JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Quality Cluster Generation Using Random Projections

P.A. Gat[1*], K.S. Kadam[2]

[1] Department of Computer Science, DKTE's College, Ichalkaranji, India
[2] Department of Computer Science, DKTE's College, Ichalkaranji, India

DOI: https://doi.org/10.26438/ijcse/v7i6.933936 | Available online at: www.ijcseonline.org

Accepted: 16/Jun/2019, Published: 30/Jun/2019

***Abstract*—** Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. Clusters are obtained by using density based clustering and DBSCAN clustering. DBSCAN cluster is a fast clustering technique, large complexity and requires large parameters. To overcome of these problems uses the OPTICS density based algorithm. The algorithm requires the simply a single parameter, namely the least amount of points in a cluster which is required as input in density based technique. Using random projection improving the cluster quality and runtime.

***Keywords*—** Cluster Analysis, Random Projections, Neighbouring

I. Introduction

Cluster is group of objects that belongs to the same class. In the other words, similar objects are grouped in one cluster and dissimilar objects are grouped into the other cluster. Cluster analysis is the most familiar and powerful unsupervised techniques. These techniques are used in data processing. This is a useful approach to arranging input data sets into a set of semantically consistent sets of a limited range of similarities. Clustering involves looking up structures in unlabeled data sets. A cluster is a set of items that are alike between them. The idea of clustering is based on a group of objects of information found in data related to an object. It means that objects are similar to each other and similar to other groups. In data mining cluster analysis is a vital study area. It has its own unique positioning and does not require data analysis and processing. It can be shown that no complete optimal criteria can be independent of the ultimate goal of clustering.

The clustering algorithms will be divided into several types, namely hierarchical algorithm, graph-based algorithm, density-based algorithm, partition algorithm, grid based algorithm. Along with these algorithms types, Density-based algorithms are well-known and simple implementation algorithm. The other two basic benefits of these algorithms are that they can identify clusters of different shapes and size. Density-based algorithms for distinguishing dense regions that are measured separately from low-density regions.

Density-based algorithms have become a flexible and well-organized technique for discovering clusters of high quality and possibly irregular shapes. Clustering is a significant operation for knowledge extraction. Clustering is a significant operation for knowledge extraction. Its purpose is to assign objects to groups such that objects within a group are more alike than objects across different groups.

The main requirements of clustering algorithm are scalability, dealing with different types of attributes, discovering clusters with arbitrary shapes, minimum requirements for domain information to determine input parameter, ability to handle noise and outlier, high dimensionality, interpretability and availability. Compare to same density-based technology, the new clustering method achieves logical acceleration while providing a logical guarantee for cluster quality Moreover, to set parameters is not difficult. The new method provides a complete analysis of algorithms and comparison with existing density-based algorithms.

II. Related Work

Ester M, Kriegel H-P, Sander J, Xu X. [1] Proposed DBSCAN algorithm is capable to handle local density variation within the cluster. It detects the cluster of different shapes and size from large amount of data which contains noise and outliers. DBSCAN discovers clusters with random shapes and sizes, which detect the occurrence of outliers in data. The drawbacks of these systems are its execution time (O (nlogn)) and its awareness to the user's permanent density parameters.

Author Ankerst M, Breunig MM, Kriegel H-P, Sander J. [2] Proposed OPTICS overcomes several of these limitations of introducing an inconsistent density and requiring the setting of only a single parameter. If data has changeable density then OPTICS algorithm used to find good clusters. It outcomes the objects in an exacting order. The drawback of system is optics algorithm which expects some kind of density decline to find cluster borders.

Alexander Hinneburg, Daniel A. Keim [3] proposed a new algorithm for clustering in large multimedia databases i.e. called DENCLUE that can handle noise. In this approach, they are able to find non-spherical shaped clusters using local density function. They evaluated performance of DBSCAN with DENCLUE which shows that DENCLUE is more superior to DBSCAN.

Hinneburg A, Gabriel H-H. [4] Proposed Denclue algorithm employs a cluster model based on kernel density estimation. A cluster is defined by a local maximum of the estimated density function. Data points are assigned to clusters by hill climbing, i.e. points going to the same local maximum are put into the same cluster. A disadvantage of Denclue 1.0 is that the used hill climbing may make unnecessary small steps in the beginning and never converges exactly to the maximum, it just comes close.

Imran Khan, Joshua Zhexue Huang[5] proposed an ensemble clustering method for high dimensional data which uses random projection to generate subspace component data sets. In comparison with popular fast-map sampling and fast-map projection, random projection preserves the clustering structure of the original data in the component data sets so that the performance of ensemble clustering is improved significantly. This paper represent two methods to measure preservation of clustering structure of generated component data sets. The comparison results have shown that Random projection preserved the clustering structure better than fast-map sampling and fast-map projection.

