Neurocomputing 000 (2017) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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

A clustering algorithm using skewness-based boundary detection

Xiangli Li, Qiong Han*, Baozhi Qiu

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

ARTICLE INFO

Article history:
Received 7 April 2016
Revised 17 December 2016
Accepted 4 September 2017
Available online xxx

Communicated by zhi yong Liu

Keywords:
Skewness
Boundary degree
Clustering algorithm
Clustering boundary

ABSTRACT

Clustering analysis has been applied in all aspects of data mining. Density-based and grid-based clustering algorithms are used to form clusters from the core points or dense grids to extend to the boundary of the clusters. However, deficiencies are still existed. To find out the right boundary and improve the precision of the cluster, this paper has proposed a new clustering algorithm (named C-USB) based on the skew characteristic of the data distribution in the cluster margin region. The boundary degree calculated by skew degree and the local density are used to distinguish whether a data is an internal point or non-internal point. And the connected matrix is constructed by removing the neighbor relationships of non-internal points from the relationships of all points, then the clusters can be formed by searching from the connected matrix towards internal of the clusters. Experimental results on synthetic and real data sets show that the C-USB has higher accuracy than that of similar algorithms.

1. Introduction

Clustering refers to a process to discover the internal structures of data or the potential data models in a dataset [1–3] by data partitioning. Thanks to the outstanding capability of discover clusters of different shapes and sizes along with outliers, density-based [4–6] and grid-based [7] clustering technology are widely applied to the fields of health care [8], information security [9], internet [10] and etc [11–15].

Data points are divided into core points, boundary points and noise points by the DBSCAN algorithm [16], and a cluster is formed when the data is expanding from the core points outwards the clustering boundary. As those methods are susceptible to parameter changes, different parameters may lead to different data dividing and clustering results. IS-DBSCAN [17], ISB-DBSCAN [18] and others [19–21] are proposed by making use of the nearest neighbor relationship instead of the neighborhood density, which effectively reduce the influence of the parameters on the algorithm. However, for the multi-density datasets, the clustering results are not always favorable because neighbor relationships can misjudge the boundary points.

Grid clustering technique divide grids into high-density ones and low-density ones with compressing expression and clusters are formed when high-density grids are connected. Grid-based clustering technologies, such as CLIQUE [22], MGM-GA [23] and etc [24,25], are efficient because grid clustering is formed with the extension of grid cells. Such an approach can be efficient, however, in the clusters forming process, if a dense grid is adjacent to a sparse grid (we called boundary grid), which probably contains noises, the algorithm is of low clustering accuracy.

Boundary points not only play a significant role in expansion-based clustering algorithms, but also in other fields of data mining. The PAC-Bayes boundary theory, a theoretical framework, combines Bayes theory [26–28] with minimum structural analysis principle of random classifier, obtaining the most generalized risk boundary. The algorithms derived from PAC-Bayes boundary are actually the “average” of Hypothesis Space, thus achieving a better classification performance [29–31]. Support Vector Machine (SVM) [32–34] also uses boundary points to improve performance. Furthermore, on the occasion of supervised, Compression Nearest Neighbor (CNN) [35] can extract the neighboring data boundary points from different classes, and it can also used to reduce the number of support vectors in the SVM algorithm, which is helpful to reduce training costs [36–39]. Besides, the study on boundary is also contributing to discover interesting models in data [40–42]. For instance, in the medical field, the clustering boundary may represent a group of people, who carry virus but not affected. With regard to handwriting recognition, the clustering boundary may stand for handwriting images which are easily misjudged to be other characters.

2. Motivation

From the standpoint of density-based clustering technology, clusters refer to the dense regions separated by sparse regions. For

* Corresponding author.
E-mail address: qhanzzu@163.com (Q. Han).

https://doi.org/10.1016/j.neucom.2017.09.023
0925-2312/

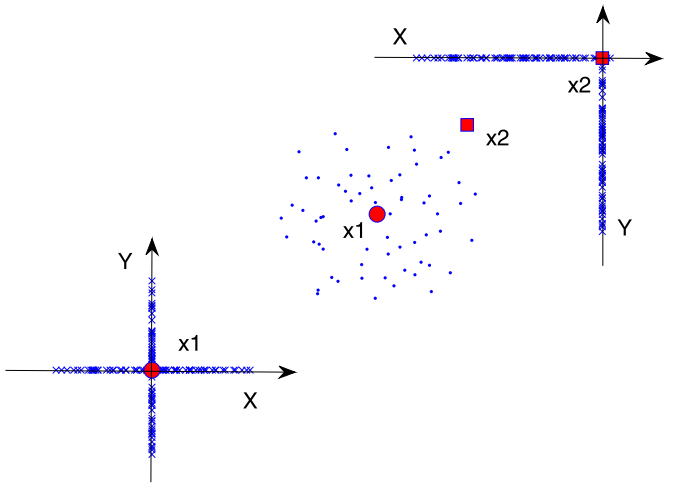

Fig. 1. Point x_1, x_2 and distribution of points after mapping on coordinate axis.

grid-based clustering technology, clusters are formed by the connected high-density grid units. Both of the technologies are adopting the expansion-based method to form clusters from core points (dense grids) to boundary points (boundary grids). And the similar clustering pattern makes it possible for the combination of them to solve cluster problems [43–45]. Boundary points, as the terminal condition of expansion, are of great importance to both technologies. In density-based clustering algorithm, the boundary points are identified by density only. However, in real datasets, margin density is probably equal to or larger than that of the internal region. Grid-based clustering algorithm determines boundaries by statistics of points within a single grid, which always causes the internal points or noise points which located nearby the boundary grid to be misjudged as boundary points.

