Improved System For Clustering Using Outward Statistical Testing On Density Metrics

Ashvita A. Jadhav
Department of Computer Engineering, Rajashree Shahu
School of Engineering and Research, JSPM NTC,
Pune, INDIA.
Email id- ash7.jadhav@gmail.com

Asst.Prof.V.S.Gaikwad
Department of Computer Engineering Rajashree Shahu
School of Engineering and Research, JSPM NTC,
Pune, INDIA
vilasgaikwad11@gmail.com

Abstract— **Clustering is the process of organizing objects into groups whose members are similar in some way and is very important technique in data mining as it has applications spread widely, for example marketing, biology, pattern recognition etc. Various algorithms have been proposed, published, implemented for clustering like first published by Rodriguez and Liao. But this algorithm is dependent and sensitive to specified parameters and also faces difficulties in identification of ideal problems. Second published by G. Wang and Q. Song. But in this algorithm observed that with increase in value of K for K-density evaluation after certain limit, the performance of the system starts dropping down and accuracy is not better in terms of Olivetti face dataset so impact on performance of the system. To avoid this problem to proposed a new method that will identify clustering centers automatically via statistical testing. To rectify this problem another research that uses outward statistical testing to detect cluster centres automatically which uses K-density concept that calculates ideal value of K for K-density evaluation with respect to the dataset that user provides. Using this method avoids the dropping down of performance and makes system even more robust.**

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

I. INTRODUCTION

Clustering is an important technique of examining data mining, which divides a set of objects into several groups in such a way that objects in same group are more similar with each other in some sense than with the objects in other groups [1]. RLClu is only based on the distance (or similarity) between objects. Secondly, as the density created clustering, it defines the clustering centers as the items with maximum local density, and can detect the non-spherical clusters [2]. Clustering is necessary when no labeled data are available regardless of whether the data are binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, multimedia, or mixtures of the above data types. Data are called static if all their feature values do not change with time, or change negligibly [7].

In this paper to propose an improved Statistical Test based Clustering Algorithm (STClu). This method is based on the notion of density. The basic idea is to keep improving the given cluster until the density in the neighborhood surpasses some threshold, i.e., every cluster identified must contain the least number of points. In this efforts to propose the algorithm Statistical Test based Clustering (STClu). In this algorithm, first, evaluate ideal value of K then define a new metric to evaluate the local density of each object and density based distance. Also, object centrality is calculated for every 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. Also, to avoid dropping down of performance of the system with increase in value of K, the system calculates ideal value of K with the help of dataset that user provides. This makes the system less dependent and more robust as it avoids the taking value K as a parameter from user.

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 depend 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 conveyed 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. EXISTING SYSTEM

In existing clustering algorithm that can detect the clustering centers automatically via statistical testing. Specifically, the proposed 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. Afterwards, it finds the objects with very 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.

IV. PROPOSED SYSTEM

To avoid the existing system problem we have proposed a new method that will identify clustering centers automatically via statistical testing. To rectify this problem another research that uses outward statistical testing to detect cluster centres automatically which uses K-density concept that calculates ideal value of K for K-density evaluation with respect to the dataset that user provides. Using this method avoids the dropping down of performance and makes system even more robust. In this system, to proposes a new clustering algorithm in an effective way that can detect the clustering centers automatically via statistical testing.

Identification of objects with extremely large value for centrality metric is done in next step. This is carried out by applying outward statistical testing algorithm. In this step, the objects with higher value of centralities are detected as cluster centres. The outward statistical testing algorithm calculates ideal number of cluster centres automatically as well. Once cluster centres are identified, the remaining i.e. non-cluster centers objects are grouped into the cluster with nearest cluster centers. So basically it is divided into steps as given below,

1) K-Density Evaluation

In this step, K-density is calculated where K is calculated automatically using the data that user provides which is more robust as compare to other clustering systems.

2) Density Based Distance Calculation

Using the K density calculated for every object, density based distance for every object is calculated.

3) Object Centrality

Using these two metrics to evaluate centrality for every object. The product of K-density and density based distance represents centrality of that object.

4) Centers Detection

Outward statistical testing algorithm is applied in order to detect cluster centers on centrality's long tailed distribution.

5) Clustering

This is the last step of the system. The non-clustering centers objects are clustered to a cluster with nearest cluster centres.

A. System Architecture

The following proposed system architecture shows the system flow of the proposed framework of automatic clustering on Density Metrics, which consists of the following steps. First calculate of k value. Second calculate the object density and density based distance. After found these two distance then find cluster centers identification with their centers.
Finally, identifies cluster centers then automatically clustered the data from datasets.

