Feature selection for large-scale data sets in GrC

Jiye Liang
Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education,
School of Computer and Information Technology, Shanxi University,
Taiyuan, China
Email: ljy@sxu.edu.cn

***Abstract*—Granular computing, as an emerging computational and mathematical theory which describes and processes uncertain, vague, incomplete, and mass information, has been successfully used in knowledge discovery. At present, granular computing faces the challenges of consuming a huge amount of computational time and memory space in dealing with large-scale and complicated data sets. Feature selection, a common technique for data preprocessing in many areas such as pattern recognition, machine learning and data mining, is of great importance. This paper focuses on efficient feature selection algorithms for large-scale data sets and dynamic data sets in granular computing.**

***Keywords*-Large-scale data sets; dynamic data sets; feature selection; rough set theory; granular computing**

I. INTRODUCTION

Granular computing (GrC) has played an important role in information granulation of human reasoning [15], [30]. It is motivated by the practical needs for simplification, clarity, low cost, approximation, and tolerance of uncertainty [31], [28]. Three basic issues in GrC are information granulation, organization, and causation. Many models and methods under these issues have been proposed, and appeared in many related fields such as interval analysis, rough set theory, cluster analysis, machine learning and data compression [16], [29]. Feature selection, a common technique for data preprocessing in many areas including machine learning, cluster analysis, pattern recognition and data mining, has hold great significance [2], [9], [17]. With the rapid development of information technology, feature selection faces the challenges of consuming a huge amount of computational time and memory space in dealing with large-scale and complicated data sets.

Among various approaches to select useful features, a special theoretical framework is Pawlak's rough set model [20], [21]. One can use rough set theory to select a subset of features that is most suitable for a given recognition problem [4], [5], [7], [18], [25]. In this paper, based on rough set theory, an accelerator is developed to improve the time efficiency of a heuristic search process [22]. Experiments show that accelerated algorithms outperform their original counterparts. However, for the very large-scale data sets, accelerated computational time is still very long. To overcome this deficiency, we propose an efficient rough feature selection algorithm for large-scale data sets, which is stimulated from a multi-granulation view [11]. Experiments indicate that, this algorithm can find a valid feature subset and is more efficient than the accelerated algorithms. In addition, for dynamic data sets, an efficient group incremental rough feature selection algorithm based on information entropy is also developed in this paper. When multiple objects are added to a decision table, the algorithm aims to find a new feature subset in a much shorter time. Experiments show that the algorithm is effective and efficient.

II. PRELIMINARY KNOWLEDGE

The information entropy from classical thermodynamics is used to measure out-of-order degree of a system. Information entropy is introduced in rough set theory to measure uncertainty of a given data set, which have been widely applied to devise heuristic feature selection algorithms [11], [13], [22], [23], [26]. Shannon's entropy [24], complementary entropy [12], [13] and combination entropy [23] are three representative entropies which have been mainly used to construct feature selection algorithms in rough set theory. The definitions of these three entropies are defined as follows.

Definition 1: Let $S = (U, C \cup D)$ be a decision table and $B \subseteq C$. Then one can obtain the condition partition $U/B = \{X_1, X_2, \cdots, X_m\}$ and decision partition $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. Based on these partitions, Shannon's conditional entropy of B relative to D is defined as

$$H(D|B) = -\sum_{i=1}^{m} \frac{|X_i|}{|U|} \sum_{j=1}^{n} \frac{|X_i \cap Y_j|}{|X_i|} \log(\frac{|X_i \cap Y_j|}{|X_i|}).$$

Definition 2: Let $S = (U, C \cup D)$ be a decision table and $B \subseteq C$. Then, one can obtain the condition partition $U/B = \{X_1, X_2, \cdots, X_m\}$ and decision partition $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. Based on these partitions, the complementary conditional entropy of B relative to D is defined as

$$E(D|B) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|Y_i \cap X_j|}{|U|} \frac{|Y_i^c \cap X_j^c|}{|U|},$$

978-1-4673-2311-6/12/$31.00 ©2012 IEEE

where Y_i^c and X_j^c are complement sets of Y_i and X_j respectively.

