] we discuss the Wisdom Granular Computing (WGC) as a basic methodology for Perception Based Computing (PBC). By wisdom, we understand an adaptive ability to make judgements correctly to a satisfactory degree (in particular, correct decisions) having in mind real-life constraints. We propose Rough-Granular Computing (RGC) as the basis for WGC. Wistech can be treated as a general framework for solving the above discusses challenging problems. In particular, the novelty of the proposed approach for the discovery of process models from data and domain knowledge lies in combining, on one side, on a number of novel methods of granular computing for wistech developed using the rough set methods and other known approaches to the approximation of vague and complex concepts such as functionals of changes.

6 Conclusions

We have discussed the methodology based on discernibility and Boolean reasoning for efficient computation of different entities including reducts and decision rules. We have presented the main properties and functionality of RSES system. We have also outlined a promising research direction based on approximate Boolean reasoning. Exemplary applications of the developed methods used for
solving problems in pattern recognition, machine learning, and data mining are included. Finally, we have listed some challenging problems.

References


