Heuristic Attribute Reduction Based on Neighborhood Knowledge Granularity

Wen Yang¹, Lei Wang²*, Chao Liu¹ and Qiangqiang Zhong¹

¹School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China
²Jiangxi Provincial Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing, Nanchang Institute of Technology, Nanchang 330099, China
Email: ezhoulei@163.com

Abstract. The classical attribute reduction algorithm is not suitable for the neighborhood decision information system of numerical attribute. Since any feature subset in real feature space can be approximated by neighborhood information particles, the concept of knowledge granularity is extended to neighborhood rough set from the perspective of the granularity computation. In this paper, knowledge granularity, attribute importance and heuristic algorithm based on the granularity of knowledge in neighborhood rough set are studied from the point of neighborhood relation matrix. The feasibility of heuristic algorithm is analyzed on UCI data sets. Experiments achieve better classification accuracy and lower attribute reduction results than the existing algorithms.

Keywords. Neighborhood rough set; knowledge granularity; feature set; attribute reduct.

1. Introduction

Pawlak put forward rough set theory in 1982, which is a practical tool for extracting important data information in classification learning [1]. At present, reduction algorithms on rough set are generally classified into two categories: (1) one is based on the perspective of algebraic, which mainly includes attribute reduction algorithm based on positive domain [2] and discernibility matrix [3]; (2) the other is based on the viewpoint of information theory [4]. However, Pawlak rough set is processes information system attribute reduction with containing numerical data, it is often used to convert numerical data into discrete data, while the discretization process causes information loss.

In order to avoid the above situation, Lin et al. proposed the concept of neighborhood information system by means of interior point and closure in topology [5]. Based on the theory of Lin, Hu et al. systematically gave the definition of neighborhood rough set, by which combined the neighborhood decision system with Pawlak rough set [6]. Liu et al. give a fast attribute reduction algorithm, which used hash table to divide objects before reducing operations, when computing the neighborhood of a certain object[7]. It was only compared with the adjacent object, which greatly reduced the computing time of the positive domain. It was difficult and costly to implement that this optimization methods in the extended rough set model. From the perspective of granular computing, some scholars have designed a variable cost reduction algorithm based on multi-granularity rough set [8].

Granular computing is a comprehensive subject combining artificial intelligence and rough set systematically [9]. By putting forward the concepts of knowledge granularity and attribute importance and their computing methods, Miao et al. proposed a reduction algorithm of the classic rough set in granular computing [10]. Wang et al. constructed a matrix-based method for calculating knowledge
granularity [11]. By contrast, there is less research on the granularity of neighborhood knowledge in neighborhood decision system.

As indicated above, there were no research about the attribute reduction algorithm that combining knowledge granularity with neighborhood rough set. Accordingly, we proposed an algorithm NGDAR. Firstly, NGDAR combines the granular structure model of granular computing with neighborhood granular. Then, this paper defines the knowledge granularity and uses the attribute significance by the matrix as heuristic information. As we know, attribute reduction is related to the neighborhood threshold and the size of conditional attribute set on neighborhood rough set. NGDAR algorithm can obtain shorter reduction and higher classification accuracy without affecting the calculation of positive domain.

The rest of this article is as follows. The basic definition of neighborhood rough set is briefly introduced in Section 2. The neighborhood rough set algorithm based on knowledge granularity is introduced in Section 3. In Section 4, the classification accuracy and calculation time of the algorithm are analyzed by experiments on UCI data set. Finally, the outlook is summarized in Section 5.

2. Preliminaries
In this section, we introduce some important definitions on neighborhood rough set. A detailed introduction is described in the references.

Generally, a decision system can be loosely described as \(<U, C \cup D>\), where \(U = \{x_1, x_2, \ldots, x_n\}\) is the set of object space, \(C = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}\) is the feature space, and \(D\) is the decision attribute subset.

2.1. Neighborhood granule
In neighborhood rough set, the granule is used to represent similarity relations in feature space.

**Definition 1.** Given arbitrary object \(x_i \in U\) and set \(B \subseteq C\), the neighborhood \(n_\delta^B(x_i)\) of \(x_i\) in the subspace \(B\) is defined as [6]

\[
n_\delta^B(x_i) = \{x_j \in U \mid \Delta_B(x_i, x_j) \leq \delta\}
\]

where \(\Delta_B(x_i, x_j)\) is a metric function of \(U\). Normally, metric function can be implemented with \(p\)-norm, Frobenius norm, etc. Assume that \(x_i\) and \(x_j\) are two objects in feature space \(C\), \(f(x_i, \alpha_k)\) denotes the value of object \(x_i\) in the \(k\)th dimension \(\alpha_k\), then a general metric, named Minkowsky distance, is defined as

\[
\Delta_B(x_i, x_j) = \left( \sum_{k=1}^{m} \left| f(x_i, \alpha_k) - f(x_j, \alpha_k) \right|^p \right)^{1/p}
\]

where \(p\) can be valued 1, 2 and \(\infty\) in equation (2).