Definition 3: Let $S = (U, C \cup D)$ be a decision table and $B \subseteq C$. Then, one can obtain the condition partition $U/B = \{X_1, X_2, \cdots, X_m\}$ and decision partition $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. Based on these partitions, the combination conditional entropy of B relative to D is defined as

$$CE(D|B) = \sum_{i=1}^{m}(\frac{|X_i|}{|U|}\frac{C^2_{|X_i|}}{C^2_{|U|}} - \sum_{j=1}^{n}\frac{|X_i \cap Y_j|}{|U|}\frac{C^2_{|X_i \cap Y_j|}}{C^2_{|U|}}).$$

where $C^2_{|X_i|}$ denotes the number of pairs of objects which are not distinguishable from each other in the equivalence class X_i.

For convenience, a uniform notation $ME(D|B)$ is introduced to denote the above three entropies. For example, if one adopts Shannon's conditional entropy to define the attribute significance, then $ME(D|B) = H(D|B)$. In [11], [22], [26], the attribute significance is defined as follows (See Definitions 4-5).

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

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

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

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

III. An accelerator for feature selection

In rough set theory, feature selection (also called attribute reduction) aims to retain the discriminatory power of original features. It plays an important role in many areas including pattern recognition, machine learning and data mining. In the last two decades, many techniques of attribute reduction have been developed. Skowron proposed a discernibility matrix approach to obtain all attribute reducts of an information system [25]. Kryszkiewicz proposed an approach to computing the minimal set of attributes that functionally determine a decision attribute[10]. In addition, to conceptualize and analyze various types of data, researchers have generalized Pawlak's classic rough set model. The generalized rough set models include neighborhood rough set model [14], fuzzy rough model [1], decision-theoretic rough set model [27], variable precision rough set model (VPRS)[32] and dominance rough set model [3]. Attribute reduction based on these generalizations was also redefined. To improve the time efficiency, researchers have also developed many heuristic attribute reduction algorithms which can generate a single reduct from a given table[11], [22], [26]. However, above algorithms are computationally time-consuming for large-scale data sets. To overcome this deficiency, we developed an accelerated framework, which can be used to accelerate a heuristic process of feature selection[22].

Theorem 1: Let $S = (U, C \cup D)$ be a decision table, $X \subseteq U$ and $P = \{R_1, R_2, \cdots, R_n\}$ be a family of attribute sets with $R_1 \preceq R_2 \preceq \cdots \preceq R_n$ ($R_i \in 2^C$). Given $P_i = \{R_1, R_2, \cdots, R_i\}$, we have

$$POS^U_{P_{i+1}}(D) = POS^U_{P_i}(D) \cup POS^{U_{i+1}}_{R_{i+1}}(D),$$

where $U_1 = U$ and $U_{i+1} = U - POS^U_{P_i}(D)$.

Theorem 1 implies that the target decision D can be positively approximated by using granulation orders P on the gradually reduced universe. This mechanism motivates the idea of the accelerator for improving the computing performance of a heuristic attribute reduction algorithm.

Based on Theorem 1, we concentrate on the rank preservation of significance measures of attributes, which can be studied in the following theorem.

Theorem 2: Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $U' = U - POS^U_B(D)$. For $\forall a, b \in C - B$, if $Sig^{outer}(a, B, D, U) \geq Sig^{outer}(b, B, D, U)$, then $Sig^{outer}(a, B, D, U') \geq Sig^{outer}(b, B, D, U')$.

Based on Theorem 1 and 2, a common accelerated feature selection algorithm is developed in the following.

Algorithm 1. A common accelerated feature selection algorithm (FSA)

Input: Decision table $S = (U, C \cup D)$;
Output: One reduct red.

Step 1: $red \leftarrow \emptyset$;
Step 2: Compute $Sig^{inner}(a_k, C, D, U)$, $k \leq |C|$;
Step 3: Put a_k into red, where $Sig^{inner}(a_k, C, D, U) > 0$;
Step 4: $i \leftarrow 1$, $R_1 = red$, $P_1 = \{R_1\}$ and $U_1 \leftarrow U$;
Step 5: While $ME^{U_i}(red, D) \neq ME^{U_i}(C, D)$ Do
{Compute the positive region $POS^U_{P_i}(D)$,
$U_i = U - POS^U_{P_i}(D)$,
$i \leftarrow i + 1$,
$red \leftarrow red \cup \{a_0\}$, where $Sig^{outer}(a_0, red, D, U_i) = max\{Sig^{outer}(a_k, red, D, U_i), a_k \in C - red\}$,
$R_i \leftarrow R_{i-1} \cup \{a_0\}$,
$P_i \leftarrow \{R_1, R_2, \cdots, R_i\}$ };
Step 6: return red and end.

