A Feature Selection Method for Online Hybrid Data Based on Fuzzy-rough Techniques

Ye Yuling
(Yichang Testing Technique Research Institute, Research Center for Simulation & Information, Yichang 443003, China)

Abstract

Data reduct based on rough set theory was an effective feature selection method, however, classic rough set theory cannot deal with hybrid data and can't applied to online systems either. So the rough set model based on fuzzy equation relation was improved to reduct the hybrid systems. The entropy was used to measure the discernibility power of the information and the definition of relative reduct was improved, and the notion of sequential reduct was proposed to deal with real online systems. A complete algorithm was proposed and applied to several UCI data. Experiments show that sequential reduct algorithm is an effective feature selection method for real online systems.

1. Introduction

With the development of the information acquirement and storage ability, more and more large databases were constructed. These databases built for many purposes and consist of large quantity of attributes. However, for a specific purpose there are many redundant and irrelevant attributes, and these attributes not only increased the cost of the storage and management of data but also confused the algorithm of data mining and lead to a low learning precision. Two methods called feature selection [1] and feature extraction were presented to pre-process the data before mining [2]. Our work focused on feature selection which is to select effective attributes for a specific purpose from the whole attribute set.

Rough set theory introduced by Pawlak in 1981 has come to many achievements both in theory [3] and application [4]. Attribute reduction based on rough set theory became an effective feature selection method [5]. However, classical rough set theory was theorized based on equivalence relation and can only apply to the systems with discrete or nominal data. However, in the real world data always come with not only discrete data but continuous data and even fuzzy data, called hybrid data [6]. In order to apply the rough set theory to reduct hybrid data systems, some methods were proposed to discrete the continuous attributes [7]. But Shen [8] pointed out that this process will lead to some information loss of the original systems. Fuzzy theory was introduced to rough set theory too and fuzzy rough set theory [9] and rough fuzzy theory [10] were proposed. Hu [11] constructed the rough set model based on fuzzy equivalence relation and proposed a greedy reduction algorithm, experiments shown that it is an effective method for feature selection. However, the notion of relative reduction is not reasonable and the definition is constructed for static systems. Many real systems are online and they are dynamic systems with hybrid data. For an online system the significance of an attribute for a specified purpose is changing, and the notion of reduction in the static systems is no longer appropriate. Looking for effective reduct algorithm for a static system is a hot research field in the classic rough set theory; however, little work has done to the sequential reduct for online systems. In this paper, we proposed the notion of sequential reduct based on Hu's rough set theory which was constructed on fuzzy equivalent relation. An online sequential reduction algorithm was proposed, and some experiments on UCI databases show that the method is effective for the online noisy hybrid systems.

The rest of the paper is organized as follows: the notion of sequential reduct was presented in section 2. The sequential reduct algorithm for the fuzzy rough set model is proposed in section 3. Some experiments and analysis were presented in the section 4 and section 5 comes with the conclusion.

2. Sequential reduct

In the rough set theory, the notion of information system is denoted as $S=(U,A,f,V)$. Where $U=\{x_1,x_2,\cdots,x_n\}$ is a nonempty and finite set

978-0-7695-3571-5/09 $25.00 © 2009 IEEE
DOI 10.1109/GCIS.2009.55

called universe. $A=\{a_1,a_2,\cdots,a_m\}$ is an attribute set. $f:U\times A\rightarrow V$ is a mapping function and $V=\{V_1,V_2,\cdots,V_m\}$. V_i is the value field of x_i on attribute a_i . If $A=C\bigcup D$ and $C\bigcap D=\phi$, C is called condition attribute set and D is called decision attribute set, and S is called decision system.

