Future Generation Computer Systems 89 (2018) 166–177

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

CANF: Clustering and anomaly detection method using nearest and farthest neighbor

Azadeh Faroughi, Reza Javidan *

Computer Engineering and IT Department, Shiraz University of Technology, Shiraz, Iran

HIGHLIGHTS

- A New nearest and farthest neighbor selection method is proposed.
- Our method computes radius of data subgroups based on variance of data points.
- It doesn't need to consider all data to compute the density.
- Anomaly detection is possible in more complex cases.
- It has low time complexity.
- Setting a parameter causes the model to fit to various datasets.
- PCA is used in proposed clustering estimator to reduce dimensions of dataset.
- It is tested on synthetic and real datasets and its feasibility is demonstrated.
- The new approach reduces the time complexity from $O(n^2)$ to $O(n \log n)$.

ARTICLE INFO

Article history:
Received 3 November 2017
Received in revised form 2 April 2018
Accepted 18 June 2018
Available online 26 June 2018

Keywords:
Nearest neighbor density estimator
Farthest neighbor
Subgroups
Anomaly detection
Clustering
Principal component analysis (PCA)

ABSTRACT

Nearest-neighbor density estimators usually do not work well for high dimensional datasets. Moreover, they have high time complexity of $O(n^2)$ and require high memory usage, especially when indexing is used. These problems impose limitations on applying them for small datasets. In order to overcome these limitations, we proposed a new method called CANF which stands for clustering and anomaly detection using nearest and farthest neighbors. This method calculates distances to nearest and farthest neighbor nodes to create dataset subgroups. Therefore, computational time complexity is of $O(n \log n)$ and space complexity is constant. In each iteration of subgroup formations, outlier points of subgroups are detected. After subgroup formation, a proposed assembling technique is used to derive correct clusters. CANF uses a new parameter to detect clusters which are not easily separable. Many experiments on synthetic datasets are carried out to demonstrate the feasibility of CANF. Furthermore, on real-world datasets we compared this algorithm to similar algorithms in anomaly detection task and in clustering task namely LOF and DBSCAN, respectively and the results showed significantly higher accuracy of the CANF, especially in high dimensions. Moreover, to overcome high dimensional datasets problems, Principal Component Analysis (PCA) is used in the clustering method, which preprocesses high-dimensional data. The results showed the effectiveness of the proposed method both for clustering as well as anomaly detection applications.

1. Introduction

The main purpose of clustering is to find the structure of unlabeled datasets [1]. Data should be partitioned into clusters in a way that intra-clustering similarity becomes greater and inter-clustering similarity becomes less. Unlike classification methods in which each data is assigned to precaution groups, in clustering there is no prior information about the class membership of the data, and in fact, clusters are extracted from data information itself. Data clustering can be used in some applications such as marketing, biology, analysis, classification of network traffic, image processing, time series forecasting, machine learning, pattern recognition and natural language processing [2–7] and it is

* Corresponding author.
E-mail addresses: A.faroughi@sutech.ac.ir (A. Faroughi), javidan@sutech.ac.ir (R. Javidan).

https://doi.org/10.1016/j.future.2018.06.031
0167-739X/

potentially useful in other fields [8–12]. Clustering presents some challenge such as selecting a suitable clustering algorithm that can handle a huge number of dimensions and distributed data. Some of the density-based clustering and anomaly detection methods are based on nearest neighbor density estimators. The time complexities of these methods are $O(n^2)$, because they need to find the nearest neighbor for every data in the dataset. Consequently, utilizing these methods are impractical when the dataset is large [13–15]. There are other methods that use k-nearest neighbor method to find the nearest neighbor in the dataset [15,16]. Although, these methods reduce the time complexity to $O(n \log n)$, these indexing algorithms have high memory requirement and this speedup only occurs in datasets with few dimensions [17].

