THE DEVELOPMENT OF THE GENERALIZATION ALGORITHM BASED ON THE ROUGH SET THEORY

Marina Fomina, Alexey Kulikov, Vadim Vagin

Abstract: This paper considers the problem of concept generalization in decision-making systems where such features of real-world databases as large size, incompleteness and inconsistence of the stored information are taken into account. The methods of the rough set theory (like lower and upper approximations, positive regions and reducts) are used for the solving of this problem. The new discretization algorithm of the continuous attributes is proposed. It essentially increases an overall performance of generalization algorithms and can be applied to processing of real value attributes in large data tables. Also the search algorithm of the significant attributes combined with a stage of discretization is developed. It allows avoiding splitting of continuous domains of insignificant attributes into intervals.

Keywords: knowledge acquisition, knowledge discovery, generalization problem, rough sets, discretization algorithm.

ACM Classification Keywords: H.2.8 Database Applications: data mining; I.2.6 Learning: knowledge acquisition; B.2.4 High-Speed Arithmetic: algorithms.

1. Introduction

Many enterprises in the various areas create and maintain huge databases with information about their activity. However without the productive analysis and generalization such streams of the "raw" data are useless. Due to the application of methods for information generalization in decision making systems, the construction of the generalized data models and processing of large arrays of experimental data are possible. There are sources of such large dataflows in many areas. Application domains of methods for generalization include marketing, medicine, the space researches and many others. Common for these data is that they contain a great many of the hidden regularities, which are important for the strategic solutions making. However, the discovery of these regularities lays outside the human possibilities mainly because of large and permanently increasing size of the data. Therefore the methods for generalization and computer systems implementing these methods are used to derive such regularities.

Concept generalization problem under redundant, incomplete or inconsistent information is very actual. The purpose of this paper is to consider opportunities of the using the rough set theory for solution of a problem of generalization, and to propose the methods improving work of known algorithms. The new discretization algorithm of continuous attributes and the search algorithm of the significant attributes which essentially increase an overall performance of algorithms for generalization will be proposed.

2. Statement of the Generalization Problem

For the description of object we will use features a_1, a_2, …, a_k , which are further called attributes. Each object x is characterized by a set of given values of these attributes: x={v_1, v_2, …,v_k}, where v_i is value of the *i-th* attribute. Such description of an object is called *feature description*. For example, the attributes may be a color, a weight, a form, etc.

Let we have a training set U of objects. It contains both the positive examples (which are concerning to interesting concept) and the negative examples. The concept generalization problem is the construction of the concept allowing the correct classifying with the help of some recognizing rule (*decision rule*) of all positive and negative objects of training set U. Here the construction of the concept is made on the basis of the analysis of a training set.

Let's introduce the following notions related with set U. Let U = {x_1, x_2, …, x_n} is a non-empty finite set of objects. A = {a_1, a_2, …, a_k} is a non-empty finite *set of attributes*. For each attribute the set V_a is defined which refers to the *value set* of attribute *a*. We will denote given value of attribute *a* for object $x\in U$ by $a(x)$. At the decision of a generalization problem often it is necessary to receive the description of the concept, which is specified by value of one of the attributes. We will denote such attribute *d* and call it *decision* or *decision attribute*. The attributes which are included in *A* are called *conditional attributes*. The decision attribute can have some values though quite often it is binary. The number of possible values of a decision attribute *d* is called the rank of the decision and is designated as $r(d)$. We will denote the value set of the decision by $V_d =\{ v_1^d ,v_2^d ,...,v_{r(d)}^d \}$. The decision attribute *d* defines the partition of *U* into classes $C_i = \{x\in U\colon d(x)= v_i^d \}$, $1\le i\le r(d)$.

Generally the concept generated on the basis of training set *U* is an approximation to concept of set *X*, where the closeness degree of these concepts depends on the representativeness of a training set, i.e. how complete the features of set *X* are expressed in it.

