IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-1 E-ISSN: 2347-2693

Automatic Clustering Based on Outward Statistical Testing Using Advanced Density Metrics

Ashvita A. Jadhav[1*], V.S.Gaikwad[2]

[1*]Department of Computer Engineering, Rajashree Shahu School of Engineering and Research, JSPM NTC, Pune, INDIA
[2]Department of Computer Engineering, Rajashree Shahu School of Engineering and Research, JSPM NTC, Pune, INDIA

Available online at: www.ijcseonline.org

Received:25/12/2016 Revised: 08/01/2017 Accepted: 22/01/2017 Published: 31/01/2017

Abstract— Clustering is the process of organizing objects into several groups whose members are similar in some way and is very important technique in data mining as it has applications spread extensively example marketing, biology, pattern recognition etc. Various algorithms have been proposed, published, implemented for clustering like the one published by Rodriguez and Liao but this algorithm is dependent and sensitive to specified parameters and also faces difficulties in identification of ideal problems. Another one published by G. Wang and Q. Song this algorithm do not list all possible numbers of the nearest neighbors and the accuracy is not better in terms of Olivetti face data set then impact of this on performance. This paper overcome the problem faces by above algorithm so, here proposes a new clustering method that will identify cluster centers automatically via statistical testing. Here first define a new metric to evaluate the local density of an object which is named K-density and second metric is define to evaluate the distance of an object to its neighbors with higher density. Then, product of these two metrics is used to evaluate the centrality of each object. After analyzing the distribution of these metrics further transformed the clustering center identification into a problem of extreme-value detection from a long-tailed distribution Finally, apply outward statistical testing method to detect the clustering centers automatically and then completed the clustering process by assigning each of the rest objects to the cluster that contains its nearest neighbor with higher K-density.

Keywords- Clustering, Clustering Center Identification, Long-tailed Distribution, Outward Statistical Testing

I. INTRODUCTION

Clustering algorithms attempt to classify elements into categories, on the basis of their similarity. Several different clustering strategies have been proposed, but no consensus has been reached even on the definition of a cluster [2]. In this work used the Distance based method for clustering. This method is based on the notion of density. The fundamental thought is to keep developing the given group the length of the thickness in the neighbors surpasses some limit, i.e., for every information point inside a given bunch, the sweep of a given group needs to contain no less than a base number of focuses.

The goal of clustering is to identify structure of a data set by objectively organizing data into several groups where the within-group-object similarity is minimized and the between-group-object dissimilarity is amplified. Clustering is vital when no marked data are accessible paying little heed to whether the data are binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, interactive media, or blends of the above data sorts. Data are called static if all their element values don't change with time, or change insignificantly. The group of clustering analyses has been performed on static data. If not all, clustering programs developed as an independent program or as part of a large suite of data analysis or data mining software to date work only with static data [7].

In this effort the algorithm Statistical Test based Clustering. In this procedure, first, here define a new metric to evaluate the local density of each object. Then, employ an outward statistical test method to identify the clustering centers automatically on a centrality metric constructed based on the new local density and new minimum density-based distance. Cluster is gathering of items that has a place with a similar class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. Clustering is the process of making a group of abstract objects into classes of similar objects. To employ metric, Normalized Mutual Information, this has been widely used to evaluate the performance of the clustering algorithm for sensitive analysis.

Clustering is considered as an unsupervised learning in data mining. Unsupervised learning means that only the intrinsic structure of the data defines the groups of data. The trouble here is that for the most part there is no from the earlier data about the structure of the information set or potential parameters, Thus, to solve this problem, many methods make assumptions to fit the data. Then assumptions are made according to the distribution of data set or the shape of data

set [8]. Clustering problems arise in many different applications, such as data mining and knowledge discovery [9].

Data mining is an interdisciplinary subfield of computer science and the analysis step of the ”Knowledge Discovery in Databases” process and it is the approach of discovering patterns in large data sets involving methods at the artificial intelligence, statistics, machine learning, and database systems. Data mining is an important part of knowledge discovery, is defined as the automated method of finding previously unknown, nontrivial, and useful information from databases.

The amount of the data seems to increase every single day for the majority domains related to information processing, and the need to find a way to mine and get knowledge from databases is still crucial. Data Mining defines the automated extraction procedures of hidden predictive information from databases. Data mining problems addressed by intelligent systems are: pattern recognition, prediction, classification, clustering etc.

