A Technical Insight into Clustering Algorithms & Applications

Nandita Yambem[1], and Dr A.N.Nandakumar[2]

[1] Research Scholar ,Department of CSE, Jain University,Bangalore, India
[2] Professor ,Department of ISE , New Horizon College of Engineering,Bangalore, India.

---***---

Abstract - *Today is the age of data which is being produced at a tremendous rate. Cluster analysis divides data into groups for the purposes of summarization or improved understanding to assist in decision making. Researchers in data mining and machine learning faces challenges due to many factors which can make use of this large data to extract useful knowledge. As a part of my ongoing Research work I have undertaken an exhaustive technical survey to cover the key aspects of Clustering like - the need, clusters types and various clustering algorithms, Clustering applications etc*

Key Words: **Clusters, Outliers, Clustering Algorithms-partitioning, hierarchical, density, grid based, model based, high dimensional data, constraint based clustering.**

1. INTRODUCTION

A *cluster* is a collection of objects which are "similar" between them and "dissimilar" to the objects belonging to other clusters.Intra class cluster the objects are highly similar and distance between objects are minimized .Interclass cluster the objects are dissimilar to objects in other classes and distance between the objects in the dissimilar clusters are maximized[1][2].

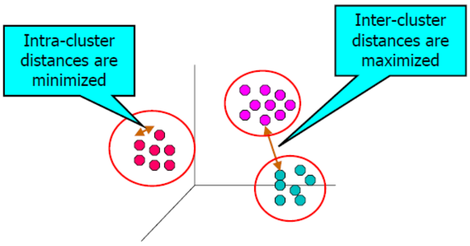
Intra-cluster distances are minimized
Inter-cluster distances are maximized

Fig:1 Clusters

Clustering is an unsupervised learning process of partitioning or grouping physical or abstract objects into classes/subsets of similar objects. The partitioning is done with the help of clustering algorithm. So, clustering is useful and lead to the discovery of previously unknown groups within the data [2].

1.1 Need For clustering [2]

- **Scalability**

  Clustering algorithms which are highly scalable are needed to handle large databases containing millions and billions of objects

- **Dealing with different types of attributes**

  Applications may require clustering techniques for complex data types other data types such as graphs, sequences, images, and documents in addition to data type binary, nominal (categorical), and ordinal data, or mixtures of these data types

- **Discovery of clusters with arbitrary shape**

  Basic clustering algorithm determines clusters based on Euclidian or Manhattan distance measures which find spherical clusters with similar size and density. However, a cluster can be of any shape, so, it is important to develop algorithms which detect arbitrary shape clusters.

- **Minimal requirements for domain knowledge to determine input parameters**

  The clustering result may be sensitive to parameters (desired number of clusters) as inputs provided to domain knowledge to clustering algorithm. For high dimensional data sets, parameters are hard to determine .Hence specifying requirement of domain knowledge is cumbersome and makes difficult to control the quality of clustering.

- **To deal with noise and outliers**

Clustering algorithms can be sensitive to real-world data sets contain outliers and/or missing, unknown, or erroneous data or inaccurate data generally called noise may produce poor quality clusters. Therefore, robust clustering methods are needed to handle noise and outliers.

- **Incremental clustering and insensitive to order of input records**

  In many applications, a new clustering algorithm has to be recomputed as some clustering algorithms cannot incorporate incremental updates into existing clustering structures. Hence, depending on the order in which the objects are presented clustering algorithms may return dramatically different clustering, given a set of data objects. Hence, incremental clustering algorithms and algorithms that are insensitive to the input order are needed.

- **High dimensionality**

  Most clustering algorithms can handle low dimensional data sets involving two or three dimensions. But finding clusters of data objects is challenging in a high-dimensional space considering that data can be very sparse and highly skewed.

- **Incorporation of user-specified constraints**

  Most real-world applications may need to perform clustering under various kinds of constraints. A challenging task is to find data groups with good clustering behavior that satisfy specified constraints.

- **Interpretability and usability**

  The clustering results need to be interpretable, comprehensible, and usable to the users. Hence clustering may be needed to be tied with a specific semantic interpretations and applications. so Study of how an application goal may influence the selection of clustering features and clustering methods is important.

