Data Mining Process Using Clustering: A Survey

Mohamad Saraee
Department of Electrical and Computer Engineering
Isfahan University of Techno1ogy, Isfahan , 84156-83111
saraee@cc.iut.ac.ir

Najmeh Ahmadian
Department of Electrical and Computer Engineering
Isfahan University of Techno1ogy, Isfahan , 84156-83111
Najmeh_2020@yahoo.com

Zahra Narimani
Department of Electrical and Computer Engineering
Isfahan University of Techno1ogy, Isfahan , 84156-83111
z.narimani@gmail.com

Abstract

Clustering is a basic and useful method in understanding and exploring a data set. Clustering is division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Interest in clustering has increased recently in new areas of applications including data mining, bioinformatics, web mining, text mining, image analysis and so on. This survey focuses on clustering in data mining.

The goal of this survey is to provide a review of different clustering algorithms in data mining. A Categorization of clustering algorithms has been provided closely followed by this survey. The basics of Hierarchical Clustering include Linkage Metrics, Hierarchical Clusters of Arbitrary and Binary Divisive Partitioning is discussed at first. Next discussion is Algorithms of the Partitioning Relocation Clustering include Probabilistic Clustering, K-Medoids Methods, K-Means Methods. Density-Based-Partitioning, Grid-Based Methods and Co-Occurrence of Categorical Data are other sections. Their comparisons are mostly based on some specific applications and under certain conditions. So the results may become quite different if the conditions change.

KeyWords: clustering, partitioning, unsupervised learning, hierarchical clustering

1. INTRODUCTION

We are living in a world full of data. One of the means in dealing with these data is to group them into a set of clusters.

Most researchers consider a cluster by the internal homogeneity and the external separation [1], [2], i.e., patterns in the same cluster should be similar to each other, while patterns in different clusters should not. Both the similarity and the dissimilarity should be examinable in a clear way.

The problem of clustering has interested for several decades, with surveys [3], [4] ,[5] and papers (X-means [6], Gmeans [7], CLARANS [8], CURE [9], CLIQUE [10], BIRCH [11], DBSCAN [12]). The hard part of clustering is recognizing a good group of

clusters and data points to label as outliers and thus ignore from clustering.

Categorization of clustering algorithms isn't straightforward and groups below overlap. We provide a classification closely followed by this survey. The basics of hierarchical clustering Hierarchical Clusters of Arbitrary and Binary Divisive Partitioning are presented in section 2 and subsection. Algorithms of the Partitioning Relocation Clustering include Probabilistic Clustering, K-Medoids Methods, K-Means Methods are surveyed in the section 3 .Partitioning algorithms of the second type are surveyed in the section Density-Based-Partitioning. (see section 4) Density-Based Connectivity and Density Functions are two subsections of this section. Some algorithms first separate the clustering space into a finite number of cells (segments) and then perform the required operations on the quantized space. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters. We discuss them in the section Grid-Based Methods. (Section 5) Categorical data is connected with transactional databases. The concept of a similarity alone is not sufficient for clustering such data. The idea of categorical data co-occurrence comes to rescue. The algorithms are surveyed in the section 6.

2. Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [13]. An agglomerative clustering starts with one point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of 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.

To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such derived proximity measure is called a linkage metric.

Major inter-cluster linkage metrics [14] include single link, average link, and complete link. Linkage metrics-based hierarchical clustering suffers from time complexity.

COBWEB is the popular hierarchical clustering algorithm for categorical data. It has two very important qualities. "First, it utilizes incremental learning. Instead of following divisive or agglomerative approaches, it dynamically builds a dendrogram by processing one data point at a time. Second, COBWEB belongs to conceptual or model-based learning. This means that each cluster is considered as a model that can be described intrinsically, rather than as a collection of points assigned to it. COBWEB.s dendrogram is called a classification tree"[15]. Tree is potentially updated (by an insert /split /merge/create operation) and decisions are based on an analysis of a category utility [16]

Chiu [17] presented another conceptual or model-based approach to hierarchical clustering that contains several different useful features, such as the extension of BIRCH-like preprocessing to categorical attributes, outliers handling, and a two-step strategy for monitoring the number of clusters. It also involves both numerical and categorical attributes and constitutes a blend of Gaussian and multinomial models.

