Indian Journal of Science and Technology, Vol 8(22), DOI: 10.17485/ijst/2015/v8i22/79131, September 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Shared Nearest Neighbour Density based Clustering Approach on a Proclus Method to Cluster High Dimensional Data

S. Gayathri[1*], M. Mary Metilda[2] and S. Sanjai Babu[1]

[1]Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; gay_sri123@yahoo.com, ssanjai2000@gmail.com
[2]Queen Mary's College, Chennai - 600004, Tamil Nadu, India; metilda_dgvc@yahoo.co.in

Abstract

Background/Objective: A high dimensional data is a dataset that ranges from a few to a hundreds of dimensions. Clustering such datasets needs an efficient algorithm such as Proclus but the algorithm has a drawback of ignoring cluster with small data points. So the proposed paper gives an ensemble of clustering that combines technique of two clustering algorithms to achieve a quality cluster of even small data points. **Methods/Statistical Analysis:** The research paper adapts a novel method of implementing a density based approach over a Proclus algorithm to cluster even small data points. These combined algorithms are tested using synthetic datasets. The Proclus algorithm is modified at a specific point where the density based algorithm is implemented. **Findings:** The results of the proposed algorithm are found to contain more clusters than mere Proclus algorithm does. The results is as such because in Proclus clustering the data point whose size is small are ignored so that only clusters with large number of data points can exists. However after the involvement of the shared nearest neighbor density based algorithm even the small data points are clustered which paves way for a more accurate and an efficient clustering process especially in a high dimensional data. **Applications/Improvements:** The application is a combination of two efficient algorithms but implemented in a simple way thereby reducing the complexity of the algorithm. The proposed technique can be applied on all high dimensional datasets irrespective of their sizes and shapes.

Keywords: Density based Approach, High Dimensional Data, Proclus, SNN Algorithm

1. Introduction

Clustering technique has reached its level of importance and usage at the dawn of e-business marketing. As known clustering is grouping of meaningful collection of objects that have similar characteristics so that a classified data can be furnished. A varied form of clustering has been used and implemented at many fields like information retrieval, biology, climate, business, medicine and psychology. The greater the distance between the groups, greater the cluster is distinct. Partitioning a single database into a classified group of clusters forms the basic function of clustering. In the research paper, clustering is implemented on a high dimensional data. A high dimensional dataset is defined as a dataset which has more number of dimensions. As the dimensions are high the data are scattered throughout the dataset which results in a scarcity of data for clustering, called the curse of dimensionality. To overcome this setback many algorithms has been proposed by researchers to achieve an efficient clustering.

The research paper uses one such algorithm called Proclus. In a dataset consisting of, high dimensional data as the range of dimensionality increases the last neighborhood of point is supposed to be almost same as close as its closest neighbor for a wide range of distance of distance functions and data distribution which forms the main motivation for projected clustering[1]. The sparsity of data in a high dimensional space poses a major challenge in the clustering of data. Clusters normally may exist in different

*Author for correspondence

portions of a dataset called subspace which can be considered as the subsets of these attributes. In a cluster the similarity measures between two points are done by the distance functions. There are many varieties of distance functions. The most scopeful dimensions are selected by eliminating irrelevant and the repeated data points. The distance function indeed improves their performance by speeding up the clustering algorithm[2]. Proclus acts well in the case of high number of representative points because Proclus is biased over the cluster which is hyper-spherical in shape. Proclus clustering, used in the research paper is a kind of subspace clustering which uses many phases of data manipulation to get the required cluster. The repeated occurring algorithms pick up the initial k-medoids from the high dimensional mining data to get a reduced dataset. Repeating this step reduces the size of the dimension in a dataset. Each time the potential data are separated from the outliers. When a new data point is found it is clustered with similar data and the noisy ones are rejected.

Proclus is an optimum algorithm for a high dimensional but has its own cons. The Proclus algorithm rejects the cluster with smaller data points completely which may also contain important clusters. So it is really important to overcome such drawbacks which enhances the efficiency of the program. To achieve the enhancement of the algorithm a density based algorithm called the SNN algorithm is implemented. Normally a density base algorithm works on highly dense points. When the clustered data is really dense the cluster of data becomes unclear as all the data seems to be very near and of same distance. So in this algorithm the distance between the two points are calculated. The distance is called the Euclidean distance. All the data points which are within this Euclidean distance are clustered and the rest of the data points that are more than this distance are considered as noisy data or outliers. The algorithm that the research paper uses is a type of density based algorithm called the Shared Nearest Neighbor network algorithm or the SNN algorithm. SNN algorithm works on the principle of similarity measures within the neighbors. If point x is close to y and if they are close together to a set of points z then it can be said that x and y are close with greater confidence since their similarity is confirmed by the points in set z[3]. If there are two clusters of data then the distance between the points that are nearest to each other are calculated. The next step is, it checks the neighbors that share the data points with them. The neighbors which share the common distance are calculated and clustered. No matter the data points are small or big, it does cluster them. This algorithm works on the leftover small cluster of Proclus algorithm. This has increased the efficiency of the clustering for the high dimensional data.

