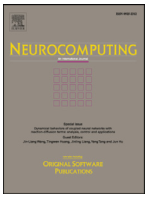
Contents lists available at ScienceDirect

Neurocomputing

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



An efficient data reduction method and its application to cluster analysis

Jianpei Wang, Shihong Yue*, Xiao Yu, Yaru Wang

School of Electrical Engineering and Automation, Tianjin University, 300072, Tianjin, China

ARTICLE INFO

Article history:
Received 28 August 2016
Revised 20 January 2017
Accepted 22 January 2017
Available online xxx

Communicated by Jun Yu

Keywords:
Data reduction
Dimensionality reduction
Sample reduction
Clustering validity

ABSTRACT

Data reduction plays a very important role in the data mining field, but the existing methods have not been able to efficiently identify all major features which are hidden in the large datasets. On some occasions, they even cause the loss of the original key features. In this paper, a new efficient measure was developed to reduce a given dataset and to uncover the major features by multiplying the defined absolute density with the defined local density of any data. These two kinds of densities were estimated with a fast grid-based bisecting method. To test its performance on feature reduction and sample reduction, a group of feature-different synthetic datasets and 24 benchmark datasets were used as examples and the clustering accuracy, runtime and separability among clusters were used as measurements. The results strongly proved the proposed method could fast reduce a dataset and identify the most important key features. Additionally, it also can effectively determine the optimal number of clusters by suppressing the noisy data and enhancing the separation among clusters.

1. Introduction

New information technologies and data mining techniques have led to an explosion in assessable data, resulting in two fundamental problems: high dimensionality and massive data [1–3]. High dimensionality often means the prevalence of a large number of useless/ineffective features that tend to keep key features hidden, while redundant data not only take up storage and memory space, but also affect classification accuracy and decision correctness. Data reduction under these conditions becomes a necessary technique for discovering key properties wherein irrelevant or unimportant information is reduced from a massive dataset. The main operating principles and methods behind data reduction are dimensionality reduction and sample reduction [4–6].

In the past decades, data reduction techniques have been mainly focused on the feature reduction that can be categorized as linear and nonlinear groups. The most widely used feature reduction method is the Principal Component Analysis (PCA) [7]. It projects the data into a lower-dimensional space where the samples can be separated by a linear classifier (under the condition of that the samples can be separated by density in the input space). Thus the usage of PCA is based on the assumption of the linear separability of the input data. In the case of datasets with non-linear separability, an efficient method is the Locally Linear Embedding (LLE) [8]. LLE has recently been proposed as a general method for dimensionality reduction of high-dimensional nonlinear datasets. Besides these, more recently, some novel dimensionality reduction algorithms have been proposed for image clustering and multi-view learning [9–11]. Based on the patch alignment framework, a novel semi-supervised dimensionality reduction was introduced for multi-view data analysis [9]. The hypergraph was used to integrate different relationships from different views to overcome the over-simplified assumption of pairwise relationships among data in most methods. After that, a sparse patch alignment framework [10] was proposed for the embedding of data underlying in multiple manifolds, which adopt an optimization strategy for constructing local patches instead of a fixed neighborhood size. For the semi-supervised multi-task feature selection problems, a manifold regularized multi-view feature selection method was presented in [11]. It exploited the label information, label relationship, data distribution, as well as correlation among different kinds of features simultaneously. More review can be found in [12–14].

However, as the amount of information increases exponentially, reducing the total amount of massive data becomes necessary and inevitable. Sample reduction provides an efficient solution to this problem. The goal of sample reduction is to separate the unimportant details from the essential structures hidden in the data, while the structure of the data can be shown by various clusters. Thus, the clusters at various levels are a mostly visual approach, and are intended to maximize insight into a dataset, as well as uncover the

* Corresponding author.
E-mail address: shyue1999@tju.edu.cn (S. Yue).

http://dx.doi.org/10.1016/j.neucom.2017.01.059
0925-2312/© 2017 Elsevier B.V. All rights reserved.

structure of the data. In general, any useful algorithm is based on a set of requirements that are intended to advance the use of the data reduction as a data structure discovery technique; the technique must also be practical in terms of computational effort with the following characteristics:

(i) It always is validated for arbitrary data structures such as density-different, size-different, and irregular clusters after data reduction.

(ii) It must not be affected by partially overlapped clusters. Inversely, the reduced dataset should have better separation among clusters than the original dataset. Thus the overlapped parts among clusters must be reduced as data reduction proceeds.

(iii) It should be practically unsupervised and not utilize parameters defined for unrealistic prior information.

(iv) It should be non-iterative. An iterative algorithm usually is time-consuming against the linear time complexity. Therefore, a partitioning dataset is favorable as far as clustering efficiency is considered.