3. Basic Notation of the Rough Set Theory

The rough set theory has been proposed in the beginning of 80th years of the last century by the Polish mathematician Z. Pawlak. Later this theory was developed by many researchers and was applied to the decision of various tasks. We will consider how the rough set theory can be used to solve concept generalization problem (also see [1-8]).

In Pawlak's works [1, 9] the concept of an information system has been introduced. An *information system* is understood as pair $\mathrm{S} = (U,A)$, where $U = \{x_1, x_2, ..., x_n\}$ is a non-empty finite set of objects named *training set* or *universe*, and $A = \{a_1, a_2, ..., a_k\}$ is a non-empty finite *set of attributes.* A *decision table* (or *decision system*) is an information system of the form $\mathrm{S} = (U,A\cup \{d\})$, where $d\notin A$ is a distinguished attribute called *decision* or *decision attribute*, *A* is a set of *conditional attributes*.

Let us introduce the *indiscernibility* or *equivalence relation* on the training set *U*: $IND(A)\subseteq U\times U$. We will say, that if $(x, y) \in IND(A)$ then *x* and *y* are indiscernible by values of attributes from *A.* A set of equivalence classes of relation *IND(A)* is denoted by $\{ X_1^A\ X_2^A\ , ..., X_m^A\}$. Then we can approximately define set *X* using attribute values by the constructing of the lower and upper approximations of *X*, designated by $\underline{A}X$ and $\overline{A}X$ respectively. As a *lower approximation* of set *X* we will understand the union of equivalence classes of an indiscernibility relation which belongs to *X*, i.e. $\underline{A}X = \bigcup \{X_i^A \mid X_i^A \subseteq X\}$. And as an *upper approximation* of set *X* we will understand the union of equivalence classes which part of objects belongs to *X*, i.e. $\overline{A}X = \bigcup \{X_i^A \mid X_i^A \cap X \neq \varnothing\}$. The set $U\setminus \overline{A}X$ will consist of *negative objects* for *X*. A set $POS_A(d) = \underline{A}C_1\cup\ldots\cup\underline{A}C_{r(d)}$ includes objects, which are guaranteed concerning to one of the decision classes, and this set is called *positive region* of the decision system S.

Rough set X is formed by pair $< \underline{A}X, \overline{A}X >$. If upper and lower approximations of *X* are equal then *X* is an ordinary set.

The equivalence relation can be associated not only with the full set of conditional attributes *A* but also with any attribute subset $B\subseteq A$. Further this relation is denoted as $IND(B)$ and is called a *B-indiscernibility relation*. Formally the *B*-indiscernibility relation is defined as follows: $IND(B) = \{(x, y) \in U\times U\colon \forall a\in B\ (a(x) = a(y))\}$.

Thus two objects belong to same equivalence class, if they cannot be discerned by the given subset of attributes. The concepts of *B*-upper and *B*-lower approximations based on $IND(B)$ are similarly introduced.

Since it is not always possible to find a single-valued decision for all objects of a decision system, we will introduce notion of a generalized decision. We will define function $\partial_B\colon U\to \mathrm{P}(V_d)$ which is called a *generalized decision* of S on a set of attributes $B\subseteq A$, as follows: $\partial_B(x) = \{v \in V_d : \exists x' \in U\ (x' IND(B) x\ \wedge\ d(x') = v)\}$. The generalized decision ∂_A of a system S is simply called the generalized decision of S. Instead of ∂_A we also will write ∂_{S}. The decision table S is *consistent*, if $|\partial_A(x)|=1\ \forall x\in U$, otherwise S is *inconsistent*.

Since not all conditional attributes are equally important, some of them can be excluded from a decision table without loss of the information contained in the table. The minimal subset of attributes $B\subseteq A$ which allows to keep the generalized decision for all objects of a training set, i.e. $\partial_B(x) = \partial_A(x)\ \forall x \in U$, is called a *decision-relative reduction* of a table $S = (U, A \cup \{d\})$. In the sequel, when considering decision tables, instead of a decision-relative reduction we will use a *reduction*.

Now let us consider the methods for concept generalization.