2.1. Hierarchical Clusters of Arbitrary Shapes

The hierarchical agglomerative clustering algorithm CURE (Clustering Using REpresentatives) was introduced by Guha[18]. CURE will find clusters of different shapes and sizes, and it is insensitive to outliers. It also uses two stages

.The first one is data sampling. Second is data partitioning in p partitions. Therefore fine granularity clusters are constructed in partitions first. A characteristic of CURE is that it represents a cluster by a fixed number c of points scattered around it. The distance between two clusters used in the agglomerative process is measured by the minimum of distances between two scattered representatives. Single and average link closeness is replaced by representatives.

Selecting representatives scattered around a cluster makes it possible to cover non-spherical shapes. As before, agglomeration continues until requested number k of clusters is achieved.

2.2. Binary Divisive Partitioning

In information retrieval, and document clustering applications binary taxonomies are very important. Linear algebra methods, based on *singular value decomposition* (SVD) are used for this purpose [19] and SVD application resulted in the PDDP algorithm (Principal Direction Divisive Partitioning) [20].

This algorithm cut in half data in Euclidean space by a hyperplane that passes through data centroid with the largest singular value. The dividing hyperplane is orthogonal to a line connecting two centroids. The k-way splitting is also possible if the k largest singular values are considered. This way results in a binary tree and is safe for categorize documents. Hierarchical divisive bisecting k-means was proven [21] to be preferable for document clustering. The problem in PDDP or 2-means is which cluster must split. Strategies are: (1) split each node at a given level, (2) split the cluster with highest cardinality, and, (3) split the cluster with the largest intra-cluster variance. All three strategies have problems. For more information about it, see [24].

3. Partitioning Relocation Clustering

In this section we survey data partitioning algorithms, which divide data into several subsets.

Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate data, this results in high quality clusters.

3.1. Probabilistic Clustering

One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found. More specifically, probabilistic models assume that the data comes from a mixture of several populations whose distributions and priors we want to find. In the *probabilistic approach*, data is considered to be a sample independently drawn from a mixture model of several probability distributions [22]. we associate the cluster with the corresponding distribution.s parameters such as mean, variance, etc. Each data point carries not only its (observable) attributes, but also a (hidden) cluster ID (class in pattern recognition). Each point x is assumed to belong to one and only one cluster. Probabilistic Clustering can be modified to handle recodes of complex structure and it can be stopped and resumed with sequence of data. It also results in easily interpretable cluster system.

3.2. K-Medoids Methods

In k-medoids methods a cluster is represented by one of its points. “When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid”[15].

The algorithm PAM (Partitioning Around Medoids) and the algorithm CLARA (Clustering LARge Applications) are two early versions of k-medoid methods [23]. CLARANS (Clustering Large Applications based upon RANdomized Search) is further development in spatial databases clustering algorithms. [24]

k-medoids has two advantages. It covers any attribute types and since peripheral cluster points do not affect them, it is lesser sensitive to outliers.

3.3. K-Means Methods

The k-means algorithm [25] is the most popular clustering tool used in scientific and industrial applications. "The name comes from representing each of k clusters C_j by the mean (or weighted average) c_j of its points, the so-called centroid"[15].

The algorithm X-means [26] speeds up the iterative process. It searches for the best k in the process itself. "X-means tries to split a part of already constructed cluster based on outcome of BIC criterion (Bayesian Information Criterion) "[27]. This way gives a much better initial guess for the next iteration.

It has the good geometric and statistical sense for numerical attributes and doesn't work well with categorical attributes and can be negatively affected by a single outlier.

4.Density-Based Partitioning

An open set in the Euclidean space can be divided into a set of its connected components. A cluster, introduced as a connected dense component, grows in any direction that density leads. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes. Also this provides a natural defense against outliers.

Spatial data clustering is used for metric space [28]. There are two major approaches for density-based methods. The first approach holds density to a training data point and is reviewed in the sub-section Density-Based Connectivity. The second approach holds density to a point in the attribute space and is explained in the sub-section Density Functions. It includes the algorithm DENCLUE.

