www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 19941-19944
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.15

A Review on Dynamic Clustering: Using Density Metrics

Megha S.Mane[1], *Prof.N.R.Wankhade*[2]
[1]PG student, Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Nashik, Maharashtra, India
megha.mane175@gmail.com
[2]Head and Associate Professor, Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Nashik, Maharashtra, India
nileshrw_2000@yahoo.com

Abstract: Clustering of high dimensional dynamic data is challenging problem. Within the frame of big data analysis, the computational effort needed to perform the clustering task may become prohibitive and motivated the construction of several algorithms or the adaptation of existing ones, as the well known K-means algorithm. One of the critical problem in k-means, k-menoid, k-means or other clustering algorithms required to pre-assigned no. of k which cannot detect non-spherical clusters. With the existing RLClu algorithm needs users to pre-assign two minimum thresholds of the local density and the minimum density-based distance. Clustering is the process of data classification when none prior knowledge required for classification. To overcome these problems STClu clustering algorithm is proposed. In this algorithm a new metric is defined to evaluate the local density of each object, which shows better performance in distinguishing different objects. Furthermore, an outward statistical test method is used to identify the clustering centers automatically on a centrality metric constructed based on the new local density and new minimum density-based distance. Dynamic clustering is an approach to get and extract clusters in real time environments. It has much application such as, data warehousing, sensor network etc. Therefore there is need of such technique in which the data set is increasing in size over time by adding more and more data.

Keywords: Clustering, clustering center identification, long-tailed distribution, outward statistical testing

1. Introduction

Clustering is a commonly used tool that aims to identify similar samples in a dataset. It can be viewed both from the machine learning and the statistics point of view, with classical algorithms pertaining to each domain. When considered from the algorithmic standpoint, the complexity of the clustering problem is known to be NP hard, even for the usual K-means, when the number of clusters is not fixed. Clustering is widely used in various applications such as, image segmentation, time series analysis, information retrieval, spatial data analysis and biomedical research. There have challenging tasks with this clustering approach such as, variety and scale of data increases rapidly and user have only prior knowledge about the data. Generally, there are number of clustering algorithms are available such as, k-means, k-menoids, k-means++ etc are the centroid based algorithms in which user required to specify ‘k’ in advance where ‘k’ is nothing but the number of clusters. The requirement of the parameter k specified in advance is considered as one of the critical drawbacks of this kind of algorithms. K-means algorithm is useful in non-convex boundaries and performs empirically very well. Hierarchical algorithms are another type of algorithm which also known as similarity based clustering methods. It recursively finds the nested cluster either in agglomerative mode or divisive mode. Hierarchical algorithms are useful in variety of searching methods because they naturally create a tree-like hierarchy which can be leveraged for the search process. Another is distribution based clustering algorithm such as, EM algorithm to estimate the parameters of the mixture model, so as to obtain the clustering results. It appoints the fixed number of Gaussian distributions to approach the distribution of the objects. However, this algorithm also required to preassign number of k cluster which usually difficult for large real world datasets. Density-based algorithms define the clusters as areas of higher density than the remainder of the dataset. The advantage of density-based algorithms is that they do not need to specify the number of clusters in advance, and can detect the outliers of the dataset. However, they have limitations in handling high-dimensional data like text. Because the feature space of high-dimensional data is usually sparse, density-based algorithms such as DBSCAN have difficulty to distinguish high-density regions from low-density regions.

Rodriguez and Laio proposed RLCLu Clustering algorithm that can efficiently combines all beneficial features of above algorithms. Specifically, RLClu is only based on the distance or similarity between objects. Second, as the density based clustering, it defines the clustering centers as the objects with maximum local density, and can detect the non-spherical clusters. There are two metrics defined in RLCLu algorithm namely, local density and minimum density based distance. The local density metric plays a critical role in RLClu but is sensitive to a preassigned parameter, cutoff distance, when the data set is small. Whereas, in identification of clustering center required users to preassign two minimum thresholds of the local density and the minimum density-based distance. Also RLCLu is sensitive to some ‘k’ parameters and it also suffers from the parameter setting problem. Therefore, STCLu algorithm exhibits the better performance than the RLCLu algorithm. The proposed algorithm defines a metric to evaluate local density of each object. Furthermore, a method known as, outward statistical test is appointed to automatically identify the center of clustering. More specific, STCLu algorithm obtains the representation of objects in low dimensional region.

