A Time-Reduction Strategy to Feature Selection in Rough Set Theory

Hongxing Chen, Yuhua Qian, Jiye Liang, Wei Wei, and Feng Wang

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006, Shanxi, China
chx@sxu.edu.cn, jinchengqyh@126.com, ljy@sxu.edu.cn, weiwei@sxu.edu.cn, sxuwangfeng@126.com

Abstract. In rough set theory, the problem of feature selection aims to retain the discriminatory power of original features. Many feature selection algorithms have been proposed, however, quite often, these methods are computationally time-consuming. To overcome this shortcoming, we introduce a time-reduction strategy, which can be used to accelerate a heuristic process of feature selection. Based on the proposed strategy, a modified feature selection algorithm is designed. Experiments show that this modified algorithm outperforms its original counterpart. It is worth noting that the performance of the modified algorithm becomes more visible when dealing with larger data sets.

Keywords: Ordered decision table, consistency, fuzziness.

1 Introduction

Feature selection, also called attribute reduction, is a common problem in pattern recognition, data mining and machine learning.

In the last twenty years, many techniques of attribute reduction have been developed in rough set theory. The concept of the β-reduct proposed by Ziarko provides a suite of reduction methods in the variable precision rough set model [1]. An attribute reduction method was proposed for knowledge reduction in random information systems [2]. Five kinds of attribute reducts and their relationships in inconsistent systems were investigated by Kryszkiewicz [3], Li et al. [4] and Mi et al. [5], respectively. By eliminating some rigorous conditions required by the distribution reduct, a maximum distribution reduct was introduced by Mi et al. in [5]. In order to obtain all attribute reducts of a given data set, Skowron proposed a discernibility matrix method [6], in which any two objects determine one feature subset that can distinguish them. According to the discernibility matrix viewpoint, Qian et al. [7] and Shao et al. [8] provided a technique of feature selection for interval ordered information systems and incomplete ordered information systems, respectively. The above feature selection methods are usually time consuming and intolerable to process large-scale data. To support efficient feature selection, many heuristic attribute reduction methods have been developed in rough set theory, cf. [9-18]. Each of these methods

P. Wen et al. (Eds.): RSKT 2009, LNCS 5589, pp. 111–119, 2009.

preserves a particular property of a given information system. For convenience, we review only four representative heuristic attribute reduction methods. Hu and Cercone proposed a heuristic feature selection method, called positive-region reduction, which keeps the positive region of target decision unchanged [9]. As Shannon's information entropy was introduced to search reducts in the classical rough set model [15], Wang et al. used its conditional entropy to calculate the relative attribute reduction of a decision information system [16]. Liang et al. defined new information entropy to measure the uncertainty of an information system [12] and applied the entropy to reduce redundant features [13]. Qian and Liang in [14] presented combination entropy for measuring the uncertainty of information systems and used its conditional entropy to obtain a feature subset.

In this paper, we are not concerned on how to construct a heuristic function for feature selection. Our objective is to focus on how to improve the time efficiency of a heuristic attribute reduction algorithm. We employ a new rough set framework, which is called positive approximation. The main advantage of this approach stems from the fact that this framework is able to characterize the granulation structure of a rough set using a granulation order. Based on the positive approximation, we develop a common strategy for improving the time efficiency of a heuristic feature selection, which provides a vehicle of making algorithms of rough set based feature selection techniques faster. By incorporating the strategy into each of the two representative heuristic attribute reduction methods (positive reduction and Shannon's entropy based reduction), we construct their modified versions. Numerical experiments show that each of the modified methods can choose the same feature subset as that of the corresponding original method while greatly reducing computing time.

2 Preliminaries

In this section, we will review several basic concepts and positve approximation in rough set theory. Throughout this paper, we suppose that the universe U is a finite nonempty set.

Let U be a finite and non-empty set called the universe and $R \subseteq U \times U$ an equivalence relation on U, then $K = \langle U, R\rangle$ is called an approximation space [19, 20]. Given an approximation space $K = \langle U, R\rangle$ and an arbitrary subset $X \subseteq U$, one can construct a rough set of the set on the universe by elemental information granules in the following definition:

$$\begin{cases} \underline{R}X = \cup\{[x]_R | [x]_R \subseteq X\}, \\ \overline{R}X = \cup\{[x]_R | [x]_R \cap X \neq \emptyset\}, \end{cases}$$

where $\underline{R}X$ and $\overline{R}X$ are called R-lower approximation and R-upper approximation with respect to R, respectively. The order pair $\langle \underline{R}X,\ \overline{R}X\rangle$ is called a rough set of X with respect to the equivalence relation R.

