Applied Mathematics and Nonlinear Sciences

<https://www.sciendo.com>

Attribute Reduction Method Based on Sample Extraction and Priority

Biqing Wang^{1,2†}¹Endicott College of International Studies, Woosong University, Daejeon 300718, Republic of Korea²School of Mathematics and Computer, Tongling University, Tongling 244000, Anhui, China

Submission Info

Communicated by Juan Luis García Guirao

Received November 28th 2020

Accepted January 31st 2021

Available online April 12th 2021

Abstract

Attribute reduction is a key issue in the research of rough sets. Aiming at the shortcoming of attribute reduction algorithm based on discernibility matrix, an attribute reduction method based on sample extraction and priority is presented. Firstly, equivalence classes are divided using quick sort for computing compressed decision table. Secondly, important samples are extracted from compressed decision table using iterative self-organizing data analysis technique algorithm (ISODATA). Finally, attribute reduction of sample decision table is conducted based on the concept of priority. Experimental results show that the attribute reduction method based on sample extraction and priority can significantly reduce the overall execution time and improve the reduction efficiency.

Keywords: rough sets; sample extraction; priority; attribute reduction; algorithm

1 Introduction

Rough sets theory is a mathematical tool to deal with vague and uncertain knowledge presented by Pawlak in 1982 [1]. It can mine the hidden knowledge from the mass data and has been successfully applied in fields such as data mining, pattern recognition, decision support system [2, 3], etc. In recent years, it is also used in big data [4], information sharing [5], and other fields. Attribute reduction is one of the core contents of the Rough sets theory. It has been proved to be a NP-Hard problem to find all reductions or a minimal reduction in Rough Sets [6]. So, the common method is to adopt the heuristic algorithm. Normally there are two main classes, the algorithms that are based on the discernibility matrices and the algorithms that are based on positive region.

Because the attribute reduction algorithms based on discernibility matrix is succinct and easy to understand, it has attracted a great deal of attention from scholars [7]. But in this kind of algorithms the discernibility matrix need be constructed and there are a lot of repeat elements in the discernibility matrix. In disposing big datasets,

[†]Corresponding author.Email address: ycwbq@163.com

this kind of algorithms not only cost a great deal of memory space, but also waste much computer time in reduction. Because the discernibility matrix need not be constructed, the attribute reduction algorithms based on positive region attract more and more attention from scholars [8]. But the time complexity of this kind of algorithms are still far from satisfactory in processing large data sets [9].

If unnecessary elements are removed from decision table and only essential elements are extracted for attribute reduction, high efficiency of reduction algorithms can be reached. Hart first puts forward the concept of the sample extraction. There are some typical algorithms, such as the CNN [10], the RNN [11], etc. Numerous sample extraction algorithms can be broadly divided into two kinds: classification competence enhancement approaches and classification competence preservation approaches [12]. At present, many scholars have done lots of correlative investigations to the sample extraction.

In the paper [9], condition attributes are clustered firstly. Then the cluster are randomly sampled. Finally, attribute reduction is conducted on the sample data set. The literature [13] proposes an attribute reduction algorithm that is based on sample extraction. There are three steps in this algorithm. First, important instances are selected from the set of objects. Second, discernibility matrix is constructed by the use of the selected instances. Finally, all reductions of decision table are computed. In [14], an attribute reduction algorithm based on granular computing and discernibility ability is presented. This algorithm uses stratified sampling in statistics to divide original large dataset into several sample subsets and conduct attribute reduction on each subset. In the end, weight fusion was made to get final reduction result of original large dataset according to reduction results of all the subset. In [15], it is stated that useless samples will be filtered to reduce memory space and running time during reduction calculation. The algorithm called condensation rule based on fuzzy rough sets (FRSC) is introduced in [16]. In rough sets theory, the samples near the decision boundary are often these samples that have smaller lower approximation. Based on this characteristic, FRSC uses lower approximation as a basis for sample extraction.

Most sample extraction algorithms use random sampling or sampling methods based on a certain criteria. But the reasons for that and how to achieve optimum sample effect have not been fully interpreted in theory and practice. According to the above scholars' researches, an attribute reduction method based on sample extraction and priority is presented in this paper. Because the samples near the classification boundary have great uncertainty, they can provide more information on classification and play an important role in classification. So, in this paper, the samples near the classification boundary are extracted and used for attribute reduction. Experiments illustrate the validity and efficiency of this method.