The organization of the paper is as follows. Section 2 contains the materials and methods used for the research work namely an ensemble of Proclus and SNN algorithm. The implementation of the algorithm is discussed in section 3. Section 4 explores about the results and discussion and finally, Section 5 concludes the research work.

2. Materials and Methods

Clustering high dimensional data demands complex process and varied methods. The proposed algorithm involves a method of combination of two important methods of clustering which are projected clustering and density based clustering. To mine a high dimensional data not all projected clustering methods can be used rather it requires an efficient algorithm. Proclus is one such efficient algorithm that works well on high dimensional data. This algorithm has its own pros and cons. The drawback of this algorithm is over come by combining a density based method called the SNN algorithm with the Proclus algorithm. A high dimensional data set, an ensemble of two algorithms and a WEKA tool kit are essential materials for the achievement of this paper. Many research have been made in the field of high dimensional data mining using projected clustering and density based clustering that are explained in Marwan Hassani et al[4], Irene Ntousi et al[5], Rahmat Widia Sembiring et al[6], B.A.Tidke et al[7] and Adrino Moreira et al[8]. The papers involve a combined process of density based clustering and projected clustering in high dimensional data. An incremental mining technique has also been proposed based on the SNN algorithm by Sumeet Singh and Amit Awekar[9]. This algorithm is implemented in collections of documents like nuggets in[10] by Levent Ertöz, Michael Steinbach and Vipin Kumar.

2.1 Proclus

Proclus is considered as the most efficient algorithm when compared to other approaches[11]. There are two kinds clustering algorithms depending on the way they cluster the data points and they are partitioned clustering and hierarchal clustering. Proclus is a partitioned clustering algorithm based on the idea of k-medoids. The main idea behind the method is to iteratively compute a scopeful medoid for each cluster also a high dimensional data is parted down to many subspaces using Proclus which

is also a top-down subspace algorithm. Proclus normally uses a three phase approach which consists of initialization, iteration and refinement of clusters. Selection of a set of medoids forms the initialization process. A medoid is the most centrally located point in a cluster. The process is a partition method that minimizes the sum of dissimilarities between each object and its corresponding points. The basic aim of k-medoids clustering algorithm is to find k-clusters in n object by first arbitrarily selecting a representative object which is the medoid for each and every cluster k forms the input parameter. A greedy algorithm is used to select a set of potential medoids that are very far from each other which begins the initialization process. There should be atleast an instance or medoid representing each cluster. The huge data sets are turned into a reduced dataset by selecting specific potential medoids from the reduced dataset. The iteration phase is processed by repeatedly replacing bad medoids with new randomly close medoids to determine whether the quality of the cluster has improved. A cluster is said to be a quality cluster if the average distance between the cluster and objects are small. For every medoid a dimension is chosen so as to create a subspace and the data points are clustered at the different subspace. Once when a subspace is chosen for every medoid an average Manhattan distance is calculated to assign points for every medoid to create a new cluster. Manhattan Segmental Distance is defined relative to a dimension. The Manhattan Segmental Distance between the point x1 and x2 for the dimension D is defined as:

$$d_D\left(x_1 x_2\right) = \frac{\Sigma_{i \in D} \left| x_{1,i} - x_{2,i} \right|}{\left| D \right|}$$

The next comes the refinement phase that calculates new dimensions for each medoid based on the cluster formed and then assigns new points to medoids removing the outliers. Since the input has to be the average number of dimensions for the cluster it is important that the subspace must be similar in size. Finally the cluster can be defined as a group of instances which is associated with subspaces and medoids. The process also produces outliers and non-overlapping partitions. As every single step has its own way of handling clusters and dimensions the Proclus method is found faster than clique especially in larger data sets.

2.2 SNN Algorithm

Shared Nearest Neighbor is density based algorithm. As known density based algorithms clusters the data depending on the density of points. The region that has the high density of points determines the existence of clusters and the regions with very low density are considered as outliers or noise. As Proclus these density base algorithm is also efficient in mining large data sets with different shapes and sizes.