II RELATED WORK

W. E. Donath [14] To show the effect of the maximum degree of any node being limited, and it is also shown that the right-hand side is a concave function of *U.* Lars Hagen [13] To present theoretical analysis showing that the second smallest eigenvalue of the Laplacian yields a new lower bound on the cost of the optimum ratio cut partition. M. Ester [12] they present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. P. S .Bradley [11] they used polyhedral distance, the problem can be formulated as that of minimizing a piecewise-linear concave function on a polyhedral set which is shown to be equivalent to a bilinear program: minimizing a bilinear function on a polyhedral set. T. Kanungo [9] to achieve faster clustering and better separation of clusters. V. Estivill-Castro [10] to achieve a system that can detect arbitrary shapes of clusters of different size and density. O.Dongquan Liu [8] to design a clustering method that can handle such irregular data sets and generate all values of parameters automatically. T. Warren Liao [7] this paper surveys and summarizes previous works that investigated the clustering of time series data in various application domains. B. Nadler [6] they present both synthetic examples and real image segmentation problems where various spectral clustering algorithms fail. In contrast, using this coherence measure finds the expected clusters at all scales. C.-P. Lai [3] to propose a novel approach named 2LTSC for clustering the time series by considering whole time series in the first level and the sub sequences in the second level to counter the failure to provide well rounded information. W. C. Xiankun Yang [5] to define arbitrary shape of clusters in spatial clustering, achieving fast and effective clustering without any need of knowing priori distribution. Donald C. Wunsch [4] to provide biomedical researchers with an overview of the status quo of clustering algorithms. A. Rodriguez and A.Laio [2] to achieve characterization of cluster centers with the help of density.

III PROBLEM STATEMENT

A number of clustering algorithms have been proposed based on different clustering mechanisms. In this existing clustering algorithm that can detect the clustering centers automatically via statistical testing result show data sets for example A-Set, S-Set, Shape Sets, High-dimensional data Sets, Real World Data Sets and calculate the Normalized Mutual Information with respect to STClu under each possible K. The impact of the number of nearest neighbors K on the performance of STClu in terms of NMI (Normalized Mutual Information). To overcome this problem to implement the new clustering algorithm that can detect the clustering centers automatically via statistical testing and also improve the accuracy in terms of Olivetti Face Data.

IV MOTIVATION

Clustering is one of the research places in the field of data mining and has extensive applications in practice it has been widely used in different disciplines and applications. It also has developed many algorithms on clustering but in this idea to overcome the shortages of identification of the “ideal” number of clusters and proposes a new clustering algorithm that can detect the clustering centers automatically via statistical testing.

V EXISTING SYSTEM

In existing algorithm first defines a new metric to measure the density of an object that is more robust to the preassigned parameter, further generates a metric to evaluate the centrality of each object. After that it identifies the objects with extremely large centrality metrics as the clustering centers via an outward statistical testing method. Finally, it groups the remaining objects into clusters containing their nearest neighbors with higher density. To find out the clustering center using following steps.

1)Metric Extraction.
2) Clustering center identification.
3) Object clustering.

A. Existing System Architecture

The following figure illustrates the system flow of the existing framework of automatic clustering on Density Metrics, which consists of the following steps. First we calculate the object density and density based distance.And then find clusters with their centers.

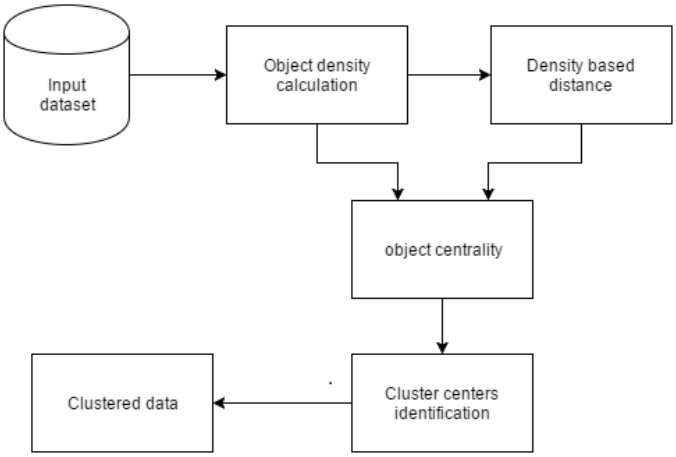

Figure1. Existing System Architechture

B. Algorithm for Existing system

Inputs: O ←{O1, O2,….., On}: A set of n objects
K: the number of nearest neighbors in K-density p̂
Output: CLU; //A set of clusters
1 RhosetMatrix ← ϕ, Deltaset ← ϕ, NNSet ← ϕ, GamaSet ← ϕ;
// Metric extraction
2 distanceMatrix ← Distancefunction (O);
3 RhoSet ← Fp̂ (distanceMatrix, k);
4 [DeltaSet, NNSet] ← Fδ (distanceMatrix, RhoSet);
5 GamaSet ← RhoSet. DeltaSet;
// clustering center identification
6 X ← sort (GamaSet, "descend");
7 R ← {Ri ← Xi, n /Xi+1, n} (1 ≤ i ≤ n-1);
8 m ← [0.1n], k ← 0;
9 while m > 2 do
10 | calculate the critical value rm according to eq.7;
11 | if Rm > rm then;
12 | | K ← m;
13 | | break;
14 | end
15 | m ← m-1;
16 end
17 Identify the objects corresponding to
{R1, R2,.... ,Rk} // object clustering