2 Proposed methodology

This paper makes some improvements on the previous attribute reduction algorithms which are based on sample extraction to compress decision table, improve sampling method and enhance the reduction efficiency. These improvements primarily reflected in the following aspects.

2.1 Data preprocessing

In some references, it is believed that sample extraction can removes redundant objects from decision table. But in fact the selected samples are far more likely to be still redundant if only the sample extraction method is used. Therefore data preprocessing must be done by using other methods before sample extraction to remove redundant objects from decision table completely. This paper proposes that U is classified according to C to solve the problem of data redundancy, namely computation U/C . That is, U is sorted by each condition attribute in order. Then, the sorted U is analysed and equivalence classes are divided. Next, non-redundant objects can be selected out. The quick sort method is used in the sorting stage.

Let $S=(U, CUD, V, f)$ be a decision table to be ordered and $U=\{u_1, u_2, \dots, u_n\}$. Let $\{u'_1, u'_2, \dots, u'_n\}$ be the set of objects which are ordered by C . H_i denotes the i -th equivalence class and H is the set of H_i . The basic

design ideas are in Algorithm 1.

Algorithm 1: Algorithm of computing U/C using quick sort.

Input: $S, U=\{u_1, u_2, \dots, u_n\}, C=\{a_1, a_2, \dots, a_k\}$.

Output: U/C .

```

for( $i=1; i \leq k; i++$ )
    sort  $U$  with quick sort by  $a_i$ ;
 $t=1; H_1 = \{u'_1\}$ ;
for( $j=2; j \leq n; j++$ )
    if( $f(u'_j, a_i) == f(u'_{j-1}, a_i)$  for  $\forall a_i$ )
         $H_t = H_t \cup \{u'_j\}$ ;
    else  $\{t++; H_t = \{u'_j\}\}$ 
return  $H$ .

```

The running time of Algorithm 1 consists of two parts: quick sort and equivalence class division, among which the main cost is quick sort. Because the total cost of $|C|$ times quick sort is $O(|C||U|\log|U|)$, the time complexity of Algorithm 1 is $O(|C||U|\log|U|)$ consequently. An example of equivalence class partition is given later. Table 1 is a redundant decision table. Among the objects, u_1 and u_3 are duplicate, and u_5 and u_6 are duplicate.

Table 1 A redundant decision table.

U	a	b	c	D
u_1	1	0	1	0
u_2	2	0	1	1
u_3	1	0	1	0
u_4	0	0	0	0
u_5	1	2	0	1
u_6	1	2	0	1
u_7	2	2	0	1

U/C is computed through Algorithm 1. First, U is sorted through condition attribute a using quick sort. After sorting, U is $\{u_4, u_3, u_1, u_5, u_6, u_2, u_7\}$. Similarly, U is sorted through other condition attributes in order. Eventually, $U/\{a, b, c\}$ is $\{\{u_7\}, \{u_6, u_5\}, \{u_4\}, \{u_3, u_1\}, \{u_2\}\}$.

On the basis of sorting, non-redundant data are chosen to form the compressed decision table. The compressed decision table is defined as follows.

Definition 1. In decision table S , $U/C = \{[u'_1]_C, [u'_2]_C, \dots, [u'_d]_C\}$, where $u'_1, u'_2, \dots, u'_d$ are certain elements in various equivalence classes. $\{u'_1, u'_2, \dots, u'_d\}$ is denoted as U' . $S' = (U', C \cup D, V', f')$ is defined as compressed decision table, where U' is the domain of S' ; V' : the set of values of all attributes in S' ; f' : the information function of S' ; f' .

According to Definition 1, generation algorithm of the compressed decision table is given below.

Algorithm 2: Generation algorithm of the compressed decision table

Input: S

Output: $S' = (U', C \cup D, V', f')$

```

 $U' = \emptyset$ ;
computing  $U/C = \{[u'_2]_C, [u'_2]_C, \dots, [u'_d]_C\}$  using Algorithm 1;
for( $i=1; i \leq d; i++$ )
     $U'_i = U' \cup \{u'_i\}$ 
return  $S'$ .

```

For Algorithm 2, time complexity is $O(|C||U|\log|U|)$. The compressed decision table of Table 1 can be got through Algorithm 2, which is shown in Table 2.