This accelerator provides an efficient accelerated strategy for heuristic feature selection based on rough set theory. Note that each of the modified algorithms can choose the same attribute reduct as its original version, which possesses the same classification accuracy. Experiments carried out on

nine UCI data sets show that these accelerated algorithms outperform their original counterparts. Furthermore, as the size of data set increases, the efficiency of accelerated algorithms is more and more obvious.

IV. An efficient rough feature selection algorithm with a multi-granulation view

In this section, according to the idea of using samples to estimate the totality, we develop a highly efficient rough feature selection algorithm from a multi-granulation view [11]. We remark that there are three key problems should be considered. The first problem is selecting sub-tables from the large-scale one, the second one is finding reduct on sub-tables, and the last one is the fusing the all the reducts on sub-tables together. A sub-table can be considered as a single small granularity; and one can estimate on this small granularity the reduct of the original table. In the process of selecting small granularity, one of the most important issues is how to determine the size of a small granularity. With the use of some concepts and formulas in statistics, we first introduce a familiar approach to determine sample size [6].

Let S be a data table (the original large-scale data table) and let the size of S be denoted by N. Then, the sample size M' is defined as [6]

$$M' = \frac{Z^2 \times \sigma^2}{E^2},$$

where, σ is the standard deviation on S, Z is $Z-$statistic under confidence intervals, and E is an acceptable tolerance error. If M' is larger than 5% of the overall size, it should be adjusted as [6]

$$M = \frac{M'N}{M'+N}.$$

Because decision tables in rough set theory are categorical data, we introduce the coefficient of unalikeability u to replace the standard deviation σ [8].

$$u = \frac{\sum_{i=1}^{|U|}\sum_{j=1}^{|U|} c(x_i, x_j)}{|U|^2},$$

where $x_i, x_j \in U$, and $c(x_i, x_j) = \begin{cases} 1, & x_i \neq x_j \\ 0, & x_i = x_j \end{cases}$. Then, we expand the definition of $c(x_i, x_j)$ into multi-dimensional data, which is denoted by $c_m(x_i, x_j)$.

$$c_m(x_i, x_j) = \sum_{k=1}^{|C|} \delta(a_k(x_i), a_k(x_j)), \tag{1}$$

with the function δ being given by

$$\delta(a_k(x_i), a_k(x_j)) = \begin{cases} 1, & a_k(x_i) \neq a_k(x_j) \\ 0, & a_k(x_i) = a_k(x_j). \end{cases}$$

Hence, for a decision table S, the coefficient of unalikeability can be redefined as

$$u_1 = \frac{\sum_{i=1}^{|U|}\sum_{j=1}^{|U|} c_m(x_i, x_j)}{|U|^2}, \tag{2}$$

and sample size is redefined as

$$M_1' = \frac{Z^2 \times u_1^2}{E^2}. \tag{3}$$

According to equations (1)-(3), one can determine the size of sub-table (small granularity) on a large-scale decision table as follows.

For a decision table, its reducts are directly related to its decision distribution. Thus, the decision distribution on a small granular space may also affect the estimated result. To ensure the decision distribution on a small granular space is close to the large-scale one, an algorithm for selecting sub-table is developed as follows.

Algorithm 2. An algorithm for selecting small granularity on a large-scale decision table

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

Output: n small granularity $S_j = (U_j, C \cup D)$ $(j = 1, 2, \cdots, n)$.