18 for i ← 1 to n do
19 | if Oi is unlabeled then
20 | | Mark Oi the label of its nearest neighbor
With higher p̂ according to NNSet;
21 | end
22 end
23 CLU ←{Clui, 1 ≤ i ≤ k};
24 return;

Clustering results on the first 100 Olivetti Face Data for existing System

(a) STClu

(b) RLClu

In STClu, in order to make a fair comparison, similar to RLClu, they follow the clustering process which performs on the K-density and the cut-off distance dc is set to 0.07 STClu algorithm identifies 46 clusters automatically. For algorithm RLClu, they set the number of clustering centers as 46 identified by STClu manually in advance [1].

VI PROPOSED SYSTEM

To implement an improved a new clustering algorithm that can detect the clustering centers automatically via statistical testing on different data sets. Here consider two metric first defined a new metric, K-density to measure the local density of each object. Then second metric is defined to evaluate the distance of an object to its neighbors with higher density.After combine product of these two metrics to evaluate the centrality of each object and found by a long tailed distribution. Finally apply clustering algorithm that can detect the clustering centers automatically.

To shown above Clustering results on the first 100 Olivetti Face Data for existing algorithm the accuracy is not better for STClu is 64.50% and RLClu is 65.89% .To overcome this problem to implementing a new clustering algorithm that can detect the clustering centers automatically via statistical testing.

VII CONCLUSION

To implement a method and improved a statistical test based clustering algorithm that can automatically identify the clustering center and further cluster the objects in an effective way. Here consider two metric first defined a new metric, K-density to measure the local density of each object. Then second metric is defined to evaluate the distance of an object to its neighbors with higher density.After combine

product of these two metrics to evaluate the centrality of each object and found by a long tailed distribution. Finally apply clustering algorithm that can detect the clustering centers automatically.

REFERENCES

[1] G. Wang and Q. Song, "Automatic Clustering via Outward Statistical Testing on Density Metrics," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 1971-1985, Aug. 1 2016.

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

[3] C.-P. Lai, P.-C. Chung, and V. S. Tseng, "A novel two-level clustering method for time Series data analysis," Expert Systems with Applications, vol. 37, no. 9, pp. 6319–6326, 2010.

[4] I. Rui Xu, Donald C. Wunsch, "Clustering algorithms in biomedical research: a review," IEEE Reviews in Biomedical Engineering, vol. 3, pp. 120–154, 2010.

[5] W. C. Xiankun Yang, "A novel spatial clustering algorithm based on delaunay triangulation," J. Software Engineering & Applications, vol. 3, pp. 141–149, 2010.

[6] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering," in Advances in Neural Information Processing Systems, 2006, pp. 1017–1024.

[7] T. Warren Liao, "Clustering of time series data-a survey," Pattern Recognition, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[8] o. Dongquan Liu, Sourina, "Free-parameters clustering of spatial data with non-uniform density," in IEEE conference on cybernetics and intelligent systems, 2004, pp. 387 – 392.

[9] 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 Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881 – 892, 2002.

[10] V. Estivill-Castro and I. Lee, "Argument free clustering for large spatial point-data sets via boundary extraction from Delaunay diagram," Computers, Environment and Urban Systems, vol. 26, no. 4, pp. 315–334, 2002.

[11] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Clustering via concave minimization," Advances in neural information processing systems, pp. 368–374, 1997.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in Proceedings of International Conference on Knowledge Discovery and Data Mining, vol. 96, no. 34, 1996, pp. 226–231.

[13] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," IEEE Transactions on Computer-aided design of integrated circuits and systems, vol. 11, no. 9, pp. 1074–1085,1992.

[14] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," IBM Journal of Research and Development, vol. 17, no. 5, pp. 420–425, 1973.

Author's Profile

Miss.Ashvita A.Jadhav, received the Bachelor of Engieering Degree in Computer Science & Engineering from Deogiri Institute of Engineering & Management Studies, Aurangabad,India in year 2014. She is currently pursuing Master of Engineering from Rajshree Shahu School of Engineering and Research, JSPM NTC, Pune, India. Her main research work focuses on Data Mining.

Mr. Vilas S.Gaikwad, received the BE Degree in Computer Science & Engineering from the Dr.BAMU Aurangabad, the M.Tech degree in Computer Science & Engineering from Walchand College of Engineering(An autonomous Institute), Sangli. He is currently working as Assistant Professor in the Department of Computer Engineering, Rajashree Shahu School of Engineering and Research, JSPM NTC, Pune, India. His research area include Image Processing and Computer Network.