Applied Soft Computing Journal 73 (2018) 623–634

Contents lists available at ScienceDirect

Applied Soft Computing Journal

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

Application of a density based clustering technique on biomedical datasets

Md Anisur Rahman *, Md Zahidul Islam

School of Computing and Mathematics, Charles Sturt University, Panorama Avenue, Bathurst, NSW 2795, Australia

HIGHLIGHTS

- Evaluation of a density based clustering technique (DenClust) on biomedical datasets.
- DenClust produces number of clusters and initial seeds using a deterministic process.
- DenClust performs better than six existing techniques on twenty biomedical datasets.
- An empirical analysis to evaluate the quality of initial seeds was also performed.
- Sign test results indicate the superiority of DenClust over the existing techniques.

ARTICLE INFO

Article history:
Received 1 August 2017
Received in revised form 3 September 2018
Accepted 8 September 2018
Available online xxxx

Keywords:
Clustering
Cluster evaluation
K-means
Data mining
Machine learning
Biomedical datasets

ABSTRACT

The detection of the number of clusters in a biomedical dataset is very important for generating high quality clusters from the biomedical dataset. In this paper, we aim to evaluate the performance of a density based K-Means clustering technique called DenClust on biomedical datasets. DenClust produces the number of clusters and the high quality initial seeds from a dataset through a density based seed selection approach without requiring an user input on the number of clusters and the radius of the clusters. The high quality initial seeds for K-Means results in high quality clusters from a dataset. The performance of DenClust is compared with six other existing clustering techniques namely CRUDAW-F, CRUDAW-H, AGCUK, GAGR, K-Means, and K-Means++ on the twenty biomedical datasets in terms of two external cluster evaluation criteria namely Entropy and Purity and one internal cluster evaluation criteria called Sum of Squared Error (SSE). We also perform a statistical non-parametric sign test on the cluster evaluation results of the techniques. Both the cluster evaluation results and statistical non-parametric sign test results indicate the superiority of DenClust over the existing techniques on the biomedical datasets. The complexity of DenClust is $O(n^2)$ but the overall execution time of DenClust on the datasets is less than the execution time of AGCUK and GAGR having $O(n)$ complexity.

1. Introduction

Clustering is an unsupervised learning technique. It groups similar records in a cluster and dissimilar records in different clusters. It extracts hidden patterns from large datasets that help in decision making processes in various fields including medical research, crime detection/prevention, social network analysis and market research [1–7]. Therefore, it is very important to produce good quality clusters from a dataset.

K-Means is one of the top ten data mining techniques because of its simplicity [8]. It is a widely used clustering technique, where the number of clusters (k) needs to be provided by a user even before the clustering process starts [9–12]. Based on a user defined number of clusters, K-Means first randomly selects k number of records as the initial seeds. It then goes through the clustering processes and produces k clusters.

However, one of the limitations of K-Means is its requirement of the user input on k. It can be difficult for a user (data miner) to estimate the correct value for k [13,14]. Another limitation of K-Means is the possibility of selecting poor quality initial seeds due to its random seed selection criteria. A set of poor quality initial seeds may produce poor quality clusters [9,15,16].

* Corresponding author.
E-mail address: arahman@csu.edu.au (M.A. Rahman).

https://doi.org/10.1016/j.asoc.2018.09.012
1568-4946/

Some existing clustering techniques called CRUDAW-H and CRUDAW-F [1,17] obtain the initial seeds and the number of clusters automatically through a deterministic process. The deterministic process of CRUDAW-H and CRUDAW-F requires a user defined threshold called r (radius of a cluster), which is then used for producing the initial seeds and the number of clusters for a dataset. However, it can be difficult for a user to correctly estimate a suitable value of r especially when the user does not have good understanding on the dataset.

However, some existing clustering techniques work on the datasets that have only numerical attributes or for datasets that have only categorical attributes while in reality many datasets have both numerical and categorical attributes [1,15,18–20]. There are techniques that can work on both numerical and categorical attributes, but some of them do not consider any similarity between categorical values while clustering the records, in the sense that if two categorical values (of an attribute) belonging to two records are different then the distance between the two records in terms of the attribute is considered to be 1 (regardless of the similarity of the values), and otherwise 0 [1,15,19–21].

K-Means is one of the popular clustering algorithms [22]. A recently proposed modified version of K-Means that selects initial seeds deterministically has been applied successfully on biomedical datasets [23]. However, the technique requires the number of cluster k as an input which might be hard for a user to estimate the correct number of clusters for a dataset [23]. Moreover, the technique does not work on a dataset having any categorical attributes.

