Knowledge-Based Systems 73 (2015) 324–334

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# A novel approach for discretization of continuous attributes in rough set theory

Feng Jiang [a,*], Yuefei Sui [b]

[a] College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, PR China
[b] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China

ARTICLE INFO

*Article history:*
Received 16 September 2013
Received in revised form 17 October 2014
Accepted 19 October 2014
Available online 25 October 2014

*Keywords:*
Rough sets
Discretization
Supervised
Multivariate
Cuts

ABSTRACT

Discretization of continuous attributes is an important task in rough sets and many discretization algorithms have been proposed. However, most of the current discretization algorithms are univariate, which may reduce the classification ability of a given decision table. To solve this problem, we propose a supervised and multivariate discretization algorithm — SMDNS in rough sets, which is derived from the traditional algorithm naive scaler (called Naive). Given a decision table $DT = (U, C, D, V, f)$, since SMDNS uses both class information and the interdependence among various condition attributes in $C$ to determine the discretization scheme, the cuts obtained by SMDNS are much less than those obtained by Naive, while the classification ability of $DT$ remains unchanged after discretization. Experimental results show that SMDNS is efficient in terms of the classification accuracy and the number of generated cuts. In particular, our algorithm can obtain a satisfactory compromise between the number of cuts and the classification accuracy.



## 1. Introduction

In recent years, rough set theory, proposed by Pawlak [43], has attracted worldwide attention of many researchers and practitioners. It has been conceived as a tool to conceptualize and analyze various types of data. Rough set theory can be used in attribute–value representation models to describe the dependencies among attributes, evaluate the significance of attributes and derive decision rules. The theory shows important applications to intelligent decision-making and cognitive sciences, as a tool for dealing with vagueness and uncertainty of information. In addition, rough set theory has been proven to be an effective tool for feature selection, many rough set-based feature selection methods have been proposed [18,24,33,48,54–58,60,61].

Many data mining methods, including decision trees and rough set theory, would work better on discretized or binarized data [8,28]. Discretization of continuous attributes is, therefore, an important task in data mining, particularly for the classification problem. Although rough set theory has been successfully applied to various fields, discretization of continuous attributes is still one of the main problems in rough sets. In general, discretization problem can be defined as a problem of searching for a suitable set of cuts (i.e. boundary points of intervals) on attribute domains [38].

Recently, the discretization of continuous attributes has gained considerable interest in rough set theory. Many traditional discretization algorithms have been applied to rough sets. Some novel discretization algorithms have also been proposed from the view of rough sets. For instance, Min et al. proposed a divide-and-conquer discretization algorithm in rough set theory, which divides the given decision table into $K$ subtables and combines the discretization schemes of subtables into the whole scheme [34]. Jia et al. proposed an immune algorithm for the discretization of decision tables in rough set theory [25]. Singh and Minz proposed a discretization approach based on clustering and rough set theory [50]. Blajdo et al. compared the results of six promising discretization approaches from the perspective of rough sets [6]. Tian et al. proposed a core-generating discretization method, which was used as the pre-processor of rough set-based feature selection [51].

Based on the information theory, Ching et al. proposed a discretization algorithm called CADD, which takes class-attribute interdependence redundancy (CAIR) as the discretization criterion

* Corresponding author. Tel./fax: +86 532 88959036.
*E-mail addresses:* jiangkong@163.net (F. Jiang), yfsui@ict.ac.cn (Y. Sui).

http://dx.doi.org/10.1016/j.knosys.2014.10.014
0950-7051/

[9]. Based on the CADD algorithm, Kurgan and Cios further proposed the CAIM (class-attribute interdependency maximization) algorithm [29]. To overcome the two drawbacks of CAIM algorithm, Tsai et al. proposed the CACC (class-attribute contingency coefficient) algorithm, which can generate a better discretization scheme [52]. However, as pointed by Yan et al., the CACC algorithm also has its own drawbacks. Therefore, based on the rough set theory and mutual information, Yan et al. proposed a novel class-attribute interdependency discretization algorithm NCAIC [59].

The current discretization methods can be classified into two categories: unsupervised and supervised [13]. Unsupervised methods, such as equal width intervals (EW) [7] and equal frequency intervals (EF) [13] algorithms, do not make use of class information in data sets. The resulting discretization schemes do not provide much efficiency when used in the process of classification, e.g., they contain more intervals than we actually need [47].

To avoid the problems of unsupervised discretization methods, various supervised methods have been proposed, such as the entropy-based algorithms [14,15], statistics-based algorithms [28,31], class-attribute interdependency algorithms [9,29,52], Naive [42], and boolean reasoning-based methods [5,36,38–40], etc.