\(n_\delta^B(x_i)\) is the information granule centring with object \(x_i\), which is called neighborhood and the size depends on the threshold \(\delta\). From the \(n_\delta^B(x_i)\), we can obtain a symmetric neighborhood relation matrix \(M_\delta^U = (m_{ij})_{m \times m}\), if \(x_j \in n_\delta^B(x_i)\) then \(m_{ij} = 1\) otherwise \(m_{ij} = 0\).

**Definition 2.** Let \(B_1 \subseteq C\) and \(B_2 \subseteq C\) be discrete attribute and continuous attribute, respectively. The neighborhood of object \(x\) in feature set \(B_1 \cup B_2\) is defined as [12]

\[
n_\delta^{B_1 \cup B_2}(x) = \{x_j \in U \mid \Delta_B(x_i, x_j) = 0, \Delta_B(x_i, x_j) \leq \delta\}
\]

2.2. Neighborhood upper-lower approximation
The definition of the neighborhood decision system is similar to classical rough set, and neighborhood relation \(N_\delta^C\) is used instead of equivalence relation, which expressed as \(NDS = <U, C \cup D, \delta>\). So the lower and upper approximation of NDS can be expressed as follows.
Definition 3. In neighborhood decision system $< U, C \cup D, \delta >$, $Y_1, Y_2, \ldots, Y_N$ are the decision subsets from 1 to $N$, $n^s_B(x_i)$ is the neighborhood of the object $x_i$ in the feature subset $B \subseteq C$. So the lower and upper approximations of decision $D$ with respect to feature subset $B$ are defined as [6]

$$N_B\delta D = \bigcup_{i=1}^{N} N_B Y_i \quad \text{(4)}$$

$$\overline{N}_B\delta D = \bigcap_{i=1}^{N} \overline{N}_B Y_i \quad \text{(5)}$$

where

$$N_B Y = \{ x_i \mid n^s_B(x_i) \subseteq Y, x_i \in U \} \quad \text{(6)}$$

$$\overline{N}_B Y = \{ x_i \mid n^s_B(x_i) \cap Y \neq \varnothing, x_i \in U \} \quad \text{(7)}$$

3. Knowledge Granularity Reduct Algorithm for NDS

This section gives the calculation formulas of knowledge granularity and significance on neighborhood rough set. Then we will construct a general heuristic algorithm in which the attribute importance calculated by matrix is used as heuristic information. Furtherly, we use the above definition to analyze the information in table 1.

3.1. Knowledge Granularity

Definition 4. In neighborhood decision system $< U, C \cup D, \delta >$, for arbitrary $B \subseteq C$, the neighborhood knowledge granularity of $B$ is defined as

$$NGD(B) = \frac{\sum_{i=1}^{n} |n^s_B(x_i)|^2}{|U|^2} = M_B^U$$

(8)

the knowledge granularity of $B$ relative to decision set $D$ is defined as

$$NGD(D \mid B) = M_B^U - M_{B \cup D}^U$$

(9)

where $|n^s_B(x_i)|$ is the number of objects in neighborhood $n^s_B(x_i)$. $M_B^U$ represents the average value of matrix $M_B^U$, and $M_B^U = (m_{ij})_{n \times n}$ is an $n \times n$ neighborhood relation matrix. Similarly, $M_{B \cup D}^U$ represents the average value of matrix $M_{B \cup D}^U$, and $M_{B \cup D}^U = (m_{ij})_{n \times n}$ be an $n \times n$ neighborhood relation matrix.

Example 1. According to information in table 1, supposing that $p=2$ in metric function $\Delta_B(x_i, x_j)$ and $\delta = 0.2$. From equations (3) and (8), we can get: $NGD(C) = M_C^U = 0.1$.