Hu [11] proposed the definition of relative reduct in the rough set model based on fuzzy equation relation and it satisfied the following two rules:

(1) $H(d\mid B)=H(d\mid C)$

(2) $\forall a\in B:H(d\mid B-a)>H(d\mid B)$

According to the definition Hu proposed the greedy relative reduct algorithm based on condition attribute's relative significance. But it is not reasonable for $SIG(a,B,d)=H(d\mid B-a)-H(d\mid B)$ is not always positive. So we improved the definition of Hu's relative reduct [11] as:

Definition 1: For a decision system $S=(U,A,f,V)$, $A=C\bigcup D$, $B\subseteq C$, B is called a reduct of C relative to D if B satisfies:

1) $H(D\mid B)=H(D\mid C)$

2) B is independent relative to D

In theory, the reduct is defined as a subset of attributes which has the same discernibility power as the full attribute set. However, it is too strict for the real systems for the databases always come with noise. So we specify a threshold δ to definition the weak reduct and weak relative reduct as:

Definition 2: For an information system $S=(U,A,f,V)$, $B\subseteq A$, $a\in B$, B is called a weak reduct if B satisfies:

1) $H(A)-H(B)\leq\delta$;

2) $\forall a\in B$: $H(A)-H(B-a)>\delta$

Definition 3: For a decision system $S=(U,A,f,V)$, $A=C\bigcup D$, $B\subseteq C$, $a\in B$, B is called a weak reduct of C relative to D if B satisfies:

1) $\left|H(D\mid B)-H(D\mid C)\right|\leq\delta$

2) $\left|H(D\mid B-a)-H(D\mid C)\right|>\delta$

The degree threshold δ should be specified carefully based on the real condition. In the high noisy systems δ should be given a larger value and else δ should be set as a small value. The real value of δ should determined by experiments.

For online systems, the number of the objects often increases with the running of the systems and the universe will become larger and larger. So we propose the following definition for the online systems:

Definition 4: For a given decision system $S=(U,A,f,V)$, $T=(U_i,A,f,V)$ is called a subsystem of S , $\forall x_i\in U_i$, $x_i\in U$.

Definition 5: Let $S_i=(U_i,A,f,V)$, $S_i\subset S_{i+1}$, $i=0,1,2,\cdots$ be an online system and F_i be a family of subsystems of S_i , $RED(T_i)$ is a reduct of system T_i , ε is a real number from the unit interval [0,1]. The $(F-\varepsilon)$ -sequential reducts of S_i is defined as:

$$SR_\varepsilon(S_i,F_i)=\begin{cases}\left\{a\in A:\dfrac{\left|\{T_i\in F_i:a\in RED(T_i)\}\right|}{\left|F_i\right|}\geq\varepsilon\right\} & i=0\\ \left\{a\in A:\sum_{j=0}^{m}\eta_j\cdot\dfrac{\left|\{T_{i-j}\in F_{i-j}:a\in RED(T_{i-j})\}\right|}{\left|F_{i-j}\right|}\geq\varepsilon,m\leq i\right\} & i>0\end{cases} \tag{1}$$

Where $\gamma=\left|\{T\in F:a\in RED(T)\}\right|/\left|F\right|$ is called the stability coefficient of a relative to F , and γ indicates the importance of a for the incremental system. $\gamma=1$ means a is very important, and $0<\gamma<1$ means a is somewhat important, and $\gamma=0$ means a is unimportant for the incremental system S . ε is the stability threshold, and its value should be specified by experiments. η_j is memory coefficient. In the online systems the significance of an attribute to the specified purpose is variant with time. So, the sequential reduct of the online system is variant too. $\eta_j\in[0,1]$, and the value indicates the effect of S_{i-j} to S_i. $\eta_j=1$ means the online system is stable and $\eta_j=0$ means that the system is randomly developed with time, else $\eta_j\in(0,1)$ indicates that the system is variant with time. Generally, the bigger j is the smaller η_j is.

3. Sequential reduct algorithm

We will build the sequential reduct algorithm for the online hybrid decision systems in this section.

The algorithm can be divided into 4 steps: subsystem selection, relation matrix computation, reduct of the subsystems and compute the sequential reduct.

In the first step, subsystem selection is to select a series subsystem from current system S to build the subsystem family " F ". There are three problems should be considered: how many samples should be selected for a subsystem, how many subsystems should be selected and how to select samples.

