Rough set theory in discretization method based on genetic algorithm

Lei Huang

School of Beijing University of technology, Beijing 100124, China

hl_lei@foxmail.com

Keywords: rough set, genetic algorithm, discretization.

Abstract. The rough set theory is a theory method. It is about incomplete research, uncertain expression of knowledge and data, knowledge and data discovery. It has been widely used in artificial intelligence, knowledge and data discovery, pattern recognition and classification, analysis and reasoning of imprecise data and find the potential knowledge, data mining. But the traditional rough set theory can only deal with discrete attributes in the database. While the vast majority of real database contains both discrete attributes and continuous attributes. Therefore, we must turn these continuous attributes into discrete. Aiming at this problem, this paper implements a discretization method based on genetic algorithm. It is using genetic algorithm to obtain the target, while keeping the original decision system under the condition of the indiscernibility relation and minimum break point set as the optimization target. Experimental results show that the discrete method using genetic algorithm is feasible.

1. Introduction

In 1982, the Polish scholar Z.Pawlak proposed rough set (Rough Set) theory. It is a mathematical method to deal with fuzzy and uncertain information theory, we can analyze the data and reasoning, discover the implicit knowledge, reveal the potential regularity. But the traditional rough set theory can only be to deal with discrete attributes in the database, and the vast majority of real database contains both discrete attributes and continuous attributes.

Data discretization essence is searching the breakpoint set, which could be divided into several interval attribute space of the information system, and makes the same area all take the same instance of the attribute value vector. Solving the optimal breakpoint of continuous attributes is an NP (Non - Deterministic Polynomial) problem, people have proposed many algorithms. Common practice is granularity and frequency division algorithm, though they are easy to implement, but they need dimension or divided artificially set parameters, and the results unsatisfactory.

For this purpose, this paper implements a discretization method based on genetic algorithm. This method is to maintain the original decision system under the condition of the indiscernibility relation, through genetic algorithm to obtain the optimal cut sets.

2. The discretization of the rough set

A decision table S = <U, R, V>, U = {x_1, x2, …, x_n} is a collection of objects. A = {a1, a2, …, an} is the condition attribute set. d is a decision attribute. For any a ∈A, there is a map $U \to V_a$, V_a is on the range. And assuming $V_a = [l_a, r_a] \subset R$. R is a set of real numbers. A breakpoint range V_a can be denoted as (a,c), among them $a \in A$, $c \in R$, $c \in R$, $V_a = [l_a, r_a]$. A breakpoint on an arbitrary set B. For an arbitrary set $D_a = \{(a, c_1^a), (a, c_2^a), ..., (a, c_n^a)\}$ on $V_a = [l_a, r_a]$,there is $k_a \in N$ and $l_a = c_0^a < c_1^a < ... < c_k^a < c_{k+1}^a = r_a$,and the $V_a = [c_0^a, c_1^a), \cup [c_1^a, c_2^a) \cup ... \cup [c_k^a, c_{k+1}^a]$ .So arbitrary breakpoints $D = \bigcup_{a=A} D_a$ on condition attribute set just a new decision table $S^D = < U, R, V^D, f^D >$, among them $f^D(x_a) = i \Leftrightarrow f(x_a) \in [\sigma_i^a, c_{i+1}^a]$, $x \in U$, $i \in \{0,1,...,k_a\}$ .Namely after discretization, the

original decision table is replaced by a new decision table, and different breakpoint rally will be the same decision table into a new decision table.

Discrete nature is to use the selected breakpoints to divide the property of spatial conditions. And space divides into a finite number of regions, each region corresponding to a discretization of objects in the decision table. The rationality of the breakpoints can be measured by the following criteria:

1) Consistency. That is, for any object $u, v \in U$, if the conditions can be distinguished by attribute A, then D can also be set breakpoints distinction.

2) Non-Simplification. That does not exist $D' \subset D$ which meet the consistency.

3) Minimum discretion. For any meet the consistency of D' breakpoint set, all have $card(d) \le card(D')$ D is the optimal set breakpoints.

3. Algorithm Design

3.1 The choice of alternate breakpoint

Before the discretization, the first to calculate the original property values set breakpoints, steps are as follows:

(1)i=1.

(2) Select properties a^i.

(3) According to the properties of a^i domain of all the objects in the universe U $(x_1,\ldots,x_n)$ in ascending order, then get $a_i(U) = \{a_i(1), a_i(2), ..., a_i(n)\}$ corresponds to an x_k .

(4) Set a breakpoint number, the following:

(i) $m = 1$;

(ii) $for(k = 1 to)$

(iii){

$$c_m^i = \frac{a_i(k) + a_i(k+1)}{2}$$

$$m = m + 1$$

}

(5) $i = i + 1$, if $i \le n$, then execute (2), else finish.

