Neurocomputing 133 (2014) 507–511

Contents lists available at ScienceDirect

Neurocomputing

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

Letters

A new approach for discretizing continuous attributes in learning systems

Deqin Yan [a,*], Deshan Liu [a], Yu Sang [b]

[a] Department of Computer Science, Liaoning Normal University, Dalian 116029, China
[b] Institute of Computing Technology, Research Institute of Exploration and Development, Liaohe Oilfield, PetroChina, No. 98, Oil Street, Panjin 124010, China

ARTICLE INFO

Article history:
Received 27 October 2013
Received in revised form
30 November 2013
Accepted 3 December 2013
Communicated by D. Wang
Available online 4 January 2014

Keywords:
Discretization algorithm
Class-attribute interdependency (CAI)
Mutual information
Rough sets

ABSTRACT

Discretization is a process to convert continuous attributes into discrete format to represent signals for further data processing in learning systems. The main concern in discretization techniques is to find an optimal representation of continuous values with limited number of intervals that can effectively characterize the data and meanwhile minimize information loss. In this paper, we propose a novel class-attribute interdependency discretization algorithm (termed as NCAIC), which takes account of data distribution and the interdependency between all classes and attributes. In our proposed solution, the upper approximation of rough sets as a prime part of the discretization algorithm is applied, and the class-attribute mutual information is used to automatically control and adjust the scope of the discretization of continuous attributes. Some experiments with comparison to five other discretization algorithms are reported, where 13 benchmarked datasets extracted from UCI database and the well-known C4.5 decision tree tool are employed in this study. Results demonstrate that in general our proposed algorithm outperforms other tested discretization algorithms in terms of classification performance.

1. Introduction

In learning system design, many algorithms have been developed with discrete attributes. Thus, it is useful to transform real valued attributes into discrete format (symbol attributes) to facilitate the existing algorithms or tools. Indeed, discretization of continuous attributes as a part of data preprocessing takes place in data mining, pattern recognition, machine learning and rough sets analysis [1]. It is obvious that the quality of discretization will directly affect the system performance on classification or discovery. Therefore, it is important to develop advanced algorithms to deal with discretization.

There exist various discretization algorithms in the literature [2–14]. These algorithms can be clustered as supervised or unsupervised types [2–5], the global scope or the local scope where various volumes of instances are used [6–9], the static or dynamic process depending on the existence of mutual influences between the attributes [10–12], and direct ways versus the incremental methods to determine the number of intervals [13,14]. In unsupervised discretization algorithms, typical strategies are to conduct the equal-width and equal-frequency discretization [3]. On the other hand, in supervised discretization algorithms, a possible solution can be obtained through optimizing a cost function, for example, the maximum entropy algorithm [4,15], class-attribute interdependency algorithm [16], and statistics-based [2,17–19]. In [16], based on information theory, a discretization algorithm (class-attribute dependent discretization, or simply CADD) was proposed, which takes class-attribute interdependence redundancy (CAIR) as an important discretization criterion on the selection of the candidate cutting-points that could lead to better correlation between class and discrete intervals. The main weakness of this algorithm is the utilization of a user-specified number of intervals during the process of initialization. Consequently, a certain type of training for the reasonable selection of intervals is needed in the algorithm. It has been noticed that the maximum entropy discretization method used in the procedure of the CADD algorithm [16] may result in the worst starting points. With class-attribute interdependency maximization (CAIM) criterion, CAIM algorithm proposed in [20] could avoid the disadvantages of CADD. However, CAIM algorithm has two drawbacks [21], i.e., simple discretization with a smaller number of generated intervals, and the use of the distribution of the major target class. To overcome these drawbacks, a class-attribute contingency coefficient algorithm (CACC) was proposed in [21].Unfortunately, the criterion employed in CACC may bring mistakes for some cases.

In principle, discretization algorithms should maximize the interdependency between discrete attribute values and class labels, whilst

* Corresponding author.
E-mail addresses: yandeqin@163.com (D. Yan), deshanliu@yeah.net (D. Liu), songyu2008bj@sina.com (Y. Sang).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.005

minimize the information loss due to the process of discretization. This paper aims to develop an advanced discretization algorithm based on the criteria used in CADD [16], CAIM [20], and CACC [21] algorithms. A novel class-attribute interdependency discretization algorithm (termed as NCAIC) is developed, where a new criterion, fully reflecting the information of interdependency between all the classes and attributes, is proposed in our NCAIC algorithm. Moreover, the class-attribute mutual information is taken in to account in the procedure of the algorithm, which could be helpful to automatically control and adjust the scope of the cutting-points. In order to show the advantages of the proposed NCAIC algorithm, 13 benchmark datasets, five existing discretization algorithms and the well-known C4.5 tool [22] are employed in this study. Simulation results are promising and demonstrate good potential of our proposed algorithm in dealing with the discretization problem.

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 details our proposed discretization algorithm. Section 4 reports the experimental results with comparisons and some discussions. Section 5 concludes this paper.

