International Journal
of
Linguistics & Computing Research

A Survey on Clustering Techniques

Girish Kumar Singh
Department of Computer Science and Application
Dr. Hari Singh Gour Central University
Sagar, Madhya Pradesh, India
gkrsingh@gmail.com

Abstract*— Data mining is a nontrivial process of extraction of interesting, implicit, potentially useful and previously unknown knowledge. There are various tasks of data mining like estimation & prediction, classification, association discovery, clustering, visualization of data and visual data exploration. Clustering is the process of division of data into groups of similar objects. Clustering is an unsupervised learning. The contributing areas of research in clustering include data mining, statistics, machine learning, spatial database technology, biology and marketing. Cluster analysis has been widely used in various disciplines such as pattern recognition, computer vision, data mining, bioinformatics, web mining, text mining, finance, market analysis and image analysis and so on.

Keywords — KDD, Data Mining, Clustering.

I. INTRODUCTION

The interest in automating the analysis of large volumes of data has been the motivation factor for several research projects in the emergent field called *Knowledge Discovery in Databases* (KDD) [1]. KDD is the process of knowledge extraction from a large mass of data with the goal of obtaining meaning to be able to interpret data, and to acquire new knowledge if any. This process is very complex because it consists of a techniques and approaches combining a number of mathematical and technical models to find patterns and regularities in the data [2].

Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data Mining can be defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [3] and "the science of extracting useful information from large data sets or databases" [4]. Although it is usually used in relation to analysis of data, data mining, like artificial intelligence, is an umbrella term used for a wide range of contexts. Data Mining involves the process of analyzing data to show patterns or relationships; sorting through large amounts of data; and picking out pieces of relative information existing in data. The common domain of application of data mining are business, finance, medicine/health to learn trends, patterns etc.

Data mining brings together the traditional areas of databases, statistics, machine learning, and human-computer interaction on common platform to achieve above mentioned objectives. The general model for data mining can be a 3-steps model as shown in Figure 1.

- *Data Preprocessing:* In data preprocessing step, the raw data are prepared for further processing. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format for more efficient and effective processing to suit the user.
- *Mining the Data:* After the preprocessing, the data are analyzed to extract implicit, previously unknown, interesting and potentially useful information.

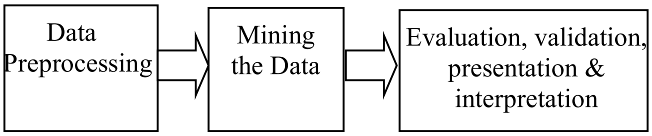

Figure 1: A three steps general Data mining

- *Evaluation, Validation, Presentation & Interpretation:* This step pertains to the results achieved in step 2. The results without any qualification bear no significance therefore the result of the data mining need evaluation and validation. Further for the appropriate utilization of the obtained results a user friendly presentation and interpretation is combined in step 3.

Data Mining has several tasks which can be roughly classified into six categories: Estimation and Prediction, Classification, Association Discovery, Clustering and Cluster Analysis, Visualization of Data, and Visual Data Exploration. Some of the most popularly applied techniques to perform data mining task are: Statistical Analysis, Decision Tree, Neural Network, Inductive Logic Programming, Clustering, Association Rule, Nearest Neighbor Technique, Genetic Algorithms, Fuzzy Logic, Rough Sets, Concept Learning and Rule-Based Reasoning.

- *Estimation and Prediction:* Estimation consists of examining the attributes of a set of entities (products, processes, samples etc.) and, based on these attribute values, assigning values to an unknown desired attribute. The term prediction is sometimes used when an estimation is done to predict the future outcome of an attribute value. A typical example of an estimation task is to use attributes that characterize a project to estimate (predict) its costs.
- *Classification:* Classification is defined as examining the attributes of a given entity and assigning it to a predefined category or class based on these attribute values.
- *Association Discovery:* Association discovery is identifying which characteristics are associated with each other in a given environment. A typical example would be to identify which characteristics of the software development team, where characteristic of the development team may be described the values of experience attributes, training attributes, domain knowledge attributes, etc. to the values of usability attributes, maintainability attributes, reliability attributes, etc.
- *Clustering & Cluster Analysis:* Clustering generally described as the task of segmenting a heterogeneous population into a set of more homogeneous subgroups. Clustering differs from classification, as it does not rely on predefined classes. Clustering segments the population into classes on the basis of the similarity between the class members. It also produces a high level description of the population applying distance measures between its elements.
- *Visualization of Data:* Data visualization is the task of describing complex information through the displays of visual data. Visualization is motivated by the need to enhance the understanding of the domain expert regarding the concept under consider.
- *Visual Data Exploration:* The visual data exploration involves inspecting large volumes of data through interactive control of visual displays. This task allows the domain experts to examine "what if" scenarios in multivariate visual displays. This task has been found to apply "visual data mining" tools with advanced user interfaces for interactive exploration.