Supervised methods can use class information to improve their performances, but most of the current supervised algorithms are univariate (or static), which discretize each condition attribute independently of other condition attributes. Due to the nature of multi-dimensional space of real world data, the interdependence among condition attributes is also important for a discretization method [35]. Given a decision table $DT = (U, C, D, V, f)$, since univariate methods do not take into account the interdependence among condition attributes in $C$, the classification ability of $DT$ may decrease after discretization. To avoid that problem, we need some discretization methods that take into account both class information and the interdependence among condition attributes. This kind of discretization method is usually called the supervised and multivariate discretization method [36,26].

In our previous work, we have presented a supervised and multivariate discretization algorithm called SMD [26]. In this paper, we aim to propose a novel discretization algorithm called SMDNS in rough sets. SMDNS is a substantial extension and improvement of SMD [26]. Compared with SMD, SMDNS has the following advantages:

(1) In SMD, given a decision table $DT = (U, C, D, V, f)$, we used a method proposed by Nguyen and Nguyen to calculate the partition $U/B$ of $U$ induced by indiscernibility relation $IND(B)$, and the time complexity for computing $U/B$ is $O(|B| \times |U| \log_2 |U|)$ [37]. In SMDNS, we use the counting sort-based method to compute $U/B$ and the time complexity for computing $U/B$ is $O(|B| \times |U|)$. Therefore, the time complexity of SMDNS is less than that of SMD, where the former is $O(|C|^2 \times |U| \log_2 |U|)$, and the latter is $O(|C|^2 \times |U| \times (\log_2 |U|)^2)$.
(2) In SMD, we first calculated the significance of each attribute in $C$ and sorted these attributes according to their significances in descending order. Then we discretized each attribute of $C$ in that order. In this paper, we find that the computation of attribute significance is not necessary, since the order in which attributes are discretized has little influence on the result of discretization and the computation of attribute significance is time consuming. Therefore, in SMDNS, we do not calculate the significance of each attribute in $C$, which can further reduce the computational cost of SMDNS. We discretize all attributes of $C$ in their original order.
(3) In SMD, to solve the problem that the classification ability of decision table $DT$ may decrease after discretization, we proposed a criterion for merging any two adjacent intervals. However, in Jiang et al. [26], we did not investigate the validity of the proposed criterion. In this paper, we modify the criterion in Jiang et al. [26] and propose a new criterion for merging any two adjacent intervals. The new criterion imposes stricter and more efficient restrictions on the merging of two adjacent intervals, aiming to improve the classification performance. Experimental results show that SMDNS can obtain a satisfactory compromise between the number of cuts and the classification performance. By using the new criterion, SMDNS can obtain better classification performance than SMD, at the cost of a few extra cuts. In Theorem 3.1, we prove the validity of the new criterion, which can guarantee that the classification ability of $DT$ remains unchanged after discretization.

Moreover, in Jiang et al. [26], the experiments were only carried on several small and low-dimensional UCI data sets, which are insufficient to demonstrate the effectiveness of SMD algorithm in solving the complicated real-world problems, especially in dealing with the huge and high-dimensional data. In this paper, the KDD-99 data set is also used to evaluate the performance of SMDNS [27]. The KDD-99 data set is a common benchmark for evaluation of intrusion detection techniques [2]. Due to the large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, data mining approaches have been widely used in intrusion detection systems [30]. Hence, discretization of continuous attributes is also an important issue for intrusion detection.

As a supervised and multivariate discretization algorithm, SMDNS is derived from Naive. Naive is a univariate and supervised algorithm [42], which does not take into account the interdependence among condition attributes in a given decision table. When using Naive to discretize a given decision table, the classification ability of the decision table may decrease after discretization, and the cuts generated by Naive are usually more than we actually need.

To solve the problems of Naive, given a decision table $DT = (U, C, D, V, f)$, SMDNS simultaneously considers all attributes in $C$, that is, takes into account the interdependence among various attributes in $C$. In SMDNS, we iteratively merge adjacent intervals of continuous values according to a given criterion, which can guarantee that the cardinality of $C$-positive region of $D$ remains unchanged. After discretization, the classification ability of $DT$ remains unchanged and the cuts generated by SMDNS are much less than those generated by Naive.

The remainder of this paper is organized as follows. Section 2 presents some basic notions that are relevant to this paper. SMDNS algorithm is presented in Section 3. Experimental results are given in Section 4. Finally, conclusions are presented in Section 5.

## 2. Basic notions

**Definition 2.1** (*Information system*). An information system is a quadruple $IS = (U, A, V, f)$, where [44,53]:

(1) $U$ is a non-empty and finite set of objects;
(2) $A$ is a non-empty and finite set of attributes;
(3) $V$ is the union of attribute domains, i.e. $V = \bigcup_{a \in A} V_a$, where $V_a$ denotes the domain of attribute $a$;