4. Methods of the Rough Set Theory

Generally a work of the algorithm based on a rough set theory consist of the following steps: search of equivalence classes of the indiscernibility relation, search of upper and lower approximations, search of a reduction of the decision system and constructing a set of decision rules. Moreover discretization is applied to processing attributes with a continuous domain. In the case of the incomplete or inconsistent input information the algorithm builds two systems of decision rules, one of them gives the certain classification, the second gives the possible one. Further we will consider the most labour-consuming steps: search of reduction and discretization making.

4.1. The Problem of Search of Reduction

Let's consider the process of search of a reduction that is very important part of any method used the rough set approach. Quite often an information system has more than one reduction. Each of these reductions can be used in procedure of decision-making instead of a full set of attributes of original system without a change of dependence of the decision on conditions that is characteristic for original system. Therefore the problem of a choice of the best reduction is reasonable. The answer depends on an optimality criterion related to attributes. If it is possible to associate with attributes the cost function which expresses complexity of receiving attribute values then the choice will be based on criterion of the minimal total cost. Otherwise as a rule the shortest reduction is chosen. However the complexity of a search of such reduction consists in that the problem for checking whether exist a reduction, which length is less than some integer *s*, is NP-complete. The problem of searching for a reduction with minimal length is NP-hard [10].

Thus the problem of a choice of relevant attributes is one of the important problems of machine learning. There are several approaches based on rough set theory to its decision.

One of the first ideas was to consider as the relevant attributes those attributes which contain in intersection of all reductions of an information system.

Other approach is related to dynamic reductions [2], i.e. conditional attribute sets appearing "sufficiently often" as reductions of sub-samples of an original decision system. The attributes belonging to the "most" of dynamic reductions are considered as relevant. The value thresholds for "sufficiently often" and "most" should be chosen for a given data.

The third approach is based on introduction of the notion of significance of attributes that allows by real values from the closed interval [0,1] to express how important an attribute in a decision table.

4.2. Discretization Making

The stage of discretization is necessary for the most of modern algorithms for generalization. The discretization is called a transformation of continuous domain of attributes in a discrete one. For example, the body temperature of the human being which is usually measured by real numbers can be divided into some intervals, corresponding to the low, normal, high and very high temperature. The choice of suitable intervals and partition of continuous domains of attributes is a problem, whose complexity grows in exponential dependence on the number of attributes to which discretization should be applied.

Let's give formal definition of a considered discretization task. Let $S = (U, A \cup \{d\})$ is a consistent decision system. We will assume that the domain of any attribute $a \in A$ is a real interval, i.e. $V_a = [l_a, r_a) \subset R$. Any pair of the form $p^a = (a, c)$ where $a \in A$ and $c \in R$, we will call *cut* on areas V_a. For each attribute $a \in A$ a set $P_a = \{\ [c_0^a, c_1^a), [c_1^a, c_2^a), ..., [c_{s_a}^a, c_{s_a+1}^a)\ \}$ where s_a is some integer, $l_a = c_0^a < c_1^a < ... < c_{s_a}^a < c_{s_a+1}^a = r_a$ and $V_a = [c_0^a, c_1^a) \cup [c_1^a, c_2^a) \cup ... \cup [c_{s_a}^a, c_{s_a+1}^a)$, we will call *partition* of a domain V_a. It is easy to notice, that the partition P_a is uniquely defined by C_a = { c_1^a c_2^a , …, $c_{s_a}^a$ }, which is called *set of cuts* of V_a. Therefore in the sequel we often will name P_a by a set of cuts and write down as P_a = $\{a\}\times C_a = \{(a,c_1^a),(a,c_2^a),...,(a,c_{s_a}^a)\}$. Then full set of cuts P can be presented as $P = \bigcup_{a\in A}\{a\}\times C_a$.

Any set of cuts P on the basis of an original decision system $\mathrm{S}=(U,A\cup\{d\})$ determines a new decision system $\mathrm{S}^P=(U,A^P\cup\{d\})$, where A^P={a^P: $a\in A$} and $a^P(x)=i \Leftrightarrow a(x)\in [c_i^a,c_{i+1}^a)$ for any object $x\in U$ and $i\in$\{0, …, s_a\}. A decision table S^P is called *P-discretization* of the table S . Our purpose is that during discretization to construct such set of cuts P.