Normally in density based algorithm the data points are clustered taking a Euclidean distance between the points and the points equal to or greater than the Euclidean distance are clustered. Now the major difference between the SNN and the density based algorithm is it determines the similarity between the points by looking at the number of the nearest neighbors[12] that any two points share. By using these similarity measures the density is defined as the sum of those similarities of the nearest neighbors of a point. In general points that are high in density becomes core-points and low density points become noise points. The three main input parameters of the SNN algorithm is k which denotes the neighbor's list size, Eps which denotes the threshold density and the minpts that define the core-points.has to be connected.

Normally in an SNN algorithm if the two points are connected using a distance measurement it cannot be just used with SNN algorithm the main condition is, it has to contain neighbors on each side. This condition is called k-neighbors sparsification. The SNN algorithm can be handled in two ways:

- The weights of the links or the distance between the two points share can be used to cluster the neighboring data points.
- A graph can be plotted specifying the distance between the two points.

The strength of the link between the two points can be defined as follows. If a and b are the two points between the links then the weights of the link can be measured as

$$\text{Dis}(a, b) = \Sigma(k + 1 - m) * (k + 1 - n), \text{ where } am = bn$$

Here k is the nearest neighbor list size and m and n are the positions of the shared nearest neighbor in the list of a and b.

There are basically two kinds of approach for an SNN algorithm which are core-approach and figure-approach. In the paper the core approach is followed. The process of core approach follows a series of steps. First it identifies the k-nearest neighbor for each points after which it calculates the number of nearest neighbor that the two

points share. Then the SNN density is calculated by the number of neighbor that share Eps distance or more neighbors. The core-points are detected from the data that if the density of the points contain minpts. The other points are considered as noise points.

3. Algorithm: Proclus – SNN Ensemble

Though Proclus is an efficient algorithm it has a drawback of rejecting the cluster with small points completely. This becomes a setback when the rejected cluster contains important data points. So to overcome the issue the research paper proposes an approach over the clustered data. In the iteration phase of the Proclus there occurs a repeated replacement of bad medoids by new medoids from the reduced data set. This is where the clusters with small points are rejected as they consider the clusters with large number of points that are more important and must be taken into account. The cluster that is at a greater distance also contains points that are necessary but since they cover very small number of data points it is neglected.

At this step, the SNN-Similarity measurement is implemented which calculates the SNN-density by applying the Eps distance between the two points that share their neighbor. Even when the number of points is small instead of ignoring it the SNN algorithm, again applies a similarity measure of medoids technique to find the core-points and cluster them[13]. Once they are clustered it gives the value of a proper cluster that cannot be ignored. As you can see in Graph 2 the small data points are clustered using the Eps distance and the SNN density is calculated with the two points that share their neighbors. The resultant cluster is also taken into account to create a clustered database. As long as the iteration phase takes place the SNN algorithm also finds cluster with small points thereby over coming drawback of Proclus.

3.1 Algorithm

Step 1: Initialization, Find a set of potential K-medoids.
Step 2: Ensure each cluster has at least one instance in the selected set.
Step 3: Iteration, select random set of k-medoids to reduce the dataset.
Step 4: Check the average distance between the instance and the medoids using the distance method

$$d_D(x_1, x_2) = \frac{\Sigma_{i \in D} |x_{1,i} - x_{2,i}|}{|D|}$$

Step 5: Clustered data which has more data points leaving the small data points that are necessary.
Step 6: Implementing SNN density based algorithm find the Eps distance between the two points.
Step 7: Find the nearest neighbor the two points share.
Step 8: Cluster the points that are equal or less than Eps distance.
Step 9: The Refinement phase - remove the outliers.

4. Results and Discussions

Observations from the resulting graph, reveals the number of clusters are more in the resulting graph than the previous Proclus graph output. This is because there are many small data points that are clustered into different groups in the proposed algorithm. There is an algorithm called the ORCLUS which is an extended version of the Proclus algorithm. The main reason for this extension is to cluster and count even the small number of representative data points. So it is the same reason the SNN algorithm is used in the research paper. Therefore it is reasonable to justify the consideration of the density based SNN algorithm over the ORCLUS algorithm. The ORCLUS algorithm normally consists of three phases as assign clusters, subspace determination and merge[14]. In the assign phase the database is well partitioned into k cluster by assigning every single point to its closest seed. The distance between the data point and its seed which is the initial point from the database is measured on the subspace.

