AN EXTENDED ROUGH SETS APPROACH TO ANALYSIS OF CUDT

Hua Wenjian and Liu Zuoliang
The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an Shaanxi 710077,China

Abstract: Classical Rough Sets Theory (CRST) is thought to be an effective mathematical approach to discovering rules from Decision Table (DT). Every entry in DT must be unique and certain qualitative value of attribute. However, there are always heterogeneous entries in DT from complex decision problem, that is, entries with a continuous quantitative attribute, unknown entries or multi-valued entries, and these types of entries often occur in same DT. The DT with these entries named Continuous Uncertain Decision Table (CUDT) cannot be analyzed directly by CRST. Fortunately, by modeling those three types of entries in CUDT with Fuzzy Sets theory (FST), we found that CUDT can be transformed into a special DT called Extended Decision Table (EDT) in which each entry is associated with a membership degree. An extended CRST is proposed to transform the CUDT into EDT and to calculate the lower approximations and the boundaries of decision concepts in EDT.

Key words: continuous uncertain decision system; extended Rough Sets Approach; Approximation of concepts

1. INTRODUCTION

In CRST, Decision system (DS) is a 4-tuple $\langle U, A \cup D, V_{A \cup D}, g \rangle$, where U is a finite set of objects; A and D are called as condition attributes and decision attributes respectively, such that $A \cap D = \varnothing$; $V_{A \cup D} = \bigcup v_{r_i}$, $r_i \in A \cup D$, $i = 1, \cdots, |A| + |D|$; $g : U \to V_{A \cup D}$. DS can be represented as a table of two dimensions called Decision Table (DT) with unique entries and certain qualitative value of attribute. However, we are always faced with a continuous uncertain

decision table (CUDT) characterized by the heterogeneous entries: entries with multiple possible values, entries of continuous quantitative attributes and unknown entries[1]. The latter two types of heterogeneous entries can be analyzed by many methods but not based on FST[5-8]. In this paper, by fuzzy modeling of these heterogeneous entries, we proposed an approach to transforming the CUDT into EDT in which every entry is associated with a degree of possibility, and extending CRST to operate approximation of the decision concepts in EDT. Next section, modeling of heterogeneous entries is discussed. Section 3 contains the method of approximating concepts in EDT. Section 4 gives a summary and prospect of this method.

2. MODELING OF HETEROGENEOUS ENTRIES

Discretization of continuous qualitative attributes by FST actually is the procedure that the appropriate fuzzy subsets labeled by the qualitative linguistic terms on the domain V_{a_i} of a continuous attribute are defined and map a value of attribute to a linguistic term. Assume that a set of discrete linguistic terms, denoted by $LS^i = (ls_1^i, ls_2^i, \cdots, ls_m^i)$, corresponding to a continuous quantitative attribute $a_i \in A$ given by the experts. Then for $\forall ls_k^i \in LS^i$ $(1 \le k \le m)$, there is a fuzzy subset $\mu_k^i : V_{a_i} \to (0,1]$ on V_{a_i} , where membership degree $\mu_k^i(a_i(u_j))$ means the degree of possibility that entry $[u_j, a_i]$ is discretized to ls_k^i . But we find that the supports of the fuzzy subsets often overlap, which results in a value of attribute in the overlapped part corresponding to at least two linguistic terms. We can describe discrete result of any entry $[u_j, a_i]$ in this situation as follows:

$$\{(ls_1^i, \mu_1^i(a_i(u_j))), (ls_2^i, \mu_2^i(a_i(u_j))), \cdots, (ls_m^i, \mu_m^i(a_i(u_j)))\}, \quad (1)$$

where $\mu_k^i > 0$, if the supports of fuzzy subsets of linguistic terms overlap; otherwise, not all μ_k^i more than 0. There is another situation we have to consider that entry may be a fuzzy number $\tilde{a}_i(u_j)$. We can describe this entry by analogy with the above formula (1) as formula (1'):

$$\{(ls_1^i, \mu_1^i(\tilde{a}_i(u_j))), (ls_2^i, \mu_2^i(\tilde{a}_i(u_j))), \cdots, (ls_m^i, \mu_m^i(\tilde{a}_i(u_j)))\}. \quad (1')$$

Calculating the membership degree in (1) and (1') is as follows.