Rest of the paper is organized in eight sections. Section 2 discusses basic concepts of clustering and section 3 gives the classification of clustering algorithms. Section 4, 5, 6 and 7 explain four major types of clustering algorithm namely hierarchical, partitioning, grid based and density based clustering algorithms. Lastly paper has been concluded in section 8.

II. CLUSTERING

Clustering is a useful technique for the discovery of data distribution and patterns in the underlying data. The goal of clustering is to discover both the dense and the sparse regions in the data set. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [5]. The contributing areas of research include data mining, statistics, machine learning, spatial database technology, biology and marketing. Cluster analysis has been widely used in various disciplines such as pattern recognition, computer vision and data mining [6].

Clustering is a challenging field of research. The specific applications pose their respective special requirements. The following are typical requirements of clustering algorithm in data mining [5]:

i. *Scalability:* Many clustering algorithms work well on small data sets containing hundreds of data objects; however, the same algorithm may not work equally efficiently on a large database with millions of objects. Clustering on a *sample* of a given large data set may lead to biased results. Therefore scalable clustering algorithms are needed.

ii. *Ability to deal with different types of attributes:* Most algorithms are designed to cluster interval–base data. However, application may require clustering of data of other types, such as binary categorical, ordinal data, spatial data or mixture of these types of data.

iii. *Discovery of clusters of arbitrary shape:* Mostly the clustering algorithms determine clusters based on Euclidean or Manhattan distance measure. Algorithms using such distance measures tend to find spherical clusters with similar size and density. However a cluster could be of any shape. A clustering algorithm to detect clusters of arbitrary shape may be useful for many real applications.

iv. *Minimal requirement for domain knowledge to determine input parameters:* A number of clustering algorithms require the users to input certain parameters. A set of clustering results corresponds to the input parameters. In the absence of the domain knowledge parameters are often hard to determine, especially for the high dimensional databases. This on one hand burdens the users and also makes it difficult to control the quality of clustering.

v. *Ability to deal with noisy data:* Most real world databases contain outliers or missing, unknown or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

vi. *Insensitivity to order of input records:* Some clustering algorithms are sensitive to the order in which data is input: for example the same set of data when presented with different ordering to such an algorithm, may yield different clusters. Developing algorithms that are not sensitive to the order of input is also a challenge in this field.

vii. High dimensionality: A database or data warehouse may have high dimensionality. Many clustering algorithms work with good efficiency for low dimensional dataset. Human eyes are good at judging the quality of clustering for up to three dimensions. It is challenging to cluster the objects in high dimensional space, especially considering that such data can be very sparse and highly skewed.

viii. Constraint based clustering: Real-world applications may need to perform clustering under various kinds of constraints.

ix. Interpretability and usability: Users expect clustering result to be interpretable, comprehensible, and usable. That is, clustering may need to be tied up with specific semantic interpretation and applications. It is important to study how an application may influence the selection of clustering methods.

III. CLASSIFIACTION OF CLUSTERING ALGORITHM

The clustering methods can be classified basically into partitioning, hierarchical, grid-based and density-based. Following is a detailed classification of clustering algorithms given by Pavel Berkhin [7]:

- Hierarchical Methods
  - Agglomerative Algorithms
  - Divisive Algorithms
- Partitioning Relocation Methods
  - Probabilistic Clustering
  - K-medoids Methods
  - K-means Methods
- Density-Based Partitioning Methods
  - Density-Based Connectivity Clustering
  - Density Functions Clustering
- Grid-Based Methods
- Methods Based on Co-Occurrence of Categorical Data
- Other Clustering Techniques
  - Constraint-Based Clustering
  - Graph Partitioning
  - Clustering Algorithms and Supervised Learning
  - Clustering Algorithms in Machine Learning
- Scalable Clustering Algorithms
- Algorithms For High Dimensional Data
  - Subspace Clustering
  - Co-Clustering Techniques

In next sections four major class of clustering algorithm has discussed.

IV. HIERARCHICAL METHOD

Hierarchical algorithms create a hierarchical decomposition of the database *D*. The hierarchical decomposition is represented by a dendrogram, a tree that iteratively splits *D* into smaller subsets until each subset consists of only one object. In such a hierarchy, each node of the tree represents a cluster of *D*. The dendrogram can either be created from the leaves up to the root (agglomerative approach) or from the root down to the leaves (divisive approach) by merging or dividing clusters at each step [8] [9]. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more of the most similar clusters and this process continue until some stopping criterion is achieved. One example of a stopping condition is the critical distance d_{min} between all the clusters of *D*. A divisive clustering starts with a single cluster containing all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved. Advantages of hierarchical clustering include:

- Flexibility regarding the level of granularity
- Ease of handling any form of similarity or distance
- Applicability to any attribute types

Disadvantages of hierarchical clustering are related to:

- Vagueness of termination criteria
- Most hierarchical algorithms do not revisit (intermediate) clusters once constructed.