As such, there are primarily two classes of sample reduction methods: sampling and vector quantization [12–15]. Sampling is a method that is generally used to quickly reduce data in an investigated dataset; it is popular due to its simplicity and easy execution, as random sampling (RS) [16]. The disadvantage of this method is that it is not capable of recovering data structures and features, which can result in loss of key information in the data reduction process. Another class of data reduction method involves the use of learning vector quantization (LVQ) algorithms [17]. LVQ is different from sampling methods in that the learning process is intuitive and involves simple implementation. Originally, the class of methods is designed to overcome the limitation of RS methods by greatly adding the runtime of data reducing process. But the class of methods is not completely unsupervised and is limited in its ability to recover key features in the data reduction process [18].

To resolve the above problems, in this paper, we proposed a new data reduction method based on a fast grid-based bisecting (GB) method [19]. By reformulating the GB algorithm, a new measure was calculated by optimally multiplying the absolute density and local density of the data. Even though the absolute density in our method is similar to other density measures used in many other clustering algorithms, e.g., DBSCAN [20], but we proposed a new method that could estimate the absolute density much faster in an unsupervised way. Most importantly, our method does not need any user-specified threshold. Therefore it can be used to analyze any data. Then a local density is also proposed to overcome the problem that the existing absolute density cannot abstract the local clustering structure of the data. By multiplying the two densities, the proposed new data reduction method not only efficiently recovers the major feature, but also rapidly reduces the amount of data and greatly suppresses noisy data. Its non-iterative and grid-based data process has a linear time complexity. Thus it can greatly facilitate the computational loads of the data reducing process.

We further apply the proposed method to an important problem in cluster analysis, that is, quantitative validation of the used clustering results, including determining the optimal number of clusters and their configurations. In past decades, research has focused on various validity indexes. Some extensive and good overviews of clustering algorithms can be found in the literature [21–25]. Estimating the correct number of clusters by cluster validity indexes in cluster analysis is highly susceptible to data structure, separation among clusters and noisy data, so the correctness of the estimated number of clusters is difficult to be guaranteed. In this paper, the validity index is used to estimate the number of clusters in the reduced datasets. A group of synthetic datasets and 24 benchmark datasets are used to validate the proposed method.

Table 1
C-MEANS algorithm.

Algorithm: *C*-MEANS
Input: A dataset X of n data; The number of clusters c.
Output: c clusters of n data.
Method:

```
1. Randomly choose $c$ data objects in $X$ as the initial cluster centers, $v_1$, $v_2$, …,$v_c$;
2. Repeat;
3. (Re)assign any data to the most similar cluster based on a chosen similar measure;
4. Update cluster centers;
5. Stop if a convergence criterion is met;
6. Otherwise, go to step 2.
```

Merits: simplicity, easy operation and realization.
Time complexity: $O(ctn)$.

2. Related work

In this section, we introduced five related algorithms, including two clustering algorithms, *C*-MEANS and GB, and four data reduction algorithms, RS, LBG, PCA and LLE, as well as five typical cluster validity indexes as following.

2.1. *C*-MEANS and GB algorithms

Assume that $X=\{x_1, x_2, \ldots, x_n\}$ is a set of n data in a D-dimensional data vector space, and C_1, C_2, …, C_c are c sets, then the belongingness of the jth data vector in X to the ith set can be represented by a binary membership function as

$$u_{ij} = \begin{cases} 1, & x_j \in C_i \\ 0, & x_j \notin C_i \end{cases}, \quad i = 1, 2, \ldots, c, \quad j = 1, 2, \ldots, n. \tag{1}$$

If any data vector in X is only assigned to a unique set, then all vectors must be assigned into c sets. This is called as a hard partitioning of X, satisfying

$$\begin{aligned} &X = C_1 \cup C_2 \cup \cdots \cup C_c, \\ &C_i \cap C_j = \phi,\ i, j = 1, 2, \ldots, c \end{aligned} \tag{2}$$

Currently, the class of partitional clustering algorithms is designed based on objective function optimization, amongst which the *C*-MEANS algorithm [22] is well accepted and the most used clustering method because of its simplicity, robustness, and easy application. The detailed steps of the *C*-MEANS algorithm and its time complexity are shown in Table 1, where c is the number of clusters, t is the totally iterative times, and any cluster center (clustering prototype) is computed by

$$v_j = \sum\nolimits_{j=1}^{n} u_{ij} x_j \sum\nolimits_{j=1}^{n} u_{ij},\ \ i = 1, 2, \ldots, c \tag{3}$$

In [19] a fast grid-based bisecting (GB) algorithm has been applied to partition any dataset into a group of grids. GB algorithm successively bisects every edge of any grid in each dimension by turns; the dataset is then partitioned into 2, 2^2,…, 2^J grids after J times, where J is the maximal bisecting times when no grid contains two points. Fig. 1 shows the bisecting process by a simple example by a dataset of 26 data points.

Table 2 shows the general bisecting process of GB. The time complexity of GB is linearly $O(n)$ while *C*-MEANS is $O(ctn)$. So GB is much faster than *C*-MEANS (see Table 1).

2.2. Four typical data reduction algorithms

Four typical data reduction algorithms, two for sample reduction and the other two for dimensionality reduction, are introduced below.

(1) Two sample reduction methods. Random sampling (RS) is a method of selecting n units out of all N samples to make