There are two kinds of attributes for a classification problem, which can be characterized by a decision table $S = (U, C \cup D)$ with $C \cap D = \emptyset$, where an

element of C is called a condition attribute, C is called a condition attribute set, an element of D is called a decision attribute, and D is called a decision attribute set. Assume the objects are partitioned into r mutually exclusive crisp subsets $\{X_1, X_2, \cdots, X_r\}$ by the decision attributes D. Given any subset $B \subseteq C$ and R_B is the equivalence relation induced by B, then one can define the lower and upper approximations of the decision attributes D as

$$\begin{cases} \underline{R_B}D = \{\underline{R_B}X_1, \underline{R_B}X_2, \cdots, \underline{R_B}X_r\}, \\ \overline{R_B}X = \{\overline{R_B}X_1, \overline{R_B}X_2, \cdots, \overline{R_B}X_r\}. \end{cases}$$

Denoted by $POS_B(D) = \bigcup_{i=1}^{r} \underline{R_B}X_i$, it is called the positive region of D with respect to the condition attribute set B.

We define a partial relation $\preceq$ on the family $\{U/B \mid B \subseteq C\}$ as follows: $U/P \preceq U/Q$ (or $U/Q \succeq U/P$) if and only if, for every $P_i \in U/P$, there exists $Q_j \in U/Q$ such that $P_i \subseteq Q_j$, where $U/P = \{P_1, P_2, \cdots, P_m\}$ and $U/Q = \{Q_1, Q_2, \cdots, Q_n\}$ are partitions induced by P, $Q \subseteq A$, respectively [12]. In this case, we say that Q is coarser than P, or P is finer than Q.

In the following, we review a particular set-approximation approach called positive approximation [21], in which a target concept is approximated by a positive granulation world. These concepts and properties will be helpful to understand the notion of a granulation order and set approximation under a granulation order.

Definition 1.[21] Let $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \cdots, R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq \cdots \succeq R_n$ $(R_i \in 2^A)$. Let $P_i = \{R_1, R_2, \cdots, R_i\}$, we define P_i-lower approximation $\underline{P_i}(X)$ and P_i-upper approximation $\overline{P_i}(X)$ of P_i-positive approximation of X as

$$\begin{cases} \underline{P_i}(X) = \bigcup\limits_{k=1}^{i} \underline{R_k}X_k, \\ \overline{P_i}(X) = \overline{R_i}X, \end{cases}$$

where $X_1 = X$ and $X_k = X - \bigcup_{j=1}^{k-1} \underline{R_j}X_j$ for $k = 2, 3, \cdots, n$, $i = 1, 2, \cdots, n$.

Definition 2.[21] Let $S = (U, C \cup D)$ be a decision table, $P = \{R_1, R_2, \cdots, R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq \cdots R_n$ $(R_i \in 2^C)$ and $U/D = \{X_1, X_2, \cdots, X_r\}$. Lower approximation and upper approximation of D with respect to P are defined by

$$\begin{cases} \underline{P}D = \{\underline{P}X_1, \underline{P}X_2, \cdots, \underline{P}X_r\}, \\ \overline{P}D = \{\overline{P}X_1, \overline{P}X_2, \cdots, \overline{P}X_r\}. \end{cases}$$

$\underline{P}D$ is also called the positive region of D with respect to the granulation order P, denoted by $POS_P^U(D) = \bigcup_{k=1}^{r} \underline{P}X_k$.

Theorem 1. Let $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \cdots, R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq \cdots \succeq R_n$ $(R_i \in 2^C)$. Let $P_i = \{R_1, R_2, \cdots, R_i\}$, then $POS_{P_{i+1}}^{U}(D) = POS_{P_i}^{U}(D) \cup POS_{R_{i+1}}^{U_{i+1}}(D)$, where $U_1 = U$ and $U_{i+1} = U - POS_{P_i}^{U}(D)$.

3 A Time-Reduction Strategy to Feature Selection in Rough Set Theory

For efficient feature selection, many heuristic attribute reduction methods have been developed in rough set theory, see [9-18]. For convenience, we only focus on the two representative attribute reduction methods.

Given a decision table $S = (U, C \cup D)$, one can obtain the condition partition $U/C = \{X_1, X_2, \cdots, X_m\}$ and the decision partition $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. Through these denotations, in what follows we review four types of significance measures of attributes.

Hu and Cercone proposed a heuristic feature selection method, called positive-region reduction (PR), which keeps the positive region of target decision unchanged [9]. In this method, the significance measures of attributes are defined as follows.