Step 1: Compute the size of small granularity M_1 (according to Algorithm 1);

Step 2: Compute $U/D = \{D_1, D_2, \cdots, D_r\}$, and the decision attribute value proportions $p_i = |D_i|/|U| (i = 1, 2, \cdots, r)$;

Step 3: Compute the numbers of each decision attribute value in the small granularity $m_i = [M_1 \times p_i] (i = 1, 2, \cdots, r)$ (function $[\cdot]$ is the rounding function);

Step 4: Select the first granularity S_1 on U, $U_1 \leftarrow \emptyset$:
for $(i = 1;\ i \leq r\ ;\ i++)$
{ Select m_i objects from D_i randomly, which is denoted by X;
$U_1 \leftarrow U_1 \cup X$;
}

Step 5: Select granularity S_j repeatedly, $j \leftarrow 2$:
Given threshold α $(0 < \alpha < 1)$;
while $(|U - \bigcup_{k=1}^{j-1} U_k| > M_1)$
{ $U_j \leftarrow \emptyset$;
Step 5.1: Select αM objects from table S_{j-1}:
{ Compute $U_{j-1}/D = \{D_1', D_2', \cdots, D_r'\}$;
for $(i = 1;\ i \leq r\ ;\ i++)$
{ Select αm_i objects from D_i' randomly, which is denoted by X';
$U_j \leftarrow U_j \cup X'$;
}
}
Step 5.2: $U'' = U - \bigcup_{k=1}^{j-1} U_k$, and select $(1-\alpha)M$ objects from U'':

{ Compute $U''/D = \{D''_1, D''_2, \cdots, D''_r\}$;
for $(i = 1; i \leq r\ ; i++)$
{ Select $(1-\alpha)m_i$ objects from D_i randomly, which is denoted by X'';
$U_j \leftarrow U_j \cup X''$;
}
}
$j \leftarrow j + 1$;
}

Step 6: $n \leftarrow j - 1$ and end.

For a large-scale decision table, from Algorithm 1 and Algorithm 2, we obtain a group of estimates to the reduct. By fusing together these estimates, we get a valid feature subset for the large-scale decision table. An efficient rough feature selection algorithm is proposed as follows.

Algorithm 3. An efficient rough feature selection algorithm(EFSA)

Input: A large-scale decision table $S = (U, C \cup D)$
Output: Feature subset Red

Step 1: Select n small granularity according to Algorithm 2 from S: $S_1 = (U_1, C \cup D)$, $S_2 = (U_2, C \cup D)$, $\cdots$, $S_n = (U_n, C \cup D)$;
Step 2: $Red \leftarrow \emptyset$;
$for\ (j = 1; j \leq n; j++)$
{ Compute the attribute reduct red_j of table $S_j = (U_j, C \cup D)$ using Algorithm 1;
$Red = Red \cup red_j$;
}
Step 3: return Red and end.

Algorithm 3 introduces a framework that is dividing and fusing on a large-scale data set. Based on this framework, by employing other reduction algorithms to find reduct on a sub-table, one can also construct appropriate efficient algorithms. Experiments indicate that, compared with the accelerated algorithm (FSA), algorithm EFSA can find a valid feature subset in a much shorter time.

In addition, algorithm EFSA can also handle some large-scale data sets that are very difficult to deal with on a PC because of the high computational time. This is a very important contribution of this algorithm. In [11], two data sets (*Poker-hand* and *Covtype*) have been employed to illustrate this conclusion. The sizes of these two data sets are 1025010 and 581012, and the numbers of features of them are 10 and 54. By using existing reduction algorithms, these two data sets are too large in scale to get the feature subset within 100 hours. However, the computational time of algorithm EFSA on these two data sets are just 0.35 hours and 6.6 hours. Therefore, this algorithm has made an important contribution to deal with large-scale data sets in applications.

V. A GROUP INCREMENTAL APPROACH TO FEATURE SELECTION

In practice, the rapid development of data processing tools has led to the high speed of dynamic data updating. Thus many real data in applications may be generated in groups instead of one by one. To address this issue, this section introduces a group incremental feature selection algorithm, which aims to deal with multiple objects at a time instead of repeatedly.

Given a decision table, when multiple objects are added, Theorems 3-5 introduce the group incremental mechanisms of three entropies respectively.