Qi Xianting et al. [6] proposed a paper in which a sub-part of the density-based representative algorithms is the density-based spatial clustering of applications with noise (DBSCAN) which has been utilized in certain fields because of its property of detecting the clusters which are of various shapes as well as sizes. When high dimensional data is present in certain applications that is when the algorithm stays no more stable. For the purpose of resolving this issue, an enhanced DBSCAN algorithm which is based on feature selection (FS-DBSCAN) is put forth. This algorithm is provided on various real world datasets and the various series of simulations are achieved.

Schneider J, Vlachos M. [7] proposed two fast density-based clustering algorithms based on random projections. Both algorithms demonstrate one to two orders of magnitude speedup compared to equivalent state-of-art density based techniques, even for modest-size datasets. We give a comprehensive analysis of both our algorithms and show runtime of O(dN log2 N), for a d-dimensional dataset. The algorithm can be viewed as a fast modification of the OPTICS density-based algorithm using FOPTICS and parameterless algorithm. The FOPTICS algorithm uses a simple definition of density combined with sampling, and the parameterless algorithm identifies areas separating cluster. These algorithms take more time complexity for execution. To overcome these problems need to use SOPTICS algorithm which gives the quality cluster using random projections.

III. PROPOSED WORK

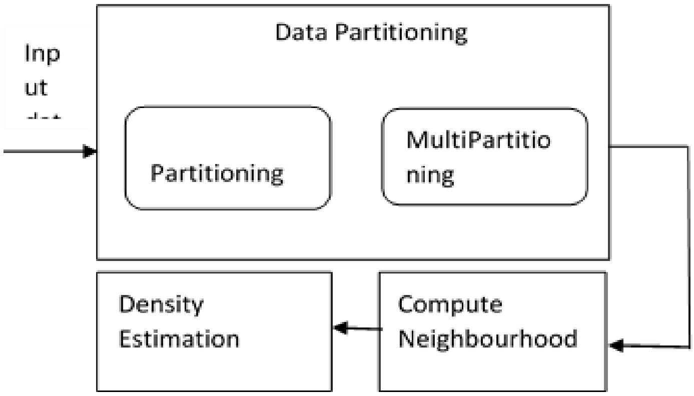

Fig 1- System Architecture

The fig 1 shows the system architecture of Quality cluster generation using random projection. We present scalable density based clustering algorithm using random projections. These methodology achieves a speed up as compared to other density based algorithm. Our algorithm require the single parameter, i.e. the minimum number of points in a cluster which as input in density based technique. The system consists of three main module first is pre-processing-partitioning and second is neighboring, third is density estimation. The first modules contain two main algorithms one is partitioning and another is multi-partitioning using these algorithms to enhance the quality clustering. The neighboring module computes the neighbor using random projection technique and improves the runtime preserving the cluster quality. The density estimation module discovers the neighborhood of each object to estimate local densities.

IV. METHODOLOGY

1. Pre-process-data partitioning:

The density-based clustering algorithm consists of two phases. The first partition information in that the close point is placed in the equivalent partition. The second phase uses these partitions to calculate only the distance or density inside the equivalent partition pair. For a partitioning start with the entire point set. This point set split it into two parts until the size of point set is at most $minSize$+1, where $minSize$ is a parameter of the algorithm. To splits the points, the predictable values of points are selected consistently at arbitrary. All points with a predictable value lesser than that of the point chosen represent one part and remainder the other part. In principle, one could also split based on distance i.e. pick a point randomly on the projection line that lies between the projected point of minimum and maximum value.

In multi-partitioning, perform the different random projections on a density based algorithm. Formally the multi-partitioning algorithm chooses a sequence of line τ:=($L0$,$L1$,…….) of CL log$N$ random lines. It projects the points on each random line Li in the sequence τ giving set of projected values for each line Li. The sequence of all these sets of projected values β:=($L0$,ρ,$L1$,ρ…..) .The points S are split into two disjoint sets $S0$ 0 and $S1$ 0 using the value rs:=$L0$.The set $S0$ 0 contains all points $P \in S$ with smaller projected value than the number rs. For line $L1$ consider set $S0$ 0 and split it into sets $S0$ 1 and $S1$ 1.Then similar process is used on $S1$ 0 to obtain sets $S21$ and $S3$ 1.The recursion ends set S contains fewer than $minSize$+1 points. The union of all sets of points resulting from any partitioning for any of the projection sets $\tau \in \beta$.