It is obvious, that is possible to construct many of sets of cuts. Therefore there is a question how among them to find set with the minimal number of elements. For this purpose we will introduce the following concepts.

Two sets of cuts P and P' we will regard as equivalent, if $\mathrm{S}^P = \mathrm{S}^{P'}$. We will say that set of cuts P is consistent with S , if generalized decisions of systems S and S^P are equal, i.e. $\partial_{\mathrm{S}}(x)=\partial_{\mathrm{S}^P}(x)\ \ \forall x\in U$. The consistent set of cuts P^{irr} is *irreducible* in S if any its own subset is not consistent with S . Finally, the consistent set of cuts P^{opt} we will call *optimal* in S if it has the minimal cardinality among sets of cuts which are consistent with S .

The problem of finding optimal set of cuts P for the given decision system S is NP-complete [11]. This fact clearly speaks about importance of development of effective heuristic algorithms for search of suboptimal set of cuts.

The general approach of the most of discretization algorithms is based that any irreducible set of cuts of a decision table S is a reduction of other decision table $\mathrm{S}^*=(U^*,A^*\cup\{d^*\})$ constructed on a basis of S as follows [11].

Let $\mathrm{S}=(U,A\cup\{d\})$ be an original decision table. An arbitrary attribute $a\in A$ defines sequence $v_1^a<v_2^a<...<v_{n_a}^a$, where $\{v_1^a,v_2^a,...,v_{n_a}^a\}$ = $\{a(x): x\in U\}$ and $n_a\le n$. Objects of new decision table S^* are all pairs of objects of S with different decisions, and the set of conditional attributes is defined as cuts of attribute domains of an original decision table, i.e. A^* = $\bigcup_{a\in A}\left\{p_i^a : p_i^a=(a,c_i^a),\ \text{where}\ c_i^a=(v_i^a+v_{i+1}^a)/2,\ 1\le i\le n_a-1\right\}$.

These attributes are binary. Set A^* is named an *initial set of cuts*. We will speak, that the cut $p_i^a=(a,c_i^a)$ *discerns* objects x and y of different decision classes, if *min*(a(x), a(y)) < c_i^a < *max*(a(x), a(y)). A value of the new attribute corresponding to a cut p_i^a for pair (x, y) is equal to 1 if objects x and y are discerned by this cut, and 0 otherwise. Moreover a new object $\perp$ for which all conditions and the decision d^* are 0 is added to the objects of a new decision table. For all other objects of a new decision table the new decision value is equal to 1. Reductions of a new decision table S^* determine all irreducible sets of cuts of an original decision table S .

On the basis of this general layout the heuristic algorithms finding a suboptimal set of cuts are developed. Often the discretization algorithm based on straightforward implementation of Jonson's strategy [8,12] is used. Computational complexity of this algorithm is equal to $O(|P|\cdot kn^3)$. It does its inapplicable for processing large databases. Thus, the main problem of discretization stage of continuous attributes is its high computational complexity. Now we propose the effective modification that solves this problem.

4.3. The Modification of the Discretization Algorithm

Our algorithm is directed towards the decreasing of time and memory consumption. It is based on the Jonson's strategy and extension of idea of iterative calculation of number of pairs of objects, discerned by a cut. This idea has been offered in [4], however, originally, it is applicable only when some restrictions on the decision table are imposed. This idea is based on assumption that there is a close relation between two consecutive cuts. So, for example, it is possible to notice, that in each row of the table S^* all the cells with value 1 are placed successively within one attribute. Therefore some pairs of objects are discerned by both consecutive cuts, and changes in the number of discernible pairs of objects can be only due to objects which attribute values lay between two these cuts. In [4], the situation, when no more than one object lies in this interval, is considered. We generalize this idea on a case of the arbitrary number of such objects. Thus, our algorithm extends idea of iterative calculating number of pairs of objects discerned by a cut to an arbitrary decision table.