The point with the minimum distance is assigned to the cluster. The second comes the determination of subspace in the database. As there are number of dimensions to find the dimensionality of any cluster is essential. The dimensions of the subspace are done by calculating the covariance matrix of the cluster and selecting the orthogonal Eigen values. As the iteration proceeds on the value of the dimensionality reduces more for every iteration to iteration. Once the subspace is determined the merge phase takes place. The merge phase reduces the number of cluster during the iteration. To achieve this, the closest clusters are needed to be merged. When the merging is done successfully the cluster with small data points are clustered together so that it is counted and not rejected. The algorithm comes to an end when the emerging process over the iteration has reduced the

number of clusters. Though Orclus is considered as an efficient algorithm it has complexity issues. The subspace dimensionality reduces the dimensionality partitioning the cluster again merging it costs more complexity which impacts on the performance. So an efficient density base clustering method called the Shared nearest neighbor is used. There are several techniques like Chameleon, CURE and DBSCAN which are better distance measurement associated density based techniques. They do not go quite well with the high dimensional data[13].

The SNN density based algorithm on the other is well organized to run on high dimensional data. Normally the direct similarity measurement cannot be trusted in high dimensional data because in high dimensional dataset the data are arranged sparse and therefore the average of similarity between the points is very low. As in Proclus the high dimensional data are already reduced to a low dimensional data due to the iteration phase the sparsity of the data is reduced and therefore the SNN algorithm can be implemented with confidence. It gradually gets along with the Proclus algorithm. It finds the distance between two points they share among them. This algorithm does not need a complex step to find the cluster and merge them instead it just finds the neighbor points and clusters them. It is also used to find the relative density[15]. So it is verified that Proclus with SNN algorithm is efficient on a non-complex approach of clustering high dimensional data.

The Figure 1 represents the high dimensional data before clustering. Figure 2 represents the same high dimensional data after Proclus clustering and Figure 3 represents the implementation of the SNN density based clustering algorithm.

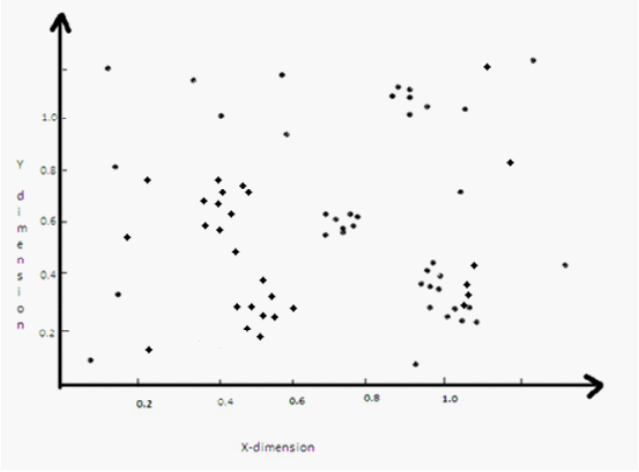

Figure 1. A High dimensional dataset before clustering.

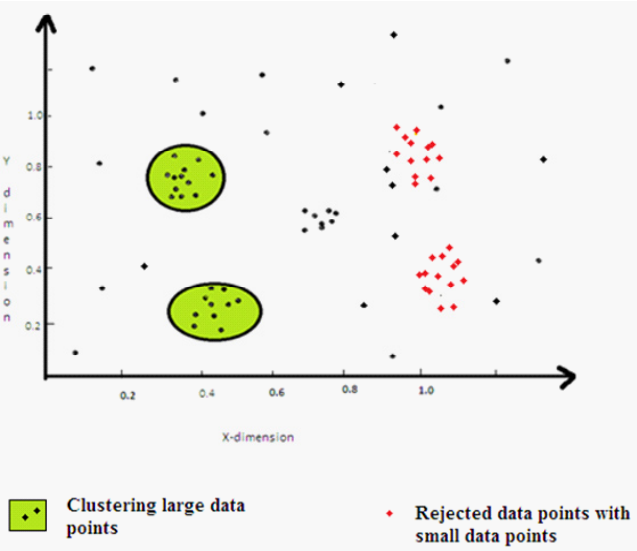

Figure 2. Graph denoting clustering large number of data points and Rejecting small data points.

