Adaptive Neighborhood Size Selection Method for Neighborhood Rough Sets Postprint

Authors: Peng Xiaoran, Liu Zunren, Ji Jun

Date: 2018-05-20T00:00:00+00:00

Abstract

The effectiveness of neighborhood rough set applications depends on the selection of the neighborhood size parameter #1;. In attribute reduction algorithms based on neighborhood rough sets, existing #1; selection methods are generally point-wise, relying solely on human experience to specify a particular value. This approach fails to consider the specific characteristics of the actual problem when selecting #1;, thus limiting the practical applicability of the algorithm. To address this limitation, an adaptive #1; selection method is proposed. Its key feature is that instead of specifying a fixed value for #1;, it specifies an interval for #1; values. Within this interval, the most appropriate #1; value is automatically selected using a fitness function that incorporates the characteristics of both the dataset and the classifier itself. Experimental results demonstrate that, compared with point-wise #1; selection methods, the adaptive approach can identify attribute sets with fewer attributes and higher classification accuracy. The experiments verify that this method can further enhance the practicality of attribute reduction algorithms based on neighborhood rough sets.

Full Text

Preamble

Adaptable Method for Determining Neighborhood Size of Neighborhood Rough Set

Peng Xiaoran[1], Liu Zunren[2], Ji Jun[2]
(1. College of Data Science & Software Engineering; 2. College of Computer Science & Technology, Qingdao University, Qingdao, Shandong 266071, China)

Abstract: The effectiveness of neighborhood rough set applications depends heavily on the selection of the neighborhood size parameter . When employing attribute reduction algorithms based on neighborhood rough sets, existing methods for determining typically adopt a point-value approach—that is, specifying a fixed value based solely on human experience. Such methods fail to consider the specific characteristics of the problem at hand, limiting the practical utility of these algorithms. To address this limitation, we propose an adaptive method for determining . The key innovation of this method is that instead of specifying a single value, it defines an interval for . Within this interval, a fitness function that incorporates the intrinsic properties of both the dataset and the classifier automatically selects the most appropriate value. Experimental results demonstrate that compared with point-value methods, our adaptive approach yields attribute subsets with fewer attributes and higher classification accuracy, thereby significantly enhancing the practicality of attribute reduction algorithms based on neighborhood rough sets.

Keywords: neighborhood rough set; neighborhood size; attribute reduction; classification

---

0 Introduction

The performance of attribute reduction algorithms based on neighborhood rough sets critically depends on the selection of the neighborhood size parameter [5,8-16]. An appropriate value enables the algorithm to obtain high-quality attribute reduction sets, while an inappropriate value may lead to suboptimal or even poor results. For the reduction sets produced by these algorithms, we desire both minimal attribute count and high effectiveness. Effectiveness is measured by the classification accuracy achieved when a classifier uses the reduction set on the dataset—higher accuracy indicates greater effectiveness of the reduction set. Consequently, the discussion of selection is both important and meaningful.

Rough set theory posits that knowledge is granular, representing an ability to classify objects in a universe [1]. The classical Pawlak rough set [2] employs equivalence partitioning and equivalence classes to enable granular computing, but this approach only applies to discrete variables. However, real-world applications often involve numerical data, which has constrained the broader application of rough set theory.

To address this limitation, Zadeh [3] introduced the concepts of information granulation and granular computing. Building upon these ideas, Lin [4] proposed the neighborhood model. Hu [5] developed the neighborhood rough set model [6-7] based on neighborhood granulation and rough approximation, which can handle numerical data and significantly extends the applicability of rough set theory. However, unlike the classical Pawlak rough set, the introduction of neighborhood granulation means that the effectiveness of attribute reduction algorithms based on neighborhood rough sets is highly sensitive to the parameter.