Let $x \in U$; A , a attribute set, $a_i \in A$, $1 \le i \le |A| = n$; V_i is value domain of a_i , $V_i \subset R$, and R is real number set; $\tilde{v}_{ih}$ is the h th fuzzy subset on V_i , and its membership degree is $\mu_{\tilde{v}_{ih}}(a_i(x))$, $h = 1, \cdots, m_i$, where m_i is the number of fuzzy subsets on V_i ,a linguistic term of $\tilde{v}_{ih}$ labeled accordingly by v_{ih} . Let $\pi(\bullet, \tilde{v}_{ih})$ denote the membership degree that $\bullet$ is discretely described as v_{ih} . We can get that if $a_i(x)$ is crisp, then

$$\pi(a_i(x), \tilde{v}_{ih}) = \sup_{a_i(x) \in V_i} \{\min[1, \mu_{\tilde{v}_{ih}}(a_i(x))]\} = \mu_{\tilde{v}_{ih}}(a_i(x)) \ ; \quad (2)$$

and if $\tilde{a}_i(x)$ is a fuzzy number on V_i , then

$$\pi(\tilde{a}_i(x), \tilde{v}_{ih}) = \sup_{a_i(x)\in V_i}\{\min[\mu_{\tilde{a}_i(x)}(a_i(x)), \mu_{\tilde{v}_{ih}}(a_i(x))]\} = \mu_{\tilde{v}_{ih}}(\tilde{a}_i(x)) . \quad (3)$$

Complementarity for unknown entries is actually a procedure of obtaining the most possible entries according to other known entries under the same attribute. We are sure that the unknown entry must be the element of the set of linguistic terms given by the analysts. The set of linguistic terms is equal either to the set or to the subset of discrete linguistic terms of the attribute of the unknown entry. We assume that entry $[u_j, a_i]$ is unknown. The idea of complementarity for unknown entry is formally declared by FST as below. There is a fuzzy set $\mu_i : LS^i \to (0,1]$ meaning that for every ls_k^i, membership degree $\mu_i(ls_k^i)$ indicates the possible degree of complementarity for $[u_j, a_i]$ with ls_k^i. The entry $[u_j, a_i]$ after complementarity can be:

$$\{(ls_1^i, \mu_1^i(ls_1^i)), (ls_2^i, \mu_2^i(ls_2^i)), \cdots, (ls_m^i, \mu_m^i(ls_m^i))\}. \quad (4)$$

The multi-valued entries generally occur in decision entries of some military command decision problems we are faced with[1]. Transforming the multi-valued entries into single-valued entries is such a process that the most possible decision value can be selected from the alternative decision values. Let d_i denote the ith decision attribute, and $S^i = \{s_1^i, s_2^i, \cdots, s_m^i\}$ denote discrete linguistic terms of d_i; Assume that $[u_j, d_i]$ is a multi-valued entry, and $[u_j, d_i]$ is represented as $s_1^i, s_2^i, \cdots, s_m^i$. The idea of a multi-valued entry description similar with the above complementarity is formally described as follows.

There is a fuzzy set on S^i, $\mu_i : S^i \to (0,1]$, for every $s_k^i \in S^i$ ($1 \le k \le m$), $\mu_i(s_k^i)$ meaning the possible degree that $[u_j, d_i]$ takes s_k^i. Accordingly, $[u_j, d_i]$ could be represented as:

$$\{(s_1^i, \mu_i(s_1^i)), (s_2^i, \mu_i(s_2^i)), \cdots, (s_m^i, \mu_i(s_m^i))\} \quad (5)$$

For DT of CRST, there is only one membership degree equal to 1 in formula (5), and the other all are 0. But for CUDT, there are at least two membership degrees more than 0 in formula (5).

Definition 1. Unified description of Heterogeneous Entry (HE). Let W_i be a set of linguistic terms of attribute $a_i \in A$, where $W_i = \{w_{ih}\}$, $h = 1, \cdots, m_i$, $m_i = |W_i|$; If the value domain of a_i is continuous quantitative, fuzzy discretization of a_i make the $w_{ih} \in W_i$ the label of fuzzy subset $\tilde{w}_{ih}$ on the value domain of a_i; if the value domain of a_i is discrete, any entry of a_i can take the element of W_i directly. So the unified description of heterogeneous entry in CUDT is