Theorem 3: Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_m\}$ and $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. The Shannon's conditional entropy of D with respect to B is $H_U(D|B)$. Suppose that U_X is an incremental object set, $U_X/B = \{M_1, M_2, \cdots, M_{m'}\}$ and $U_X/D = \{Z_1, Z_2, \cdots, Z_{n'}\}$. We assume that $(U \cup U_X)/B = \{X'_1, X'_2, \cdots, X'_k, X_{k+1}, X_{k+2}, \cdots, X_m, M_{k+1}, M_{k+2}, \cdots, M_{m'}\}$ and $(U \cup U_X)/D = \{Y'_1, Y'_2, \cdots, Y'_l, Y_{l+1}, Y_{l+2}, \cdots, Y_n, Z_{l+1}, Z_{l+2}, \cdots, Z_{n'}\}$. Then, the new Shannon's conditional entropy becomes

$$H_{U\cup U_X}(D|B) = \frac{1}{|U| + |U_X|}(|U|H_U(D|B) + |U_X|H_{U_X}(D|B)) - \Delta,$$

where $\Delta = \sum_{i=1}^{k}(\sum_{j=1}^{l}(\frac{|X_i \cap Y_j|}{|U|+|U_X|} log \frac{|X_i||X'_i \cap Y'_j|}{|X'_i||X_i \cap Y_j|} + \frac{|M_i \cap Z_j|}{|U|+|U_X|} log \frac{|M_i||X'_i \cap Y'_j|}{|X'_i||M_i \cap Z_j|}) + \sum_{j=l+1}^{n} \frac{|X_i \cap Y_j|}{|U|+|U_X|} log \frac{|X_i|}{|X'_i|} + \sum_{j=l+1}^{n'} \frac{|M_i \cap Z_j|}{|U|+|U_X|} log \frac{|M_i|}{|X'_i|})$.

Theorem 4: Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_m\}$ and $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. The complementary conditional entropy of D with respect to B is $E_U(D|B)$. Suppose that U_X is an incremental object set, $U_X/B = \{M_1, M_2, \cdots, M_{m'}\}$ and $U_X/D = \{Z_1, Z_2, \cdots, Z_{n'}\}$. We assume that $(U \cup U_X)/B = \{X'_1, X'_2, \cdots, X'_k, X_{k+1}, X_{k+2}, \cdots, X_m, M_{k+1}, M_{k+2}, \cdots, M_{m'}\}$ and $(U \cup U_X)/D = \{Y'_1, Y'_2, \cdots, Y'_l, Y_{l+1}, Y_{l+2}, \cdots, Y_n, Z_{l+1}, Z_{l+2}, \cdots, Z_{n'}\}$. Then, the new complementary conditional entropy becomes

$$E_{U\cup U_X}(D|B) = \frac{1}{(|U \cup U_X|)^2}(|U|^2 E_U(D|B) + |U_X|^2 E_{U_X}(D|B)) + \Delta,$$

where $\Delta = \sum_{i=1}^{k}(\sum_{j=1}^{l} \frac{|X_i \cap Y_j||M_i - Z_j| + |M_i \cap Z_j||X_i - Y_j|}{(|U \cup U_X|)^2} + \sum_{j=l+1}^{n} \frac{|X_i \cap Y_j||M_i|}{(|U \cup U_X|)^2} + \sum_{j=l+1}^{n'} \frac{|M_i \cap Z_j||X_i|}{(|U \cup U_X|)^2})$.

Theorem 5: Let $S = (U, C \cup D)$ be a decision table, $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_m\}$ and $U/D = \{Y_1, Y_2, \cdots, Y_n\}$. The combination conditional entropy of D with respect to B is $E_U(D|B)$. Suppose that U_X is

an incremental object set, $U_X/B = \{M_1, M_2, \cdots, M_{m'}\}$ and $U_X/D = \{Z_1, Z_2, \cdots, Z_{n'}\}$. We assume that $(U \cup U_X)/B = \{X'_1, X'_2, \cdots, X'_k, X_{k+1}, X_{k+2}, \cdots, X_m, M_{k+1}, M_{k+2}, \cdots, M_{m'}\}$ and $(U \cup U_X)/D = \{Y'_1, Y'_2, \cdots, Y'_l, Y_{l+1}, Y_{l+2}, \cdots, Y_n, Z_{l+1}, Z_{l+2}, \cdots, Z_{n'}\}$. Then, the new combination conditional entropy becomes