Existing methods for determining generally adopt a point-value approach, where is set to a specific value based on empirical judgment [5,9-16]. For instance, Hu [5] specified = 0.125 in comparative experiments with other algorithms; Liu [9] set to the standard deviation of the normalized dataset; Duan [11] assigned different values for different datasets; and Chen [15] used = 0.1 in their comparative studies. However, it is crucial to recognize that different datasets and classifiers possess distinct characteristics. Some datasets contain minimal noise, while others are highly noisy. Furthermore, an attribute set that performs well with one classifier may yield poor results with another. Therefore, relying solely on empirical point-value methods without considering these specific circumstances inherently limits the practical utility of attribute reduction algorithms.

To overcome these limitations, this paper proposes an adaptive method for determining in neighborhood rough sets. Rather than specifying a single value, this method defines a range for and designs a fitness function that evaluates each candidate within this interval by incorporating the intrinsic properties of both the dataset and the classifier. The value with the highest fitness is automatically selected as the final result. Additionally, by adjusting the weights in the fitness function, the algorithm can better satisfy practical application requirements, further enhancing its utility.

---

1 Preliminaries

1.1 Neighborhood Granulation

Definition 1 (Metric Calculation). Given an n-dimensional real space , for any two points x = (x , x , ···, x) and x = (x , x , ···, x) in the space, a metric calculation d(x , x) is defined on . The Euclidean distance is commonly used:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{n} (x_{ip} - x_{jp})^2}$$

Definition 2 (Neighborhood Particle). In real number space, let U = {x , x , ···, x } be a non-empty finite set of samples, called the universe. For a sample x U, the -neighborhood information particle generated by x , denoted as (x), is defined as:

$$\delta(x_i) = \{x_j \mid x_j \in U, d(x_i, x_j) \leq \delta\}$$

where 0. We call (x) the neighborhood particle of x .

1.2 Decision Table and Approximations

Definition 3 (Decision Table). A decision table is defined as a quadruple DT = (U, C, D, V, f), where U is the universe; C is the set of condition attributes;

D is the set of decision attributes, with C D = ; V is the value domain; and f is the information function.

Definition 4 (Upper and Lower Approximations). Given a decision table DT = (U, C, D, V, f), let D divide U into N equivalence classes: D = {D , D , ··· , D }. The lower and upper approximations of the decision attribute set D with respect to the condition attribute set B C are defined as:

$$\underline{N_B}D = \bigcup_{i=1}^{N} \underline{N_B}D_i, \quad \overline{N_B}D = \bigcup_{i=1}^{N} \overline{N_B}D_i$$

where:

$$\underline{N_B}D_i = \{x \mid \delta_B(x) \subseteq D_i, x \in U\}$$

$$\overline{N_B}D_i = \{x \mid \delta_B(x) \cap D_i \neq \emptyset, x \in U\}$$

The positive region of decision attribute set D with respect to condition attribute set B is defined as:

$$Pos_B(D) = \underline{N_B}D$$

The boundary region is:

$$BN_B(D) = \overline{N_B}D - \underline{N_B}D$$

The negative region is:

$$NEG_B(D) = U - \overline{N_B}D$$

Definition 5 (Attribute Reduction). Given a decision table DT = (U, C, D, V, f), for any attribute subset B C, if Pos_B(D) = Pos_C(D) and for any a B, Pos_{B-{a}}(D) < Pos_B(D), then B is called an independent attribute subset of C. If B is independent and Pos_B(D) = Pos_C(D), then B is called an attribute reduct of C.

2 F2HARNRS Algorithm and Value Analysis

Greedy strategies are characterized by their ability to find optimal or near-optimal solutions in relatively short time. Hu [5] combined this greedy 思想 with neighborhood rough sets to construct a fast forward heterogeneous attribute reduction algorithm based on neighborhood rough sets (F2HARNRS), which has been widely applied and studied [8-11]. Our discussion of selection is built upon this algorithm.

The F2HARNRS algorithm uses the number of positive region samples under the original attribute set as its greedy objective. Its specific strategy is: initialize the attribute reduction set as an empty set (where the current positive region is empty), then iteratively select the attribute that adds the most samples to the current positive region, and add it to the reduction set. This process continues until either all samples to be tested are included in the current positive region, or no remaining attribute can increase the number of samples in the positive region. A key optimization leverages the property that newly added attributes cannot cause samples already in the positive region to become non-positive, allowing the algorithm to compute the positive region only for samples not yet determined to be positive. The algorithm is shown as Algorithm 1.

