Data & Knowledge Engineering 95 (2015) 23–38

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

Design of computationally efficient density-based clustering algorithms

Satyasai Jagannath Nanda *, Ganapati Panda

Department of Electronics and Communication Engineering, Malaviya National Institute of Technology Jaipur, Rajasthan 302017, India
School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Odisha 751013, India

ARTICLE INFO

Article history:
Received 1 September 2012
Received in revised form 5 May 2014
Accepted 24 November 2014
Available online 29 November 2014

Keywords:
Clustering, classification, and association rules
Mining methods and algorithms
DBSCAN
Fast DBC
Physical action datasets
Seismic catalog of Japan

ABSTRACT

The basic DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm uses minimum number of input parameters, very effective to cluster large spatial databases but involves more computational complexity. The present paper proposes a new strategy to reduce the computational complexity associated with the DBSCAN by efficiently implementing new merging criteria at the initial stage of evolution of clusters. Further new density based clustering (DBC) algorithms are proposed considering correlation coefficient as similarity measure. These algorithms though computationally not efficient, found to be effective when there is high similarity between patterns of dataset. The computations associated with DBC based on correlation algorithms are reduced with new cluster merging criteria. Test on several synthetic and real datasets demonstrates that these computationally efficient algorithms are comparable in accuracy to the traditional one. An interesting application of the proposed algorithm has been demonstrated to identify the regional hazard regions present in the seismic catalog of Japan.

1. Introduction

Density based clustering algorithms classify a dataset into groups based on the crowdedness of points in the multidimensional space. They consider clusters as high-density areas in the data space separated by regions of low density. Recent literature [1–4] reflects that these algorithms have become popular among the data mining researchers due to their potential to detect clusters of arbitrary shape and ability handle outliers present in a dataset. Further these algorithms do not need a priori information about the number of clusters or their shapes as input parameters, as the number of clusters evolve with the progress of algorithm [5–7].

First Ester et al. [8,9] have proposed the density-based clustering algorithm called DBSCAN (Density-Based Spatial Clustering of Applications with Noise) which uses minimum number of input parameters and useful to cluster large spatial databases. An incremental version of DBSCAN (incremental DBSCAN) is reported by Ester et al. [10]. It is used for incremental updates of a clustering result after insertion of a new object or deletion of an existing object from the database. The incremental algorithm yields the same result as the DBSCAN. Sander et al. [11] have suggested a generalized DBSCAN (GDBSCAN), which can cluster point as well as spatially extended objects (spatial and nonspatial attributes). A parallel version of DBSCAN (PDBSCAN) is designed by Xu et al. [12]. The DBCluC (Density-Based Clustering with Constraints) is proposed by Zaiane and lee [13] by employing the DBSCAN to cluster spatial data in the presence of obstacles. The VDBSCAN algorithm [14] detects cluster with varied density as well as automatically selects several values of input threshold distance parameter (*Eps*) for different densities. Other parameters (ex. k) of VDBSCAN are automatically generated based on the characteristics of the datasets [15]. The VDBSCAN has the same time complexity as that of DBSCAN. The DBSCAN fails to identify clusters with local density variation (different densities that exist within a cluster). In order to overcome this limitation Ram

* Corresponding author at: Department of Electronics and Communication Engineering, Malaviya National Institute of Technology Jaipur, Rajasthan 302017, India.

http://dx.doi.org/10.1016/j.datak.2014.11.004
0169-023X/

et al. [16] have suggested DVBSCAN which is capable to handle clusters separated by the sparse region as well as clusters with variable density. The time complexity of this algorithm is higher compared to that of the DBSCAN.

Density based algorithm OPTICS (ordering points to identify the clustering structure) is introduced by Ankerst et al. [17] which computes an augmented cluster ordering for automatic and interactive cluster analyses. Both DBSCAN and OPTICS analyze the neighborhood of an object, and both depend on spatial index structure such as R*-tree [18] or X-tree [19] to efficiently process the neighborhood queries. Since the efficiency of answering these queries with these indices reduces with increase in the number of dimensions, the two methods are inefficient for high-dimensional data. Therefore another algorithm DENCLUE (DENsity based CLUstEring) proposed by Hinneburg et al. [20], models the overall point density analytically as the sum of influence functions the data points. This density based algorithm provides a compact mathematical description of arbitrarily shaped clusters in high-dimensional datasets and is significantly faster than its counterpart algorithms. A hybrid clustering algorithm BRIDGE [21] integrates the k-means and DBSCAN. The BRIDGE enables the DBSCAN to handle very large databases and simultaneously increase the performance of kmeans by discarding noise. The CUBN proposed by Wang and Wang [22] integrates the density based and distance-based clustering. It finds border points using the erosion operation and then starts clustering the border and the inner points according to the nearest distance. Ma and Zhan [23] proposed an adaptive density based clustering (ADBC) algorithm that uses adaptive strategy (based on spatial object distribution) for neighbor selection to improve clustering accuracy. The ST-DBSCAN algorithm [24] has the potentiality to discover clusters with non-spatial, spatial and temporal values of the objects. The DBSCAN fails to detect noise points of varied density but this algorithm assigns density factor to each cluster and removes the lacunae.

