IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol.7, No.2, Mar-April 2017

AN INTELLIGENT APPROACH FOR VARIOUS WIDTH CLUSTERING

Gift Lee Jones. J
Assistant Professor,
Dept Of Computer Science,
St.Joseph's Institute Of Technology.

Sinduja. S
UG Scholar,
Dept Of Computer Science,
St.Joseph's Institute Of Technology.

Shrividya. R
UG Scholar,
Dept Of Computer Science,
St.Joseph's Institute Of Technology.

***Abstract*—The approach based on various width clustering method has been used in many engineering and scientific applications to group data into clusters efficiently. Though this method has proved to form well-defined clusters based on their domain, in some cases this method tends to increase the number of clusters to a greater extent which drops down the purpose of clustering. In this paper we propose an additional feature to the existing system which would reduce the number of clusters keeping the benefits of various width clustering. The key here is to avoid small clusters by initializing minimum threshold for the clusters beyond which the data node will get merged with the existing cluster and does not form a new small cluster. It also calculates the distance between the two clusters. If the distance is small,merging takes place,i.e.,the small cluster combines with the neighbor cluster,else if the distance is large, the new small cluster is kept as it is,avoiding the formation of very large cluster.**

***Keywords*-Fixed Width Clustering, Various Width Clustering, threshold, dataset, performance.**

I. INTRODUCTION

Clustering is generally used to group the scattered data in a particular dataset into blocks or clusters which makes it easy to identify and access data. It is said to reduce the time taken to access the data.

In general,a cluster is a group or block of similar data that is usually grouped based on the characteristics of the data.

In initial stages of clustering applications,a method known as "FIXED WIDTH CLUSTERING" was used which later was found to have high level of overlapping. To overcome this issue,a method known as "VARIOUS WIDTH CLUSTERING" was introduced,where the threshold of each cluster varies according to the characteristics of the data in the data set.

In the existing method of various width clustering,the clusters are formed according to structure of data available. But it is found that in few cases,this method tends to form large number of clusters which inturn reduces the performance.

To overcome this special case,we introduce a special condition where we initialize a minimum threshold beyond which the nearest neighbor of the small cluster would be found and the data set would be merged with it.

It is expected to reduce the number of clusters to a greater extent and thus improve the performance of the existing Various width clustering algorithm.

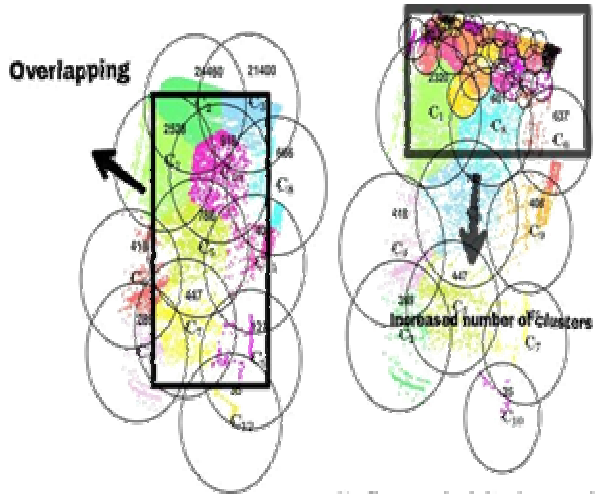

Figure 1 *: Fixed Width Clustering And Various Width Clustering*

Here there are 3 major components used

- Cluster-Width Learning
- Partitioning
- Merging

Cluster-width learning: For a dataset D,the function NNk(Hi) is used to find the K-Nearest neighbors where Hi is the selected object. The value of k is considered to be 50% of the data set which assures a large cluster. Random object H={H1,H2,....,Hr} are selected from the dataset for which the radius (r) of the K-Nearest Neighbor is calculated. The average here is taken as the global width.

Partitioning: Initially the the threshold value is got from the user. Initially this method clusters the data set that is widely scattered using a large width. Secondly,the cluster

threshold which exceeds the user defined threshold are subdivided into a number of clusters based on their based on their density.