Based on huge data analysis, internal points are founded to be surrounded by their neighboring points, while the neighboring points of boundary point are always located in one side of it. Fig. 1 shows that, x_1 and x_2 are taken as reference points respectively, and the points around them are mapped into X and Y axis. After the mapping of central x_1, other points can be found to symmetrically spread on the two sides of x_1 in different axis after mapping. However, when x_2 is regarded as the reference point, other points are located in just one side of x_2 in X or Y axis only after mapping; thus, the mappings of point x_2 lie in a skew pattern. Based on the above features, this paper proposes a new skewness-based measurement method to separate boundary objects from a dataset. Unlike the density-based or grid-based method, boundary points in this method are determined by the distribution of their neighbors, which can effectively avoid the effects of density and neighboring radius.

3. Algorithm

3.1. Skewness and boundary degree

Suppose a dataset $X = \{x_i | x_i \in R^{m*n}; i = 1, 2, ..., n; m, n \in N\}$.

Definition 1 ($Skewness\, S_c(x_p)$)**.** $Skewness\, S_c(x_p)$ is used to measure the data skew distribution, defined as follows:

$$S_c(x_p) = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k}(x_{ij} - x_{pj})^2}{k} \qquad x_{ij} \in N_{k-dist}(x_p). \tag{1}$$

Among which $N_{k-dist}(x_p)$ [46] refers to k nearest neighbor of x_p, $x_p \in X$.

Skewness has been widely utilized in the field of data statistics and analysis [47–49]. This paper aims to study whether the distribution of neighboring points of x_p are skew or not. Therefore, x_p itself is taken as the reference point. Left skewness or right skewness has no impact on the study, so the definition of $S_c(x_p)$ is changed to a non-directional value.

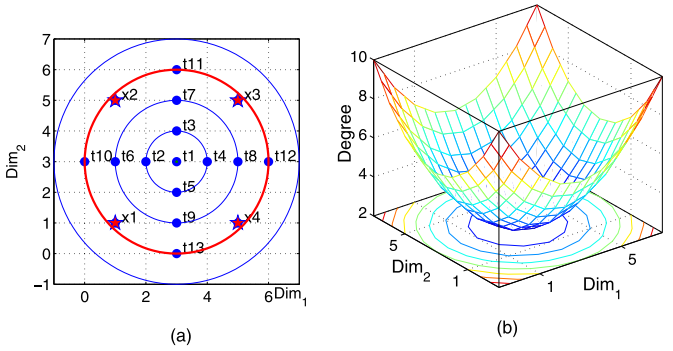

Fig. 2. Boundary degree testing model. (a) Points $x_1 - x_4$ and Positions $t_1 - t_{13}$ (b) 3-dim view of boundary degree.

Definition 2 (Local density $D_{local}(x_p)$)**.** The *local density* refers to the compactedness of point x_p and its neighboring points. *Local density* is the reciprocal of summing up standard deviations drawn from different dimensions of x_p and its neighboring points. The definition is as follow:

$$D_{local}(x_p) = \frac{1}{\sum_{j=1}^{m}\sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_{ij} - \bar{x})^2}} \qquad x_{ij} \in N_{k-dist}(x_p). \tag{2}$$

$\bar{x}$ refers to the average value of x_p and its k nearest neighbors. In the sparse areas, the value of $D_{local}(x_p)$ is relatively small. In the dense areas, the value of $D_{local}(x_p)$ is relatively large.

Definition 3 (Boundary Degree)**.** *Boundary degree* refers to the degree of boundary of data points. The *boundary degree* of x_p is calculated by *skewness* $Sc(x_p)$ multiplying *local density* $D_{local}(x_p)$. The definition is as follow:

$$Boundary\, degree = S_c(x_p)D_{local}(x_p) = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k}(x_{ij} - x_{pj})^2}{k\sum_{j=1}^{m}\sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_{ij} - \bar{x})^2}}. \tag{3}$$

Where, $\bar{x}$ refers to the average value of x_p and its k nearest neighbors. The smaller the *boundary degree* of data point x_p is, the closer x_p gets to the central position of a cluster. But when the *boundary degree* gets larger, x_p will get closer to or be on the cluster margin position of a cluster.

3.2. Discussions on boundary degree

To clearly and quantitatively study the changes of boundary degree, this paper has examined the features of boundary degree with modeling. Suppose that x_1, x_2, x_3, x_4 are the neighboring points of point p and point p is movable. Fig. 2(a) shows that, t_1 is located in the central position of x_1, x_2, x_3, x_4. Expand t_1 to get three groups of positions, including (t_2, t_3, t_4, t_5), (t_6, t_7, t_8, t_9) and (t_{10}, t_{11}, t_{12}, t_{13}). The circumradius of the three groups is apart with a distance unit from one to another. (t_{10}, t_{11}, t_{12}, t_{13}) are located on the margin position of a circumcircle formed by x_1, x_2, x_3, x_4. In accordance with the skewness definition, the value of boundary degree should increase from t_1 to t_{13}. Upon calculation, the boundary degree value at relevant positions is shown in Table 1, while the 3-dimensional graph of boundary degree is shown in Fig. 2(b).

From Table 1, the value of boundary degree of t_1 is the smallest. Based on the position of t_1, $t_2 - t_5$ move outwards with a

Please cite this article as: X. Li et al., A clustering algorithm using skewness-based boundary detection, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.023