In the DBSCAN [8], the process of merging the preliminary clusters, obtained just after determining the core points, consumes a lot of time. This merging criterion determines the distance between all the points in two different clusters and compares it with the threshold distance to find out whether the two clusters should be merged.

In this paper instead of using the individual points, we propose the use of mean of the associated patterns to merge the primary clusters. This simple idea drastically reduces the computational complexity of the algorithm. Further, a new density based clustering algorithm is proposed considering correlation coefficient as the evaluation and merging criteria. The computational complexity of the proposed correlation algorithm is further reduced by applying new merging conditions. The performance evaluation of the proposed algorithms is carried out using two synthetic, two low dimensional (*Iris*, *Small Soybean*) and four high dimensional Vicon physical action datasets.

The cluster analysis of seismic events provides vital information about the identification of seismotectonic patterns present in a specific geographic region. This information extracted from huge amount of seismic events is helpful for evaluating regional seismic hazards, determining aftershocks related to main quakes, rise of seismic activities prior to a large earthquake and constructing empirical earthquake forecast methodology. First Omori Law [33] provides the empirical formulation which describes the clustering of aftershocks in time. Barenblatt et al. [34] have proposed that most earthquakes come in clusters (clusters especially consist of main-shock and its aftershocks). Initially Gasperini and Mulargia [35] have applied cluster analysis to determine the complex tectonic activities of Italian territory. Ogata et al. [36,37] have explored the clustered nature of earthquakes using Epidemic Type Aftershock Sequence (ETAS) model. Dzwinel et al. [38] have dealt the multidimensional scaling and visualization of earthquake clusters over space and time. In another communication Vecchio et al. [40] have analyzed the statistical properties of earthquake clustering and have observed that the clustered events are locally Poissonian which signify the presence of correlation in the events of quakes. The ETAS model has been suggested by Zhuang et al. [39] to stochastically cluster the earthquakes present in Japan catalog. Zaliapin et al. [41] used cluster analysis to identify the aftershocks in the southern California region. The fuzzy clustering has been suggested by Ansari et al. [42] to identify the seismotectonic provinces present in the Iranian plateau. Recently Nanda et al. [43] proposed a tri-stage cluster identification model to categorize the hazardous aftershocks present in California and Indonesian Catalog.

Being inspired by the recent trend of research in the cluster analysis of earthquake catalogs, in this paper we have employed the popular density based algorithm DBSCAN and have suggested the Fast DBC algorithm to effectively cluster the seismic events of Japan Catalog. It is observed that the result automatically evolved clusters obtained with the new approach, effectively classifies the regional hazard regions considering the magnitude and depth of seismic events.

The rest of the paper is organized as follows. Section 2 deals with the basic density based algorithm the DBSCAN. The computationally efficient DBSCAN algorithm, use of correlation coefficients in density based approach along with its computationally efficient version is proposed in Section 3. The performance evaluation of the new algorithms on synthetic and real life datasets is carried out in Section 4. The application of proposed algorithms to analyze the seismic catalog of Japan is presented in Section 5. The concluding remarks of the paper are outlined in Section 6.

2. Basic density based clustering algorithm: DBSCAN

The DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a fundamental density based approach for cluster analysis proposed by Ester et al. [8]. The objective is to determine cluster of arbitrarily shape. In this approach the density for a particular point in the dataset is represented by counting the number of points present in a specified radius (ϵ) of that point. The density also includes the point itself. The algorithm is very popular as it is simple to implement and requires only two input parameters (ϵ and minimum number of objects in a cluster *Mn_Pts*). Further, the algorithm supports the user to determine an appropriate value for input parameters. The algorithm can easily be described with the help of some fundamental definitions outlined in sequel:

Definition 1. Distance neighborhood of a point — The ϵ neighborhood of a point y in the dataset $P_{N\times D}$ is defined as

$$N_\epsilon(y) = \{z \in P : d(y,z) \leq \epsilon\} \tag{1}$$