Merging: In the process of partitioning,large clusters leads to the formation of several small cluster which might lie inside another cluster. The purpose of this merging is to combine these small ones that lie inside another cluster to a one whole cluster thereby increasing the performance.

II. SYSTEM ARCHITECTURE

In some cases,when the data is highly scattered,there tends to be formation of several small clusters which contains very few data nodes inside it. Formation of these small clusters tend to put down the purpose of grouping up of data(clustering). The purpose of this paper is to make the existing method to perform well in the above case as well.

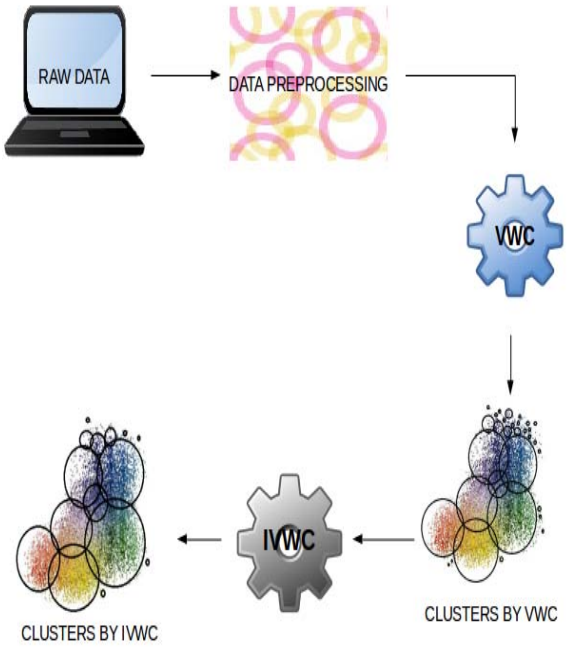

Figure 2 : System Architecture

Application of this is a two step process. Firstly,we initialize a minimum threshold for the clusters.After the formation of clusters using the various width clustering,a condition statement is included which checks for the size of the cluster. If the threshold of the cluster is below the initialized minimum threshold,the nearest neighboring cluster is found and this cluster is merged with the neighboring cluster.

Here arises another drawback. What if the neighbor cluster lies far away from the current small cluster?

Merging these two clusters would result in the formation of a very large cluster which is not actually needed. To handle this situation,the distance between the centroid of the current small cluster and the neighboring cluster is found. If this distance is smaller than α(maximum allowable distance),this small cluster is merged with neighbor cluster. Or else,if the distance between the centroids is large,the small cluster is left as it is.

And hence,including these conditions to the existing system would reduce the number of clusters to a greater extent which leads to the increase in performance without compromising with the quality of clusters.