Table 1. Neighborhood decision information system.

| U   | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | d   |
|-----|-------|-------|-------|-------|-------|-----|
| $x_1$ | 0.21  | 0.63  | 0.31  | 0.50  | 0.44  | 1   |
| $x_2$ | 0.63  | 0.48  | 0.78  | 0.15  | 0.64  | 1   |
| $x_3$ | 0.47  | 0.21  | 0.35  | 0.22  | 0.63  | 1   |
| $x_4$ | 0.51  | 0.21  | 0.31  | 0.47  | 0.52  | 2   |
| $x_5$ | 0.45  | 0.75  | 0.61  | 0.45  | 0.37  | 2   |
| $x_6$ | 0.51  | 0.42  | 0.27  | 0.16  | 0.56  | 1   |
| $x_7$ | 0.36  | 0.22  | 0.42  | 0.52  | 0.68  | 3   |
| $x_8$ | 0.34  | 0.82  | 0.06  | 0.18  | 0.41  | 1   |
| $x_9$ | 0.31  | 0.61  | 0.74  | 0.15  | 0.40  | 1   |
| $x_{10}$ | 0.45  | 0.58  | 0.18  | 0.25  | 0.83  | 3   |
3.2. Attribute Significance

**Theorem 1.** In decision system \( <U, C \cup D, \delta > \). Let \( M_C^{U}, M_{C-\{a\}}^{U}, M_{C-D}^{U}, M_{(C-\{a\}) \cup D}^{U} \) are the neighborhood relation matrices for \( N_C, N_{C-\{a\}}, N_{C-D}, N_{(C-\{a\}) \cup D} \), respectively. The importance of \( a \) in \( C \) is defined as

\[
\text{sig}^{\text{in}}(\alpha, C, D) = \frac{M_C^{U} - M_{C-\{\alpha\}}^{U} + M_{C-D}^{U}}{M_C^{U}}
\]

(10)

**Proof.** From Definition 4. We have

\[
\text{sig}^{\text{in}}(\alpha, C, D) = \frac{M_C^{U} - M_{C-\{\alpha\}}^{U} + M_{C-D}^{U}}{M_C^{U}}
\]

Theorem 2. In neighborhood decision information system \( \text{NDS} = <U, C \cup D, \delta > \). Let feature subset \( P \subseteq C, \forall \beta \in C - P \), the significance measure of \( \beta \) in \( P \) is defined as

\[
\text{sig}^{\text{out}}(\beta, P, D) = \frac{M_P^{U} - M_{P-\beta}^{U} + M_{P \cup \{\beta\}}^{U}}{M_P^{U}}
\]

(11)

proves the same.

**Definition 5.** In decision system \( <U, C \cup D, \delta > \). For any feature subset \( B \subseteq C \), if \( B \) is a relative reduct of the decision information system, the following two conditions must be satisfied:

1. \( \text{NGD}(D | B) = \text{NGD}(D | C) \)
2. \( \forall \alpha \in B, \text{NGD}(D | B - \{\alpha\}) \neq \text{NGD}(D | C) \)

3.3. Reduction Algorithm

According to the neighborhood knowledge granularity model proposed in Subsection 3.1, in this paper, a heuristic reduction algorithm (NGDAR, Algorithm 1) is proposed, which takes the importance degree of attributes by matrix as an important formula. The above formula is given in Subsection 3.

**Algorithm 1.** Heuristic attribute reduction algorithm based on knowledge granularity for NDS (NGDAR).

```
Input: A neighborhood decision system NDS \( \equiv <U, C \cup D, \delta > \)
Output: A reduct \( \text{red} \) on \( U \).
1 red \( \leftarrow \emptyset \)
2 for \( 1 \leq i \leq |C| \) do
3 calculate \( \text{sig}^{\text{in}}(\alpha_i, C, D) \)
4 if \( \text{sig}^{\text{in}}(\alpha_i, C, D) > 0 \) then
5 \( \text{red} \leftarrow \text{red} \cup \{\alpha_i\} \)
6 end if
7 end for
8 let \( P \leftarrow \text{red} \)
9 while \( \text{NGD}(D | P) = \text{NGD}(D | C) \) do
10 for \( \forall \beta_i \in C - P \) do
11 find \( \beta_0 \) such that \( \beta_0 = \max \{\text{sig}^{\text{out}}(\beta_i, P, D), \beta_i \in C - P\} \)
12 \( P \leftarrow (P \cup \{\beta_0\}) \)
13 end for
14 end while
15 \( \text{red} \leftarrow P \)
16 return \( \text{red} \), end
```
In our NGDAR algorithm, the complexity of time that computing neighborhood knowledge granularity is $O(|U|^2)$, and the time complexity of that computing kernel attribute is $O(|C||U|^2)$. Then, the total complexity is about $O(|U||C|^2 + |C||U|^2)$.