In this study, we present a novel clustering technique called DenClust that works on biomedical datasets having categorical and/or numerical attributes. DenClust produces the number of clusters k and the high quality initial seeds through a deterministic process without requiring a user input on any parameters such as k and r. The density based seed selection approach of DenClust works well on biomedical datasets. However, DenClust also works well on other kinds of datasets. The initial seeds produced by DenClust are the centers of a dense region and they are expected to represent the natural clusters. Therefore, the initial seeds are expected to be of high quality. Our empirical analysis presented in Section 4.8 to evaluate the quality of initial seeds also supports this expectation. Our experimental results also indicate the superiority of DenClust.

In this study, we mainly aim to evaluate the performance of DenClust on biomedical datasets. We collect twenty (20) biomedical datasets from the UCI Machine learning repository and Bioinformatics [6,7,24]. In the experimental evaluation of this paper, we apply DenClust on these biomedical datasets whereas in the conference version of DenClust paper [25] we used only three datasets. To evaluate the performance of DenClust, we implement DenClust and six (6) other existing clustering techniques namely CRUDAW-F and CRUDAW-H [1,17], K-Means [10], K-Means++ [26], AGCUK [19] and GAGR [20]. We compare DenClust with the existing techniques using two external cluster evaluation criteria namely Entropy and Purity and one internal cluster evaluation criteria called Sum of Square Error (SSE) [1,27,10].

The experimental results indicate that the performance of DenClust is better than the existing techniques in terms of three cluster evaluation criteria on the biomedical datasets used in this study. Moreover, we perform a statistical non-parametric sign test on the cluster evaluation results of the techniques that suggests statistical significance of DenClust over the existing techniques on these biomedical datasets. We also present an empirical analysis on the quality of the initial seeds produced by DenClust as well as an empirical analysis on the T value (a user defined parameter), for the considered datasets. The complexities and the execution time of the techniques are also presented in this study.

The structure of the paper is as follows. In Section 2, we discuss some existing clustering techniques. DenClust is presented in Section 3. The experimental results on biomedical datasets and discussions are presented in Section 4. We provide some concluding remarks in Section 5.

2. Literature review

2.1. Dataset

In this study, we consider a dataset D having n records $D = \{R_1, R_2, \ldots, R_n\}$, and m attributes $A = \{A_1, A_2, \ldots, A_m\}$. The attributes of a dataset can be categorical and/or numerical.

2.2. K-Means and K-Means++

K-Means requires a user to input the number of clusters k. It then randomly selects k records as the initial seeds from a data set [10,11]. All other records are assigned to the nearest seeds to form the initial set of clusters. Based on the records in each cluster, K-Means re-calculates the seed of each cluster [10,28]. All records of the dataset are assigned again to different clusters in such a way that a record is assigned to the cluster, the seed of which has the minimum distance with the record. The process continues until one of the termination conditions (user defined number of iterations or a minimum difference between the values of the objective function in two consecutive iterations) are satisfied. K-Means++ is a modified version of K-Means, where the initial seeds of the clusters are selected using a probabilistic approach [26]. However, the number of clusters k in K-Means++ is a user defined parameter. K-Means and K-Means++ also do not work on a dataset that has both categorical numerical attributes.

2.3. Bisecting-K-Means (BKM)

Bisecting K-Means (BKM) is a variation of K-Means which also selects initial seeds randomly [29]. At the beginning, it considers the whole dataset as one cluster and then it divides the whole dataset into two partitions using K-Means. From the two partitions, BKM picks one partition as a cluster and the remaining partition is considered as a dataset for partitioning using K-Means once again. BKM again applies K-Means to partition the dataset (remaining partition) to two sub-partitions. The process of partitioning the records continues until it reaches the desired number of clusters. From the two partitions, one partition is selected for further division based on the size of the cluster (partition) or the similarity of the records within the cluster. In BKM, for some cases refinement may be required on the initial clusters to produce the final clusters from a dataset. Moreover, another limitation of BKM is that it requires a user input on the number of clusters.

2.4. Basic Farthest Point Heuristic (BFPH) and New Farthest Point Heuristic (NFPH)

Basic Farthest Point Heuristic (BFPH) [30] requires the number of clusters k as a user input. It then randomly selects a record as the first initial seed. However, unlike K-Means, the other seeds are selected deterministically. The record having the maximum distance with the first seed is selected as the second seed. For the selection of the third seed, the distance between a record and its nearest seed is used. The record having the maximum distance (with its nearest seed) is considered as the third seed. The seed selection process continues until BFPH produces the user defined number of initial seeds or runs out of records. The initial seeds are given to K-means to produce the final clusters.

New Farthest Point Heuristic (NFPH) [30] also requires the number of clusters as an input. However, it selects all seeds (including