Postprint of Neighborhood Rough Set Attribute Reduction Algorithm Based on Matrix Preservation Strategy

Authors: Gao Yang, Liu Zunren, Ji Jun

Date: 2018-10-11T00:00:00+00:00

Abstract

Attribute reduction is of great significance for data processing. In attribute reduction algorithms based on neighborhood rough sets, positive region computation serves as a crucial basis for ensuring their effectiveness and constitutes the primary factor influencing their time overhead. To reduce the algorithm' s time overhead, this paper improves the positive region computation of the existing FHARA algorithm by adopting a retention strategy, which utilizes matrix retention to preserve the squares of metric computation values, thereby transforming the originally n-dimensional computation into a one-dimensional computation and reducing the computational time for each metric calculation. Based on this approach, a neighborhood rough set attribute reduction algorithm based on matrix retention strategy is proposed, and the algorithm is validated through multiple UCI datasets. Compared with existing algorithms, experimental results demonstrate that for most datasets, the proposed algorithm can effectively and more rapidly obtain attribute reductions.

Full Text

Preamble

Title: Neighborhood Rough Set Attribute Reduction Algorithm Based on Matrix Reservation Strategy

Authors: Gao Yang, Liu Zunren, Ji Jun
(College of Computer Science & Technology, Qingdao University, Qingdao, Shandong 266071, China)

Abstract: Attribute reduction is of great significance for data processing. In attribute reduction algorithms based on neighborhood rough sets, positive region calculation serves as the essential foundation for ensuring effectiveness and

constitutes the primary component of computational time overhead. To reduce algorithmic time costs, this paper improves upon the positive region calculation of the existing FHARA algorithm by adopting a reservation strategy. By using a matrix to preserve the squares of computed metric values, the original n-dimensional computation is reduced to one-dimensional computation, thereby decreasing the computation time for each metric calculation. Building upon this improvement, we propose a neighborhood rough set attribute reduction algorithm based on the matrix reservation strategy and validate it through multiple UCI datasets. Compared with existing algorithms, experimental results demonstrate that for most datasets, the proposed algorithm can obtain attribute reductions more efficiently and rapidly.

Keywords: neighborhood rough set; positive region; attribute reduction; fast algorithm

0 Introduction

With the rapid development of information technology, people face not only the problem of data volume explosion but also the more critical issue of high data dimensionality. When processing high-dimensional data, the "curse of dimensionality" phenomenon is widespread. Therefore, attribute reduction is highly meaningful for datasets with massive data volumes, as it can mitigate the impact of dimensionality disasters.

Rough set theory has been widely applied to attribute reduction in data processing. Classical Pawlak rough sets are defined based on equivalence partitions and equivalence classes, ensuring the execution of granular computing. However, this approach is only suitable for discrete variables. When dealing with numerical data types, which are prevalent in real-world applications, numerical data must be discretized. This processing alters the original attribute properties of the data, causing information loss, and different discretization methods lead to different processing results, which severely restricts the application of rough set theory.

To address this problem, Zadeh proposed the concept of information granulation and granular computing, providing a basic framework for granular computation. Lin introduced the concept of neighborhood models based on information granulation and granularity. Hu Qinghua et al. proposed neighborhood information systems and neighborhood decision table models based on neighborhood granulation and rough approximation of basic neighborhood information particles. Ultimately, the neighborhood rough set model method proposed through various studies combines the equivalence approximation of classical rough sets with neighborhood approximation, enabling simultaneous support for both numerical and discrete data types and expanding the application scope of rough set theory.

However, unlike classical Pawlak rough sets, the neighborhood rough set model defines δ-neighborhoods between samples. When calculating the positive region of neighborhood rough sets, it is necessary to traverse all samples and determine the δ-neighborhood relationships between samples through metric calculations. Consequently, the computational load in neighborhood real number space is much greater than that in classical discrete space, leading to excessive time overhead in attribute reduction algorithms based on neighborhood rough sets.

To reduce time overhead, Hu et al. proposed the F2HARNRS (fast forward heterogeneous attribute reduction based on neighborhood rough sets) algorithm based on forward greedy strategy. Subsequently, Liu et al. improved the positive region calculation of this algorithm and proposed the faster FHARA (fast hash attribute reduct algorithm), which reduced the positive region computation time overhead of F2HARNRS.

Building upon the above research, this paper improves the FHARA algorithm by using a matrix to preserve metric calculations between samples, enabling one-dimensional metric computation after dimensionality increase. This reduces the computational load of positive region calculation. Through comparison with the FHARA algorithm, experimental results demonstrate that the proposed algorithm can obtain attribute reductions more rapidly.

1.1 Neighborhood Rough Set

Definition 1: Given an n-dimensional real number space ^{n}, for any two points x and x in the space, define d(x , x) as a metric calculation on ^{n} satisfying:

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

Definition 2: For a given non-empty finite set U = $\{x_1, x_2, \cdots, x\}$ on real number space, where U is the universe. For any sample x U, its δ-neighborhood is defined as:

$$\delta(x_i) = \{x_j \mid x_j \in U \wedge d(x_i, x_j) \leq \delta\}, \quad \delta \geq 0$$

The δ-neighborhood information granule generated by x is called the neighborhood particle of x .

1.2 Neighborhood Decision System

Definition 3: For a quadruple NDT = U, A, V, f , where U is the universe; A = C D, with C being the condition attributes and D being the decision attributes, and C D = ; V is the value domain of the information function f; f: U × A → V is a mapping, then this quadruple is called a neighborhood decision system.