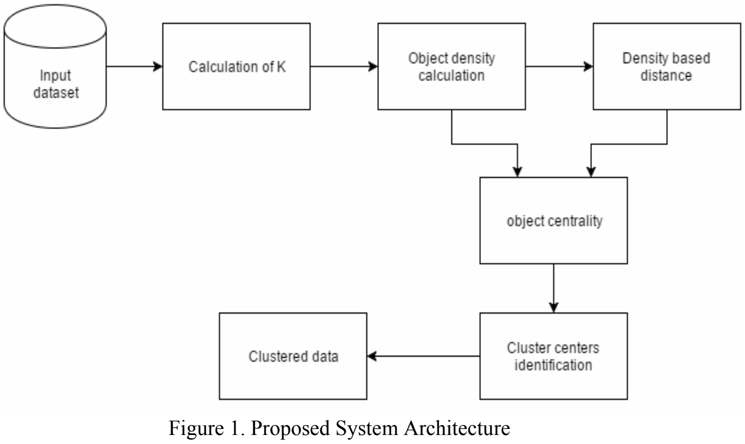

Figure 1. Proposed System Architecture

Although, the existing system showed that it can identify cluster centers along with ideal number of cluster centers automatically. The performance of system drops down once value of K is increased over a certain limit. To rectify this drawback we have introduced a technique that calculate ideal value of K with respect to the dataset that user provides. This avoids the accepting value of K as a parameter from user. Also improved accuracy in terms of Olivetti dataset which makes system much more robust and resistant to dropping down of performance.

B. Algorithm for Proposed System

Input: Set O of n objects.
Step 1:
Function calculate **K** (n)
// This function returns ideal value of k min 0 and max n
Step 2:
Distancefunction (O) // Calculating distance
Step3:
Calculate k-density using formula, K-density x.

$$x_i = \frac{K}{\sum_{j=1}^{K} d_{i,j}} \qquad \#$$

Step 4:
Calculate density based distance using formula,
New minimum density-based distance y

$$y_i = \min_{j \neq i \,\square\, x_i < x_j} (d_{i,j})$$

Step 5:
Define centrality by product of these two metrics
Step 6:
Sort objects by values of their centralised in descending order
Step 7:
Apply outward statistical testing to detect cluster centers
Step 8:
For each non-cluster center Oi do Mark Oi the label of its nearest neighbor with higher k-density.

C. Mathematical Model for Proposed System

1. Input → dataset
O → A set of n objects
K: the number of nearest neighbors in K-density x;
2. Output → Clusters
Let us Consider S= {x, y, γ, C, $d_{i, j, K}$}
Where x=local density,
y=minimum density based distance,
Ci =C1,C2,…..Cn,
where C is set of clustering centers,
K=nearest neighbor,
di, j =denotes the set of distances between object Oi and its K nearest neighbors.

V. EXPERIMENTAL SETUP AND RESULT

To performed clustering on various datasets. The types of datasets used are 2 dimensional datasets and real world dataset. Experimental results shows that the system evaluates ideal value of K and then evaluates K-density and density based distances which are used further to calculate the centrality of objects. An outward statistical testing algorithm is applied in order to identify the cluster centers and then non clustering center objects are clustered successfully.

Evaluated the results of the proposed system against that of existing systems on Olivetti image dataset. The results are as shown in the below table 1.

Table 1. Comparison between existing system and proposed system.

System	Existing System	Proposed System
Accuracy	64.50%	65.74%

The Following Figure 2 shows performance of existing and proposed system in the graph,

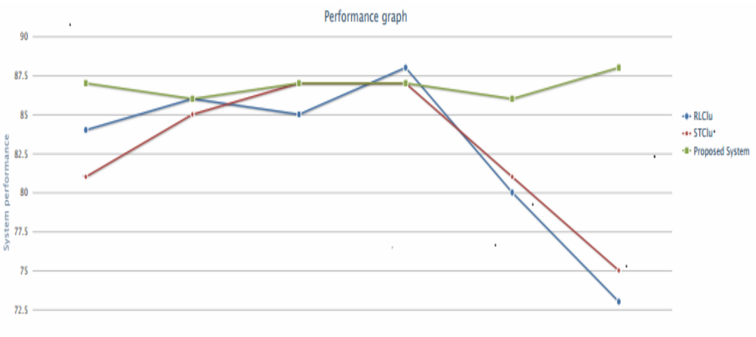

Figure 2. Graphical Representaion of Performance

VI. CONCLUSION AND FUTURE WORK

A system with improved outward statistical testing method is introduced. This system is able to calculate the ideal value of K for evaluation of K-density with the help of given dataset as input which is later used for evaluation of K-Density, density-based distance and object centrality. An outward statistical testing algorithm is applied on centrality metric to identify the cluster centers automatically. Now that have developed a system that can identify ideal value of K according to data user provide. This avoids breaking down of performance on increase in value of K. Also improved accuracy in terms of Olivetti image dataset which makes system much more robust and resistant to dropping down of performance.

For future research can make the algorithm flexible in such a way that it can cluster the data provided irrespective of its type or aspects. But for that purpose distances among all the objects in given data should be calculated in advance.

ACKNOWLEDGMENT

I express true sense of gratitude towards my project guide Prof.V.S.Gaikwad, of computer department for his invaluable co-operation and guidance that he gave me throughout my research, for inspiring me and providing me all the lab facilities, which made this research work very convenient and easy. I would also like to express my appreciation and thanks to our HOD Prof.R.H.Kulkarni and Director Dr.Prof.A.B.Auti and all my friends who knowingly or unknowingly have assisted me throughout my hard work.

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.

[15] A.Y.Ng,M.I.Jordan,Y.Weiss et al., “On spectral clustering:Analysis and an algorithm,” Advances in neural information processing systems, vol. 2, pp. 849–856, 2002.

AUTHORS 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.