# A Survey: Different Overlapping Clustering Approach

**Krupa A. Patel[1], Mr. M. B. Chaudhari[2]**
[1]Dept of CSE
[2]Professor, Dept of CSE
[1, 2] G.E.C., Gandhinagar, India

***Abstract-*** *In recent year huge amount of data and extreme use of that data in meaningful manner is one of the major scopes for research. Data mining is the process of extracting knowledge form dataset. Cluster analysis is decades' old concept of data mining which performs grouping of similar objects and dissimilar in another group. It is group of unsupervised data. It is used in various applications in the real world such as data/text mining, voice mining, image processing, web mining, medical data mining and many others. There are two types of clustering 1) hard 2) soft. Hard(crisp) clustering strict data point to belong to single cluster while soft(fuzzy) clustering allows one data point to belong to more than one cluster. There are many methods designed for clustering. This paper discuss about variants of overlapping clustering.*

***Keywords***- Data mining, Cluster analysis, overlapping clustering

## I. INTRODUCTION

Clustering is an unsupervised classification which deals with finding a structure in a collection of unlabelled data. The main objective of clustering is to find natural groupings among objects. It organizes data object in such that has higher intra cluster similarity and lower inter cluster similarity. Clustering method has many real time application in the fields like medical domain for disease prediction, medical image segmentation, biological sequence analysis; Market basket analysis, Pattern Recognition, Text segmentation and Social Media Analysis.

Clustering methods can be categorized according to the following criteria [1]:

1. Type of input data: To deal with different types of input data such as numerical, categorical and mixed, different clustering methods are used.
2. Type of proximity measures: Different types of similarity measures are defined to deal with different type of input data, some of them are Euclidean distance, Manhattan distance etc.
3. Type of generated cluster: In this category two types of clustering methods are defined one is Exclusive (Non-Overlapping) another one is Overlapping.
4. Type of clustering strategy used: In term of cluster strategy, clustering methods are divided into six groups.

| Clustering approach | Description | Example |
|---|---|---|
| Partitioning | Divide in various number of groups and then evaluate using any approach | K-means<br>K-medoids |
| Hierarchical | Create a hierarchical decomposition of the set of data using divisive or agglomerative method | BIRCH |
| Density-Based | Based on density and connectivity functions | DBSCAN |
| Grid-Based | Based on a multiple-level granularity structure | STING |
| Model-Based | Model is hypothesized for each cluster to find the best fit of data for a given model | EM |
| Graph-Based | Based on graph theory and connectivity function | DClusterR |

This paper is organized in four sections as: Section 2 presents overview of Overlapping clustering. Section 3 presents different overlapping clustering techniques and finally section 4 mentions a brief conclusion with future scope.

## II. OVERLAPPING CLUSTERING

Most of the clustering methods generate exclusive clusters i.e. one sample can belong to single cluster. However, almost real-world datasets cannot be fully explained using such strict constraint. For ex., in the field of medicine, various diseases share some common overlapping symptoms such as fever is common symptom in typhoid, malaria, viral infection and many others. Similarly, in social media analysis there can be an actor who belongs to multiple communities. Hence, this clustering method has become exceedingly popular since it is able to identify clusters where one data object can belong to multiple clusters.

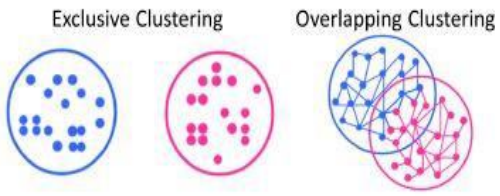


Figure 2.1: Pictorial representation of Exclusive vs. Overlapping Clustering

Fuzzy clustering can allow data objects to belong to multiple clusters with different degree of membership value. If data point have very less degree than it considers as outlier or noise and it will be ignore.