We suppose online system S is worked from system S_0. The method to compute the sequential reduct for S_0 is not the same as S_i ($i > 0$). For S_0, in the proposed algorithm, $5N_S$ subsystems are randomly selected from S_0:

(1) N_S with the size of 100% of the S_0, which made up of the subsystem family SF_0;

(2) N_S with the size of 90% of the S_0, which made up of the subsystem family SF_{-1};

(3) N_S with the size of 80% of the S_0, which made up of the subsystem family SF_{-2};

(4) N_S with the size of 70% of the S_0, which made up of the subsystem family SF_{-3};

(5) N_S with the size of 60% of the S_0, which made up of the subsystem family SF_{-4}.

For S_i ($i > 0$), N_S subsystems are selected from S_i with the size of N_D, which made up of the subsystem family SF_i. Subsystems are made up of all the newly increased N_δ samples and randomly selected $N_D - N_\delta$ samples. For an online system S the number of samples is increased continuously, but the ability of storage and computing is limited, on the other hand the history samples are not always as important as the new samples, so we limited the number of the samples in the S to N_D by removing the oldest samples.

In the second step, relation matrix computation is to compute the relation matrix for each attribute. No matter samples $\{x_i\}_{i=1}^n$ are described by nominal attributes, numeric features or fuzzy variables, the relation between the samples can be denoted by a relation matrix: $M(R) = (r_{ij})_{n\times n}$. Hu proposed the method in [11].

The third step is to compute the relative reducts of subsystems in F. We improved the relative significance of attribute in the decision systems as follows:

Definition 6: For a decision system $S = (U, A, f, V)$, $A = C \bigcup D$, $B \subseteq C$, $a \in B$, the significance of attribute a in attribute set B relative to D is defined as:

$$SIG(a,B,D) = \left| H(D \mid B) - H(D \mid B - a) \right| \quad (2)$$

Based on the above measures, a greedy algorithm for computing relative reduct can be constructed.

Algorithm 1: algorithm for calculating relative reduct for a hybrid decision system.

```
Input: $DS = (U, C \bigcup D)$ and the threshold $\delta$
Output: one relative reduct $RED$ of $DS$
Step 1: $\forall a \in C$ and $d \in D$: compute the relation matrix;
Step 2: $RED = C$, $H_0 = H(D \mid RED)$;
Step 3: for each $a_i \in RED$, Compute:
        $H_i = H(D \mid RED - a_i)$
Step 4: choose attribute $a_k$ which satisfies:
        $| H_k - H_0 | = \min_i (| H_i - H_0 |)$
Step 5: if $| H_k - H_0 | \le \delta$, $RED = RED - a_k$, $H_0 = H_k$ goto step 3; Else end and return $RED$
```

The last step is to compute the sequential reduct $SRED$ of the online system. For S_0, the attribute a is added to the $SRED_0$, if it satisfies the definition (1) with the specified ε. As an intermediate result, for SF_i the $NF_i = \left|\{T \in SF_i : a \in RED(T)\}\right|$, ($i = 0, -1, -2, -3, -4$) should be saved.

For S_i ($i > 0$), $NF_i = \left|\{T \in SF_i : a \in RED(T)\}\right|$ will be computed first, then $SRED_i$ can be computed by formulate:

$$SRED_i = \left\{ a \in A \;\middle|\; \frac{\sum_{j=1}^{5} \eta_j NF_{i-6+j,a}}{5N_s} \ge \varepsilon \right\} \quad (3)$$

Then we only need to compute N_s relative reduct after every increase.

Table 1 **Specification of simulation systems**

system	abbreviation	samples	classes	Condition attributes		
				total	numeric	nominal
Wisconsin Diagnostic Breast Cancer	WDBC	569	2	31	31	0
Wisconsin Prognostic Breast Cancer	WPBC	194	2	34	34	0
Dr. Detrano's database Cleve	Cleve	197	2	13	6	7
Sonar, Mines vs. Rocks	Sonar	138	2	60	60	0
Johns Hopkins University Ionosphere database	Ionosphere	351	2	35	35	0
Australian Credit Approval	Credit	690	2	14	6	8
Heart disease dataset	Heart	270	2	14	7	7