Table 2 Compressed decision table.

U	a	b	c	D
u_2	2	0	1	1
u_3	1	0	1	0
u_4	0	0	0	0
u_5	1	2	0	1
u_7	2	2	0	1

2.2 Sample extraction

After removing redundant objects in the previous step, to improve the efficiency of attribute reduction, we try to further reduce the number of objects by sampling. In some papers, instances are selected by K -Means. The value of K needs be determined artificially beforehand and cannot be changed during the whole computation process. Furthermore, the value of K often cannot be precisely estimated for large data sets with high dimensionality. But, iterative self-organising data analysis technique algorithm (ISODATA) makes improvement, which is just aiming at this question. Thus, ISODATA is chosen to extract samples in this paper. Sample extraction algorithm based on ISODATA is provided below. Let λ be the number of clusters. N_i is the number of objects in the i -th cluster. D_{ij} is the distance between object u_j in the i -th cluster and cluster centre. The average distance from all objects in the i -th cluster to cluster centre is D_i . The sampling distance is mD_i .

Algorithm 3: Sample extraction algorithm based on ISODATA.

Input: $S' = (U', \cup D, V', f')$

Output: $S'' = (U'', \cup D, V'', f'')$

$U'' = \emptyset$

cluster the objects in U' using the ISODATA;

for($i=1; i \leq \lambda; j++$)

 for($j=1; j \leq N_i; j++$)

 if($D_{ij} > mD_i$)

$U'' = U'' \cup u_{ij}$;

return S'' .

In Algorithm 3, m is a constant and its reasonable value can be determined by the following experimental research. As time complexity of ISODATA is $O(|U'|^2)$ [17] and time complexity of the later dual loop is $O((|U'|N_i \times N_i) = O(|U'|))$ time complexity of sample extraction algorithm is $O(|U'|^2) = O(|U|^2)$.

Through sample extraction, objects that have an important influence on the reduction are selected, which greatly reduces the number of objects and improves the efficiency of reduction.

2.3 Attribute reduction

After sample extraction, attribute reduction can be conducted. Some literature reduce the attributes using the discernibility matrix. However, the discernibility matrix method shows some disadvantages that lead to its lower efficiency and thus largely limit the wider applications, such as it occupies a large amount of RAM and has high time complexity. The attribute reduction algorithms based on positive region have better time and space complexity because they do not need to establish the discernibility matrix. According to the concept of positive region, an attribute reduction algorithm based on priority is put forward in this paper. Then the effect of this algorithm is illustrated by the time complexity analysis and experiments.

This paper introduces the concept of priority to measure the discernibility ability. The definition of priority is presented here.

Definition 2. For decision table S , attribute set $P \subseteq C$ divides U in k equivalence classes. The number of elements

of each equivalence classes are $|E_1|, |E_2|, \dots, |E_k|$, respectively. The priority of P is defined as

$$pri(P) = \sum_{i=1}^k \left(|E_i| \frac{|E_i|}{|U|} \right) \quad (1)$$

Particularly, $pri(\emptyset) = |U|$.

It can be seen from the above definition that the more equivalence classes are divided and the more balanced, the stronger the ability of the attribute set to distinguish the object set.

Definition 3. For decision table S , the priority of attribute a with regard to set of attributes $P \subseteq C$ is defined as $pri(P \cup \{a\})$.

Clearly, the smaller the priority is, the larger the discernibility ability of attribute set is. Theorem 1 is given on this basis.

Theorem 1. In S , for $R \subseteq C$ and $\forall a \in R$, R is an attribute reduction of C if and only if $pri(R) = pri(C)$ and $pri(R - \{a\}) > pri(C)$.

Proof. According to the above definitions, Theorem 1 holds clearly.

According to the above theoretical foundation, the attribute reduction algorithm which is based on priority is given below.

Algorithm 4: Attribute reduction algorithm based on priority

Input: $S'' = (U'', \cup D, V'', f'')$

Output: an attribute reduction R

$R \leftarrow \emptyset$;

calculate $pri(C)$;

while($pri(R) > pri(C)$)

 { calculate priority of each attribute in C with regard to R ;

 choose the attribute a that has the smallest priority;

$R = R \cup \{a\}$, $C = C - \{a\}$;

return R .