For some cut $p_t^a = (a, c_t^a) \in A^*$ for the attribute *a* where $a \in A$ and $1 \le t \le n_a$, and some subset $X \subseteq U$ we introduce the following notation: $W^X(p_t^a)$ is a number of pairs of objects from *X* discerned by a cut p_t^a; $l^X(p_t^a)$ and $r^X(p_t^a)$ is the number of objects from *X*, which have a value of the attribute *a* less (more) than c_t^a; $l_q^X(p_t^a)$ and $r_q^X(p_t^a)$ is the number of objects from *X*, which have a value of the attribute *a* less (more) than c_t^a and belong to the *q-th* decision class, where *q* = 1, …, *r*(*d*); $N^X(p_t^a, p_{t+1}^a)$ is the number of objects from *X*, values of the attribute *a* which lay in an interval (c_t^a, c_{t+1}^a); $N_q^X(p_t^a, p_{t+1}^a)$ is the number of objects from *X*, values of attribute *a* which lay in an interval (c_t^a, c_{t+1}^a) and belonging to the *q-th* decision class, where *q* = 1, …, *r*(*d*).

Now we formulate two our theorems which underlie proposed discretization algorithm. The first theorem will allow us to derive value $W^X(p_{t+1}^a)$ from $W^X(p_t^a)$, where p_t^a and p_{t+1}^a are two consecutive cuts of a domain of the attribute *a*.

Theorem 1. Let set $X \subseteq U$ consists of $N^X(p_t^a, p_{t+1}^a)$ objects which values of the attribute *a* belongs to an interval (c_t^a, c_{t+1}^a). Then

(*a*) $l_q^X(p_{t+1}^a) = l_q^X(p_t^a) + N_q^X(p_t^a, p_{t+1}^a) \ \forall q = 1, \dots, r(d)$;

(*b*) $r_q^X(p_{t+1}^a) = r_q^X(p_t^a) - N_q^X(p_t^a, p_{t+1}^a) \ \forall q = 1, \dots, r(d)$;

(*c*) $$W^X(p_{t+1}^a) = W^X(p_t^a) + N^X(p_t^a, p_{t+1}^a) \cdot \left(r^X(p_t^a) - l^X(p_t^a)\right) -$$
$$-\sum_{i=1}^{r(d)} N_i^X(p_t^a, p_{t+1}^a) \cdot \left(r_i^X(p_t^a) - l_i^X(p_t^a)\right) + \sum_{i=1}^{r(d)} \left(N_i^X(p_t^a, p_{t+1}^a)\right)^2 - \left(N^X(p_t^a, p_{t+1}^a)\right)^2 .$$

Let's consider a case when during discretization we have a set of cuts $P \subseteq A^*$ that defines equivalence classes X_1, X_2, …, X_m of the indiscernibility relation *IND*(A^P) of table S^P, and also two consecutive cuts p_t^a and p_{t+1}^a of the attribute *a*. Then we can calculate value $W_P(p_{t+1}^a)$ from $W_P(p_t^a)$ as follows:

Theorem 2. Let there are *K* equivalence classes $X_{\alpha_1}, X_{\alpha_2}, \dots, X_{\alpha_K}$ to each of which belongs $N^{X_{\alpha_i}}(p_t^a, p_{t+1}^a) \ge 1$ objects which values of attribute *a* are within an interval (c_t^a, c_{t+1}^a). Then

$$W_P(p_{t+1}^a) = W_P(p_t^a) + \sum_{i=1}^{K} \Bigg[N^{X_{\alpha_i}}(p_t^a, p_{t+1}^a) \cdot \left(r^{X_{\alpha_i}}(p_t^a) - l^{X_{\alpha_i}}(p_t^a)\right) - \sum_{q=1}^{r(d)} N_q^{X_{\alpha_i}}(p_t^a, p_{t+1}^a) \cdot \left(r_q^{X_{\alpha_i}}(p_t^a) - l_q^{X_{\alpha_i}}(p_t^a)\right) +$$
$$\sum_{q=1}^{r(d)} \left(N_q^{X_{\alpha_i}}(p_t^a, p_{t+1}^a)\right)^2 - \left(N^{X_{\alpha_i}}(p_t^a, p_{t+1}^a)\right)^2 \Bigg].$$

Now we present steps of our algorithm. We will name its *GID* (Generalized Iterative algorithm for Discretization).