2. Related Work

T. Kanungo, D.M. Mount, et al. [1], proposed k-means clustering is Lloyd's algorithm. It is also known as filtering algorithm. For implementation of this algorithm it required kd-tree data structure. Kd-tree data structure stores the multidimensional data points in it. A kd-tree is a binary tree, represents a hierarchical subdivision of the point set's bounding box with the help of axis aligned splitting hyper planes. Each individual node of the kd-tree is linked with a closed box, known as a cell. The proposed algorithm is simple and easy to understand and for implementation, it only required a kd-tree be built once for the given data points. With proposed algorithm better efficiency is achieved due to the data points do not vary throughout the computation. Hence, this data structure does not need to be recomputed at each stage. Since there are typically many more data points than “query” points (i.e., centers), the relative advantage provided by preprocessing in the above manner is greater.

G. Dong, M. Xie[2], proposed image segmentation for segmentation of color image based on neural networks. The proposed solution provides complete solution for both the supervised and unsupervised segmentation of color images. It can systematically address the problem of color image segmentation. It includes uniformity in color representation, color reduction and clustering in unsupervised segmentation and color learning in supervised segmentation. They have implemented an unsupervised segmentation using SOM-based color reduction, and SA-based color clustering. The supervised segmentation is achieved by HPL learning and pixel classification. It has ability to produce optimal result segmentation with fewer computational costs

T. Warren Liao [3], proposed time series clustering has been shown effective in providing useful information in various domains. They have organized the domain into three categories. These categories are depending upon either time or frequency domain. In this features are directly extracted from raw data/indirectly with models constructed from the raw data. Mainly, the fundamentals of time series clustering studies are highlighted. In paper [4], N. Jardine and C. J. V. Rijsbergen introduced HAC clustering algorithm is introduced. It more aggressive optimization for low level computations and it benefits from the most pair wise similarity. In this higher dimensions problem is encounter in KNN classification. In this they avoid dense centroids. The clustering approach for large vocabularies complete-link clustering can be more efficient than an unoptimized implementation of GAAC.

A.K. Jain, M.N. Murty et al.[5], P. Berkhin [6], Anil K. Jain [7], addressed the problem of many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness in exploratory data analysis. As everyone knows that clustering is the combinatorial problem having and differences in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur. They were presents an overview of pattern clustering methods from a statistical pattern recognition perspective. The goal is in terms of providing advice and references to fundamental concepts attainable to the broad community of clustering practitioners [5].

P. Berkhin [6] introduced clustering is a division of data into groups of similar objects. Data modeling keeps clustering in historical perspective rooted in mathematics, statistics, and numerical analysis. Generally, clustering techniques do not distinguish between the two: neither noise nor abnormalities fit into clusters. Multiple ways are available to learn descriptive learning of managing outliers. A.K. Jain [7], introduced, the connectivity based clustering assumes that the objects close to each other are more possible to be in the same cluster than the objects far away from each other; this kind of clustering algorithms usually organizes the objects as a hierarchical structure but does not produce a unique partition, and still needs users to preassign a distance threshold to generate appropriate clusters.

P. S. Bradley, O. L. Mangasarian et al. [8] and D. Arthur and S. Vassilvitskii[9], proposed approach for assigning points to clusters. It is based on simple concave minimization model. K-Median Algorithm is very simple and efficient which quickly locates the useful stationary point. Utility of the proposed algorithm lies in its ability to handle large databases and hence would be a useful tool for data mining. Comparing it with the k-Mean Algorithm, we have exhibited instances where the k-Median Algorithm is superior, and hence preferable. D. Arthur and S. Vassilvitskii[9], referred to simply as “k-means,” Lloyd’s algorithm begins with k arbitrary “centers,” typically chosen uniformly at random from the data points. Each point is then assigned to the nearest center, and each center is recomputed as the center of mass of all points assigned to it.