Definition 3. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in B$. The significance measure of a in B is defined as

$Sig_1^{inner}(a, B, D) = \gamma_B(D) - \gamma_{B-\{a\}}(D)$,

where $\gamma_B(D) = \frac{|POS_B(D)|}{|U|}$.

Definition 4. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in C - B$. The significance measure of a in B is defined as

$Sig_1^{outer}(a, B, D) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D)$.

As Shannon's information entropy was introduced to search reducts in classical rough set model [15], Wang et al. used its conditional entropy to calculate the relative attribute reduction of a decision information system [16]. This reduction method is denoted by SCE. This conditional entropy reads as

$$H(D|B) = -\sum_{i=1}^{m} p(X_i) \sum_{j=1}^{n} p(Y_j|X_i) \log(p(Y_j|X_i)),$$

where $p(X_i) = \frac{|X_i|}{|U|}$ and $p(Y_j|X_i) = \frac{|X_i \cap Y_j|}{|X_i|}$. Using the conditional entropy, the definitions of the significance measures are expressed in the following way.

Definition 5. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in B$. The significance measure of a in B is defined as

$Sig_2^{inner}(a, B, D) = H(D|B - \{a\}) - H(D|B)$.

Definition 6. Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in C - B$. The significance measure of a in B is defined as

$Sig_2^{outer}(a, B, D) = H(D|B) - H(D|B \cup \{a\})$.

All the definitions used above are used to select an attribute in a heuristic feature selection algorithm. For a given decision table, the intersection of all attribute reducts is said to be indispensable and is called the core. Each attribute in the core must be in every attribute reduct of the decision table. The core may be an empty set. The above two kinds of significance measures can be used to find the core attributes. The following theorem is of interest with this regard.

Theorem 2. Let $S=(U, C\cup D)$ be a decision table and $a \in C$. If $Sig_{\Delta}^{inner}(a, C, D) > 0$ ($\Delta = \{1, 2\}$), then a is a core attribute of S in the context of type Δ.

In a heuristic feature selection algorithm, based on the above theorem, one can find an attribute reduct by gradually adding selected attributes to the core attributes. For more clear representation, we denote the significance measure of an attribute by $Sig_{\Delta}^{outer}(a, B, D, U)$ ($\Delta = \{1, 2\}$), which denotes the value of the significance measure on the universe U.

Hence, we can construct an improving forward search algorithm based on the positive approximation, which is formulated as follows. In this general algorithm framework, we denote the evaluation function (stop criterion) by $EF^U(B, D) = EF^U(C, D)$. For example, if one adopts Shannon's conditional entropy, then the evaluation function is $H^U(B, D) = H^U(C, D)$. That is to say, if $EF^U(B, D) = EF^U(C, D)$, then B is said to be an attribute reduct.

Algorithm 1. A general improved feature selection algorithm based on the positive approximation (FSPA)

Input: Decision table $S = (U, C \cup D)$;

Output: One reduct *red*.

Step 1: $red \leftarrow \emptyset$; //*red* is the pool to conserve the selected attributes

Step 2: Compute $Sig^{inner}(a_k, C, D, U)$, $k \leq |C|$;

Step 3: Put a_k into *red*, where $Sig^{inner}(a_k, C, D, U) > 0$;// These attributes form the core of the given decision table

Step 4: $i \leftarrow 1$, $R_1 = red$, $P_1 = \{R_1\}$ and $U_1 \leftarrow U$;

Step 5: While $EF^{U_i}(red, D) \neq EF^{U_i}(C, D)$ Do

{Compute the positive region of positive approximation $POS_{P_i}^U(D)$,

$U_i = U - POS_{P_i}^U(D)$,

$i \leftarrow i + 1$,

$red \leftarrow red \cup \{a_0\}$, where $Sig^{outer}(a_0, red, D, U_i) = max\{Sig^{outer}(a_k, red, D, U_i), a_k \in C - red\}$,

$R_i \leftarrow R_i \cup \{a_0\}$,

$P_i \leftarrow \{R_1, R_2, \cdots, R_i\}$ };

Step 6: return *red* and end.

