Information Sciences 271 (2014) 65–81

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Quick attribute reduct algorithm for neighborhood rough set model

Liu Yong [a,b,*], Huang Wenliang [c], Jiang Yunliang [d,e], Zeng Zhiyong [a]

[a] Institute of Cyber Systems and Control, Zhejiang University, 310027 Hangzhou, China
[b] State Key Laboratory of Industrial Control and Technology, Zhejiang University, 310027 Hangzhou, China
[c] China United Network Communication Corporation, 100032 Beijing, China
[d] Huzhou Teachers College, 313000 Huzhou, China
[e] Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 25 June 2013
Received in revised form 7 January 2014
Accepted 11 February 2014
Available online 20 February 2014

Keywords:
Rough set
Neighborhood rough set
Attribute reduct
Bucket

ABSTRACT

In this paper, we propose an efficient quick attribute reduct algorithm based on neighborhood rough set model. In this algorithm we divide the objects (records) of the whole data set into a series of buckets based on their Euclidean distances, and then iterate each record by the sequence of buckets to calculate the positive region of neighborhood rough set model. We also prove that each record's θ-neighborhood elements can only be contained in its own bucket and its adjacent buckets, thus it can reduce the iterations greatly. Based on the division of buckets, we then present a new fast algorithm to calculate the positive region of neighborhood rough set model, which can achieve a complexity of $O(m|U|)$, m is the number of attributes, $|U|$ is the number of records containing in the data set. Furthermore, with the new fast positive region computation algorithm, we present a quick reduct algorithm for neighborhood rough set model, and our algorithm can achieve a complexity of $O(m^2|U|)$. At last, the efficiency of this quick reduct algorithm is proved by comparable experiments, and especially this algorithm is more suitable for the reduction of big data.

1. Introduction

Attribute reduct, originally presented by Pawlak's in rough set theory [1,2], is a quite useful data preprocessing technique. The core idea of attribute reduct is to obtain the sub attribute set (or feature set) which can keep the same discriminability comparing with the full attribute set. It is also mentioned as a semantic-preserving dimension reduction [3,4], thus the attribute reduct can be implemented in feature selection [5–11], data mining [12–14] and data compression [15], etc.

The classical attribute reduct algorithms are established on the equivalence approximate space and only compatible for discrete data set. They need to scatter the records when processing *continuous numerical data*, denoted by numerical data in the following paper, this will lead to losing of information (including the neighborhood structure information and order structure information in real spaces) [3,16], so the reduct of the numerical data set are strongly related with the methods of scatting. To overcome this drawback, many extensions of classical rough set theory and their corresponding definitions on attribute reduct have been presented, such as fuzzy rough set [4,17–19], tolerance approximate models [20], similarity rough approximate model [21], dominance approximation relation model [22], covering approximation model [23–26]

[*] Corresponding author at: Institute of Cyber Systems and Control, Zhejiang University, 310027 Hangzhou, China. Tel.: +86 138 0571 9977.
E-mail address: cckaffe@hotmail.com (L. Yong).

http://dx.doi.org/10.1016/j.ins.2014.02.093
0020-0255/© 2014 Elsevier Inc. All rights reserved.

and neighborhood granular model [27,28]. Among all the extensions, neighborhood rough set model [5,16,29,30] can be regarded as a specified implementation of the neighborhood granular model. The neighborhood rough set model can process both numerical and discrete data set via the θ-neighborhood set,[1] which will not break the neighborhood structure and order structure of data set in real spaces. Although the attribute reduct of neighborhood rough set model has been successfully implemented in many applications, e.g. feature selection [16], classifier [30], rule learning [31], etc., it still suffers from the low computation efficiency of attribute reduct on neighborhood rough set model, especially in computing the neighborhood of each record, which is a quite usual and inevitable operation in reduct algorithm of neighborhood rough set model.

