Information Sciences 270 (2014) 73–91

Contents lists available at ScienceDirect

Information Sciences

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

An effective discretization method for disposing high-dimensional data

Yu Sang [a], Heng Qi [a], Keqiu Li [a,*], Yingwei Jin [b], Deqin Yan [c], Shusheng Gao [d]

[a] School of Computer Science and Technology, Dalian University of Technology, No. 2, Linggong Road, Dalian 116023, China
[b] School of Management, Dalian University of Technology, No. 2, Linggong Road, Dalian 116024, China
[c] School of Computer and Information Technology, Liaoning Normal University, No. 850, Huanghe Road, Dalian 116029, China
[d] 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 15 September 2012
Received in revised form 20 November 2013
Accepted 15 February 2014
Available online 28 February 2014

Keywords:
Feature discretization
High-dimensional data
Dimension reduction
Local Linear Embedding (LLE)

ABSTRACT

Feature discretization is an extremely important preprocessing task used for classification in data mining and machine learning as many classification methods require that each dimension of the training dataset contains only discrete values. Most of discretization methods mainly concentrate on discretizing low-dimensional data. In this paper, we focus on discretizing high-dimensional data that frequently present the nonlinear structures. Firstly, we present a novel supervised dimension reduction algorithm to map high-dimensional data into a low-dimensional space, which ensures to keep intrinsic correlation structure of the original data. This algorithm overcomes the deficiency that the geometric topology of the data is easily distorted when mapping data that present an uneven distribution in high-dimensional space. To the best of our knowledge, this is the first approach to solve high-dimensional nonlinear data discretization with a dimension reduction technique. Secondly, we propose a supervised area-based chi-square discretization algorithm to effectively discretize each continuous dimension in the low-dimensional space. This algorithm overcomes the deficiency that existing methods do not consider the possibility of being merged for each interval pair from the view of probability. Finally, we conduct the experiments to evaluate the performance of the proposed method. The results show that our method achieves higher classification accuracy and yields a more concise knowledge of the data especially for high-dimensional datasets than existing discretization methods. In addition, our discretization method has also been successfully applied to computer vision and image classification.

1. Introduction

Discretization is one of the preprocessing methods used frequently in data mining, machine learning and knowledge discovery [50,12,20], which has been generally used to slice the value domain of each continuous dimension into a finite number of intervals associated with a discrete value. The significance of discretization methods derives from interest in extending to continuous dimension classification methods, such as decision trees, rough set theory, Bayesian classifier or Bayesian networks, which are used for dealing with discretized data. Discretization can also facilitate the interpretation of the obtained results and to improve the accuracy of classification tasks [29,48]. It has been used for some applications such as medical diagnosis [34,38] as a necessary preprocessing step.

[*] Corresponding author. Tel.: +86 411 84709242.
E-mail address: keqiu@dlut.edu.cn (K. Li).

http://dx.doi.org/10.1016/j.ins.2014.02.113
0020-0255/© 2014 Elsevier Inc. All rights reserved.

Existing discretization methods, such as chi2-based heuristic algorithms [24,30,44,43,8], class-attribute interdependency algorithms [11,26,46,31], entropy-based methods [16,25] and correlation-based discretization methods [3,10,32] are proposed to find good partition of each continuous dimension of a dataset. Presently, there are some works addressing the discretization of high-dimensional data. Entropy-based methods perform well on high-dimensional data regarding both the discretization intervals and classification accuracy. Ferreira et al. [17] proposed an incremental supervised FD technique based on recursive bit allocation. The method can achieve the highest mutual information with the class label after discretization. Mehta et al. [32] proposed a correlation preserving discretization algorithm based on principal component analysis (PCA), which can discretize high-dimensional data by considering the correlation structure of the data. However, high-dimensional data now in the real-world frequently present the nonlinear structures; therefore, it is still a challenging task to study more efficient discretization methods for nonlinear high-dimensional data.

In this paper, we present a novel high-dimensional data discretization method. The main contributions of this paper are summarized as follows:

1. We present a supervised dimension reduction algorithm for data discretization, which effectively maps high-dimensional data into a lower intrinsic dimensional space. This method keeps the intrinsic correlation structure of the original data and overcomes the deficiency of the geometric topology of data, which is easily distorted when mapping data that presents an uneven distribution in high-dimensional space.
2. We propose a supervised area-based discretization algorithm to effectively discretize each continuous dimension in the low-dimensional space. This algorithm overcomes the deficiency of chi2-based methods using the change of chi-square as merging criterion to discretize the data without considering the possibility of being merged for each interval pair from the view of probability.
3. We conduct the experiments results on real-world and synthetic datasets to evaluate the performance of the proposed method by comparison with some popular discretization methods. The experimental results show that the proposed method outperforms existing methods over the performance metrics considered. Furthermore, we also apply our proposed method to computer vision and image classification.

The remainder of this paper is organized as follows. We introduce related work in Section 2. We present our proposed method in Section 3. Experiments and performance evaluation are presented in Section 4. Finally, we summarize our work and conclude this paper in Section 5.

2. Related work

Existing discretization methods mainly focus on disposing low-dimensional data. Liu et al. [29] and Tsai et al. [46] present a taxonomy of discretization methods including several main axes: bottom-up vs. top-down, and supervised vs. unsupervised [15] and so on. Top-down methods, such as class-attribute interdependency discretization, start from the initial interval and recursively split it into smaller intervals. Bottom-up methods, such as chi2-based discretization, begin with the set of single value intervals and iteratively merge adjacent intervals.

In the unsupervised methods, continuous ranges are divided into subranges by the user specified width (range of values) or frequency (number of instances in each interval). There are not many unsupervised methods available in the literature which may be attributed to the fact that discretization is commonly associated with the classification task. Unsupervised methods provide no class information, such as equal-width and equal-frequency [15], kernel density estimation (KDE) [6], tree-based density estimation (TDE) [42]. The equal-width and equal-frequency methods can be implemented with a low computational cost, and the EFB method [49] with naive Bayes (NB) classification produces good results. The KDE and TDE methods are state-of-the-art unsupervised top-down methods, which use density estimators to select the best cut-points and automatically adapt subintervals to the data. They determine the discretized number of intervals by the cross-validated log-likelihood.

Supervised methods provide class information with each feature value and they are much more sophisticated, such as entropy-based discretization method, class-attribute interdependency methods, and the chi2-based heuristic algorithms. Entropy-based method proposed by Fayyad and Irani [16] recursively selects the cutpoints on each target feature to minimize the overall entropy and determines the appropriate number of intervals by using the minimum description length principle (MDLP). Kononenko's method [25], a variant of Fayyad and Irani's method, analyzes the biases of eleven measures for estimating the quality of the multi-valued features. The values of these measures tend to linearly increase with the number of values of a feature. Whereas, Kononenko introduces a function based on the MDL principle whose value slightly decreases with the increasing number of feature's values.

Class-attribute interdependency methods are distinguished top-down supervised discretization methods with the objective to maximize the interdependency between the class and the continuous-valued feature and to generate a possibly minimal number of discrete intervals. The chi2-based methods are famous bottom-up supervised discretization methods based on statistical independence. The chi-square statistic is used to determine whether the current interval pair is to be merged or not. These methods trade off the number of intervals with the number of inconsistent instances and control the process of discretization by introducing inconsistency with the aim to control the degree of misclassification.