Computing the significance measure of an attribute $Sig^{inner}(a_k, C, D, U)$ is one of key steps in FSPA, and its time complexity is $O(|U|)$. Hence, the time complexity of computing the core in Step 2 is $O(|C||U|)$. In Step 5, we begin with the core and add an attribute with the maximal significance into the set in each stage until finding a reduct. This process is called a forward reduction algorithm whose time complexity is $O(\sum_{i=1}^{|C|} |U_i|(|C| - i + 1))$. Thus the time complexity of FSPA is $O(|U||C| + \sum_{i=1}^{|C|} |U_i|(|C| - i + 1))$. However, the time complexity of a classical heuristic algorithm is $O(|U||C| + \sum_{i=1}^{|C|} |U|(|C| - i + 1))$. Obviously, the time complexity of FSPA is much lower than that of each of classical heuristic attribute reduction algorithms. Hence, one can draw a conclusion that the general feature selection algorithm based on the positive approximation (FSPA) may largely reduce the time consumption for feature selection from decision tables.

Table 1. Data sets description

	Data sets	Samples	Features	Classes
1	Mushroom	5644	22	2
2	Tic-tac-toe	958	9	2
3	Dermatology	358	34	6
4	Kr-vs-kp	3196	36	2
5	Breast-cancer-wisconsin	683	9	2
6	Backup-large.test	376	35	19

4 Experimental Analysis

Many heuristic feature selection methods have been developed for symbolic data [9, 13-16]. The two heuristic algorithms mentioned in the above part are very representative. The objective of the following experiments is to show the time efficiencies of the proposed general framework for selecting a feature subset. The data used in the experiments are outlined in Table 2, which were all downloaded from UCI Repository of machine learning databases.

In this section, in order to compare the above two representative feature selection algorithms (PR, SCE) with the modified ones, we employ six UCI data sets from Table 2 to verify the performance of time reduction of the modified algorithms, which are all symbolic data. For uniform treatment of all data sets, we remove the objects with missing values.

For any heuristic feature selection algorithm in rough set theory, the computation of classification is the first key step. For convenient comparison, we still adopt original classification algorithm with the time complexity $O(|C||U|^2)$. In what follows, we apply each of the original algorithms along with its modified version for searching attribute reducts. To distinguish the computational times, we divide each of these six data sets into twenty parts of equal size. The first part is regarded as the 1st data set, the combination of the first part and the second part is viewed as the 2nd data set, the combination of the 2nd data set and the third part is regarded as the 3rd data set, $\cdots$, the combination of all twenty parts is viewed as the 20th data set. These data sets can be used to calculate time used by each of the original feature selection algorithms and the corresponding modifications one and show it vis-a-vis the size of universe. These algorithms are run on a personal computer with Windows XP with Pentium(R) D 3.4GHz processor and 1.00GB memory. The software being used is Microsoft Visual Studio 2005 and Visual $C\#$.

In the sequence of experiments, we compare PR with FSPA-PR on the six real world data sets shown in Table 2. To be concise, we only display the experimental results of the first two data sets in Table 2, which are also shown in Figures 1, 2. In each of these figures, the x-coordinate pertains to the size of the data set (the 20 data sets starting from the smallest one), while the y-coordinate concerns the computing time. Comparison SCE with FSPA-SCE on the first two data sets are shown in Figures 3, 4.

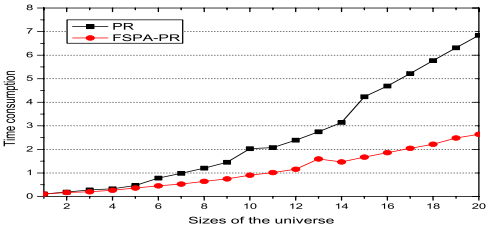

Fig. 1. Times of PR and FSPA-PR versus the size of data (data set Dermatology)

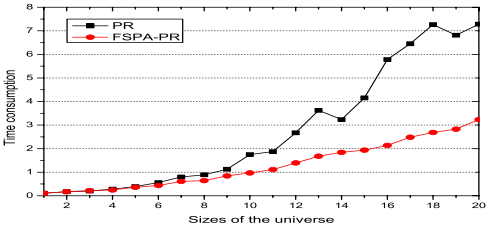

Fig. 2. Times of PR and FSPA-PR versus the size of data (data set Backup-large)

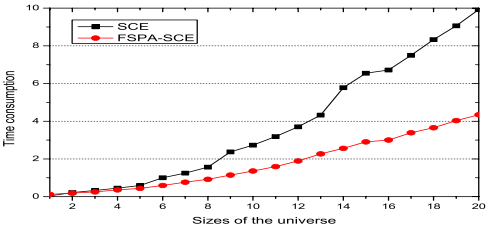

Fig. 3. Times of SCE and FSPA-SCE versus the size of data (data set Dermatology)

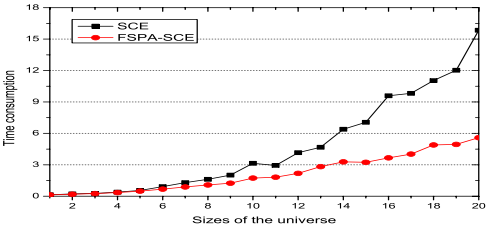

Fig. 4. Times of SCE and FSPA-SCE versus the size of data (data set Backup-large)

It is easy to note from Figures 1-4 that the computing time of each of these two algorithms increases with the increase of the size of data. As one of the important advantages of the FSPA, as shown in Figures 1-4, we see that the modified algorithms are much more faster than their original counterparts. Furthermore the differences are profoundly larger when the size of the data set increases.