## III. DIFFERENT OVERLAPPING TECHNIQUES

There are many overlapping techniques other than fuzzy some of them are as below:

| Sr. | Technique | Description |
|---|---|---|
| 1 | FCM | It is first overlapped variant of K-means also called as fuzzy K-means which is based on the concept of fuzziness. Here data points are assigned to a particular cluster with membership degree between [0,1]. A data with highest membership for the particular cluster is assigned to that cluster.<br>In case highest degree of membership value is not unique than assign to arbitrary cluster. Small degree membership value object can consider as outlier or noise.[1] |
| 2 | OKM | Overlapping K-means uses heuristic approach to assign data points to one or more clusters by determining set of possible sorted assignment. Distances from each data points and clusters centroid are calculated and assignment of data points to multiple clusters is done by sorting the clusters from nearest to farthest.[11] |
| 3 | R-OKM | It extends OKM method. Objective function of R-OKM minimizes the distance between each observation and corresponding weight assigned to it.[6] |
| 4 | WOKM | It is the extended version of OKM and Weighted K-means which includes weighting factor into the objective function of OKM and this weighting factor is used to cluster the data points more appropriately and distance weight is assigned by the feature weights too.[11] |
| 5 | PCM | It overcomes limitation of FCM. This relaxes the column sum constraint so that the summation of every column satisfies the looser constraint. All element of the k-th column can be any number between [0,1], and at least one of them is positive. In this case the value should be interpreted as the *typicality* of relative to cluster rather than its membership in the cluster.[7] |
| 6 | PFCM | It combines FCM and PCM techniques.[8] |
| 7 | OPC | It accepts cluster k and s threshold as input. Than create two distance table and similarity table for preprocessing work. Calculate percentile for the assignment of similarity level(<5 % than level 0 else 1) |
| 8 | MOOKE | Divided in two parts 1) standard K-means process and generate matrix table 2) create membership table and compare it with matrix table generated by k-means run to *convlist*, it consider as threshold to allow object that belong to more than one cluster.[3] |
| 9 | NEOKM | It is replica of K-means. It manages assignment of data over the overlapped area of clusters and ignores outlier. There are two parameters that control the assignment. Two phases 1) Same as k-means to allow single assignment 2)allowed to assign multiple cluster[2] |
| 10 | KHM-OKM | It is improved OKM. As OKM relay on random initial parameter KHM-OKM uses harmonic mean value as a initial centroid point and then perform OKM on that result value.[4] |

## IV. CONCLUSTION AND FUTURE WORK

Despite of number of limitations of overlapping clustering partitioning algorithm it is widely used in clustering analysis and in prediction analysis. Its scope is not limited in one particular domain. Many improvements have been done to overcome the existing limitations of FCM to improve its performance. This survey show that FCM or OKM can be improved further to make it more accurate and applicable for different application.

## REFERENCES

[1] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis:Methods for Classification, Data Analysis and Image Recognition,Wiley, 1999.

[2] Y. Chen, H. Hu. An overlapping Cluster algorithm to provide nonexhaustive clustering. Presented at European Journal of Operational Research. pp. 762-780, 2006.

[3] S. Baadel, F. Thabtah, and J. Lu. Multi-Cluster Overlapping K-means Extension Algorithm. In proceedings of the XIII International Conference on Machine Learning and Computing, ICMLC'2015. 2015.

[4] Sina Khanmohammadi, Naiier Adibeig, Samaneh Shanehbandy. An Improved Overlapping k-Means Clustering Method for Medical Applications", Elsevier 2016

[5] Tanawat Limungkura, Peerapon Vateekul. Enhance Accuracy of Partition-based Overlapping Clustering by Exploiting Benefit of Distances between Clusters, 2016 International Conference on Knowledge and Systems Engineering

[6] Chiheb-Eddine ben N'Cir, Guillaume Cleuziou, Nadia Essoussi. Identification of Non-Disjoint Clusters with Small and Parameterizable Overlaps, IEEE 2013

[7] Said Baadel, Fadi Thabtah, Joan Lu. Overlapping Clustering: A Review, SAI Computing Conference 2016

[8] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek. A Possibilistic Fuzzy c-Means Clustering Algorithm, IEEE 2013

[9] B.Durgadevi, Dr.S.Rajalakshmi. Performing Age Group Clustering in Breast Cancer Datasets Using FCM Algorithm, IJESRT 2013

[10] Manisha Goyal, Mr. M.B. Chaudhary, Ms. Pinal Patel, A Survey: Different Improvements and Integrated Approaches of K-means. IJIRT 2018

[11] Argenis Aroche, José Francisco Martínez-Trinidad, José Arturo Olvera-López, Airel Perez-Suarez Study of Overlapping Clustering Algorithms Based on Kmeans through FBcubed Metric, Springer 2014.