Definition 4: For a given neighborhood decision system NDT = U, A, V, f , D divides U into N equivalence classes: U/D = {D$_1$, D$_2$, ···, D_N}. For any B C, define the lower and upper approximations of decision attribute D with respect to B 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

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

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

Definition 5: Based on Definition 4, the boundary region of decision attribute set D with respect to B is defined as:

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

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

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

The dependency degree of decision attribute D on B is defined as:

$$\gamma_B(D) = \frac{|Pos_B(D)|}{|U|}$$

Definition 6: Given a finite set C, if a B, γ_{B-{a}}(D) < γ_B(D), then B is called an independent attribute subset; if γ_B(D) = γ_C(D) and B is independent, then B is called an attribute reduct of C.

For a given dataset, designing and utilizing effective algorithms to delete redundant attributes and find minimal attribute reducts is an NP-Hard problem.

2.1 Introduction to F2HARNRS and FHARA Algorithms

Greedy strategy can obtain optimal or near-optimal solutions in relatively short time. Hu et al. first defined the contribution of condition attributes to classification based on the dependency function, called attribute significance, which serves as an evaluation metric for attribute set importance. They then constructed a forward greedy attribute reduction algorithm based on neighborhood rough sets, namely the F2HARNRS algorithm.

Definition 7: Given a neighborhood decision system NDT = U, C D, V, f , for any attribute a C - B, define the attribute significance of a relative to set B as:

$$SIG(a, B, D) = \gamma_{B\cup\{a\}}(D) - \gamma_B(D)$$

This is equivalent to:

$$SIG(a, B, D) = |Pos_{B\cup\{a\}}(D)| - |Pos_B(D)|$$

The basic idea of the F2HARNRS algorithm is: initialize the reduct set as empty, where the dependency degree of decision attributes on the set is 0. Each iteration calculates the attribute significance of all remaining attributes and selects the attribute with maximum significance (i.e., the attribute that maximally increases the number of samples in the current positive region) to add to the reduct set. This process continues until the significance of all remaining attributes becomes 0, meaning all samples are included in the current positive region and the dependency value remains unchanged upon adding new attributes. The output set yields a dependency degree of 1, indicating the positive region equals the universe. This algorithm preserves the most significant attributes, effectively ensuring the core is not reduced. Notably, when new attributes are added, samples originally belonging to the positive region will not become non-positive region samples. Therefore, during algorithm execution, positive region calculations only need to be performed on samples not yet determined to be in the positive region, reducing sample judgment frequency.

The positive region calculation of the F2HARNRS algorithm can be represented as shown in [Figure 1: see original paper]. Sample x needs to perform metric calculations with all samples in the universe, with time complexity O(||U||^2).

Subsequently, Liu et al. improved the positive region calculation method of the F2HARNRS algorithm and proposed the faster FHARA algorithm.

The positive region calculation of the FHARA algorithm can be represented as shown in [Figure 2: see original paper]. The universe is partitioned into a finite collection {B_0, B_1, ···, B_k}, where x_0 is a special sample in universe U defined as x_0 = {x | a C, a(x) = min[a(x)]}. If sample x B_q, then the δ-neighborhood only exists within B_q. Therefore, x only needs to perform metric calculations with samples in its own set and adjacent sets, reducing time complexity to O(||U|| · k).

2.2 FARBMRS Algorithm

Analysis of the above algorithm' s maximum computational load reveals that metric calculations at different dimensions are independent. For example, the metric calculation between samples x = (x $_1$, x $_2$) and x = (x $_1$, x $_2$) in 2-dimensional space is:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

After dimensionality increase, the metric calculation between x = (x $_1$, x $_2$, x $_3$) and x = (x $_1$, x $_2$, x $_3$) in 3-dimensional space is:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2}$$

These two metric calculations are clearly related:

$$d(x_i, x_j) = \sqrt{d(x_i, x_j)^2 + (x_{i3} - x_{j3})^2}$$

Therefore, if we preserve the previous metric calculation d(x , x)^2, then after dimensionality increase, the metric calculation between samples only requires one-dimensional computation (x $_3$ - x $_3$)^2, rather than the original three-dimensional computation.

Similarly, extending to n dimensions, when samples increase to (n+1) dimensions, the FHARA algorithm' s metric calculation requires (n+1)-dimensional computation. By adopting a reservation strategy to preserve metric calculations in n-dimensional space, only one-dimensional computation is needed when increasing to (n+1) dimensions. Based on this analysis, we improve the positive region calculation of the FHARA algorithm and propose the Fast Attribute Reduction Based on Matrix Reservation Strategy (FARBMRS) algorithm.

For the above analysis, we propose the following reservation strategy:

Improvement: Let the current attribute reduct set be red. Before calculating the significance of attribute a C - red, first perform metric calculations between samples not yet determined to be in the positive region and the required samples under red, and use a matrix dist[U][U] to preserve the squares of the computed metric values. Then, when calculating the significance of attribute a relative to red after dimensionality increase, we only need to retrieve the corresponding values from the matrix and add the one-dimensional metric calculation value.

The improved positive region calculation is shown in Algorithm 1.

Algorithm 1: Positive Region Calculation

Input: NDT = U, C D, V, f , red, dist
Output: Positive region Pos.

Step 1: For each x U

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