1.2 CLUSTER TYPES

1. Well-separated clusters
2. Center-based clusters
3. Contiguous clusters
4. Density-based clusters
5. Property or Conceptual
6. Described by an Objective Function

1) **Well-Separated Cluster Definition[3]**:

- A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.
- Sometimes a threshold is used to specify that all the points in a cluster must be sufficiently close (or similar) to one another.
- Well-separated clusters do not need to be globular, but can have any shape.

Figure 2: Three well-separated clusters of 2 dimensional points

2) **Center-based/Prototype-based Cluster Definition[3]**:

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster called a centroid, than to the center of any other cluster.

- Such clusters tend to be globular.

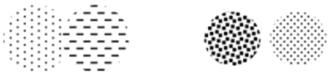

Figure 3: Four center-based clusters of 2 dimensional points

3) **Density-based definition[3]**:

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.

- Clusters are irregular or intertwined, and when noise and outliers are present.

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.

- Used when the clusters are irregular or intertwined, and when noise and outliers are present.

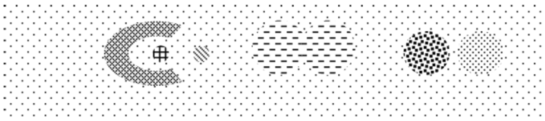

Figure 5: Six dense clusters of 2 dimensional points.

4) **Shared Property or Conceptual Clusters [3]**

- Finds clusters that share some common property or represent a particular concept

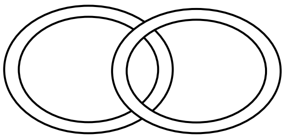

Figure 6: Two Overlapped Circles

5) **Clusters Defined by an Objective Function[3]**

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function.
- Can have global or local objectives.
- Hierarchical clustering algorithms typically have local objectives
- Partitional algorithms typically have global objectives
  - A variation of the global objective function approach is to fit the data to a parameterized model.
- Parameters for the model are determined from the data.
- Mixture models assume that the data is a 'mixture' of a number of statistical distributions.
- Map the clustering problem to a different domain and solve a related problem in that domain
- Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
- Clustering is equivalent to breaking the graph into connected components, one for each cluster.
- Want to minimize the edge weight between clusters and maximize the edge weight within clusters

2. CLUSTERING APPLICATIONS

Clustering finds wide applications in [4][3][7]

- Spatial data analysis
- Economic Science(Market research)
- Land use
- Insurance
- City-planning
- Earth-quake studies
- Economic Science
- Image processing
- Medical diagnosis
- Text Mining
- Document classification
- Cluster Weblog data to discover groups of similar access patterns
- Pattern Recognition.
- Spatial Data Analysis
- Create thematic maps in GIS by clustering feature spaces

3.CLUSTERING METHODS

Table 1: Clustering methods classification

Clustering methods	Partitioning method	k-means
		k-medoids
		k-modes
		PAM
		CLARANS
		CLARANS
		FCM
	Hierarchical method	BIRCH
		CURE
		ROCK
		Chameleon
		Echidna
	Density based method	DBSCAN
		OPTICS
		DBCLASD
		DENCLUE
	Grid Based method	Wave-cluster
		OptiGrid
		STING
	Model Based method	EM
		COBWEB
		CLASSIT
		SOMS
	Clustering High Dimensional Data	CLIQUE
		PROCLUS
	Constraint Based Clustering	COP K-means
		PCKmeans
		CMWK-means

3.1 Partitioning method [4][5][8]:

- Various partitions are constructed and then evaluated them by some criterion.
- Finds mutually exclusive clusters of spherical shape
- Clustering distance based
- May use mean or medoid to represent cluster center
- Effective for small to medium-size data sets

3.2 Hierarchical method [4][5][8]:

- A hierarchical decomposition of the set of data (or objects) is created using some criterion.
- Cannot correct erroneous merges or splits
- May incorporate other techniques like micro clustering or consider object " linkage"

3.3 Density-based [4][5][8]

- Clustering based on density (local cluster criterion), such as density-connected points
- Discover arbitrary shaped clusters
- Clusters are dense regions of objects in space that are separated by low density regions
- Cluster density: Each point must have a minimum number of points within its " neighborhood"
- Handle noise
- One scan
- May filter out outliers
- Need density parameters as termination condition

3.4 Grid-based [4][5][8]

- Clustering based on a multiple-level granularity structure
- A multi resolution grid data structure
- Fast processing time (typically independent of the number of data objects, yet dependent on grid size).

