Information Retrieval Based on Decompositions of Rough Sets Referring to Fuzzy Tolerance Relations

Chen Wu, Jun Dai
[1] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003,China

Abstract
Various expanded rough set models based on tolerance relations enlarge the application fields of rough set theory. Through generating tolerance relations to fuzzy tolerance relations and combining with dominance relations, a tolerance class of a fuzzy tolerance relation is further decomposed into a positive fuzzy tolerance class, a negative fuzzy tolerance class and a purely fuzzy tolerance class. Then the paper gives out relative definitions of lower approximation and upper approximation respectively. Furthermore it proposes the decomposition of rough sets of fuzzy tolerance relations. This method can cope with continuous attributes effectively. Information retrieval, as one of its example applications, is introduced to show its meaningfulness.

Key words: *Information retrieval,rough set, fuzzy set, fuzzy tolerance relation, dominance relation*

1. Introductions

Z.Pawlak put forward rough sets theory, as a new mathematic tool in dealing with uncertainty, vagueness, non-deterministivity in information processing ([1,2]). In practice, people have discovered that indiscernibility relations based on which rough sets work are much rigorous. Some people propose to extend the original rough set models. For example, some put forward rough sets based on tolerance relation by using tolerance relation instead of indiscernibility relation; others suggest rough sets based on dominance relation by using dominance relation instead of indiscernibility relation. All these enlarge application scopes of rough set theory. Literature ([3]) decomposes tolerance class into positive tolerance class, negative tolerance class and pure tolerance class by introducing dominance relations. However, firstly, there is a large amount of fuzzy information, secondly, the capability of rough sets in dealing with the continuous attributes in the information system is limited because continuous attributes are usually processed by discrediting them into binary system in rough sets and it may bring errors by using this method apparently for most of continuous attributes are fuzzy attributes. This article decomposes the tolerance class of a fuzzy tolerance relation into positive fuzzy tolerance class, negative fuzzy tolerance class and pure fuzzy tolerance class by extending tolerance relations to fuzzy tolerance relations and introducing dominance relations. This method can deal with continuous attributes effectively. One example of information retrieval is used to demonstrate its application usefulness and values.

Manuscript received March 5, 2009
Manuscript revised March 20, 2009

2. Basic concepts

A. Basic theories of rough sets

Definition 1. Suppose R is an indiscernibility relation on universe U (i.e. an equivalence relation, with reflexivity, symmetry and transitivity). For any $X \subseteq U$, we can define following subsets of U : Lower approximation: $R_*(X)=\cup \{[x]|[x] \subseteq X\}$;
Upper approximation: $R^*(X)= \cup \{[x]|[x] \cap X \neq \varnothing\}$;Boundary: $Bn_R(X)= R^*(X) - R_*(X)$.Wherein [x] is the equivalence class where element x is in. $<R_*(X), R^*(X)>$ is called the rough sets based on indiscernibility relation R.

B. Basic theories of fuzzy sets

Here the fuzziness mainly means the illegibility during the transitional phases between the differences of objective things; the fuzzy concepts mean no unambiguous extensions. According to the requirement of general set theory, an object belongs to a set or doesn't belong to it, it should be the only one case. Such sets can't deal with idiographic fuzziness. L. A. Zadeh presented fuzzy sets theories, which use precise mathematical language to describe the fuzziness.

Definition 2. A is a fuzzy subset on universe U, for any $u \in U$, $\mu_A(u) \in [0,1]$ is called membership degree for u to A. The following mapping $\mu_A: U \rightarrow [0,1]$ or $u \mapsto \mu_A(u)$ is called A's membership function. The fuzzy subset is only described by membership function. F(U) denotes the class consisted of all fuzzy subsets on U.

Definition 3. If A, B are two fuzzy sets on U, then the union and intersections of A and B and the complimentary set of A are all fuzzy sets. Their membership functions are respectively defined as follows:

$(A \cup B)(x)= \max(A(x), B(x))$
$(A \cap B)(x)= \min(A(x), B(x))$
$A^c(x)=1-A(x)$

The union and intersections of two fuzzy sets may be generalized to any more fuzzy sets.

C. Fuzzification of continuous attributes

Transform every numerical value attribute into corresponding fuzzy set, transform attributes value into membership degree. Membership functions may presented by users or experts in the domain, however, in many cases it's impractical, we may transform the original database into corresponding fuzzy membership matrix directly by using fuzzy C-mean Clustering Algorithm, wherein C denotes the amounts of fuzzy sets divided by numerical attributes.

We can also use triangle membership function , π membership function, etc.. Suppose a is a continuous attribute, the domain of a may expressed as $V=\{V_a(u):a \in A,\ u \in U\}$. According to the size of the domain and the distribution of the values of attributes, a is fuzzified as k semantic variables Y_i(i=1,2, $\cdots$,k). Every semantic variable Y_i has a membership function, to ensure the integrity of the distribution, the values of the intersection between two neighbor membership functions are 0.5, in addition, k centers of fuzzy distribution m_i may decided by using Kohonen self-organization network map algorithm.