3.2 Alternate breakpoint coding.

This article directly to the breakpoint code selection state, rather than coding of breakpoint value itself. Method is as follows: the decision tables all breakpoints are encoded into a chromosome, which use "1" to select the breakpoint says, "0" said don't select the breakpoint, the length of the chromosome for the breakpoint number after. Assuming the decision table have four attributes {a1, a2, a3,}, the alternate breakpoints, respectively{{1.2,1.5,1.6},{0.3,0.5,0.9},{4.2,3.5,5.2,2.8}}, To change the decision table may be encoded as a breakpoint set 101011011110. The chromosome length is the sum of all the attributes to alternate breakpoint number, in this case, the length of 12. Its corresponding location "1" means to select the breakpoint, "0" said don't select the breakpoint. According to the chromosome coding breakpoints for decision table {{1.2,1.6},{0.5,0.9},{4.3,5.2}}.

3.3 Alternate breakpoint coding.

Discretization of continuous attributes decision table value is the premise of not change the decision table of indiscernibility relation (after discretization of the upper and lower approximation set with equal discretization before), after the discretization of the decision table with the original with the same indiscernibility relation, only in this discretization, under the premise of can ensure after discretization of decision table is equivalent to the original decision table. Therefore, the discretization problem can be regarded as optimization problem with constraints. The optimization goal is the minimum cut sets, constraint condition is to keep the indiscernibility relation. Therefore, the individual fitness function can be expressed as (1):

$$f(x) = \exp(-N_{incinsist}) / N_{cuts} \quad (1)$$

Among them $N_{incinsist}$ is the inconsistent decision table after discretization object number. N_{cuts} is that all the total number of remaining breakpoint properties. And molecules are caused to the individual decision table indiscernibility relation change degree of punishment, namely individual penalty term for infeasible solution.

3.4 Indiscernibility

Definitions For a subset of attributes $P \subseteq A$ and $x, y \in U$,if $\forall a \in P$,have $a(x) = a(y)$ namely only according to the attribute sets P provides information ,can apart from A to B, this is we call x, y in P is a indiscernibility attribute subset. Denoted:

$$ind(P) = \{(x, y) \in U \times U \mid \forall a \in P, a(x) = a(y)\} \quad (2)$$

The rough set theory is that knowledge is based on the ability of object classification, much fewer objects can be divided into category, is the indiscernibility relation between them.

4. Experimental test

As shown in table 1 of the decision table, d values for decision. And is a and b the range of values. For the theory of domain.

Table 1 Decision table before test

Numble	U	a	b	d
1	x_1	0.8	2	1
2	x_2	1	0.5	0
3	x_3	0.8	2	1
4	x_4	1	0.5	0
5	x_5	0.8	2	1
6	x_6	1	0.5	0
7	x_7	1.3	3	0

In table 1, assuming that $V_a = [0, 2)$, $V_b = [0, 4)$,the value of a, b to $a(U) = \{0.8,1,1.3,1.4,1.6\}$ and $b(U) = \{0.5,1,2,3\}$.

In the experiment, Maximum genetic algebra for 300 generations. The rest of the parameters use the default Settings.

5. Results analysis

New decision table 2 are obtained by the above method after discretization for breakpoints for: Breakpoint for attribute 2 is {1.15, 1.5}. Breakpoint for attribute 3 is 1.5.

Table 1 the result

Numble	U	a	b	d
1	x_1	1	2	1
2	x_2	1	1	0
3	x_3	2	2	1
4	x_4	2	1	0
5	x_5	2	2	1
6	x_6	3	2	0
7	x_7	2	1	0

The experimental data by above knowable:

(1) While maintaining the original decision system under the condition of the indiscernibility relation, through the above methods, Breakpoint for attribute 2 is {1.15, 1.5}. Breakpoint for attribute 3 is {1.5}.

(2) Using genetic algorithm to get the segmentation points less, resulting from the reduction of decision table attributes is less, Simple rules is obtained, and its coverage instance scope is bigger.

6. Summary

Because of the complexity of the practical problems, the use of any kind of discrete algorithms are very difficult to ensure that the results of all the problems solved when finally obtained is optimal, and the computational efficiency of the algorithm to solve large-scale problems also determine whether the premise, therefore, discretization algorithm new exploration is still very important.

This paper presents a discrete genetic algorithm and rough set theory is proposed. The algorithm minimum break point set as the objective function, and keeps the original information system indistinguishable relationship unchanged constraints using genetic algorithm to find the optimal set breakpoints, when dealing with small-scale data, more advantages.

Of course, this method there is still much room for improvement. For example, due to genetic algorithm is a stochastic optimization method, sometimes only give a suboptimal solution globally, it is difficult to obtain the optimal solution within the global scope, and global optimization process takes a lot of time, high time complexity. How it can be further improved, the focus of future research will do it.

References

[1]. Lark H. X. Discovery in Databases：An Attribute-oriented Rough Set Approach. University of Regina. Science, 1998, 20(3).

[2]. Pawlak Z; Grzymala-Bausse J. Glowinski R Rough Sets. 2005(30).

[3]. Hole, RC. Very Simple Classification Rules Reform Well on Most Commonly Used Datasets. Machine Learning, 1993(11):63~90.

[4]. Kerber, R Chi merge. Discretization of Numeric Attribute. Proc.of Tenth Nation Conference on Artificial Intelligence, 1992(3):123~128.