3.5 Model-based [4][5][8]

- Hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.
- Assume the data generated from K probability distributions
- Typically Gaussian distribution Soft or probabilistic version of K-means clustering
- Need to find distribution parameters.
- This method locates the clusters by clustering the density function. It reflects the spatial distribution of the data points.
- This method also automatically determines the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods

3.6 Clustering High Dimensional Data [2]

- A data object may be described by 10 or more attributes. Such objects are referred to as a high-dimensional data space.
- Clustering high-dimensional data is the search for clusters and the space in which they exist.
- Two major kinds of methods:
  - *Subspace clustering approaches* searches for clusters existing in subspaces of the given high-dimensional data space, where a subspace is defined using a subset of attributes in the full space
  - *Dimensionality reduction approaches* try to construct a much lower-dimensional space and search for clusters in such space. A method may construct new dimensions by combining some dimensions from the original data.

3.7 Constraint Based Clustering [6]

- Constraint based clustering is performed by the incorporation of user or application-oriented constraints.
- Refers to the user expectation or the properties of desired clustering results.
- Provide us with an interactive way of communication with the clustering process.
- Constraints can be specified by the user or the application requirement
- Constraints on individual objects
- Constraints on selection of clustering parameters
- Constraints on distance or similarity functions
- User-specified constraints on the properties of individual clusters
- Semi-supervised clustering based on "partial" supervision.

CONCLUSIONS

Clustering is a compelling process in Data mining that leads to logical segregation of huge volumes & variety of data to speeds up the data search process. Cluster Analysis groups the objects, called as a cluster/s, such that objects that are similar to each other in a given cluster and dissimilar to the objects in other cluster. A number of factors make it difficult for researchers to extract knowledge from these larger clusters & the factors which lead to the need for clustering and different cluster types and classification of clustering methods. Several Researchers till date, have attempted to develop & improvise clustering methods & algorithms for Data mining Analysis. As a part of my Research I have presented the basics & popular Research work in this domain.

REFERENCES

[1] Clustering Techniques: A Brief Survey of Different Clustering Algorithms, Deepti Sisodia Technocrates Institute of Technology, Bhopal, India, Lokesh Singh Technocrates Institute of Technology, Bhopal, India Sheetal Sisodia Samrat Ashoka Technological Institute, Vidisha, India Khushboo saxena Technocrates Institute of Technology, Bhopal, India, International Journal of Latest Trends in Engineering and Technology (IJLTET) ,2012.

[2] Jiawei Han and Michheline Kamber, Data mining concepts and techniques-a reference book

[3] Introduction to Data Mining, Pang-ning Tan, Michael Steinbach, Vipin Kumar

[4] Clustering and its Applications, L.V. Bijuraj, Proceedings of National Conference on New Horizons in IT - NCNHIT 2013

[5] A Comparative Study of Various Clustering Algorithms in Data Mining, S.Saraswathi, Dr. Mary Immaculate Sheela, IJCSMC, Vol. 3, Issue. 11, November 2014

[6] Survey on Clustering Techniques of Data Mining, Namrata S Gupta, Bijendra S.Agrawal, Rajkumar M. Chauhan, Asst. Prof. Smt. BK Mehta IT Centre (BCA College), Palanpur, Gujarat, INDIA, Principal, CCMS, Vadu, Gujarat, INDIA, Foreman Instructor I.T.I. Amirgadh Ex. Asst. Professor BCA College Palanpur, Gujarat, INDIA, AIJRSTEM, 2015

[7] Clustering Techniques and Applications to Image Segmentation, Liang Shan, shan@cs.unc.edu

[8]Comparing Clustering Algorithms, www.cise.ufl.edu/~jmishra/clustering

[9] Data Clustering: Theory, Algorithms, and Applications, Guojun Gan, Chaoqun Ma and Jianhong Wu, SIAM series