3. Decomposition of rough sets based on tolerance relation

In practical applications, some attributes, especially described by numerical value, often bring uncertainty and fluctuation because of no specific definition to the values of attributes. And now using indiscernibility relation of traditional rough sets to describe them "strictly" or "accurately" is obviously unreasonable. The article presents the rough set models which use tolerance relation instead of indiscernibility relation.

Definition 4. Let SIM be the tolerance relation on U (is reflexive, symmetric) and $X \subseteq U$. The lower and upper approximation sets of X are defined as follows:

$SIM_*(X)= \cup \{[x]_s | [x]_s \subseteq X\}$

$SIM^*(X)= \cup \{[x]_s | [x]_s \cap X \neq \varnothing\}$

where $[x]_s =\{y|\ y \in U$ and $(x, y) \in SIM\}$ is the tolerance class to which x is belong to.

In the following, dominance relation is introduced to further decompose tolerance relation. Dominance relation between x and y, denoted by xD_py, means that the value of every attribute in attribute set P of x is at least as well as those of y. where x, $y \in U$, $D_p^+(x)=\{y \in U: yD_px\}$, $D_p^-(x)=\{y \in U: xD_py\}$. The definition of dominance and theorems may consult literature ([5]). Decomposes the tolerance class of a tolerance relation into positive tolerance class, negative tolerance class and pure tolerance class by introduce dominance relations. And then define lower approximation, upper approximation and boundary, may consult literature([4]).

4. Rough sets based on fuzzy tolerance relations and their decompositions

Definition 5. $\forall$ x,y $\in$ U , $\forall$ q $\in$ Q, define fuzzy relation R : U × U $\rightarrow$ [0,1] as $xRy=\{(x,y) \in U \times U \geq \alpha \mid \mu_R(x,y) \geq \alpha\}$ in the formula, $\mu_R(x,y) \geq \alpha <=> | \mu_q(x)- \mu_q(y) | \leq 1-\alpha$, for $\forall$ q $\in$ Q; α is a fixed constant, called threshold. We use the simple method comparing distances of membership degree to solve the measurement of fuzzy tolerance relations. You may also use other methods. It is easily proved that fuzzy tolerance relations are reflexive and symmetric, but not transitive. So R is a fuzzy tolerance relation. The calculation about fuzzy tolerance degree may refer to literature ([6]).

$[x]_{FS}$ is the class based on fuzzy tolerance relation R , where *x* is in, and $x \in U$, i.e.

$[x]_{FS} =\{y \in U \mid q \in Q, \mu_R(x,y) \geq \alpha\}$

Suppose $X \subseteq U$, the lower approximation and upper approximation based on fuzzy tolerance relation approximate universe (X, R) are respectively described as follows:

$FS_*(X)=\{ [x]_{FS} \mid [x]_{FS} \subseteq X\}$

$FS^*(X)=\{ [x]_{FS} \mid [x]_{FS} \cap X \neq \varnothing\}$

The bounder of X is:

$Bn_{FS}(X)= FS^*(X) - FS_*(X)$

The related properties of rough sets based on fuzzy tolerance relation may consult some literatures.

Now we introduce dominance relation to rough sets based on fuzzy tolerance relations to further decompose a class into positive fuzzy tolerance class, negative fuzzy tolerance class and pure fuzzy tolerance class.

In the fuzzy tolerance class of *x*, those , decided by *x* ,whose membership degrees of all attributes are lower than those of *x* are negative fuzzy similar with x, denoted by $[x]_{FS}^-$; those, deciding x ,whose membership degrees of all attributes are high than those of x, are positive fuzzy similar with x, denoted by $[x]_{FS}^+$; those whose membership degrees of some attributes are high than those of x and some lower, are neither negative fuzzy similar with x nor positive fuzzy similar with x and are called pure similar with x, denoted by $[x]_{FS}^0$. That is

$[x]_{FS}^-=[x]_{FS} \cap Dp^-(x)$

$[x]_{FS}^+=[x]_{FS} \cap Dp^+(x)$

$[x]_{FS}^0=[x]_{FS}-[x]_{FS}^- -[x]_{FS}^+ \cup \{x\}$

Then we can define upper and lower approximations respectively according to them as follows:.

$FS_*^+(X)=\{ [x]_{FS}^+ \mid [x]_{FS}^+ \subseteq X\}$

$FS^{*+}(X)=\{ [x]_{FS}^+ \mid [x]_{FS}^+ \cap X \neq \varnothing\}$

$FS_*^-(X)=\{ [x]_{FS}^- \mid [x]_{FS}^- \subseteq X\}$

$FS^{*-}(X)=\{ [x]_{FS}^- \mid [x]_{FS}^- \cap X \neq \varnothing\}$

$FS_*^0(X)=\{ [x]_{FS}^0 \mid [x]_{FS}^0 \subseteq X\}$

$FS^{*0}(X)=\{ [x]_{FS}^0 \mid [x]_{FS}^0 \cap X \neq \varnothing \}$

5. A example of information retrieval

An information system is given in Table 1.By using fuzzy C-mean Clustering Algorithm to fuzzily process attributes A,B,C,D of Table 1, Table 1 is converted into Table 2.