4.1. Density-Based Connectivity

The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise)[29] targeting low-dimensional spatial data is the major representative in this category. Two input parameters ε and MinPts are used to define:

1) An ε-neighborhood $N_{\varepsilon}(x) = \{y \in X \mid d(x, y) \leq \varepsilon\}$ of the point x

2) A core object (a point with a neighborhood consisting of more than MinPtspoints)

3) A concept of a point y density-reachable from a core object x (a finite sequence of core objects between x and y exists such that each next belongs to an ε-neighborhood of its predecessor)

4) A density-connectivity of two points x, y (they should be density-reachable from a common core object).

So defined density-connectivity is a symmetric relation and all the points reachable from core objects can be considered as maximal connected components presenting as clusters. The points that are not connected to any core point are outliers (they are not wrapped by any cluster). The non-core points inside a cluster represent its boundary and core objects are internal points. There are any limitations on the dimension or attribute types because processing is out of data ordering.

One problem is by considering two parameters ε and MinPts, there is no straightforward way to fit them to data.

Other representative algorithms are GDBSCAN [30], OPTICS [31], and DBCLASD [32].

4.2. Density Functions

Hinneburg & Keim [33] compute density functions defined over the underlying attribute space instead of computing densities pinned to data points. They introduced the algorithm DENCLUE (DENsity-based CLUstEring). It has a firm mathematical foundation Along with DBCLASD that uses a density function. DENCLUE focus on local maxima of density functions called density-attractors and uses a hill-climbing technique for finding them. It finds center-defined clusters and arbitrary-shape clusters that are defined as continuations along sequences of points whose local densities are more than threshold ξ.Also the algorithm can be considered as a grid-based method and it applied in high dimensional multimedia and molecular biology data.

5. Grid-Based Methods

In the previous section vital concepts of density, connectivity, and boundary were described. Another concept of them is to inherit the topology from the underlying attribute space. To limit the search combinations, multi-rectangular segments are considered. Since some binning is for numerical attributes, methods partitioning space are frequently called grid-based methods.

Our attention moved from data to space partitioning. Data partitioning is induced by points' membership in segments resulted from space partitioning, while space partitioning is based on grid-characteristics accumulated from input data. Grid-based clustering techniques are independent of data ordering. In contrast, relocation methods and all incremental algorithms are very sensitive to data ordering. While density-based partitioning methods work best with numerical attributes, grid-based methods work with attributes of different types.

BANG-clustering [34] improves the similar hierarchical algorithm GRIDCLUST [35]. Grid-based segments summarize data. The segments are stored in a special BANG-structure that is a grid-directory integrating different scales. Adjacent segments are neighbors. Nearest neighbors is a common face has maximum dimension. The density of a segment is a ratio between number of points in it and its volume. From the grid-directory, a dendrogram is directly calculated.

"The algorithm WaveCluster works with numerical attributes and has an advanced multi-resolution"[36]. WaveCluster is based on ideas of signal processing. It applies wavelet transforms to filter the data. It has also High quality of clusters, Ability to work well in relatively high dimensional spatial data, and successful handling of outliers.

6. Co-Occurrence of Categorical Data

Categorical data frequently relates to the concept of a variable size transaction that is a finite set of elements called items from a common item universe. For example, market basket data is this form. Every transaction can be presented in a point-by-attribute format, by enumerating all items j. traditional clustering methods, based on similarity measures, do not work well. Since categorical/transactional data is important in customer profiling, assortment planning, web analysis, and other applications, different clustering methods founded on the idea of co-occurrence of categorical data have been developed. The algorithm ROCK (Robust Clustering algorithm for Categorical Data) [19] has many common aspects with the algorithm CURE (section Hierarchical Clustering): (1) it is a hierarchical clustering, (2) agglomeration continues until specified number k of clusters is constructed, and (3) it uses data sampling in the same way as CURE does.

The algorithm SNN (Shared Nearest Neighbors) [37] is based on a density approach with the idea of ROCK. SNN uses similarity matrix by only keeping K-nearest neighbors and resulting is in complexity $O(N^2)$.