$$[x, a_i] = \{(w_{ih}, \pi_{ih}(x)) \mid w_{ih} \in W_i \text{ and } \pi_{ih}(x) > 0\}, \quad (6)$$

denoted by $HE(x, a_i)$, where $\pi_{ih}(x)$ means degree of possibility that $x \in U$ is described by the h th linguistic term of the i th attribute,

$$\pi_{ih}(x) = \begin{cases} \pi(\tilde{a}_i(x), \tilde{w}_{ih}) & a_i \text{ is continuous quantitative}, \tilde{a}_i(x) \text{ is fuzzy} \\ \pi(a_i(x), \tilde{w}_{ih}) & a_i \text{ is continuous quantitative}, a_i(x) \text{ is crisp} \\ \pi(a_i(x), w_{ih}) & a_i \text{ is discrete} \end{cases} . \quad (7)$$

3. EXTENED ROUGH SET APPROACH

A HE information system $\tilde{I}_{HE} = \langle U, A, W, \tilde{\varsigma}_{HE} \rangle$ means that every entry can be regarded as the complex of the pairs of linguistic terms and possibility degrees on a_i, denoted by

$$[x, a_i] = \{(w_{ih}, \pi_{ih}(x))\},\ \pi_{ih}(x) > 0,\ 1 \le h \le m_i, m_i = |W_i|. \tag{8}$$

There is a suggestion of subdivision of uncertain dataset proposed by Salido, Murakami and Bodjanova respectively[2,3]. Their method substitutes the complex (a_i, w_{ih}) for a_i, where $1 \le h \le m_i$, $m_i = |W_i|$ is the number of linguistic terms. Their approach causing the larger number of attributes may be too computational complexity during the operation of reduction and rule induction[4]. However, our approach declared below could avoid this despite of demanding much computational space.

Definition 2. Let $P \subseteq A$ be the attribute set describing $x \in U$. W_i is the linguistic term set of the ith attribute $q_i \in P$, and $m_i > 1$ means the number of element of W_i more than 1. $[x, q_i] = \{(w_{ih}, \pi_{ih}(x))\}$, $\pi_{ih}(x) > 0$, let $\Pi_i(x)$ denote the set consisting of all degrees of possibility $\pi_{ih}(x)$ in pair $(w_{ih}, \pi_{ih}(x))$ of $[x, q_i]$. If x^j is a subobject of x, then $[x^j, q_i] = (w_{ih}, \pi_{ih}(x^j))$, $\pi_{ih}(x^j) \in \Pi_i(x)$. If only one attribute in P has multi-valued entries, then $j = 1, \cdots, p_i$, $p_i = |\{h \mid \pi_{ih}(x) > 0\}| \le m_i$; if more than one attribute in P have multi-valued entries, then $j = 1, \cdots, \prod_{i:p_i>0} p_i$, $p_i = |\{h \mid \pi_{ih}(x) > 0\}| \le m_i$. So, there must be a vector of x^j related to P:

$$((w_{1h}, \pi_{1h}(x^j)), \cdots (w_{ih}, \pi_{ih}(x^j)) \cdots, (w_{nh}, \pi_{nh}(x^j))). \tag{9}$$

And P possible degree $\pi_P(x^j)$ that P describe x^j is modeled by fuzzy T-norm operator as:

$$\pi_P(x^j) = \min\{\pi_{1h}(x^j), \pi_{2h}(x^j), \cdots, \pi_{nh}(x^j)\},\ n = |P|. \tag{10}$$

Definition 3. Let p be the number of x^j, and P possible degree that P describe x^j is $\pi_P(x^j)$, then the relative coefficient is

$$\eta_P(x^j) = \frac{\pi_P(x^j)}{\sum_{t=1}^{p} \pi_P(x^t)}, \tag{11}$$

which means the scale of the part of x occupied by x^j related to P.

Definition 4. Extended Information Table is a 5-tuple $\tilde{I} = \langle U', \eta, A, W, \tilde{\varsigma} \rangle$, where U' is the finite set of all subobjeccts; η is the set of relative coefficients, i.e. $\eta = \{\eta_A(x')\}$, $x' \in U'$; A is the finite set of condition attributes; $W = \bigcup_{i:a_i \in A} W_i$, W_i is set of qualitative value domain of $a_i \in A$, or called as the set of qualitative linguistic terms of a_i; $\tilde{\varsigma}$ is called as the extended information function, such that $x' \in U', a_i \in A$, $\tilde{\varsigma}(x', a_i) = (w_{ih}, \pi_{ih}(x'))$.