Tab.1：Continuous information system

No.	A	B	C	D
1	0.21	6.98	0.40	16.37
2	0.27	8.00	0.23	13.81
3	0.26	7.55	0.28	14.97
4	0.20	6.35	0.15	12.32
5	0.38	7.45	0.05	15.90
6	0.25	6.25	0.03	15.95
7	0.37	9.02	0.07	15.61
8	0.17	4.45	0.20	14.60
9	0.18	6.73	0.17	10.93
10	0.74	10.13	0.83	14.61

Let α =0.3,we can obtain some fuzzy tolerance class as follows:{1,3}, {2,3}, {3,1,2}, {4,9}, {5,7}, {6,8}, {7,5,10}, {8,6}, {9,4}, {10,7}. Then we decompose each [x] into negative, positive and pure classes:

Tab.2：Fuzzy information system

No.	A	B	C	D
1	1	1	1	1
2	0.8	0.6	0.5	0.7
3	0.9	0.8	0.8	1
4	1	1	0.1	0
5	0	0.5	0	1
6	0.9	1	0.1	1
7	0	0.3	0	1
8	0.9	0.9	0.3	0.9
9	1	1	0.1	0
10	0	0	0	0.9

$[1]_{FS}^-=\{1,3\}$, $[1]_{FS}^+=\{1\}$, $[1]_{FS}^0=\{1\}$;
$[2]_{FS}^-=\{2\}$, $[2]_{FS}^+=\{2,3\}$, $[1]_{FS}^0=\{2\}$;
$[3]_{FS}^-=\{2,3\}$, $[3]_{FS}^+=\{1,3\}$, $[3]_{FS}^0=\{3\}$;
$[4]_{FS}^-=\{4,9\}$, $[4]_{FS}^+=\{4,9\}$, $[4]_{FS}^0=\{4,9\}$;
$[5]_{FS}^-=\{5,7\}$, $[5]_{FS}^+=\{5\}$, $[5]_{FS}^0=\{5\}$;
$[6]_{FS}^-=\{6\}$, $[6]_{FS}^+=\{6\}$, $[6]_{FS}^0=\{6,8\}$;
$[7]_{FS}^-=\{7,10\}$, $[7]_{FS}^+=\{7,5\}$, $[7]_{FS}^0=\{7\}$;
$[8]_{FS}^-=\{8\}$, $[8]_{FS}^+=\{8\}$, $[8]_{FS}^0=\{6,8\}$;
$[9]_{FS}^-=\{4,9\}$, $[9]_{FS}^+=\{4,9\}$, $[9]_{FS}^0=\{4,9\}$;
$[10]_{FS}^-=\{10\}$,$[10]_{FS}^+=\{7,10\}$, $[10]_{FS}^0=\{10\}$.

Then we query based on this and can obtain more correlative results. For instance, query the elements: a little weaker than and must satisfy the condition A=0.25&B=7.00. According to the condition we can get X={1,3,4,6,8,9}. And then calculate the lower approximation of X's positive fuzzy tolerance class, the result is {1,4,6,8,9}.

6. Conclusions

For one information system, mostly it occurs that some attributes are discrete and some are continuous, so we can take following method to process: for discrete attributes we can decompose them by using common tolerance relation and dominance relation, for continuous attributes we firstly fuzzily handle them and then use tolerance relation and dominance relation to cope with.

Introducing fuzzy relation to tolerance relation can extend its application scope. Introducing dominance relation to fuzzy tolerance relation can extend its correlative scope, applying on the information retrieval can extend its retrieval scope.

References：

[1] Pawlak Z.Rough Sets[J].International Journal of Computer and Information Science ,1982;341−356

[2] Slowinski R,Vanderpooten D. Similarity Relation as a Basis for Rough Approximations[C]. In Proceedings of the Second Annual Joint Conference on Information Science, Wrightsville Beach, N Carolina ,USA, also ICS Research Report 53,1995:249−250

[3]LI gang, ZHANG Xueting, Decompositions of rough sets based on similarity relations, computer engineering and applications 2004,(2)：85−87

[4]MA Zhifeng, XING Hancheng, ZHENG Xiaomei, A fuzzy query strategy based on similarity relation rough set computer engineering and applications 2000;36(2)：123−125

[5]AN Liping, CHEN Zengqiang, YUAN Zhuzhi, Multi-criteria sorting decision based on extended rough sets theory. Journals of system engineering 2004,(12)：Vol.19 No.6

[6]ZHANG Chengyi, LU Changing, On measures of similarity between fuzzy rough sets computer engineering and applications 2004,(2)：58−60