Algorithm 1: F2HARNRS

Input: Decision table DT = (U, C, D, V, f), neighborhood size
Output: Attribute reduction set red

```
red =
smp_chk = U
max_pos =
while smp_chk    do
    max_pos =
    max_i =
    for each a_k   C - red do
        Pos_i = Pos(smp_chk, red   {a_k}, D,  , U)
        if max_pos < Pos_i then
            max_pos = Pos_i
            max_i = k
        end if
    end for
    if max_pos    then
        red = red   {a_max_i}
        smp_chk = smp_chk - max_pos
    else
        break
    end if
end while
return red
```

In Algorithm 1, Pos() is the positive region calculation function, which is also the most computationally expensive part. Its function is: under the current value, calculate the positive region of decision attribute D with respect to condition attribute set B in the decision table.

The performance of the F2HARNRS algorithm is closely tied to the value. Different values lead to different algorithm effects because varying values produce different neighborhood particles (x), which result in different positive region calculations, ultimately yielding different attribute reduction sets. The reliability of these different reduction sets also varies. Reliability is primarily reflected in two aspects: the number of attributes in the reduction set and the classification accuracy achieved when a classification algorithm uses this set. In other words, we evaluate whether the reduction achieves good reduction effects on the original attribute set and whether classification accuracy is not severely compromised after reduction.

Furthermore, the F2HARNRS algorithm has two termination conditions: (1) all samples to be tested are included in the current positive region, or (2) no remaining attribute can increase the number of samples in the positive region. The algorithm terminates when either condition is satisfied.

---

3 Adaptive Value Selection Method

Based on the analysis in Section 2, the main contribution of our proposed selection method is improving the evaluation criterion from relying solely on human experience to incorporating the actual properties of the dataset and classifier. This transforms the problem of evaluating values into analyzing the reliability of their corresponding reduction sets, where reliability can be directly observed from the data.

The core idea of the adaptive selection method is: specify a range for , select multiple candidate values within this range, evaluate each candidate, and finally select the best value as the result.

3.1 Fitness Function

Considering practical circumstances, we propose a fitness function to evaluate the quality of values. The function contains two variables: length_ratio and accuracy_ratio, representing the reliability of the corresponding reduction set. Specifically, length_ratio represents the fitness value of the number of attributes in the reduction set obtained under a particular value, while accuracy_ratio represents the fitness value of the classification accuracy achieved by the classifier using this reduction set. The fitness value of is determined by the weighted sum of these two variables, as shown in Equation (1):

$$fit(\delta) = \alpha \cdot length\_ratio + \beta \cdot accuracy\_ratio \quad (1)$$

where and are weights satisfying > 0, > 0, and $+ = 1$. The weight settings can be adjusted according to application requirements by increasing the weight of the more important factor. In our experiments, we use $= 0.4$ and $= 0.6$.

For a given reduction set, fewer attributes yield higher length_ratio values, and higher classification accuracy yields higher accuracy_ratio values.

3.2 Method Steps

The steps of our selection method are:

a) Specify the interval [a, b] (a 0) for and select n candidate values within this interval.

b) For each value, compute the reduction set using the attribute reduction algorithm, and record the corresponding number of attributes and classification accuracy.

c) Normalize the two sets of records (length and accuracy) from step (b), then adjust them according to the principle that fewer attributes yield higher length_ratio and higher accuracy yields higher accuracy_ratio.

d) Substitute into Equation (1) to calculate the fitness value fit( ) for each value, and select the value with the maximum fitness as the final result.

3.3 Analysis of Value Range

As analyzed in Section 2, the reliability of reduction set red first increases and then decreases as increases, indicating that both excessively small and large values are meaningless. Generally, the interval can be selected as (0, b] based on specific circumstances. In Hu’ s experimental conclusions [5], for their selected datasets and classifiers, (0.1, 0.3] was a better interval, where most classifiers achieved good performance. In our experiments, referencing Hu’ s results, we specify the interval as [0.02, 0.4] with a step size of 0.02, obtaining 20 candidate values, which yields good results. A reasonable interval avoids unnecessary computation and improves efficiency.