In [18] we have already proposed an approach for density clustering based on finding nearest and farthest neighbors. This approach first creates some subgroups of dataset. Subgroups are implemented in a way that nearest neighbor of each sample node is selected among the set of sample nodes and then radius of each subgroup is computed according to distance between them and variance of these two subgroups, where variance at first iteration is zero. After that, the nodes that are in the region of each subgroup are assigned to nearest subgroup. Inside of each subgroup, farthest node from previous representative node is selected as the new representative node. According to these new representative nodes, regions of subgroups are expanded. Finally, some of these subgroups are merged together to form clusters. Since the number of subgroups is smaller than the size of the dataset, the algorithm has an appropriate speed. This new paper is expansion of our previous work in which, to detect anomaly, during the formation of the subgroups, the density of the border nodes of each subgroup is investigated. If density of each border node is much less than other border nodes, then the node is identified as an anomaly. The number of the border nodes in each subgroup is defined as the logarithm of the number of the nodes belonging to a subgroup. Therefore, the CANF can detect scattered and clustered anomalies. In addition, since [18] had a capability to generate at most $2 \times \log n$ clusters, the algorithm has limited performances when applied on datasets which require a larger amount of clusters to be characterized. In this new paper, it is possible to add some new sample nodes among unlabeled nodes during the subgroups formations. Therefore, the number of clusters is determined according to the condition of the problem. On the other hand, since many of the dimensions in high dimensional datasets are often irrelevant, using this algorithm to find the clusters of high dimensional data may lead to wrong results and generate a wrong number of clusters of real-world datasets, because these irrelevant high dimensions can confuse the clustering algorithm by hiding clusters in noisy data. Principal component Analysis (PCA) [19] is used to improve CANF to overcome the previously mentioned drawback.

The main contributions of this paper are:

1- A New nearest and farthest neighbor selection method is proposed to define dataset subgroups. This method differs from the ones that use k-nearest neighbor or a fixed radius to create subgroups. Our method computes radius of subgroups based on variance of data points.
2- In order to do anomaly detection during subgroups formations, we compute the ratio of the density of each border node of each subgroup to average the density of other border nodes of the same subgroup. So CANF does not need to consider all the data to compute the density and making anomaly detection possible in more complex cases and reducing low time complexity.
3- This approach uses a parameter for assembling subgroups to form main clusters. If the ratio between the density of middle point of every two subgroups and the density of these subgroups is higher than a pre-defined parameter, they are considered as the same cluster. Setting this parameter appropriately allows the model to fit various datasets
4- PCA is used in the proposed clustering estimator to reduce the dimensions of the dataset.
5- CANF is tested on synthetic datasets and its feasibility is demonstrated. In addition, the performance of CANF is evaluated on real datasets to compare its performance in anomaly detection and clustering to those of LOF and DBSCAN, respectively. The new approach reduces the time complexity from $O(n^2)$ to $O(n \log n)$ and when the dimension of the dataset is high, PCA-based method has a better performance.

The rest of this paper is organized as follows: In Section 2 some related works about common clustering methods are reviewed. In Section 3 the proposed method is described in detail and describes how the new method is applied to anomaly detection and clustering tasks. Section 4 explains the proposed algorithm based on PCA. In Section 5, experimental results on synthetic datasets and real word datasets are presented. Time and space complexity calculations for CANF are explained in Section 6. The final Section covers the conclusion.

2. Related works

Clustering is defined as unsupervised classification of data into groups or clusters. Various types of clustering algorithms have been proposed and developed in the literature (e.g., [20] and the references therein). Generally, clustering methods are divided into three main categories: partitioning approaches, hierarchical approaches, and density-based approaches.

In *partitioning approaches,* various partitions are constructed and then evaluated according to some criteria such as the minimum sum of square errors. Typical methods of these approaches are k-means [21] and CLARANS [22]. These methods are simple, and they converge to a local optimum very fast. However, the limitation of these methods is that the number of clusters must be predefined and they do not work well with clusters of different sizes and shapes.

In *Hierarchical approaches*, a hierarchical decomposition of datasets is created using some criteria. CURE [23] and CHAMELEON [24] are some methods of these approaches. These methods are suitable for clusters with different sizes and shapes, but their complexity is high, and their convergence is slow.

Density-based algorithms such as DBSCAN [25], SSN [26] and OPTICS [27] and MSC [28] are based on the connectivity and densities of the datasets. In these approaches, clusters are zones with high density of data which are separated by regions of lower density. In these methods, clusters can be arbitrarily shaped, and the number of clusters is automatically and simultaneously determined during the operation of clustering. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a dense region (*minPts*). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficient points, a cluster is started. Otherwise, the point is labeled as noise.

Some nearest neighbor algorithms like LOF [15] and DBSCAN determine a local neighborhood based on a global parameter, i.e., k or ε, and the density is calculated based on these variables. In addition, these algorithms consider the entire dataset to find the nearest neighbors, which leads to time complexity of $O(n^2)$ for n nodes. To overcome this execution complexity, some researches have focused on reducing this cost by employing different indexing methods while these algorithms need high memory.

There are some ensemble approaches (e.g., Feating [29] and $\text{Local}_{\text{Model}}$ [30]) that are specifically designed for only classification tasks and they build individual models using the entire dataset. Furthermore, both Feating and $\text{Local}_{\text{Model}}$ use a tree structure to define local regions.