For Algorithm 4, time complexity is $O(|C|^2|U''|/|C|) = O(|C|^2|U/C|) = O(|C||U|)$. When Algorithms 1–4 are all took into account, total time complexity of reduction is $O(|U|^2)$, because normally $|C| \ll |U|$ in the big data environment. That is to say, most of the time is spent on sample extraction using ISODATA. The higher the data volume is, the quicker the time increases.

3 Results and discussions

To determine the reasonable value of m and compare the time complexity of the algorithms, we use 3 data sets from UCI machine learning databases to make experiments with MATLAB. The experimental conditions are as follows: Windows 10 OS, Inter Core i7 CPU and 8 GB RAM. The basic information about the data sets is shown in Table 3. Further to ensure the accuracy of reduction, the values of attributes are pre-treated. The results are average values of five experiments.

Table 3 UCI data sets.

Data sets	$ U $	$ C $
Statlog	6435	36
Winequality	4898	11
Contraceptive	1473	9

First, experiment to explore the relationship between different sampling distances and reduction effectiveness with Statlog data set is conducted and reasonable value of m is determined. Reduction result obtained by the traditional attribute reduction algorithm based on positive region [18] is that reduction contains 10 attributes

and reduction time is 849.51 s. Reduction result obtained by the Algorithm 4 in this paper is shown in Table 4, where computing time is the sum of sampling time and reduction time, D_i is the average distance from all objects in the i -th cluster to cluster centre.

From Table 4 it can be seen that the smaller the sampling distance is, the fewer the number of samples is and the greater the difference between the number of the reduced attributes and traditional algorithm is. To bring reduction effectiveness in line with traditional algorithm as far as possible, a proper sampling distance must be chosen. Clearly, if sampling distance is D_i , namely $m=1$, Algorithm 4 achieves a good balance between the number of samples and reduction effectiveness. In this case, the number of samples is relatively small and time complexity of Algorithm 4 is low correspondingly. Reduction time accounts for just 13% of traditional algorithm. But at the same time, reduction effectiveness basically coincides with the traditional algorithm. The experimental research results on other data sets are similar to the Statlog. Hence the value of m is 1 in all the later experiments. That is to say, the sampling distance is D_i .

Table 4 The relationship between different sampling distances and reduction effectiveness for Statlog.

Sampling distance	Sample number	Attribute number of minimum reduction	Computing time (s)
$0.6D_i$	5473	9	329.61
$0.8D_i$	3965	8	195.13
D_i	2648	8	111.42
$1.05D_i$	2344	7	79.82
$1.1D_i$	2077	7	65.15
$1.2D_i$	1628	6	54.59

Next, traditional attribute reduction algorithm based on positive region and Algorithm 4 in this paper are separately used to calculate the time required for getting reduction and compare the efficiency of the algorithms. The corresponding experimental results on Winequality data set are listed. There are 11 condition attributes in database and computing time of this paper is the sum of redundancy removal time, sampling time and reduction time.

Table 5 Comparison on the effect of reduction for Winequality data set

$ U $			Traditional algorithm		Algorithm 4	
Original data set	Redundancy removal	Sampling	Reduction result	Reduction time (s)	Reduction result	Computing time (s)
4898	3961	1788	5	113.62	4	33.25

As seen from Table 5, the original Winequality data set has nearly 20% redundant data, which is a very common phenomenon in real data. At the same time, the sampling rate is 45%. So, the number of objects of the original data set has been greatly decreased after redundancy removal and sampling. Moreover, reduction of samples basically coincides with the original data set, but computing time just accounts for 29% of the original data set. Overall, it is clear that reduction efficiency is greatly improved.

The experimental results on Contraceptive data set are listed here. There are nine condition attributes in database and computing time of this paper is the sum of redundancy removal time, sampling time and reduction time.

The result shown in Table 6 is similar to that in Table 5. Redundant data provide 3% of the original Contraceptive data set and the sampling rate reaches 47%. Thus, the number of objects drops significantly. Consequently, computing time of Algorithm 4 in this paper just accounts for 33% of the original data set. This also illustrates the efficiency of the proposed approach. Meanwhile, comparing Table 5 with Table 6, it is found that the higher the data volume is, the more obvious the advantage of Algorithm 4 is.

Table 6 Comparison on the effect of reduction for Contraceptive data set.