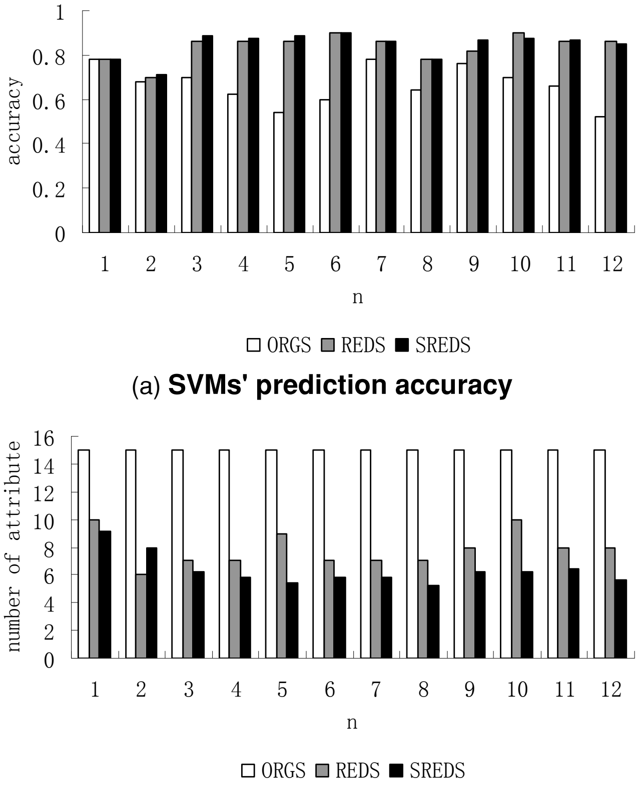
4. Simulations

The performance of the algorithm SRED is evaluated on the benchmark problems from UCI database which were shown in Table 1. All the simulations have been conducted in MATLAB 6.5 environment running on an ordinary PC with 2.66 GHz CPU.

In order to test the validity of the sequential reduct algorithm, support vector machine (SVM) was proposed to work as a validation function. "spline" function were adopted to work as the kernel function for SVMs. For every system, we choose the first N_p samples to build up S_0 , and the samples from $(i-1)N_\Delta$ to $N_D+(i-1)N_\Delta$ to build up S_i . The samples in the system S_i were divided into two set: train set ST_i and test set SS_i. SVMs were trained by ST_i and the trained SVMs were used to predict the decision of SS_i. We choose the first NT samples of S_i to build the ST_i and the rest to build SS_i. We set $\eta_0=1, \eta_1=0.9$, $\eta_2=0.8$, $\eta_3=0.7$, $\eta_4=0.6$ in every simulation, and for each problem, the result was averaged over 5 trials.

4.1 simulation 1

The simulation was based on database "Credit" which data increased chunk by chunk with the running of the system. Sequential reduct were computed after every chunk (the size of the chunk may be variant). In the algorithm $\delta=0.01$, $\varepsilon=0.25$, $N_p=100$, $N_D=100$, $N_\Delta=50$. The SVMs were trained by original system, reduct system and reduct system respectively and the performance were shown in figure 1. In which, "ORGS" indicates original system, "REDS" reduct system and "SREDS" sequential reduct system. From figure 1 we can reach the following conclusion:



(a) **SVMs' prediction accuracy**

(b) **Number of the condition attributes**

Fig. 1 **SVM's prediction accuracy and the number of condition attributes comparison on "Credit" database**

1) The number of attributes in sequential reduct system is usually less than in reduct system;

2) The SVM prediction accuracy of sequential reduct system is mostly higher than the reduct system and on both reduct and sequential reduct system SVMs have the higher prediction accuracy than on original system.