Algorithm 1. *Algorithm GID*.

Input: The consistent decision table S.

Output: Suboptimal set of cuts *P*.

Used data structures: *P* is a suboptimal set of cuts, *L* = [*IND*(A^P)] – the set of equivalence classes of an indiscernibility relation of the table S^P; *A** - a set of possible cuts.

```
1. $P := \varnothing$; $L$ := {$U$}; $A^*$: = initial set of cuts;
2. For each attribute $a \in A$ do:
       begin
         $W_P(p_0^a) := 0$;
         For each $X_i \in L$ do:
```

```
        $r^{X_i} := |X_i|$; $l^{X_i} := 0$;
      for $q$ = 1,..., $r(d)$ assign $r_q^{X_i} := \left|\{x \in X_i : d(x) = v_q^d\}\right|$; $l_q^{X_i} := 0$;
      For each cut $p_j^a = (a, c_j^a) \in A^*$ do:
        For all classes $X_{\alpha_i}$ which objects have a value of attribute $a$ from an interval $(c_{j-1}^a, c_j^a)$ to
        calculate $N^{X_{\alpha_i}}$ and $N_q^{X_{\alpha_i}}$.
      Find $W_P(p_j^a)$ according to the theorem 2.
      Count values $r^{X_{\alpha_i}}, l^{X_{\alpha_i}}$ and $r_q^{X_{\alpha_i}}, l_q^{X_{\alpha_i}}$ under the theorem 1.
    end;
```

3. Assume as $p_{\max}$ the cut with maximal value $W_P(p)$ among all cuts p from A^*.
4. Assign $P := P \cup \{p_{\max}\}$; $A^* := A^* \setminus \{p_{\max}\}$;
5. For all $X \in L$ do: if $p_{\max}$ divides the set X into X_1 и X_2 then remove X from L and add to L two sets X_1 and X_2.
6. If all sets from L consist of the objects belonging to same decision class then Step 7 otherwise go to the Step 2.
7. End.

Let's estimate computational complexity of offered algorithm. The most labour-consuming steps of algorithm are the second and the fifth.

On step 2, during calculation of number of pairs of objects discerned by a cut $p_j^a = (a, c_j^a)$ values $r^{X_{\alpha_i}}$, $l^{X_{\alpha_i}}$, $N^{X_{\alpha_i}}$ and $r_q^{X_{\alpha_i}}$, $l_q^{X_{\alpha_i}}$, $N_q^{X_{\alpha_i}}$ are changed, where q = 1, ..., $r(d)$. These operations are carried out only for those equivalence classes X_{α_i}, even which one object satisfies to the condition of belonging of value of attribute a to interval (c_{j-1}^a, c_j^a). For one such equivalence class it will be executed $3 \cdot r(d)$+3 described operations. We will designate this number as α. It does not depend on the number of objects n and the number of attributes k. The number of such equivalence classes cannot exceed the number n_j of objects which belong to them and which value of attribute a are in interval (c_{j-1}^a, c_j^a). Hence, during calculation $W_P(p_j^a)$ for one cut p_j^a it is carried out no more than $\alpha \cdot n_j$ operations. Therefore during processing all cuts of one attribute it will be executed $\sum_{j=1}^{n-1} \alpha \cdot n_j \le \alpha \cdot n$ operations. For processing the cuts of all k attributes it is required $\alpha \cdot kn$ operations. The second step repeats $|P|$ times. It means, that its total computational complexity is equal to $O(|P| \cdot kn)$.

On step 5 splitting equivalence classes is carried out. We take the worse case when finally any class consists of exactly one object. Since there are n objects then during work of the algorithm it will be executed n-1 splitting operations. Hence, computational complexity of the fifth step is $O(n)$.