$ U $			Traditional algorithm		Algorithm 4	
Original data set	Redundancy removal	Sampling	Reduction result	Reduction time (s)	Reduction result	Computing time (s)
1473	1425	665	9	12.62	8	4.13

4 Conclusions

For large data sets, redundant data expend a lot of computer run time and memory in attribute reduction. To overcome this issue, the paper puts forward an attribute reduction method based on sample extraction and priority. To start with, this method removes redundant data using quick sort. Then, the samples near the classification boundary are extracted and used for reduction calculation. According to the concept of positive region, an attribute reduction algorithm that is based on priority is proposed in this paper and used for reduction. Experiments show that this method suits attribute reduction calculation on large data sets well, specifically. With the rise of smart medical care, a large number of medical data resources have been integrated. Medical big data knowledge mining has become a research focus in the current academic field. The method proposed in this article can be used to effectively reduce the dimensionality of massive medical data and improve the efficiency of knowledge mining. This is the next application and research direction.

References

- [1] Pawlak Z, Rough sets. International Journal of Computer and Information Science, 1982, 11(5), pp. 341-356.
- [2] Bera, S., Roy, S.K. Fuzzy Rough Soft Set and Its Application to Lattice. Granular Computing, 2020, 5, pp. 217-223 .
- [3] Sun, B., Ma, W., Chen, X. et al. Multigranulation Vague Rough Set over Two Universes and Its Application to Group Decision Making. Soft Computing, 2019, 23, pp. 8927-8956.
- [4] Chelly Dagdia, Z., Zarges, C., Beck, G. et al. A Scalable and Effective Rough Set Theory-based Approach for Big Data Pre-processing. Knowledge and Information Systems, 2020, 62, pp. 3321-3386 .
- [5] Wafo Soh, C., Njilla, L.L., Kwiat, K.K. et al. Learning Quasi-identifiers for Privacy-preserving Exchanges: a Rough Set Theory Approach. Granular Computing, 2020, 5, pp. 71-84 .
- [6] Wong S K M, Ziarko W, Optimal decision rules in decision table. Bulletin of Polish Academy of Sciences, 1985, 33(11-12) , pp. 693- 696.
- [7] Liu W J, Gu Y D, Feng Y B, et al., An improved attribute reduction algorithm of decision table. Pattern Recognition and Arti?cial Intelligence, 2004, 17(1) , pp. 119-123.
- [8] Du J L, Chi Z X, Zhai W, An improved algorithm for reduction of knowledge based on signi?cance of attribution. Mini-Micro System, 2003, 24(6) , pp. 976-978.
- [9] Li Hua, Research on the Model and Algorithms of Rough Set for Multi-label Data. Ph.D. thesis, Shanxi University, Taiyuan, China, 2017.
- [10] Hart P, The condensed nearest neighbor rule. IEEE Transaction on Information Theory, 1968, 14(5) , pp. 515- 516.
- [11] Gates G W, The reduced nearest neighbor rule. IEEE Transactions on Information Theory, 1972, 18(3) , pp. 431- 433
- [12] Brighton H, Mellish C, Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery, 2002, 6(2) , pp. 153 -172.
- [13] Wang Xizhao, Wang Tingting, Zhai Junhai, An Attribute Reduction Algorithm Based on Instance Selection. Journal of Computer Research and Development, 2012, 49(11) , pp. 2305 -2310.
- [14] JI Su-Qin, SHI Hong-Bo, Lv Ya-Li, An Attribute Reduction Algorithm Based on Granular Computing and Discernibility. Pattern Recognition and Artificial Intelligence, 2015, 28(4) , pp. 327-334.
- [15] Yang Yanyan, Rough Set Based Mechanisms and Algorithms for Incremental Attribute Reduction. Ph.D. thesis, North China Electric Power University, Beijing, China, 2017.
- [16] Pan Wei. Rough Sets Model with Entropy for Multi-criteria Ordernal Decision System. Ph.D. thesis, University of Electronic Science and Technology of China, Chengdu, China, 2017.
- [17] Jin ping, Zong Yu, Jiang He, et al., Muti-space FCM Algorithm. Computer Engineering and Applications, 2007, 43(28), pp. 162-165.
- [18] Zhang Wenxiu, Wu Weizhi, Liang Jiye, et al., Rough Set Theory and Method. Beijing: Science Press, 2001.

This page is intentionally left blank