2. Neighborhood:

A neighborhood consisting of nearby points and estimate of density using preprocessing modules. Each set of data partitioning consisting of nearby points; all points in a set are neighbors of each other. Using nearby points to reduce the runtime considering evaluating all pairwise distances only for a single random projection and perform fewer random projections. In neighborhood, creation process contains sequences of nearby points $S \in \delta$ is an ordering of points projected onto a random line. For each sequence $S \in \delta$ we pick a point, i.e. a center point $Pcente$. For this center point, we add all other points' $S \backslash Pcenter$ to its neighborhood ($Pcenter$). The center $Pcenter$ is added to the neighborhood (P) of all points $P \in S \backslash Pcenter$.

3. Density Estimate:

Density estimation needs to measure the volume containing a fixed amount of points. Density estimation is a non-parametric way to estimate the probability density function of a random variable. To compute a density estimate for a point depends on its $minPts$-nearest neighbor. Therefore for a single partitioning splitting a set if it is at least of size $minSize$>= $minPts$+2 a natural lower bound. For set of size $minPts$+2 at least one point is removed by a split leaving $minPts$+1 points in a final set. For such a set there could be one or more points such that $minPts$ closest to the neighbor. In multi-partitioning fewer points are used. Each final set for partitioning is essentially a random set of generally nearby points.

V. RESULT AND DISCUSSIOIN

Firstly, the algorithms have been implemented and tested using the well-known benchmark data set. Perform testing of the application by the use of algorithms OPTICS, and SOPTICS, and finally made comparison between both of these algorithms on the same dataset has also been done.

The following results have been obtained by applying OPTICS and SOPTICS on the different data set.

Dataset	MUSK	SVM	AGGRIGATION	JAIN
OPTICS Runtime	5766ms	4560ms	5590ms	3279ms
SOPTICS Runtime	3363ms	3100ms	3201ms	2289ms
OPTICS FSCORE	0.65787	0.76348	0.83456	0.65897
SOPTICS FSCORE	0.68321	0.78903	0.86234	0.69456

VI. CONCLUSION

Density based techniques can provide the building the group for clustering algorithms. Our work contributes to density based clustering by new algorithm SOPTICS, which is a random projection based version of OPTICS algorithm.

REFERENCES

[1] Ester M, Krigel H-P, Sander J, Xu X(1996)"A density based algorithm for discovering clusters in large spatial databases either noise." In proceeding of the ACM conference knowledge discovery and data mining (KDD),pp226-231.

[2] Ankerst M, Breunig MM, Kriegel H-P, Sander J (1999) "Optics: ordering points to identify the clustering structure." In: Proceedings of the ACM international conference on management of data (SIGMOD),pp. 49–60.

[3] Alexander Hinneburg, Daniel A. Keim (1998),"An Efficient Approach to Clustering in Large Multimedia Databases with Noise [Online] Available: http://www.aaai.org.

[4] Hinneburg A, Gabriel H-H (2007) Denclue 2.0: fast clustering based on kernel density estimation. In Advances in intelligent data analysis (IDA), pp 70–80.

[5] Imran Khan, Joshua Zhexue Huang (2012)," Ensemble Clustering of High Dimensional Data With random Projection." In: Proceeding of the international conference on information and knowledge management.

[6] Qi Xianting, Wang Pan,"A density-based clustering algorithm for high-dimensional data with feature selection", 2016,IEEE.

[7] Schneider J, Vlachos M (2013) "Fast parameter less density-based clustering via random projections." In: Proceedings of the international conference on information and knowledge management (CIKM), pp 861–866.

[8] Johannes Schneider, Michail Valchos(2017) "Scalable density based clustering with quality guarantees using random projections." Published in Journal: Data Mining and Knowledge Discovery Volume 31 Issue 4, July 2017 pages 972-1005.

Authors Profile

Prachi A. Gat pursed Bachelor of Engineering from SITCOE,Yadrav, India University, in year 2017, She is currently pursuing Master of Technology from DKTE‘s TEI, (An Autonomous Institute), Ichalkaranji, India. Her research work focuses on Data Mining and Machine Learning.

Prof .K. S. Kadam, Assistant Professor of Computer Science & Engineering, at DKTE Society's Textile & Engineering Institute, Ichalkaranji ,India. He is a member of the ISTE, CSI. His current research interests include Grid and Cloud Computing, Database Engineering, System Programming, Data Mining and Warehouse, Advanced Database and Compiler Construction, Big Data Analytics.