The main problem with hierarchical clustering algorithms so far has been the difficulty of deriving appropriate parameters for the termination condition, e.g. a value of d_{min} which is small enough to separate all “natural” clusters and, at the same time large enough such that no cluster is split into two parts. Ejcluster [10] hierarchical algorithm presented in the area of signal processing automatically derives a termination condition. Ejcluster follows the divisive approach. Experiments show that it is very effective in discovering non-convex clusters. However, the computational cost of Ejcluster is $O(n^2)$ due to the distance calculation for each pair of points. This is acceptable for applications such as character recognition with moderate values for n, but it is prohibitive for applications on large databases.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [11] has the complexity O(n) using a hierarchical data structure called CF-tree for multiphase clustering. In BIRCH, a single scan of the dataset yields a good clustering and one or more additional scans can be used to improve the quality of cluster further. However, it handles only numerical data and it is order-sensitive. Also, BIRCH does not perform well when the clusters do not have uniform size and shape since it uses only the centroid of a cluster when redistributing the data points in the final phase.

Clustering Using Representatives (CURE) [12] employs a combination of random sampling and partitioning to handle large databases. It identifies clusters having non-spherical shapes and wide variances in size by representing each cluster by multiple points. The representative points of a cluster are generated by selecting well-scattered points from the cluster and shrinking them toward the centre of the cluster by a specified fraction. However, CURE is sensitive to some parameters such as the number of representative points, the shrink factor used for handling outliers, number of partitions. Thus, the quality of clustering results depends on the selection of these parameters.

RObust Clustering using linKs (ROCK) [13] is a representative hierarchical clustering algorithm for categorical data. It introduces a novel concept called "link" in order to measure the similarity/proximity between a pair of data points. Thus, the ROCK clustering method extends to non-metric similarity measures that are relevant to categorical data sets. It also exhibits good scalability properties in comparison with the traditional algorithms employing techniques of random sampling. Moreover, it seems to handle successfully data sets with significant differences in the sizes of clusters.

V. PARTITION METHOD

Partitioning algorithms had been popular clustering algorithms long before the emergence of data mining. Given a set D of n objects in a d-dimensional space and an input parameter k, a partitioning algorithm organizes the objects into k clusters such that the total deviation of each object from its cluster centre or from a cluster distribution is minimized. The deviation of an object from the cluster centre is commonly computed using a similarity function. There are many partitioning methods such as k-mean algorithm [14], EM (Expectation Maximization) [15] algorithm [16], PAM (Partition around Medoid, k-medoid) algorithm [17], CLARA [17], CLARANS [18] etc.

The partitioning algorithms generally start with an initial partition of the database and then use an iterative control strategy to optimize an objective function. Each cluster is represented by the centre of gravity of the cluster (*k*-mean algorithms) or by one of the objects of the cluster located near its centre (*k*-medoid algorithms). Partitioning algorithms use a two-step procedure. First, determine *k* representatives minimizing the objective function. Second assign each object to the cluster with its representative "closest" to the considered object. The second step implies that a partition is equivalent to a Voronoi diagram and each cluster is contained in one of the Voronoi cells. Thus the shape of all clusters found by a partitioning algorithm is convex which is very restrictive.

Kaufman and Rousseeuw in 1990 proposed CLARA (Clustering LARge Applications) which relies on sampling to handle large data sets. CLARA draws a sample in random from the data set, applies PAM to the sample, and finds the medoids of the sample. The quality of clustering at this stage is measured based on the average dissimilarity of all objects in the entire data set, and not only of those objects in the samples. Experiment reported in [17] indicates that CLARA is more efficient than PAM. The main disadvantage of the CLARA is that, one can't find the best clustering if the any sampled medoid is not among the best k medoids.

Ng & Han proposed a partitioning algorithm for spatial databases called CLARANS (Clustering Large Applications based on RANdomized Search) [18]. It is an improved k-medoid method with improved effectiveness and efficiencies. Ng & Han have also discussed methods to determine the "natural" number of clusters k. CLARANS assumes that all objects to be clustered can reside in main memory at the same time which does not hold for large databases. Furthermore, the run time of CLARANS is prohibitive on large databases.

VI. GRID BASED METHOD

The grid-based clustering approach uses a multi resolution grid structure. It quantizes the space into a finite number of cells that form a grid structure on which all the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is typically independent of the number of data objects, yet dependent on only the number of cells in each dimension in the quantized space.