TABLE 1

Cluster algorithm	Complexity	Capability of tackling high dimensional data
K-means	$O(NKd)$ (time) $O(N+K)$ (space)	No
Fuzzy *c*-means	Near $O(N)$	No
Hierarchical clustering*	$O(N^2)$ (time) $O(N^2)$ (space)	No
CLARA	$O(K(40+K)^2+K(N-K))^+$ (time)	No
CLARANS	Quadratic in total performance	No
BIRCH	$O(N)$ (time)	No
DBSCAN	$O(N \log N)$ (time)	No
CURE	$O(N_{sample}^2 \log N_{sample})$ (time) $O(N_{sample})$ (space)	Yes
WaveCluster	$O(N)$ (time)	No
DENCLUE	$O(N \log N)$ (time)	Yes
FC	$O(N)$ (time)	Yes
CLIQUE	Linear with the number of objects, Quadratic with the number of dimensions	Yes
OptiGrid	Between $O(Nd)$ and $O(Nd \log N)$	Yes
ORCLUS	$O(K_0^3+K_0Nd+K_0^2d^3)$ (time) $O(K_0d^2)$ (space)	Yes

COMPUTATIONAL COMPLEXITY OF CLUSTERING ALGORITHMS AND CAPABILITY OF TACKLING

7. CONCLUSION &FUTURE WORK

Clustering is one of the most fundamental and essential data analysis techniques. Clustering can be used as an independent data mining task to distinguish intrinsic characteristics of data or as a preprocessing step with the clustering results then used for classification, correlation analysis, or anomaly detection. These clustering algorithms (summarizing the computational complexity of some typical and classical clustering algorithms and capability of tackling high dimensional as the most important attribute are in Table1 that get from [38]) are effective in determining a good clustering if the clusters are of convex shape, similar size and density, and if the number of clusters can be reasonably estimated. In general, the disability to identify the appropriate number of clusters is one of the most fundamental shortcomings of non-hierarchical techniques.

At the preprocessing and post-processing phase, feature selection/extraction (as well as standardization and normalization) and cluster validation are as important as the clustering algorithms.

Usually, algorithms are designed with certain assumptions and favor some type of biases. So it is not accurate to say “best” in the context of clustering algorithms, although some comparisons are possible. These comparisons are mostly based on some specific applications, under certain conditions, and the results may become quite different if the conditions change.

Hierarchical clustering algorithms represent the data set into several levels of partitioning which are usually represented by a dendrogram – a tree which splits the data set recursively into smaller subsets.- Although hierarchical clustering algorithms can be very effective in knowledge discovery, the cost of creating the dendrograms is prohibitively expensive for large data sets.

Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate data, this results in high quality clusters.

Density-Based Partitioning can be divided an open set into a set of its connected components but these algorithms have own complexity.

Generally, grid-based clustering algorithms first separate the clustering space into a finite number of cells (segments) and then perform the required operations on the quantized space. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters. A solution for better results could be instead of integrating all the requirements into a single algorithm, to try to build a combination of clustering algorithms. However, the theoretical foundation of combining multiple clustering

algorithms is still in its early stages and thus more work is needed in this direction.

In addition, studding on the impact of the coordinated sub-sampling strategies on the performance and quality of object distributed clustering needed more work. The question is to determine what types of overlap and object ownership structures lend themselves particularly well for knowledge reuse.

REFERENCES

[1] P. Hansen and B. Jaumard, “Cluster analysis and mathematical programming,” Math. Program., vol. 79, pp. 191–215, 1997.

[2] Bing Liu, Yuliang Shi, Zhihui Wang, Wei Wang, Baile Shi: Dynamic Incremental Data Summarization for Hierarchical Clustering. Electronic Edition (link) BibTeX.2006

[3] Lai, Ying Orlandic, Ratko Yee, Wai Gen Kulkarni, Sachin Scalable "Clustering for Large High-Dimensional Data Based on Data Summarization Computer Science", Illinois Institute of Technology, Chicago,IL60616,U.S. 2007

[4] GUHA, S., RASTOGI, R., and SHIM, K.. "CURE: An efficient clustering algorithm for large databases". In Proceedings of the ACM SIGMOD Conference, 73-84, Seattle, WA. 1998

[5] F. Murtagh. A survey of recent advances inhierarchical clustering algorithms. The Computer Journal, 26(4):354–359, 1983.

[6] D. Pelleg and A. Moore. "X-means: Extending K-means with efficient estimation of the number of clusters". In Proceedings of the Seventeenth International Conference on Machine Learning (ICML), pages 727–734, 2000.

[7] G. Hamerly and C. Elkan. Learning the k in k-means. In Proceedings of NIPS, 2003.