L. Hagen, A.B. Kahng [10] and w. E. Donath [11], introduced methods to speedups the computationally procedure. The proposed methods are namely, EIGl and EIG1-IG. They also proposed post-processing for improvement in current results. EIGl and EIG1-IG should be applied to ratio cut partitioning for other CAD applications, especially test and the mapping of logic for hardware simulation. In final steps they represented the theoretical analysis for proposed algorithms or methods. The spectral clustering based algorithm does not make assumptions on the forms of the clusters; it utilizes the spectrum (i.e., eigenvalues) of the similarity matrix of the data to map the data into a lower dimensional space in which the objects can be easily clustered by traditional clustering techniques [11].

Rodriguez and A. Laio [12], introduced Clustering by fast search and find of density peaks. The proposed approach aimed to classify category of elements on the basic of their similarity. They have proposed an alternative approach that is similar to the k-medoids. It has its basis only in the distance between data points. Like DBSCAN and mean-shift technique, it is able to detect non-spherical clusters and to automatically identify the absolute number of clusters. The proposed algorithm has its basic assumptions that cluster centers are encompassed by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density. A decision graph allows distinguishing genuine clusters from the density ripples generated by noise. Qualitatively, only in the former case are the points corresponding to cluster centers separated by a sizeable gap in other points.

U.V. Luxburg [13], represented spectral clustering approach, It is one of the most popular modern clustering algorithms because it is simple to implement, efficient, and can be solved efficiently using standard linear algebra software, and very often outperforms traditional clustering algorithms like k-means algorithm. They derived spectral clustering from scratch and present different points of view to why spectral clustering works. To present the most common spectral clustering algorithms, and derive those algorithms from scratch by several different approaches they described different graph Laplacians and their basic properties.

C.P. Lai and P.C. Chung et al [14], suggested piecewise aggregate approximation (PAA) algorithm. Their objective is to provide judgment aspect on currently available clustering methods. It is basically used to reduce dimensions before clustering. A method introduced for two level clustering named as, 2LTSC i.e. two level time series clustering. The whole time series is represented as level-1 whereas, in level-2 subsequences of time series is represented. A time series is broadly classified in two main categories such as, 1. Complete clustering, 2. Subsequence clustering, complete clustering is performed on separate set of time series data to group similar time series within same cluster. A subsequent clustering is based on extraction of sliding window of onetime series. They have discussed about the blood pressure analysis by which they mapped their specified problem. The proposed method provides different and deeper point of views to the user for making any important decision.

G. Wang and Q. Song [15], Proposed Statistical Test based clustering i.e. STClu algorithm. The proposed algorithm addressed the issues in RLClu algorithm. For local density of each object they have evaluated a new metric which gives the better performance in distinguishing different objects. To identify the clustering centers automatically on a centrality metric an outward statistical test method is implemented by them. Specially, STClu obtains the object representation in two dimensional. It is similar to the spectral clustering approach in which for dimension reduction similarity matrix has been used.

3. Problem Formulation

From literature survey analysis, we defined the problem statement as,
"To overcome the drawback of existing clustering algorithm such as, centroid clustering algorithms such as, k-means, k-menoids, k-means++ , hierarchical algorithms, model based clustering algorithms which required to defined value of 'k' (cluster)in advance. Also to provide better solution by proposing X-means algorithm rather than using RLCLu clustering algorithm that suffers from the problem of parameter settings".

4. System Architecture

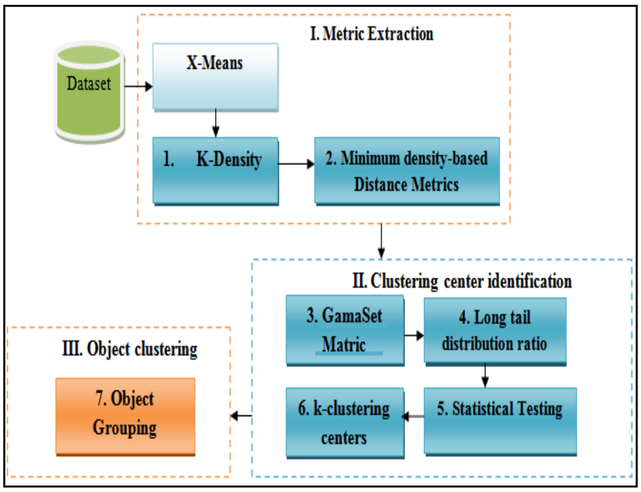

Figure 1: System Architecture

For initial cluster centroid identification we have used X-means algorithm. After X-means centroid identification we have identified K-density of every object in the given cluster. The proposed work compares the cluster creation between X-means and K-density evaluation.