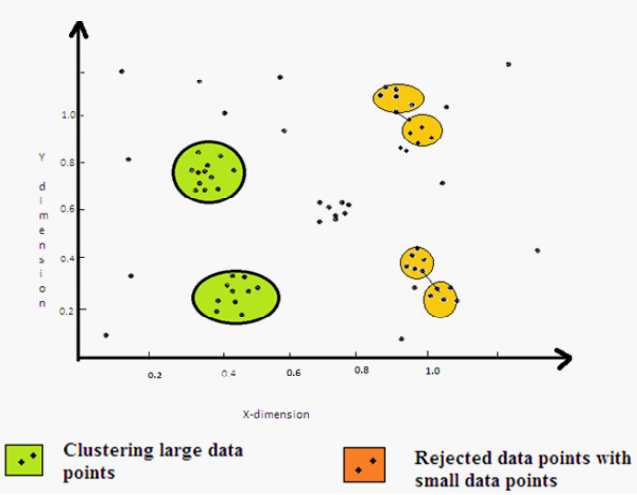

Figure 3. Clustering of small data after SNN algorithm.

The Proclus is tested using Elki cluster tool kit that allows an independent execution and evaluation of each data mining algorithm and data mining tasks. The density based algorithm is tested in WEKA tool. The experiment was tested on synthetic data sets. The data can be obtained from UCI Machine learning repository. The results contained more clusters than the usual Proclus output.

5. Conclusion

An ensemble of clustering has emerged as a prominent way to obtain efficient solutions for high dimensional

clustering. Among various techniques for handling high dimensional data projected clustering and density based clustering are used. Out of these two clustering techniques, two sub types called Proclus and SNN algorithm are chosen for our research. The drawbacks of the Proclus algorithm are overcome by the density based algorithm. The SNN algorithm implements its algorithm right at the time the small data points are rejected and left uncared. The research paper gives a combined method of two clustering algorithm and also a literature survey of many clustering technique. Our proposed approach can be highly efficient to automatically detect number of clusters with even very small data points.

6. References

1. Tidke BA, Mehta RG, Rana DP. A novel approach for high dimensional data clustering. Int J Eng Sci Adv Technology. 2012 May-Jun; 2(3):645–51.
2. Kumar KN, Kumar GN, Veera Reddy Ch. Partition algorithms – A study and emergence of mining projected clusters in high-dimensional dataset. International Journal of Computer Science and Telecommunications. 2011 Jul; 2(4):34–7.
3. Ertoz L, Steinbach M, Kumar V. A new shared nearest neighbor clustering algorithm and its applications. In Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining; 2002. p. 105–15.
4. Hassani M, Spaus P, Gaber MM, Seidl T. Density-based projected clustering of data streams. In Scalable Uncertainty Management. Springer Berlin Heidelberg. 2012; 7520:311–24.
5. Ntoutsi I, Zimek A, Palpanas T, Kröger P, Kriegel HP. Density-based projected clustering over high dimensional data streams. SDM; 2012. p. 987–98.
6. Sembiring RW, Zain JM, Embong A. Clustering high dimensional data using subspace and projected clustering algorithms. IJCSIT. 2010 Aug; 2(4):162–70.
7. Tidke BA, Mehta RG, Rana DP. A novel approach for high dimensional data clustering. Int J Eng Sci Adv Technol. 2012 May-Jun; 2(3): 645–51.
8. Moreira A, Santos MY, Carneiro S. Density-based clustering algorithms – DBSCAN and SNN. University of Minho-Portugal; 2005.
9. Singh S, Awekar A. Incremental shared nearest neighborhood or density based clustering. Indian Institute of Technology, Guwahati, Assam.
10. Ertöz L, Steinbach M, Kumar V. Finding topics in collections of documents: a shared nearest neighbor approach. Kluwer Academic Publishers; 2002.
11. Yip, Kevin Y, Ng MK, Cheung DW. A review on projected clustering algorithms. International Journal of Applied Mathematics. 2003; 13(1):35–48.
12. Yin J, Fan X, Chen Y, Ren J. High-dimensional shared nearest neighbor clustering algorithm, Fuzzy systems and knowledge discovery. Berlin, Heidelberg: Springer; 2005. p. 494–502.
13. Aguilar JS, Ruiz R, Riquelme JC, Giráldez R. SNN: A supervised clustering algorithm. Engineering of Intelligent Systems. Berlin, Heidelberg: Springer; 2001. p. 207–16.
14. Aggarwal CC, Yu PS. Finding generalized projected clusters in high dimensional spaces. ACM. 2000 Jun; 29(2):70–81.
15. Kumar KN, Naveen GK, Reddy ChV. Partition algorithms – A study and emergence of mining projected clusters in high-dimensional dataset. 2011 Jul; 2(4):34–7.