In classical rough set model, it has been proved that finding the minimal attribute reduct is NP-hard [32], so most of the reduct solutions are aimed to find a reasonable short redut. Generally speaking, there are two main kinds of attribute reduct approaches, i.e. indiscernibility matrix based methods [32–35] and significant metric function based methods [3,10,36,37]. The indiscernibility matrix based methods need to build an indiscernibility matrix, whose elements in that matrix indicate the different attributes between every two records with different decisions (labels), and then combine all the elements in that matrix to obtain the reduct. However, these methods are hard to be extended to other approximation models, e.g. fuzzy rough set model, neighborhood rough set model, etc., furthermore, their temporal and spacial computation costs are also high; The significant metric based methods normally employ a monotonic significant metric, which is related with the positive region, to test whether current selected attribute set is a reduct. The significant metric functions used in reduct algorithms can be positive region (dependency) [3,38], inconsistency [37], and entropy [39,40], etc. The candidate attribute subsets constructing policies used in significant metric function based methods include greedy searching [38], heuristic searching [10], evolutionary computation based searching [41], etc. Among all the policies, the greedy searching is commonly used due to its high efficiency. Especially the forward greedy searching can prefer to generate shorter reduct which satisfies the "shortest bias" in classifier construction, thus it is widely used in machine learning cases. In those significant metric function based methods, sorting methods are also widely adopted to decrease the computation complexities of the attribute reduct algorithms, e.g., Nguyen proposed their attribute reduct algorithm [42] by sorting decision table, Liu proposed their attribute reduct algorithm [43] by sorting the partitioning universe, Hu and Wang proposed their attribute reduct algorithm [44] by quick sorting the two dimensional table. The significant metric function based methods are easy to be implemented in those extended approximation models. However, the time-consuming of attribute reduct algorithms on those extended models will increase significantly due to the promotion of computation complexity on differing the approximation relations between records and attributes under the extended relation models, e.g fuzzy equivalence relation model [45], θ-neighborhood relation model [16].

In this paper, we present a quick reduct algorithm on neighborhood rough set model. As we know, the neighborhood rough set model defines the neighborhood relationship between every two records with θ-neighborhood relation, which concerns the Euclidean distance between those two records less than θ $(\theta > 0)$. It needs to iterate whole the sample set to obtain one record's θ-neighborhood set, thus the computation cost to obtain the θ-neighborhood sets of every records will be approximated to the square of the record number. This complexity will become unacceptable when the number of records in data set increases to a large scale, which may be the core challenge in big data problems. Current attribute reduct algorithms on the data sets with huge records focus on reducing the number of records involving in the calculation of the positive region, such as the fast reduct algorithm (F2HARNRS) in [30]. Although this records reducing based approach can decrease the records involving in the calculation of the positive region, its efficiency is highly sensitive to the distribution of the data set.[2] Thus we present a new quick attribute reduct algorithm on neighborhood rough set model, which tries to optimize the computation of θ-neighborhood on neighborhood rough set model. To the best of our knowledge, there are no attentions are paid on reducing the computation complexity of reduction algorithm via increasing the efficiency of computing θ-neighborhood.

In our approach, we present a hash based method to divide those records into a series of sequenced buckets, and we also prove that each record's θ-neighborhood elements can only exist in its own bucket and its adjacent buckets. Based on the division of buckets, we then present a new fast algorithm, F-*POS*, to calculate the positive region of neighborhood rough set model, which only needs to iterate the records in three buckets to obtain one record's θ-neighborhood and can achieve a complexity of $O(m|U|)$, m and $|U|$ are the number of attributes and records containing in the data set. Using the F-*POS* algorithm to calculate the positive region of the neighborhood rough set model, we then present a quick neighborhood rough set reduct algorithm, which can achieve a computation complexity of $O(m^2|U|)$.

The rest of the paper is organized as follows. Section 2 presents related concepts and definitions of the neighborhood rough set model. Section 3 presents our quick attribute reduct algorithm on neighborhood rough set model. Experimental analysis is given in Section 4. Conclusions come in Section 5.

2. Neighborhood rough set model

The core concept of neighborhood rough set model is to extend the equivalent approximation of classical rough set model with neighborhood approximation, which enables it to support both numerical and discrete data types. This section will only

[1] A record x_i's θ-neighborhood set includes all the records whose distances to x_i are less than θ.

[2] Some bad distribution will lead to few records being reduced in each iteration.