The computational overhead of this method primarily lies in step (b). If n candidate values need to be compared and the attribute reduction algorithm has time complexity O(red), the total time complexity of our method is O(n · red). For different attribute reduction algorithms, O(red) varies. If n is large, the interval range and step size can be appropriately adjusted.

---

4 Experimental Analysis

4.1 Experimental Environment

The UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) provides a collection of standard datasets for testing. We selected 7 numerical datasets from UCI, each providing condition attributes and decision attributes. The dataset descriptions are shown in Table 1 .

Table 1: Dataset Descriptions

Dataset	Samples	Attributes	Classes
Wine	178	13	3
WDBC	569	30	2
Sonar	208	60	2
Ionosphere	351	34	2
Credit Approval	690	15	2
German Credit	1000	24	2
WPBC	198	33	2

Experiments were conducted on a PC with an Intel(R) Core(TM) i5 CPU and 4 GB RAM, running MATLAB R2016b on Windows 7.

4.2 Relationship Between Value and Algorithm Performance

4.2.1 Relationship Between and Attribute Reduction Count Using the "F2HARNRS+SVM" approach on datasets Wine, WDBC, and Sonar, we varied from 0.04 to 1 with a step size of 0.04 (25 values total), recording the number of attributes in the reduction sets and the classification accuracy.

As shown in Figure 2 [Figure 2: see original paper], different values produce different numbers of attributes for the same dataset. Both excessively small and large values yield no results, confirming that extreme values are meaningless. As increases, the number of attributes grows until it stabilizes or becomes empty at a certain point, which we call the saturation point (indicated by dashed lines in Figure 2 [Figure 2: see original paper]).

4.2.2 Relationship Between and SVM Classification Accuracy The relationship between values and classification accuracy is shown in Figure 3 [Figure 3: see original paper]. For both classification algorithms, we randomly selected 2/3 of each class as the training set and 1/3 as the test set, repeating the process 20 times and averaging the results.

Analyzing Figures 3(a)-(c), the horizontal line represents the classification accuracy using the original attribute set, the polyline represents accuracy under different values, and the dashed line marks the saturation point. Using the saturation point as a benchmark, each figure can be divided into two parts: the first part corresponds to reduction sets under different values, while the second part corresponds to the original attribute set or an empty set. Our analysis focuses on the first part.

Three key observations emerge: First, all three polylines show an upward trend, with those in Figures 3(a) and 3(b) stabilizing after reaching a peak, indicating that reduction set effectiveness increases with and reaches its maximum in (a) and (b). Second, the saturation points differ across datasets, as do the differences between maximum and minimum accuracy values—particularly in the Sonar dataset, where this difference reaches nearly 10 percentage points. This demonstrates that each dataset has unique characteristics. Relying on empirical point-value selection for such diverse datasets is suboptimal and can be improved.

4.2.3 Experimental Conclusions Overall, as increases, the number of attributes and classification accuracy both increase. However, our goal is to obtain reduction sets with fewer attributes and higher classification accuracy. The point-value approach inherently limits the practicality of reduction algorithms in this regard.

4.3 Comparison Between Point-Value and Adaptive Selection Methods

For the point-value method, we set = 0.15. For the adaptive method, we set = 0.4, = 0.6, specify the interval as [0.02, 0.4], and use a step size of 0.02 (20 candidate values). We compare both methods across 7 datasets using SVM and 1-NN classifiers.

Table 2 shows the results for F2HARNRS+SVM.

Table 3 shows the results for F2HARNRS+1-NN.

The experimental results are related to the characteristics of the datasets, classifiers, and our experimental protocol. The key finding is that the adaptive selection method can identify better values based on actual conditions. For the same classifier, appropriate values differ across datasets; for the same dataset, appropriate values differ across classifiers.

