Knowledge-Based Systems 98 (2016) 216–225

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Clustering boundary detection for high dimensional space based on space inversion and Hopkins statistics

Baozhi Qiu, Xiaofeng Cao*

School of Information Engineering, Zhengzhou University, Zhengzhou 450001, PR China

ARTICLE INFO

Article history:
Received 21 October 2015
Revised 29 December 2015
Accepted 26 January 2016
Available online 3 February 2016

Keywords:
Clustering boundary
High dimensional space
Space inversion
Symmetry Statistics

ABSTRACT

Physicists research the symmetry of particle space through the contrast of motion law in the real space and inversion space which is created by space inversion techniques. Inspired by this theory, we propose the idea of using local space transformation and dynamic relative position to detect the clustering boundary in high dimensional space. Due to the curse of dimensionality, global space transformation approaches are not only time-consuming, but also fail to keep the original distribution characteristics. So, we inverse the space positions of the k nearest neighbors and project them on the high dimensional space coordinate system. To address the lack of statistics that can describe the uniformity of high dimensional space, we propose the Symmetry Statistics based on the Hopkins Statistics. It is employed to judge the uniformity of k nearest neighbor space of coordinate origin. Moreover, we introduce a filter function to remove some special noises and isolated points. Finally, we use boundary and filter ratios to detect the clustering boundary and propose the corresponding detection algorithm, called Spinver. Experimental results from synthetic and real data sets demonstrate the effectiveness of this algorithm.

1. Introduction

Extracting the valuable patterns [1–3] in complex information space is the key problem of data mining. The patterns will help people to understand data space, and get more valuable information. So, there is no doubt that the extraction process can be described as segmenting the information space, and different segmentation methods will lead to different pattern results. The clustering techniques [4–8], which use none experience way [9], classify the unlabeled data objects by grouping objects that are similar to each other. By analyzing the similarities and differences between different classes, it can get the stable structure of data space, i.e. clusters. In addition to this familiar pattern, we find a more interesting pattern-clustering boundary.

Clustering boundary [10], which is located at the edge of a cluster is a special pattern. Data objects within the cluster have the same class label, but there exists some differences between the interior objects and the boundary objects. Boundary objects, in many applications, may indicate special targets that need to pay close attention on, for example, people who have infected with some virus but not be attacked, side face images in the front face images, irregular handwriting signatures, the gene mutation individuals in the gene expression datasets, and etc. When scholars are delighted to find such interesting researches, they also find the traditional data mining technology cannot extract the clustering boundary. In recent years, some clustering boundary detection techniques have been proposed, such as BORDER [11,12], BRIM [13], BAND [14], BRINK [15], and etc. However, the research on clustering boundary is not so extensive as that clustering technology, especially for the high dimensional data space [16–18]. So, this study will be aimed at the clustering boundary pattern discovery in high dimensional space.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 reports the clustering boundary model. Section 4 presents experiments and performance results on a number of synthetic and real data sets. Section 5 provides an intuitive discussion on parameters analysis and scalability. Our conclusion is given in Section 6.

2. Related work

Compared with low dimensional space, high dimensional space has more complex space structure [19,20]. Because of the inherent sparsity of the data objects, the most existing clustering algorithms that based on only the similarity measures between data objects will become substantially inefficient. To solve the problem, PCA technique [21] chooses the main dimensions to represent the whole space. Then, a series of similar techniques which focus on the choice of a reasonable subspace have been proposed, including

* Corresponding author. +86 18739920964.
E-mail address: 18739920964@163.com (X. Cao).

http://dx.doi.org/10.1016/j.knosys.2016.01.035
0950-7051/

Linear Discriminant Analysis (LDA) [22], Isomap [23], Locally Linear Embedding (LLE) [24], Laplacian Eigenmaps [25], Local Preserving Projection (LPP) [26], Local Tangent Space Alignment (LTSA) [27,28], Maximum Variance Unfolding (MVU) [29], and etc. From the perspective of information theory [30], these techniques can be described as a process of information compression. However, the information compression will lose some important information, and the subspace cannot reflect the whole space structure. So, the data analysis results depend on the selection of subspace, and different subspace will lead to different patterns. To keep the complete space information, spectral clustering technique [31] transforms the whole space to a new space to finish the clustering task. Besides, we find that there exist many techniques using the method of space transformation, such as the clustering techniques using SVM [32,33] and artificial neural network [34–36], etc. SVM transforms the data space from low dimensional to high dimensional to deal with the problem of linearly inseparable. In other words, this method will analyze the data objects in a high dimensional space, but it may be caught in 'Dimension Disaster'. Artificial neural network transforms the data space to a similar brain system or a map structure system. Although this method will solve many problems in complex structure space, but it may make a simple problem more complex in the high dimensional space with a small number of samples, and make a complex problem even more complex in the high dimensional space with a large number of samples.