Some typical examples of the grid-based approach include STatistical INformation Grid-based method (STING) [19], in which the statistical information stored in the grid cellsis explored; Wave Cluster, which clusters objects using a wavelet transform method; and Clustering In QUEst (CLIQUE) [20], which represents a grid and density-based approach for clustering in high-dimensional data space.

VII. DENSITY BASED METHOD

The general idea behind the density based method is to continue to grow a cluster as long as the density in the neighborhood exceeds some threshold; that is for each data point within a given cluster, the neighborhood of given radius contains at least a minimum number of points. This approach is useful to filter out noise (outlier) and to discover clusters of arbitrary shape. Some density based methods are Density Based Spatial Clustering of Applications with Noise (DBSCAN) [21], DENsitybased CLUstEring (DENCLUE) [22], Generalized DBSCAN (GDBSCAN) [23] and Ordering Points To Identify the Clustering Structure (OPTICS) [24].

CONCLUSION

In this paper, we have presented a survey on clustering techniques. The aim of this survey is to provide very basic concepts clustering techniques. Clustering has been introduced starting with KDD and data mining. Clustering is one of the techniques in data mining to find pattern and group in the dataset. There are various types of clustering techniques and the have been discussed. There are several requirements which should be full fill by the clustering algorithms, unfortunately none of the existing algorithm satisfied the entire requirement for a clustering algorithm. Each of clustering techniques has their own advantages and disadvantages. Depending on the problem in the hand one should select the suitable clustering techniques.

REFERENCES

[1] U. Fayyad, G. Shapiro Piatetsky and P. Smyth, The KDD Process for Extracting Useful Knowledge from Volumes of Data Communication of the ACM, vol. 39, no 11, p. 27-34, November, 1996.

[2] K. M. Decker and S Focardi, Technology Overview: A Report on Data Mining, CSCS-ETH, Swiss Scientific Computer Center, 1995.

[3] S. O. Rezende, R. B. T. Oliveira, L. C. M. Felix and C. A. J. Rocha, "Visualization for Knowledge Discovery in Database", Proc. of Intl. Conf. on DataMining, September 1998, pp. 81-95.

[4] W. Frawley, G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine, fall 1992, pp. 213-228.

[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann, August 2000.

[6] A. K. Jain and R. C. Dubes "Algorithms for Clustering Data" Prentice Hall, Upper Saddle River: New Jersey, 1988.

[7] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", Grouping Multidimensional Data - Recent Advances in Clustering, pp.25-71, 2006.

[8] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[9] L. Kaufman and P.J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.

[10] J. A. Garcia, J. Fdez-Valdivia, F. J. Cortijo, and R. Molina, "A Dynamic Approach for Clustering Data", Signal Processing, Vol. 44, No. 2, pp. 181-196, 1994.

[11] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Ecient Data Clustering Method for Very Large Databases", Proceedings of ACM SIGMOD, Montreal Canada, pp. 103-114, June 1996.

[12] S. Guha, R. Rastogi, and K. Shim, "Cure: an efficient clustering algorithm for large databases," in ACM SIGMOD International Conference on the Management of Data, (Seattle, WA, USA), pp. 73–84, 1998.

[13] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes", Information Systems, Vol. 25, No. 5, pp. 345–366, 2000.

[14] J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles. Addison-Wesley, 1974.

[15] A. Dempster, N. Laird, and D. Rubin, A Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Royal Statistical Soc., Ser. B, vol. 39, no. 1, pp. 1-38, 1977.

[16] P. Bradley, U. M. Fayyad and C. Reina, "Scaling clustering algorithms to large databases" in Proceedings of 4th International Conference on Knowledge Discovery and Data Mining (California), pp. 9-15, AAAI 1998.

[17] L. Kaufman and P.J. Rousueeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[18] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining", in Proceedings of the Twentieth International Conference on Very Large Databases (VLDB'94) (Santiago, Chile), pp. 144-155, Chile, September 1994.

[19] W. Wei, J. Yang, and R. Muntz, "Sting: A statistical infromation grid approach to spatial data mining" in Proceeding of the VLDB, Athens, Greece, 1997.

[20] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", in Proceeding. of the ACM SIGMOD, pp. 94-105, 1999.

[21] 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 the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland: Oregon, pp. 226-231.

[22] A. Hinneburg, and D. A. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", in Proceeding of International Conference on Knowledge Discovery and Data Mining (KDD98), pp. 58-65, August 1998.

[23] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications", Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Vol. 2, No. 2, 1998.

[24] M. Ankerst, M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure", In Proceeding ACM SIGMOD, International Conference on Management of Data (SIGMOD'99), Philadelphia, PA, pp. 49-60, 1999.