3) The importance of an attribute to the specified purpose is varies with the running of the system. In the simulation, the anterior sequential reduct include "A2" and "A14" and not include "A12", but in the posterior

Table 2 **Parameter value and results on 7 UCI databases**

System	δ	ε	N_D	N_Δ	ORGS		REDS		SREDS	
					n	accu	n	accu	n	accu
WDBC	0.0003	0.6	100	50	30	0.9275	23.13	0.9475	20.25	0.9490
WPBC	0.0003	0.6	80	20	33	0.6000	16.60	0.5300	13.84	0.6100
Cleve	0.01	0.2	80	20	13	0.7300	8.20	0.6700	8.64	0.6720
Sonar	0.0003	0.6	80	20	60	0.7833	42.00	0.7333	35.30	0.7916
Ionosphere	0.0003	0.6	100	50	34	0.8350	27.00	0.8250	18.30	0.8500
Credit	0.01	0.25	100	50	14	0.6650	7.83	0.8367	6.32	0.8447
Heart	0.0003	0.6	100	50	13	0.7133	11.66	0.7400	9.80	0.7707
a	/	/	/	/	28.14	0.7506	19.49	0.7546	16.06	0.7840

sequential reduct “A2” and “A14” is not important any more and “A12” is included in the sequential reduct.

4.2 Simulation 2

All the databases are tested to prove the generality of the above conclusion. The value of coefficients and the simulation results are shown in the table 2.

From the table, we can generalize the conclusion of simulation 1 and further more we can get the conclusion:

1) For specified purpose there are many redundant or irrelevant attributes in the real systems which increased the uncertainty. Attribute redcut based on fuzzy-rough set model is an effective method for feature selection. Attribute reduct can decrease the number of condition attributes for a specified purposed and increase data mining algorithms’ accuracy at the same time.

2) Sequential reduct algorithm is an appropriate method for the online system feature selection. Sequential reduct algorithm is a dynamic algorithm and the highest accuracy and lest attributes proved that it is appropriate to the online systems.

5. Conclusion

Feature selection based on rough set theory can keep the system information and decrease the number of the attributes effectively. Sequential reduct algorithm brings some degree of tolerance to the noise which is effective to real online systems. In the process of sequential reduct, the most “significant” attributes of the current subsystem are selected to make up of the sequential reduct, which improved the SVM’s prediction accuracy. Compare with the reduct systems sequential reduct systems has less attributes but SVMs have higher prediction accuracy. Sequential reduct and its algorithm is an effective feature selection method.

6. References

[1] M. Dash, H. Liu. “Featrue selection for Classification”. Intelligent Data Analysis, 1997, 1(1-4): 131-156.

[2] Rajen B. Bhatt, M. Gopal. “On fuzzy-rough sets approach to feature selection”. Pattern Recognition Letters, 2005, 26(7): 965-975.

[3] Y.Y. Yao. “Constructive and Algebraic Methods of the Theory of Rough Sets”. Journal of Information Sciences, 1998, 109(1-4): 21-47.

[4] Shusaku Tsumoto. “Automated extraction of medical expert system rules from clinical databases based on rough set theory”. Information Sciences, 1998, 112(1): 67-84.

[5] Qinghua Hu, Daren Yu, Zongxia Xie. Reduction algorithm for hybrid data based on fuzzy rough sets approaches [C]// Proceedings of the third international conference on machine learning and cybernetics, Shanghai, IEEE press: 2004, 1469-1474.

[6] Miao Duo-Qian. “A New Method of Discretization of Continuous Attributes in Rough Sets”. Acta Automatica Sinica, 2001, 27(3): 296-302. (In Chinese)

[7] Richard Jensen, Qiang Shen. “Fuzzy-rough Attribute Reduction with Application to Web Categorization”. Fuzzy Sets and Systems, 2004, 141: 469-485.

[8] Qiang Shen, Richard Jensen. “Selecting informative features with fuzzy rough sets and its application for complex systems monitoring”. Pattern Recognition, 2004, 37(7): 1351-1363.

[9] Dubois D., Prade. H. Putting fuzzy sets and rough sets together [M]// Slowiniski, R. Intelligent Decision Support. Kluwer Academic, Dordrecht, 1992, 203-232.

[10] Qinghua Hu, Daren Yu, Zongxia Xie. “Information-preserving hybrid data reduction based on fuzzy-rough techniques”. Patter Recognition Letters, 2006, 27(5): 414-423.

[11] E. Hernández, J. Recasens. “A reformulation of entropy in the presence of indistinguishability operators”. Fuzzy Sets and Systems, 2002, 128 (2) 185-196.