To tackle the problems above, we hold a long-term research. We find that the space inversion [37,38] of the particle space physics and uniform distribution provide the theoretical basis for the research of high dimensional space. Space inversion is a method used to study the symmetry of particle space in the microscopic world. It reverses the spatial feature, such as the direction of forces, the direction of time, and etc. In other words, physicists take the inverse values to replace the position of each particle to establish an image space which has the similar structure in the original space. Compared with the motion law of particles in the image space and original space, scientists can judge the symmetry of particle space. Therefore, scientists propose the vector inversion about time, the geometric inversion about mathematics, the quantitative inversion about geography, and etc.

Inspired by this idea, we establish a high dimensional coordinate system for the current data points, and use the relative position to finish the detection task of clustering boundary. Unlike the traditional space transformation methods which transform the whole data space to a new space, we transform the local space to a new space, and use the dynamic relative position to replace the static position. Particularly, The specify way is that we give each data object a different positions in different local spaces, so that the relative position changes dynamically.

Another theory used in this paper is the uniform distribution. It describes the uniformity of data space based on probability statistics, and has widely used in the fields of computer science and physics, and etc. For example, researchers of data mining use the Hopkins statistics to describe the uniformity of clusters or evaluate the clustering quality of clustering analysis results. However, this statistics cannot describe the uniformity in the high dimensional data space. Generally, researchers use the dimension projection technique to analyze the space distribution with respect of each dimension. In other words, it projects data objects on certain dimensions to get the distribution of the dimensions. Compared with the dimension reduction methods, dimension oriented technique keep the whole space information, and has the characteristics of rapid calculation. After all, it costs a relatively small amount of time in one dimensional space or one dimensional array.

Though many algorithms about clustering in high dimensional space have been proposed, there are few papers which focus on clustering boundary in high dimensional space. So, these observations motive our effort to propose a clustering boundary detection algorithm based on space inversion and Hopkins statistics, called Spinver. The main contributions of this paper are as follows:

(1) propose a high dimensional space inversion technique to extract the local space features;
(2) propose a Symmetry Statistics to describe the uniformity of high dimensional data space;
(3) propose a clustering boundary detection algorithm for high dimensional data space named Spinver.

3. Clustering boundary model

In this section, we will provide the idea of space inversion and projection technique. Then, we propose the Symmetry Statistic which could describe the uniformity of high dimensional space based on Hopkins Statistics. Lastly, we develop the Spinver algorithm.

3.1. Space inversion and projection

Spherical [39] and cube sampling [40,41] are popular sampling methods used to analyze data. They all pay attention on fixed sampling window, and less attention on the data points located outside the window. So, they all belong to static sampling. More importantly, they will prejudice to the data points located at the surface of space, and cannot present the true feature of data distribution. So, in this paper we use the k nearest neighbor [42,43] as sampling method.

Given a n dimensions space S, we take x_i as the center of data space to establish a n dimensions coordinates system, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$.Then we calculate the k nearest neighbors of x_i and reverse their positions to get the relative positions. The rule of space inversion is described as follows:

$$\begin{aligned} RLocation(x_j) &= x_j - x_i \\ &= (x_{j1} - x_{i1}, x_{j2} - x_{i2}, x_{j3} - x_{i3}, \ldots, x_{jn} - x_{in}) \end{aligned} \quad (1)$$

where x_j is the k nearest neighbors of x_i and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})$. We take x_i as the original point of local space, and new coordinates are assigned to the k nearest neighbors.

Clustering boundary objects are located at the edge of clusters, and it is k nearest neighbors are not uniformly distributed. Core points are located inside clusters, and their k nearest neighbors sit around them evenly. However, different with boundary and core points, noises always have large distances between their neighbors and themselves, so noises are distributed not uniformly. Based on the analysis above, we can detect the clustering boundary by judging the local uniformity of every data point. To describe the uniformity of k nearest space, we project the data objects on each dimension. The rule of projection is described as follows:

$$PLocation(x_j) = \begin{bmatrix} x_{j1} - x_{i1} & 0 & 0 & \ldots & 0 \\ 0 & x_{j2} - x_{i2} & 0 & \ldots & 0 \\ 0 & 0 & x_{j3} - x_{i3} & \ldots & 0 \\ \vdots & \ldots & \ldots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & x_{jn} - x_{in} \end{bmatrix} \times (1 \le i \le n) \quad (2)$$

$$PLocation(x_j, d_i) = (0, 0, \ldots, x_{jd_i} - x_{id_i}, \ldots, 0)(1 \le i \le n) \quad (3)$$

where $PLocation(x_j, d_i)$is the projection coordinate of x_i on the d_i dimension. In this formula, we use the way of dimension oriented to extract the features of local space. So, the high dimensional space is divided into n one-dimensional spaces. This means that the high dimensional space is be decomposed.