Thus total computational complexity of the proposed discretization algorithm is equal to $O(|P| \cdot kn) + O(n) = O(|P| \cdot kn)$. It is less on two orders than computational complexity of Jonson's algorithm.

Also we estimate the space complexity of our algorithm. It should be noticed that it does not build the auxiliary table S^*. It is required only $k(n$-1) memory cells for a storing set of possible cuts from A^*, n cells for designating an equivalence class to which belongs each of the objects, and no more than $\alpha \cdot n$ cells for storing numbers r^{X_i}, l^{X_i}, N^{X_i} and $r_q^{X_i}$, $l_q^{X_i}$, $N_q^{X_i}$ for all equivalence classes $X_i \in L$ where $i \le n$ and q = 1,..., $r(d)$ and the value α does not depend on k and n. Hence the space complexity of our discretization algorithm is equal to $O(kn)$. It is less on the order than space complexity of Jonson's algorithm. For more details about our algorithm see [5, 6].

4.4. The Modification of Algorithm for Searching the Significant Attributes

In the majority of the algorithms which are based on the rough set theory and carrying out splitting of continuous attribute domains into finite number of intervals, the stage of discretization is considered as preparatory before search of significant attributes. And consequently at a stage of discretization there is a splitting of the domains of all continuous attributes, including insignificant. In this work the combined implementation of discretization with the search of a reduction is offered to make discretization only for those quantitative attributes which appear to be significant during search of a reduction.

Besides, as significant attributes we will consider the attributes which are included in approximate reductions with sufficiently high quality. The concept of an approximate reduction [8] represents generalization of concept of the reduction considered earlier. Any subset *B* of set *A* can be considered as an *approximate reduction* of set *A*, and value

$$\varepsilon_{(A,d)}(B)=\frac{dep(A,d)-dep(B,d)}{dep(B,d)}=1-\frac{dep(B,d)}{dep(A,d)}$$

is named a *reduction approximation error*. Here the value *dep*(*B*, *d*) represents a measure of dependence between $B \subseteq A$ and *d*: *dep*(*B*, *d*) = |$POS_B(d)$|/|*U*|. The reduction approximation error shows how precisely the set of attributes *B* approximates whole set of conditional attributes *A* (relatively *d*). Application of approximate reductions is useful while processing inconsistent and noisy data.

Thus, the developed algorithm for search of significant attributes is based on two ideas: 1) combination of discretization of quantitative attributes with the search of significant attributes, 2) search for an approximation of a reduction, but no for reduction itself. Let's name it as Generalized Iterative algorithm based on the Rough Set approach, GIRS.

4.5. Results of the Experiments

The realized experiments show that the developed algorithm allows reducing time for search of significant attributes essentially, due to combination with discretization stage and use of proposed algorithm GID.

The results of the experiments executed on 11 data sets from a well known collection UCI Machine Learning Repository [7] of the University of California are given in Table 1.

Data set	Classification accuracy				
	ID3	C4.5	MD	Holte-II	GIRS
Monk-1	81.25	75.70	100	100	100
Monk-2	65.00	69.91	99.70	81.9	83.10
Monk-3	90.28	97.20	93.51	97.2	95.40
Heart	77.78	77.04	77.04	77.2	78.72
Hepatitis	n/a	80.80	n/a	82.7	84.51
Diabetes	66.23	70.84	71.09	n/a	81.00
Australian	78.26	85.36	83.69	82.5	88.71
Glass	62.79	65.89	66.41	37.5	70.10
Iris	94.67	96.67	95.33	94.0	96.24
Mushroom	100	100	100	100	100
Soybean	100	95.56	100	100	100
Average	81.63	83.18	88.67	85.3	88.89

Table 1. Comparison of classification accuracy of the developed algorithm with other known generalization algorithms.