Definition 5. $\tilde{I} = \langle U', \eta, A, W, \tilde{\varsigma} \rangle$ is an extended information system. For $P \subseteq A$, $\forall q_i \in P$, $x' \in U'$, $y' \in U'$, $\tilde{R}_P = \{(x', y') : q_i(x') = q_i(y') = w_{ih}\}$ is the P-indiscernibility relation on U', where $q_i(x') = w_{ih}$, $w_{ih} \in W_i$.

There must exist a partition of U' : $\mathtt{A}=U'/\widetilde{R}_P=\{[x']_P : x'\in U'\}$, where $[x']_P = \{y' : (x',y')\in\widetilde{R}_P\}=\{y' : q_i(y')=q_i(x')(\forall q_i\in P)\}$. $(U',\widetilde{R}_P)$ is a Pawlak approximation space. We call $[x']_P$ in $\mathtt{A}$ as a P-equivalence class. But how do we measure degree of possibility of using w_{ih} to describe $[x']_P$.

Let $\pi_{ih}(x_i')$ be the degree of possibility that $x_i'\in[x']_P$ is described by w_{ih}. Then degree of possibility of using w_{ih} to describe $[x']_P$ is $\pi_{ih}([x']_P)=\min_{x_i'\in[x']_P}\{\pi_{ih}(x_i')\}$.

Definition 6. The relative coefficient of subobject $y'\in U'$ is $\eta_P(y')$, and then the cardinality of $[x']_P$ is $Card([x']_P)=\sum_{y'\in[x']_P}\eta_P(y')$.

Definition 7. $(U',\widetilde{R}_P)$ is a Pawlak approximation space, $[x']_P\in\mathtt{A}$. For any $Y\subseteq U'$, P-lower and P-upper approximation of Y related to $(U',\widetilde{R}_P)$ are as follows respectively:

$$\underline{\widetilde{R}}_P(Y)=\{x'\in U' : [x']_P\subseteq Y\}=\bigcup\{[x']_P : [x']_P\subseteq Y\}, \tag{12}$$

$$\overline{\widetilde{R}}_P(Y)=\{x'\in U' : [x']_P\cap Y\neq\varnothing\}=\bigcup\{[x']_P : [x']_P\cap Y\neq\varnothing\}; \tag{13}$$

P-boundary of Y related to $(U',\widetilde{R}_P)$ is:

$$\widetilde{Bn}_P(Y)=\overline{\widetilde{R}}_P(Y)-\underline{\widetilde{R}}_P(Y) \tag{14}$$

Definition 8. Accuracy of Y related to approximation space $(U',\widetilde{R}_P)$ is

$$\alpha_{\widetilde{R}_P}(Y)=\frac{Card(\underline{\widetilde{R}}_P(Y))}{Card(\overline{\widetilde{R}}_P(Y))}\ ,\ \text{where} \tag{15}$$

$$Card(\underline{\widetilde{R}}_P(Y))=\sum_{[x']_P\subseteq\underline{\widetilde{R}}_P}Card([x']_P)=\sum_{x'\in\underline{\widetilde{R}}_P}\eta_P(x'),\ Card(\overline{\widetilde{R}}_P(Y))=\sum_{[x']_P\subseteq\overline{\widetilde{R}}_P}Card([x']_P)=\sum_{x'\in\overline{\widetilde{R}}_P}\eta_P(x').$$

Definition 9. Let Ω be the classification on U', $\Omega=\{\omega_i\}$, $\omega_i\subseteq U'$, such that $\omega_i\cap\omega_j=\varnothing$, $1\le i\neq j\le n$, the lower approximation and the upper approximation of Ω are respectively:

$$\underline{\widetilde{R}}_P(\Omega)=\{\underline{\widetilde{R}}_P(\omega_1),\underline{\widetilde{R}}_P(\omega_2),\cdots,\underline{\widetilde{R}}_P(\omega_n)\}\ ,\ \overline{\widetilde{R}}_P(\Omega)=\{\overline{\widetilde{R}}_P(\omega_1),\overline{\widetilde{R}}_P(\omega_2),\cdots,\overline{\widetilde{R}}_P(\omega_n)\}\ ;$$

And P- approximation quality of Ω is:

$$\gamma_{\widetilde{R}_P}(\Omega)=\frac{\sum_{i=1}^{n}(Card(\underline{\widetilde{R}}_P(\omega_i)))}{Card(U')} \tag{16}$$

Assuming that the original decision table has been subdivided to be $\widetilde{I}$ but still with multi-valued decision entries, we can subdivide x^j into x^{jk}, called as secondary subobject. So a x^{jk} is a secondary subobject that can be actually described as a vector with both condition single-valued entry and decision single-valued entry.