2. Related work

2.1. Problem formation

Given a training dataset with M samples, where each sample belongs to only one of S classes with a continuous attribute A. A discretization scheme D on the A would discretize the continuous domain of attribute into n discrete intervals bounded by pairs of real numbers [1]

$$D:\{[d_0,d_1],(d_1,d_2],\dots,(d_{n-1},d_n]\},\tag{1}$$

where the d_0 is the minimal value and the d_n is the maximal value of the attribute A. In particular, the values in (1) are arranged in the ascending order, and these values constitute the boundary set $\{d_0,d_1,d_2,\dots,d_n\}$ for discretization scheme D, where d_i, $i=0,1,\dots,n$, are called cutting-points.

Each value of attribute A can be classified into only one of the n intervals defined in (1). The membership of each value within a certain interval for attribute A needs to be adjusted according to the change of the discretization scheme D. Also, the class variable and the discretization variable of attribute A are treated as two random variables, which will define a two-dimensional frequency matrix (named quanta matrix). For instance, as shown in Table 1, q_{ir} is the total number of continuous values belonging to the i-th class within interval $(d_{r-1},d_r]$, and C_i is the class label of the i-th class. M_{i+} represents the total number of objects belonging to the i-th class, M_{+r} is the total number of continuous values of attribute A that are within the interval $(d_{r-1},d_r]$, for $i=1,2,\dots,S$, and $r=1,2,\dots,n$.

The Shannon entropy and class-attribute information (CAI) can be presented with a given quanta matrix [16,20]

$$H(C,D|A)=\sum_{i=1}^{s}\sum_{r=1}^{n}p_{ir}\cdot\log_2\left(\frac{1}{p_{ir}}\right),\tag{2}$$

Table 1
Quanta matrix for attribute A and discretization scheme D.

Class	Interval	Class total
	$[d_0,d_1],\dots,(d_{r-1},d_r],\dots,(d_{n-1},d_n]$	
C_1	$q_{11},\dots,q_{1r},\dots,q_{1n}$	M_{1+}
$\vdots$	$:,\dots,:,\dots,:$	$\vdots$
C_i	$q_{i1},\dots,q_{ir},\dots,q_{in}$	M_{i+}
$\vdots$	$:,\dots,:,\dots,:$	$\vdots$
C_s	$q_{s1},\dots,q_{sr},\dots,q_{sn}$	M_{s+}
Interval total	$M_{+1},\dots,M_{+r},\dots,M_{+n}$	M

$$I(C,D|A)=\sum_{i=1}^{s}\sum_{r=1}^{n}p_{ir}\cdot\log_2\left(\frac{p_{ir}}{p_{i+}p_{+r}}\right),\tag{3}$$

where $p_{ir}=q_{ir}/M$ is a probability of situations when values of attribute A fall into the interval $(d_{r-1},d_r]$ and simultaneously belong to the class C_i, $p_{i+}=M_{i+}/M$ is a probability when the values belong to the class C_i and p_{+r} is a probability if the values fall into the interval.

The goal of discretization algorithms is to find an effective way to produce cutting-points [1,16,20,21]. Therefore, a good algorithm of discretization should have a reasonable criterion to decide which points should be the cutting-points. Note that in general the goodness of discretization algorithms is associated with the classification system performance.

2.2. Existing solutions

This subsection reviews some existing algorithms, aiming to get better understandings on the weaknesses of these solutions. First, we present the metrics to measure the performance of a discretization scheme. In this paper, the class-attribute interdependency (CAI) [16], reflecting the information from a continuous attribute, is used as a discretization performance evaluation metric. In CADD algorithm [16], this metric is defined as

$$CAIR(C,D|A)=\frac{I(C,D|A)}{H(C,D|A)},\tag{4}$$

where $H(C,D|A)$ and $I(C,D|A)$ are Shannon entropy and mutual information, respectively.

The CADD algorithm has the following disadvantages [16,20]:

(i) It uses a user-specified number of intervals during the initialization process of discretization intervals.
(ii) The significance test used in the algorithm requires training for the selection of a reasonable interval.
(iii) It initializes the discretization intervals by using a maximum entropy discretization method. Such initialization may lead to the worst starting points.
(iv) The over-fitting phenomena may occur.

In 2004, Kurgan and Cios proposed the CAIM discretization algorithm [20], where the criterion metric used in their algorithm is defined as

$$CAIM(C,D|A)=\frac{1}{n}\sum_{r=1}^{n}\frac{\max_r^2}{M_{+r}},\tag{5}$$

where n is the number of intervals and $\max_r$ is the maximum value among all q_{ir} values (maximum value within the r-th column of the quanta matrix).

The essential idea of CAIM algorithm is to maximize the interdependency between the continuous-valued attribute and its class labels, and also achieve the minimum number of discrete intervals possibly. In [21], CACC algorithm which is proposed with a corresponding criterion metric is defined as

$$CACC(C,D|A)=\sqrt{\frac{x}{x+M}},\tag{6}$$

where

$$x=\frac{M}{\log(n)}\left(\sum_{i=1}^{s}\sum_{r=1}^{n}\frac{q_{ir}^2}{M_{i+}M_{+r}}-1\right).$$

Compared with CAIM, the information of the data distribution is taken into consideration in this criterion. Although there is a little improvement on CAIM discretization algorithm, CACC algorithm suffers from the following drawbacks: