USING ROUGH SETS IN DATA MINING

Ing. Pavel JURKA, Doctoral Degree Programme (1)
Dept. of Intelligent Systems, FIT, BUT
E-mail: jurka@fit.vutbr.cz

Supervised by: Dr. František Zbořil

ABSTRACT

This work is about the possibilities of using new approaches and technologies in process of data mining from databases. It presents the rough sets theory, introduces the basic definitions and characteristics such as upper and lower approximation, boundary region and membership function. It presents the algorithm for discovering of decision rules.

1 INTRODUCTION

These days we use large amounts of data, which are stored (saved) in relational or latterly in object databases. These data have low information value. That's why we started to talk about the process of knowledge discovery in databases (KDD) in the beginning of the 1990s. This process is divided into a number of consequent steps.

The main parts are the data preparation, which is needed for adaptation and normalization of entering data and the data mining, which discover hidden patterns in data. The definition says: "Data mining is a nontrivial process of determination of valid, unknown and potential useful and easily understandable dependencies in data." The process of data mining uses the different knowledge of mathematics, informatics and other disciplines of science to find and analyze the relations between information. These procedures have been developed and improved to enable more exact description of acquired knowledge. The big problem in data mining is the deficiency and indeterminateness. This problem is solved by using new theories and procedures, for example fuzzy sets, genetic algorithms or rough sets.

2 ROUGH SETS

The rough sets theory was created by Z. Pawlak in the beginning of the 1980s and it is useful in the process of data mining. It offers the mathematic tools for discovering hidden patterns in data through the use of identification of partial and total dependencies in data. It also enables work with null or missing values. Rough sets can be used separately but usually they are used together with other methods such as fuzzy sets, statistic methods, genetics algorithms etc. The rough sets theory uses different approach to uncertainty. As well as fuzzy sets this theory is only part of the classic theory, not an alternative.

3 THEORY FUNDAMENTS

Suppose we are given a set of objects U called the universe and an indiscernibility relation $R \subseteq U \times U$, representing our lack of knowledge about elements of U. For the sake of simplicity we assume that R is an equivalence relation.

Suppose that X is a subset of U. We want to characterize the set X with respect to R. To this end we will need the basic concepts of rough set theory:

Lower approximation – the set of items, which can be certainly classified as items of X

Upper approximation – the set of items, which can be possibly classified as items of X

Boundary region – the set of items, which can be classified either as items of X or not

Set X is crisp with respect to R, if the boundary region of X is empty.

Set X is rough with respect to R, if the boundary region of X is nonempty.

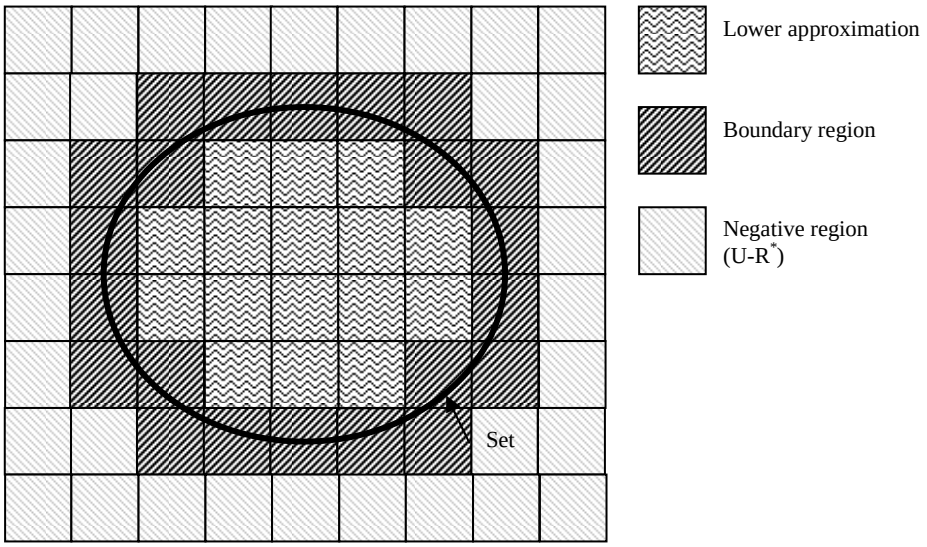

Fig. 1: *Illustration of the boundary region of rough set*

More exact definition acquires specification of the concepts of approximation and boundary region. That's why we introduce equivalence relation and equivalence classes. We define approximation as a union of equivalence classes with determined characteristics.

Let us have equivalence relation $R \subseteq U \times U$. The partition of set U is a set of nonempty subsets $\{X_1, X_2, \ldots, X_k\}$, where $X_1 \cup X_2 \cup \ldots \cup X_k = U$ *and* $X_i \cap X_j$ *for* $i \neq j$ otherwise the partition of set in reciprocally disjunctive subsets. These subsets are called classes. If we have an equivalence relation, we can say, that items in one class are reciprocally in relation and there are no relations with items in different classes. Let's identify the partition induct by relation R and items as R(x). Let's call these subsets equivalence classes. Then

Lower approximation of X can be defined as:

$$R_*(X) = \bigcup_{x \in U} \{\mathrm{R(x)} : \mathrm{R(x)} \subseteq \mathrm{X}\}$$

Upper approximation of X as:

$$R^*(X) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \emptyset\}$$

Boundary region of X as:

$$BN_R(X) = R^*(X) - R_*(X)$$

Rough sets can be also defined by **membership function**

$$\mu_X^R : U \to \langle 0,1 \rangle$$

where

$$\mu_X^R(x) = \frac{cardinality(X \cap R(x))}{cardinality(R(x))} .$$

By the help of approximations we can determine the **basic types of rough sets**

$R_*(x) \neq \emptyset \wedge R^*(X) \neq U$ set is roughly defined.