For all data sets taken into the comparison, the developed algorithm has shown classification accuracy that not concedes to other generalization algorithms, and in some cases surpasses it. Average accuracy of classification is approximately 88.9 %. It is necessary to note, that the classification accuracy received by our algorithm is much above that the classification accuracy achieved by methods of an induction of deciding trees (ID3, ID4, ID5R, C4.5) at the solving the majority of the problems. It is explained by the impossibility of representation of the description of some target concepts as a tree. Moreover it is possible to note that combining of search of significant attributes and discretization procedure is very useful. Most clearly it is visible from the results received at the decision of the Australian credit task. It is possible to explain by the presence in these data the attributes both with continuous and with discrete domains. The modification of search procedure of significant attributes is directed namely to processing of such combination.

Conclusion

We have considered the concept generalization problem and the approach to its decision based on the rough set theory. The means provided by this approach have been shown. They allow solving the problem of processing of real-world data arrays. The heuristic discretization algorithm directed towards the decreasing of time and memory consumption has been proposed. It is based on Jonson's strategy and extension of idea of iterative calculating number of pairs of objects discerned by a cut. Computational and space complexities of the proposed algorithm have linear dependence on the number of objects of decision table. Also the search algorithm of the significant attributes combined with a stage of discretization is developed. It allows avoiding splitting into intervals of continuous domains of insignificant attributes.

Bibliography

[1] Pawlak Z. Rough sets and intelligent data analysis / Information Sciences, Elsevier Science, November 2002, vol. 147, iss. 1, pp. 1-12.

[2] Bazan J. A comparison of dynamic non-dynamic rough set methods for extracting laws from decision tables / Rough Sets in Knowledge Discovery 1: Methodology and Applications // Polkowski L., Skowron A. (Eds.), Physica-Verlag, 1998.

[3] Vagin V.N., Golovina E.U., Zagoryanskaya A.A., Fomina M.V. Dostovernyi i pravdopodobnyi vyvod v intellektual'nyh sistemah (Reliable and plausible inference in intellectual systems) / Pod red. V.N. Vagina, D.A. Pospelova. Moscow, Fizmatlit, 2004. – 704 p.

[4] Nguyen S.H., Nguyen H.S. Some efficient algorithms for rough set methods / Proc. of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems, Spain, 1996, pp. 1451-1456.

[5] Kulikov A., Fomina M. The Development of Concept Generalization Algorithm Using Rough Set Approach / Knowledge-Based Software Engineering: Proceedings of the Sixth Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2004) // V.Stefanuk and K. Kajiri (eds). – IOS Press, 2004. – pp.261–268.

[6] Vagin V.N., Kulikov A.V., Fomina M.V. Methods of Rough Sets Theory in Solving Problem of Concept Generalization / Journal of Computer and System Sciences International, Vol. 43, No. 6, 2004. – pp. 878 - 891.

[7] Merz C.J., Murphy P.M. UCI Repository of Machine Learning Datasets. – Information and Computer Science University of California, Irvine, CA, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[8] Komorowski J., Pawlak Z., Polkowski L., Skowron A. Rough Sets: A Tutorial. / Rough Fuzzy Hybridization, Springer-Verlag, 1999.

[9] Pawlak Z. Rough Sets / International Journal of Information and Computer Science. 1982, 11(5), pp. 341-356.

[10] Skowron A., Rauszer C. The Discernibility Matrices and Functions in Information Systems / Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory, Kluwer, 1992, pp. 331-362.

[11] Nguyen H.S., Skowron A. Quantization of real value attributes / Second Annual Joint Conference on Information Sciences (JCIS'95) // Wang P.P. (ed.), North Carolina, USA, 1995, pp. 34-37.

[12] H.S. Nguyen, S.H. Nguyen. Discretization methods in data mining / Rough Sets in Knowledge Discovery 1: Methodology and Applications // Polkowski L. and Skowron A. (Eds.), Heidelberg, Physica-Verlag, 1998. pp. 451-482.

Authors' Information

Marina Fomina – e-mail: fominhome@mtu-net.ru

Alexey Kulikov –e-mail: kulikov@apmsun.mpei.ac.ru

Vadim Vagin – e-mail: vagin@apmsun.mpei.ac.ru

Moscow Power Engineering Institute, Krasnokazarmennaya str, 14, Moscow, 111250, Russia;