Relative coefficient of x^{jk} which means the scale of the part of x^j occupied by x^{jk} related to D is:

$$\eta_D(x^{jk})=\frac{\pi_D(x^{jk})}{\sum_{g=1}^{l}\pi_D(x^{jg})}, \tag{17}$$

where $\pi_D(x^{jk})=\min\{\pi_{1h}(x^{jk}),\pi_{2h}(x^{jk}),\cdots,\pi_{|D|h}(x^{jk})\}$.

The relative coefficient of x^j, denoted by $\eta_A(x^j)$, the relative coefficient of x^{jk}, denoted by $\eta_D(x^{jk})$, $\eta_{A,D}(x^{jk}) = \eta_A(x^j) \bullet \eta_D(x^{jk})$ is interpreted as comprehensive relative coefficient of x^{jk} which means the scale of the part of x occupied by x^{jk} related to A and D.

Definition 10. Extended Decision System is a 5-tuple $\widetilde{DT} = \langle U'', \eta, Q, W, \widetilde{\varsigma} \rangle$, where U'' is the set of secondary subobjects; η is the set of comprehensive relative coefficients, i.e. $\eta = \{\eta_{A,D}(x'')\}$, $x'' \in U''$; $Q = A \cup D$, and $A \cap D = \varnothing$, A and D are condition attribute set and decision attribute set respectively; $W = \bigcup_{i=q_i \in Q} W_i$, W_i is the set of linguistic terms on attribute q_i; $\widetilde{\varsigma}$ is satisfied with $\widetilde{\varsigma}(x'', q_i) = (w_{ih}, \pi_{ih}(x''))$, where $\pi_{ih}(x'')$ means the degree of possibility that x'' is described by the h th linguistic term of the i th q_i.

4. CONCLUSION

This paper proposed an approach based on extended Rough Sets Theory to computing the lower and upper approximation of the concepts in EDT, which can be as input to the rule induction algorithm, i.e. LEM2[4]. Some concepts, however, such as reduction before LEM2, cover and minimal rule set in LEM2, must be extended according to extended Rough Sets Theory above discussed. This is just the problem we are focusing on.

REFERENCES

1. Hua Wenjian, Liu zuoliang and Yang fan, A novel knowledge representation and induction of decision rules. *Journal of Air Force Engineering University (Natural Science Edition)* 4(6), 44-48(2003).
2. J.M Fernandez Salido, S. Murakami., Rough set analysis of a general type of fuzzy data using transitive aggregation of fuzzy similarity relations. *Fuzzy Sets and Systems* 139(3), 635-660(2003).
3. Slavka Bodjanova., Approximation of fuzzy concepts in decision making. *Fuzzy Sets and Systems* 85(1), 23-29(1997).
4. Jerzy Stefanowski, On Rough set based approaches to induction of decision rules, in: *Rough sets in knowledge discovery 1*, edited by Lech Polkowski and Andrzej Skowron. (Physica-Verlag, Heidelberg, 1998), pp.500-529.
5. Wang Guoyin, *Rough Sets Theory and Knowledge Acquisition* (Xian Jiao Tong University Publisher Press, Xian, 2001).
6. Nguyen S.H. ,Skowron A, Quantization of real value attributes: rough set and Boolean reasoning approach, in: *Second Annual Joint Conference on Information Sciences (JCIS'95)*, edited by P.P. Wang (Wrightsville Beach, North Carolina, 1995), pp.34-37.
7. Ellis J. Clarke, Bruce A. Barton., Entroy and MDL discretization of continuous variable for Bayesian belief network. *International Journal of Intelligent Systems* 15(1), 61-92(2000).
8. Zhao Weidong, Dai Weihui and Cai Bin, The discretization of continuous attributes using genetic algorithms. *Systems Engineering Theory & Practice* 23(1),62-67(2003).