$R_*(x) = \emptyset \wedge R^*(X) \neq U$ set is internally definable.

$R_*(x) \neq \emptyset \wedge R^*(X) = U$ set is externally definable.

$R_*(x) = \emptyset \wedge R^*(X) = U$ set is totally nondefinable.

Rough sets can be characterized numerically by the help of a degree of roughness, which is defined as:

$$\alpha_R(X) = \frac{cardinality(R_*(X))}{cardinality(R^*(X))}$$

Rough sets distinguish two different conceptions

Vagueness is a characteristic of the set expressed by approximation.

Uncertainty is defined as a characteristic of single items expressed by a membership function.

In the end let's resume the differences of classic, fuzzy and rough sets. Classic sets are defined simply and intuitive or axiomatically. Fuzzy sets are defined by a membership function. Rough sets are defined by approximations.

4 ALGORITHM FOR MINING DECISION RULES

4.1 DATABASE

Let us have a table with attributes as columns and single objects as lines. Then the term database is used for the couple (U,A), where U and A are finite nonempty sets called **Universe** and **Attributes**. Attributes are divided into sets C and D. $C \cup D = A$ and $C \cap D = \emptyset$. C is called conditional attributes and D is called decision attributes.

4.2 DECISION RULES

Association rules look for interesting connections among values of different attributes,. Decision rules are special cases of association rules and are used for classification analogous to the decision trees.

The general form of rule can be expressed this way:

$$if\ X\ then\ Y,$$

where X is a conditional part made up of conditional attributes. Y is a part made up of decision attributes.

4.3 ATTRIBUTES DEPENDENCY

The set of attributes D totally depends on C, if all attributes from D are uniquely determined by attributes from C. There must be a functional dependency between C and D.

Let us define a more general concept of dependency called partial dependency.

Formally we define dependency: Suppose C and D are subsets of A. Then we can say that D depends on C with the degree k (0<k<1), we identify $C \Rightarrow_k D$, if

$$k = \gamma(C, D) = \sum_{X \in U/D} \frac{cardinality(C_*(X))}{cardinality(U)}$$

- If k=1, then D is totally depended on C
- If k<1, then D is partially depended on C.
- If k=0, then D is not depended on C.

4.4 REDUCTION OF ATTRIBUTES

Reduct is the minimal set of conditional attributes that observe the degree of dependency. In a different way we can say, that it is a subset of original conditional attributes that enables us to make the same decision.

If $C \Rightarrow_k D$, then minimal subset C' from C such that $\gamma(C, D) = \gamma(C', D)$, is called as reduct C.

The reduction of attributes is the base of the rough sets theory. In general is the searching of reducts NP unfortunately a problem. That's why part of the research is aimed at searching effective algorithms that will make the calculation fast.

Importance of attributes

Suppose sets C and D and attribute $a \in C$, then we can set up an importance metric of attribute a, expressed as a modification of the degree of dependency.

$$\sigma_{(C,D)}(a) = 1 - \frac{\gamma(C - \{a\}, D)}{\gamma(C, D)}$$

$\sigma_{(C,D)}(\mathrm{Re}\,ductC) = 1$ applies for reduct C and that means, that if we remove these attributes, we cannot make the decision certainly.

4.5 ALGORITHM

Algorithm seeks for all possible decision rules built-up of conditional attributes C and decision attributes D.

1) Discretisation of quantitative attributes.
2) Creation of a reduct of conditional attributes. Arrangement of attributes according to its importance.
3) Generation of decision rules
   a) We choose the first attribute and calculate approximations of division according to D for single items.
   b) If there is no boundary region for a definite item, then the decision rule can be made with 100% possibility.
   c) If there are items with a boundary region, other attributes from C are added and the process is repeated. If data are inconsistent, then we cannot unambiguously define all rules. After using all attributes the rules are made and they have the possibility calculated as ratio of upper and lower approximation of sets.

The result of algorithm is a set of decision rules that has a 100% support and a set of rules that has a lower support. For each of these rules the support means when the rule is applicable.

5 CONCLUSIONS

The object of this work was to suggest the possibilities of alternative methods of data mining. It is about the rough sets theory and its use for the mining of decision rules data. The advantage of this method is a mathematic base of rough sets and the possibility of mathematic description of this problem. Rough sets seem to be advantageous for mining of incomplete information as well as for other algorithms.

The object of subsequent work is to suggest modification of algorithm or make a new algorithm, which minimizes or removes disadvantages. One of the possibilities is to join the advantages of rough sets with other methods. An advantageous way may be the use of fuzzy sets, which enable work with quantitative attributes.

REFERENCES

[1] Munakata, T.: Fundamentals of the New Artificial Intelligence: Beyond Traditional Paradigms. 1998.

[2] Pawlak, Z.: Data Mining - a Rough Set Perspective In PAKDD’99. 1999.

[3] Pawlak, Z.: Some Issues on Rough Sets. In Transactions on Rough Sets I. 2004.