It consists of three stages given as below:

1. Metric extraction: For clustering of 'n'- number of objects we evaluate local density and 'k'-density distance between the given object and the other objects.

2. Clustering center identification: With the help of X-means clustering we have identified initial 'k'-clusters. In this density metrics cluster creation system we used long tail distribution ratio for centroid identification. K-density of every element in cluster is calculated w.r.t. X-means centroid, gammaSets are created using K-density metrics. Object having highest value considered as clustering center. To refine clustering centers statistical testing is used and then we get the number of 'k'-cluster centers.

3. Object clustering: For clustering centers identification objects are assigned to a cluster which contains its nearest neighbor with higher local density.

5. Conclusion

There are several clustering techniques are available. Lots of computational efforts were required for accomplishment of clustering process. Basically, clustering is the process of grouping of similar objects. A well known algorithm has discussed in above section (2) such as, K-means algorithm, k-menoid etc but these having some critical issue and they required to pre-assigned no. of k which cannot detect non-spherical clusters. Existing method RLClu is similar to the connectivity and centroid based clustering and it is based on based on the distance between objects. According to our analysis RLClu is very sensitive method to some preassigned parameters and suffered from the parameter setting problem.

According our analysis from clustering algorithms, X-means can be better solution for cluster center identification. In which comparison between the cluster creation between X-means and K-density evaluation can also be shown.

Acknowledgment

I am deeply expressing my sincere gratitude to my project guide Prof. N.R.Wankhade who always supported and guided me with valuable comments. I express my immense pleasure and thankfulness to all faculty members of the Department of Computer Engineering of Late G.N.Sapkal College of Engineering, Nashik.

References

[1] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881–892, Jul. 2002.

[2] G. Dong and M. Xie, "Color clustering and learning for image segmentation based on neural networks," IEEE Trans. Neural Netw., vol. 16, no. 4, pp. 925–936, Jul. 2005.

[3] T. W. Liao, “Clustering of time series data-a survey,” Pattern Recog., vol. 38, no. 11, pp. 1857–1874, Nov. 2005

[4] N. Jardine and C. J. V. Rijsbergen, “The use of hierarchic clustering in information retrieval,” Inf. Storage Retrieval, vol. 7, pp. 217– 240, 1971.

[5] A.K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” ACM Comput. Surveys, vol. 31, pp. 264–323, 1999.

[6] P. Berkhin, “A survey of clustering data mining techniques,” in Grouping Multidimensional Data. New York, NY, USA: Springer, 2006, pp. 25–71.

[7] A.K. Jain, “Data clustering: 50 years beyond k-means,” Pattern Recog. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[8] P. S. Bradley, O. L. Mangasarian, and W. N. Street, “Clustering via concave minimization,” in Proc. Adv. Neural Inf. Process. Syst., 1997, pp. 368–374.

[9] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms, 2007, pp. 1027–1035.

[10] L. Hagen and A. B. Kahng, “New spectral methods for ratio cut partitioning and clustering,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 11, no. 9, pp. 1074–1085, Sep. 1992.

[11] W. E. Donath and A. J. Hoffman, “Lower bounds for the partitioning of graphs,” IBM J. Res. Develop., vol. 17, no. 5, pp. 420–425, 1973.

[12] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[13] U. Von Luxburg, “A tutorial on spectral clustering,” Statist. Comput., vol. 17, no. 4, pp. 395–416, 2007.

[14] C.-P. Lai, P.-C. Chung, and V. S. Tseng, “A novel two-level clustering method for time series data analysis,” Expert Syst. Appl., vol. 37, no. 9, pp. 6319–6326, 2010.

[15] G. Wang and Q. Song, “Automatic Clustering via Outward Statistical Testing on Density Metrics”, IEEE transactions on knowl. And data engineering, vol. 28, no. 8, pp. 1074–1085, Sep. 2016.

Author Profile

Megha S. Mane[1] received B.E.degree in computer engineering from Pune University, Maharashtra, India. Pursuing M.E.degree in computer engineering from Late G.N.Sapkal College of Engineering, Nashik, Maharashtra, India.

Prof.N.R.Wankhade[2] working as Head and Associate Professor, Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Nashik, Maharashtra, India.