Compared with the point-value method, the adaptive approach consistently finds attribute sets with fewer attributes and higher classification accuracy, or achieves better results on one metric while maintaining comparable performance on the other. This advantage is particularly evident because it incorporates problem-specific characteristics. For example, on the WPBC dataset using F2HARNRS+SVM, the adaptive method improves classification accuracy by nearly 11 percentage points compared to the point-value method. While the resulting reduction set contains 17 attributes (10 more than the point-value method), this is still a substantial reduction from the original 33 attributes, making it a better reduction set. Conversely, the point-value method, despite producing a smaller attribute set, suffers severe classification accuracy loss and cannot be considered a good reduction set. These results demonstrate that the adaptive selection method significantly enhances algorithm practicality.

---

5 Conclusion

The effectiveness of neighborhood rough set reduction algorithms is closely tied to the selection of neighborhood size . This paper analyzes the impact of on algorithm performance, identifies the limitations of point-value selection methods, and improves the evaluation criterion from relying solely on human experience to incorporating the intrinsic properties of datasets and classifiers. Experimental results validate the feasibility of our approach. Determining an appropriate value during the data training phase to obtain attribute reduction sets that meet practical requirements is meaningful for reducing classification workload while maintaining or even improving classification accuracy.

References

[1] Wang Guoyin, Yao Yiyu, Yu Hong. A review of rough set theory and its applications [J]. Chinese Journal of Computers, 2009, 32(7): 1229-1246.

[2] Pawlak Z, Słowiński R. Rough set approach to multi-attribute decision analysis [J]. European Journal of Operational Research, 1994, 72(3): 443-459.

[3] Zadeh LA. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic [J]. Fuzzy Sets & Systems, 1997, 90(90): 111-127.

[4] Lin TY. Granular Computing on binary relations I: Data mining and neighborhood systems [J]. Rough Sets in Knowledge Discovery, 1998(2): 107-121.

[5] Hu Q, Yu D, Liu J, Wu C. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18): 3577–3594.

[6] Wang Guoyin. Rough Set Theory and Knowledge Acquisition [M]. Xi' an: Xi' an Jiaotong University Press, 2001.

[7] Hu Qinghua, Yu Daren. Applied Rough Computing [M]. Beijing: Science Press, 2012.

[8] Du Y, Hu Q, Zhu P, Ma P. Rule learning for classification based on neighborhood covering reduction [J]. Information Sciences, 2011, 181(24): 5457-5467.

[9] Liu Zunren, Wu Gengfeng. Fast reduction algorithm for high-dimensional datasets based on neighborhood rough model [J]. Computer Science, 2012, 39(10): 268-271.

[10] Liu Y, Huang W, Jiang Y, Zeng Z. Quick attribute reduct algorithm for neighborhood rough set model [J]. Information Sciences, 2014, 271(7): 65-81.

[11] Duan Jie, Hu Qinghua, Zhang Lingjun, et al. Multi-label classification feature selection algorithm based on neighborhood rough sets [J]. Journal of Computer Research and Development, 2015, 52(1): 56-65.

[12] Chen H, Li T, Luo C, et al. Dominance-Based Neighborhood Rough Sets and Its Attribute Reduction [M]// Rough Sets and Knowledge Technology. Cham: Springer International Publishing, 2015.

[13] Guo G, Liu Z, Lou C, et al. Improving on a rapid attribute reduction algorithm based on neighborhood rough sets [C]//Proc of International Conference on Fuzzy Systems and Knowledge Discovery. 2016: 236-240.

[14] Wang C, Shao M, He Q, et al. Feature subset selection based on fuzzy neighborhood rough sets [J]. Knowledge-Based Systems, 2016, 111: 173-179.

[15] Chen Y, Zeng Z, Lu J. Neighborhood rough set reduction with fish swarm algorithm [J]. Soft Computing, 2016: 1-12.

[16] Fan A, Zhao H, Zhu W. Test-cost-sensitive attribute reduction on heterogeneous data for adaptive neighborhood model [J]. Soft Computing, 2016, 20(12): 4813-4824.

Note: Figure translations are in progress. See original paper for figures.