$$CE_{U\cup U_X}(D|B) = \frac{1}{(|U|+|U_X|)^2(|U|+|U_X|-1)}\cdot$$

$$(|U|^2(|U|-1)CE_U(D|B)+|U_X|^2(|U_X|-1)CE_{U_X}(D|B))+\Delta,$$

where $\Delta = \sum_{i=1}^{k}(\frac{|X_i||M_i|(3|X_i|+3|M_i|-2)}{(|U|+|U_X|)^2(|U|+|U_X|-1)} - \sum_{j=1}^{l}\frac{|X_i\cap Y_j||M_i\cap Z_j|(3|X_i\cap Y_j|+3|M_i\cap Z_j|-2)}{(|U|+|U_X|)^2(|U|+|U_X|-1)})$.

Based on incremental mechanisms of the three entropies, an efficient group incremental feature selection algorithm is introduced in the following.

Algorithm 5. A group incremental algorithm for reduct computation ($GIARC$)

Input: A decision table $S = (U, C \cup D)$, reduct RED_U on U, and the new object set U_X

Output: Reduct $RED_{U\cup U_X}$ on $U \cup U_X$

Step 1 : $B \leftarrow RED_U$. Compute $U/B = \{X_1^B, X_2^B, \cdots, X_m^B\}$, $U/C = \{X_1^C, X_2^C, \cdots, X_s^C\}$, $U_X/B = \{M_1^B, M_2^B, \cdots, M_{m'}^B\}$ and $U_X/C = \{M_1^C, M_2^C, \cdots, M_{s'}^C\}$.

Step 2 : Compute $(U \cup U_X)/B = \{X_1'^B, X_2'^B, \cdots, X_k'^B, X_{k+1}^B, X_{k+2}^B, \cdots, X_m^B, M_{k+1}^B, M_{k+2}^B, \cdots, M_{m'}^B\}$ and $(U\cup U_X)/C = \{X_1'^C, X_2'^C, \cdots, X_{k'}'^C, X_{k'+1}^C, X_{k'+2}^C, \cdots, X_s^C, M_{k'+1}^C, M_{k'+2}^C, \cdots, M_{s'}^C\}$.

Step 3 : If $k = 0$ and $k' = 0$, turn to *Step 4*; else turn *Step 5*.

Step 4 : Compute $ME_{U_X}(D|B)$ and $ME_{U_X}(D|C)$. If $ME_{U_X}(D|B) = ME_{U_X}(D|C)$, turn to *Step 7*; else turn to *Step 5*.

Step 5 : $while\ ME_{U\cup U_X}(D|B) \neq ME_{U\cup U_X}(D|C)\ do$
{ For each $a \in C - B$, compute $Sig^{outer}_{U\cup U_X}(a, B, D)$;
Select $a_0 = max\{Sig^{outer}_{U\cup U_X}(a, B, D),\ a \in C - B\}$;
$B \leftarrow B \cup \{a_0\}$.
}

Step 6 : For each $a \in B\ do$
{ Compute $Sig^{inner}_{U\cup U_X}(a, B, D)$;
If $Sig^{inner}_{U\cup U_X}(a, B, D) = 0$, then $B \leftarrow B - \{a\}$.
}

Step 7 : $RED_{U\cup U_X} \leftarrow B$, return $RED_{U\cup U_X}$ and end.

When multiple objects are added to the basic data set, theoretical analysis and experimental results have shown that this algorithm is effective and efficient. In particular, with the number of added data increasing, the efficiency of the group incremental feature selection algorithm become more and more obvious.

VI. Conclusions and future work

At present, feature selection for large-scale data sets is still a challenging issue in the field of artificial intelligence. In this paper, for large-scale data sets, we developed an accelerator for heuristic feature selection and an efficient rough feature selection algorithm. In addition, an efficient group incremental feature selection algorithm was also introduced for dynamic data sets.

Based on the results in this paper, some further investigations are as follows.

- Information fusion of multi-data sets or multi-granularity.
- Uncertainty measures for generalized rough set models.
- Feature selection for dynamic data sets under generalized rough set models.

Acknowledgment

This work was supported by National Natural Science Fund of China (Nos. 71031006), National Key Basic Research and Development Program of China(973) (No. 2011CB311805).