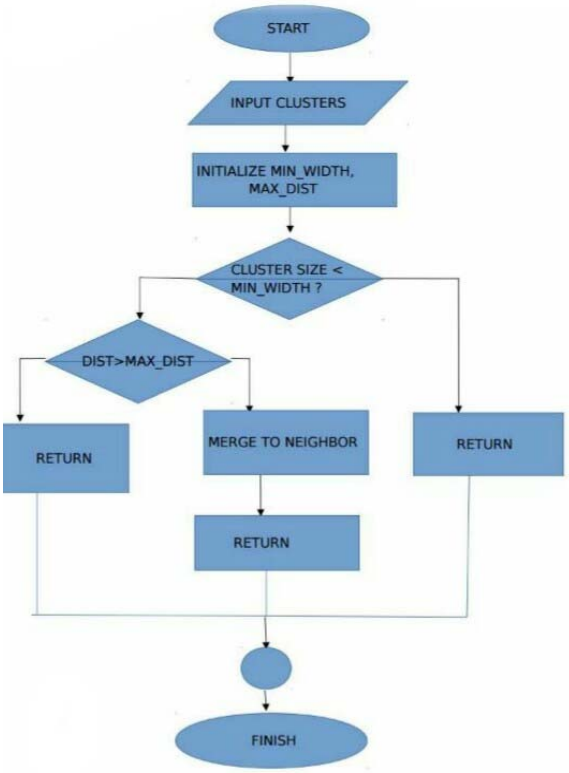

Figure 3 : *Intelligent Various Width Clustering*

III. ALGORITHM (INTELLIGENT VARIOUS WIDTH CLUSTERING)

```
Begin
Input: Clusters                 /*Result of VWC*/
Output: Clusters                /*Intelligent Clusters*/
Initialize min_width
Initialize α                    /*Maximum allowable
distance between clusters*/
for i to n                      /*For each cluster*/
if Cluster[i].size < min_width
Neigh← Closest_Cluster(Cluster[i])
Dist← Centroid_dist(Cluster[i],Neigh)
if(Dist > α)
return
```

Else
MergeClus(Cluster[i],Neigh)
End

IV. RESULT ANALYSIS

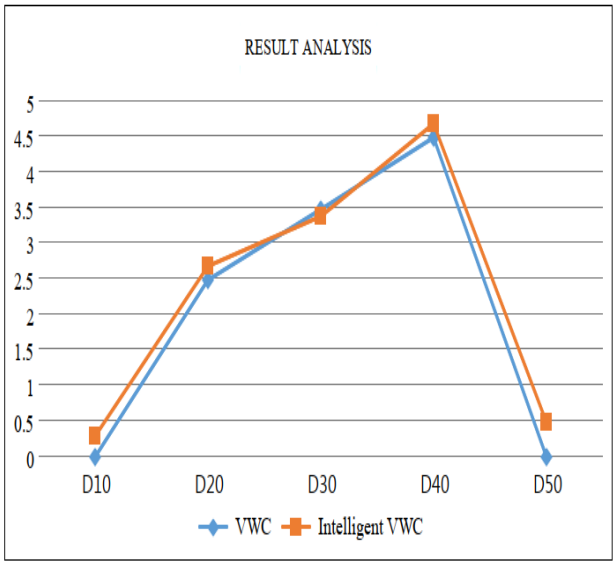

Figure 4 : *Result Analysis*

V. CONCLUSION & FUTURE SCOPE

This paper offers a method by which we could handle the performance of clustering in some special situations. This methods provides a oppurtunity to enhance the performance of Various Width Clustering algorithm by retaining the original purpose of clustering the data. By adding this to the existing system,the number of clusters can be reduced to a certain extend. It is found to increase the performance of cluster without compromising with the quality of clusters.

Though this method avoids the formation of small clusters to a certain extent,they cannot be avoided completely. In future,our challenge is to completely eliminate small clusters thereby promoting maximum performance.

REFERENCES

[1] T. Liu, A. W. Moore, and A. Gray (2006) , "New algorithms for efficient high-dimensional non-parametric classification," J. Mach. Learning Res., vol. 7, pp. 1135–1158.

[2] B. S. Kim and S. B. Park (1986) , "A fast k nearest neighbor finding algorithm based on the ordered partition," IEEE Trans. Pattern Anal. Mach. Intell., vol. TPAMI-8, no. 6, pp. 761–766, June.

[3] P. Ciaccia, M. Patella, and P. Zezula (1997), M-tree: An efficient access method for similarity search in metric spaces. In M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld,editors, Proceedings of the 23rd International Conference on Very Large Data Bases, pages 426–435,Athens, Greece, August.

[4] F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas (1996) , Fast nearest neighbor search in medical image databases. In Proceedings of the 22nd International Conference on Very Large Data Bases, pages 215–226, Mumbai, India, September.

[5] Abdul Mohsen Almalawi, Adil Fahad, Zahir Tari, Member, IEEE,Muhammad Aamir Cheema, and Ibrahim Khalil (2016), kNNVWC: "An Efficient k-Nearest Neighbors Approach Based on Various-Widths Clustering ",IEEE Transactions on knowledge and data engineering,VOL.28,NO.1,January.