5 Conclusions

To overcome the limitations of the existing feature selection schemes, in this study, based on the positive approximation, a general heuristic feature selection algorithm (FSPA) has been presented. Two representative heuristic feature selection algorithms encountered in rough set theory have been revised and modified. Note that each of the modified algorithms can choose the same feature subset as the original feature selection algorithm. Experimental studies pertaining to six UCI data sets show that the modified algorithms can largely reduce computing time of feature selection while producing the same results as those coming from the original methods. The results show that the feature selection based on the positive approximation can efficiently select a feature subset.

Acknowledgements. This work was supported by the national natural science foundation of China (No. 60773133, No. 70471003), the national high technology research and development program (No. 2007AA01Z165).

References

1. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Science 46, 39–59 (1993)
2. Wu, W.Z., Zhang, M., Li, H.Z., Mi, J.S.: Knowledge reduction in random information systems via Dempster-Shafer theory of evidence. Information Sciences 174, 143–164 (2005)
3. Kryszkiewicz, M.: Comparative study of alternative type of knowledge reduction in inconsistent systems. International Journal of Intelligent Systems 16, 105–120 (2001)
4. Li, D.Y., Zhang, B., Leung, Y.: On knowledge reduction in inconsistent decision information systems. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems 12(5), 651–672 (2004)
5. Mi, J.S., Wu, W.Z., Zhang, W.X.: Comparative studies of knowledge reductions in inconsistent systems. Fuzzy Systems and Mathematics 17(3), 54–60 (2003)
6. Skowron, A.: Extracting laws from decision tables: a rough set approach. Computational Intelligence 11, 371–388 (1995)
7. Qian, Y.H., Liang, J.Y., Dang, C.Y.: Interval ordered information systems. Computers & Mathematics with Applications 56, 1994–2009 (2008)
8. Shao, M.W., Zhang, W.X.: Dominance relation and rules in an incomplete ordered information system. International Journal of Intelligent Systems 20, 13–27 (2005)
9. Hu, X.H., Cercone, N.: Learning in relational databases: a rough set approach. International Journal of Computational Intelligence 11(2), 323–338 (1995)
10. Hu, Q.H., Xie, Z.X., Yu, D.R.: Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recognition 40, 3509–3521 (2007)
11. Hu, Q.H., Yu, D.R., Xie, Z.X.: Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognition Letters 27(5), 414–423 (2006)
12. Liang, J.Y., Chin, K.S., Dang, C.Y., Yam Richid, C.M.: A new method for measuring uncertainty and fuzziness in rough set theory. International Journal of General Systems 31(4), 331–342 (2002)

13. Liang, J.Y., Xu, Z.B.: The algorithm on knowledge reduction in incomplete information systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(1), 95–103 (2002)
14. Qian, Y.H., Liang, J.Y.: Combination entropy and combination granulation in rough set theory. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16(2), 179–193 (2008)
15. Slezak, D.: Approximate entropy reducts. Fundamenta Informaticae 53(3-4), 365–390 (2002)
16. Wang, G.Y., Yu, H., Yang, D.C.: Decision table reduction based on conditional information entropy. Chinese Journal of Computers 25(7), 759–766 (2002)
17. Wang, G.Y., Zhao, J., An, J.J.: A comparative study of algebra viewpoint and information viewpoint in attribute reduction. Fundamenta Informaticae 68(3), 289–301 (2005)
18. Wu, S.X., Li, M.Q., Huang, W.T., Liu, S.F.: An improved heuristic algorithm of attribute reduction in rough set. Journal of Systems Science and Information 2(3), 557–562 (2004)
19. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory. In: Knowledge Engineering and Problem Solving. Kluwer, Dordrecht (1991)
20. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177, 3–27 (2007)
21. Qian, Y.H., Liang, J.Y., Dang, C.Y.: Converse approximation and rule extration from decision tables in rough set theory. Computers & Mathematics with Applications 55, 1754–1765 (2008)