References

[1] Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems. 17, 191-209 (1990)

[2] Guyon, I., Elisseeff, A.: An introduction to variable feature selection. Machine Learning Research. 3, 1157-1182 (2003)

[3] Greco, S., Matarazzo, B., Slowinski, R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research. 129, 1-47 (2001)

[4] Hu, X.H., Cercone, N.: Learning in relational databases: a rough set approach. International Journal of Computational Intelligence. 11, 323-338 (1995)

[5] 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)

[6] Jia, J.P.: Principles of Statistics(the Fourth Edition), Beijing: China Renmin University Publishing (2009)

[7] Jensen, R., Shen, Q.: Fuzzy-rough sets assisted attribute selection. IEEE Transactions on Fuzzy Systems. 15, 73-89 (2007)

[8] G. Kader, M. Perry, Variability for Categorical Variables, Journal of Statistics Education 15 (2) (2007).

[9] Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence. 97, 273-324 (1997)

[10] Kryszkiewicz, M., Lasek, P.: FUN: fast discovery of minimal sets of attributes functionally determining a decision attribute. Transactions on Rough Sets. 9, 76-95 (2008)

[11] Liang, J.Y., Wang, F., Dang, C.Y., Qian, Y.H.: An efficient rough feature selection algorithm with a multi-granulation view. International Journal of Approximate Reasoning. 53, 912-926 (2012)

[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, 331-342 (2002)

[13] Liang, J.Y., Shi, Z.Z.: The information entropy, rough entropy and knowledge granulation in rough set theory. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 12, 37-46 (2004)

[14] Lin, T.Y.: Neighborhood systems and approximation in database and knowledge base systems, in: Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems, Poster Session, October, 12C15, 75-86 (1989)

[15] Lin, T.Y.: Data mining and machine oriented modeling: a granular computing approach. Applied Intelligence. 13, 113-124 (2000)

[16] Lin, T.Y.: Granular computing I: The concept of granulation and its formal model. International Journal of Granular Computing, Rough Sets and Intelligent Systems. 1, 21-42 (2009)

[17] Li, T. R., Ruan, D., Geert, W., Song, J., Xu, Y.: A rough sets based characteristic relation approach for dynamic attribute generalization in data mining. Knowledge-Based Systems. 20, 485-494 (2007)

[18] Liu, Q.: Rough sets and rough reasoning. Beijing: Science Press (2001)

[19] Mi, J.S., Wu, W.Z., Zhang, X.W.: Approaches to knowledge reduction based on variable precision rough set model. Information Sciences. 159, 255-272 (2004)

[20] Pawlak, Z.: Rough Sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Boston (1991)

[21] Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences. 177, 3-27 (2007)

[22] Qian, Y.H., Liang, J.Y., Pedrycz, W., Dang, C.Y.: Positive approximation: an accelerator for attribute reduction in rough set theory. Artificial Intelligence. 174, 597-618 (2010)

[23] 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, 179-193, (2008)

[24] Shannon, C.E.: A mathematical theory of communication. The Bell System Technology Journal. 21, 373-423, 623-656 (1948)

[25] Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems, In: R Slowiński(Eds), Intelligent decision support, handbook of applications and advances of the rough sets theory. Kluwer Academic Publisher, Dordrecht (1992)

[26] Wang, G.Y., Yu, H., Yang, D.C.: Decision table reduction based on conditional information entropy. Chinese Journal of Computer. 25, 759-766 (2002)

[27] Yao, Y.Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. Information Sciences. 178, 3356-3373 (2008)

[28] Yao, Y.Y.: Information granulation and rough set approximation. International Journal of Intelligent Systems. 16, 87-104 (2001)

[29] Yao, J.T., Yao, Y.Y.: Induction of classification rules by granular computing. Lecture Notes in Computer Science. 2475, 331-338 (2002)

[30] Zadeh. L.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy sets and systems. 90, 111-127 (1997)

[31] Zadeh. L.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information / intelligent systems. Soft Computing. 2, 23-25, (1998)

[32] Ziarko, W.: Variable precision rough set model. Journal of Computer and System Science. 46, 39-59 (1993)