4. Experiments
In this section, we will analyze the feasibility of our algorithm comparing with F2HARNRS [12]. The data set used in the experiment was downloaded from UC Irvine Machine Learning Repository and described in table 2. All experiments have been run on the same software (MATLAB R2016a) and hardware platforms (CPU: AMD Ryzen 5 4600H with Radeon Graphics 3.00GHz; RAM:16 GB; Windows 10 64bit).

We compare the classification accuracy of NGDAR algorithm with F2HARNRS. Experiment uses SVM classifier, and the results show in table 3. Assuming $\delta = 0.15$, because Hu Q H obtained the best value interval of the threshold $\delta$ in the experiment. We will compare the performance of NGDAR and F2HARNRS according to reduction results, classification accuracy and running time.

Table 2. Experimental data description.

| Data set                  | Object | Attribute | classes |
|---------------------------|--------|-----------|---------|
| 1 wine                    | 178    | 14        | 3       |
| 2 breast cancer coimbra   | 116    | 10        | 2       |
| 3 banknote authentication | 1372   | 5         | 2       |
| 4 wdbc                    | 569    | 31        | 2       |
| 5 Wireless indoor localization | 2000  | 7         | 4       |

Table 3. NGDAR and F2HARNRS on SVM classification accuracy (%).

| Data sets                          | F2HARNRS | NGDAR |
|------------------------------------|----------|-------|
| wine                               | 97.88    | 98.97 |
| breast cancer coimbra (bcc)        | 79.35    | 81.97 |
| banknote authentication (ba)       | 93.34    | 97.98 |
| wdbc                               | 96.79    | 96.77 |
| wireless indoor localization (wil) | 97.76    | 97.98 |

Table 4. Attribute selected by two reduction algorithms.

| Data sets                          | F2HARNRS | NGDAR |
|------------------------------------|----------|-------|
| Wine                               | 1,5,9,11,12,13 | 1,5,7,11,13 |
| Breast cancer Coimbra (BCC)        | 1,2,3,6,7,8,9 | 1,2,3,4,7,8 |
| Banknote authentication (BA)       | 1,2,3     | 1,3,4  |
| wdbc                               | 9,15,21,22,25 | 1,9,12,22,25,27,28 |
| Wireless indoor localization (WIL) | 1,4,5,6   | 1,3,4,5,7 |

The experiment shows that NGDAR algorithm will achieve a good classification accuracy and efficiency. Tables 3 and 4 show that NGDAR has better classification accuracy on SVM classifier under the number of reduct is almost the same, which compared with F2HARNRS in medium data set. The experiment further shows that NGDAR will save computing time on larger data sets, for example, the results of banknote authentication and wireless indoor localization, in figure 1.
Figure 1. The computational time of NGDAR and F2HARNRS on the UCI data sets.

5. Conclusion
The neighborhood rough set can manage mixed data including discrete data and continuous data, and its extension model is the main research contents of reduct algorithm. In this paper, the concept of knowledge granularity is extended to neighborhood rough set. Based on this, a new computing method is designed to compute the attribute reduction of neighborhood decision system. Experiments demonstrate the effectiveness and efficiency of the proposed algorithm. In practical application, information system is often dynamic, so the next step is to study the reduction algorithm of real-time change in the number of objects in neighborhood decision system.

Acknowledgements
This research is based on work supported by Science and Technology Project of Jiangxi Provincial Department of Education (GJJ170995) and National Natural Science Foundation of China (61562061).

References
[1] Pawlak Z 1982 Inf. Sci. 11 341-356.
[2] Hu X and Cercone N 1995 Comput. Intell 11 323-338.
[3] Skowron A, Andrzej J and Rauszer C 1992 Intelligent Decision Support (Dordrecht: Springer) pp 331-362.
[4] Wang G Y, Yu H and Yang D C 2002 J. Comput. Sci 25 759-766.
[5] Lin T Y 1998 Rough. Set. Knowl. Disc. 1 107-121.
[6] Hu Q, Yu D and Xie Z 2008 Exp. Syst. Appl. 34 866-876.
[7] Liu Y, Huang W L, Jiang Y L and Zeng Z Y 2014 Inf. Sci. 271 65-81.
[8] Liao S, Zhu Q, Qian Y and Lin G 2018 Knowl-Based Syst. 158 25-42.
[9] Miao D Q, Wang G Y, Liu Q, Liu Z Y, Yao Y Y 2007 Granular. Computing: Pst. in Now and View pp 154-160.
[10] Miao D Q and Feng S D 2002 Syst. Eng. Theory P 22 548-556.
[11] Wang L and Li T R 2013 P. R. Art. Intel. 26 447-453.
[12] Hu Q, Yu D, Liu J and Wu C 2008 Inf. Sci. 178 3577-94.