[8] R. T. Ng and J. Han." Efficient and effective clustering methods for spatial data mining". In Proc. of VLDB Conference., pages 144–155, 1994.

[9] S. Guha, R. Rastogi, and K. Shim. CURE:An efficient clustering algorithm for large databases. In SIGMOD Conference, pages 73–84, 1998.

[10] I. Jolliffe. "Principal Component Analysis". Springer Verlag, 1986.

[11] T. Zhang, R. Ramakrishnan, and M. Livny. "BIRCH: An efficient data clustering method for very large databases". In SIGMOD Conference, pages 103–114, 1996.

[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 KDD Conference, 1996.

[13] JAIN, A. and DUBES. "Algorithms for Clustering Data." Prentice-Hall, Englewood Cliffs, NJ. 1988.

[14] OLSON, C. "Parallel algorithms for hierarchical clustering." Parallel Computing, 21, 1313-1325. 1995

[15] Pavel Berkhin, "Survey of Clustering Data Mining Techniques",Accrue Software, Inc.2002

[16] CORTER, J. and GLUCK, "Explaining basic categories: feature predictability and information." Psychological Bulletin, 111, 291-303. M. 1992.

[17] CHIU, T., FANG, D., CHEN, J., and Wang, Y.. “A Robust and scalable clustering algorithm for mixed type attributes in large database environments”. In Proceedings of the 7th ACM SIGKDD, 263-268, San Francisco, CA. 2001

[18] GUHA, S., RASTOGI, R., and SHIM, K. ROCK" A robust clustering algorithm for categorical attributes". In Proceedings of the 15th ICDE, 512-521, Sydney, Australia. 1999

[19] BERRY, M.W. and BROWNE, "Understanding Search Engines: Mathematical Modeling and Text Retrieval." M.1999

[20] BOLEY, D.L." Principal direction divisive partitioning". 1998

[21] STEINBACH, M., KARYPIS, G., and KUMAR. "A comparison of document clustering techniques". 6th ACM

SIGKDD, World Text Mining Conference, Boston, MA. V. 2000

[22] MCLACHLAN, G. and BASFORD, "Mixture Models: Inference and Applications to Clustering." Marcel Dekker, New York, NY. K. 1988.

[23] KAUFMAN, L. and ROUSSEEUW,. "Finding Groups in Data: An Introduction toCluster Analysis". John Wiley and Sons, New York, NY. P. 1990

[24] NG, R. and HAN," Efficient and effective clustering methods for spatial data mining". In Proceedings of the 20th Conference on VLDB, 144-155, Santiago, Chile. J. 1994

[25] HARTIGAN,. "Clustering Algorithms". John Wiley & Sons, New York, NY. J. 1975

[26] PELLEG, D. and MOORE, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". In Proceedings 17th ICML, Stanford University. A. 2000.

[27] FRALEY, C. and RAFTERY, "A. How many clusters?. Which clustering method? Answers via model-based cluster analysis". The Computer Journal, 41, 8, 578-588. 1998

[28] HAN, J. and KAMBER, "Data Mining. Morgan Kaufmann Publishers." M. 2001.

[29] ESTER, M., KRIEGEL, H-P., SANDER, J. and XU," A density-based algorithm for discovering clusters in large spatial databases with noise". In Proceedings of the 2nd ACM SIGKDD, 226-231, Portland, Oregon. X. 1996

[30] SANDER, J., ESTER, M., KRIEGEL, H.-P., and XU, X. "Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. In Data Mining and Knowledge Discovery", 1998 2, 2, 169-194.

[31] ANKERST, M., BREUNIG, M., KRIEGEL, H.-P., and SANDER, J.. "OPTICS: Ordering points to identify clustering structure". In Proceedings of the ACM SIGMOD Conference, 49-60, Philadelphia, PA. 1999

[32] XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J. "A distribution-based clustering algorithm for mining large spatial datasets". In Proceedings of the 14th ICDE,324-331, Orlando, FL. 1998.

[33] HINNEBURG, A. and KEIM,." An efficient approach to clustering large multimedia databases with noise". In Proceedings of the 4th ACM SIGKDD, 58-65, New York, NY. D. 1998

[34] SCHIKUTA, E., ERHART, "The BANG-clustering system: grid-based data analysis". In Proceeding of Advances in Intelligent Data Analysis, Reasoning about Data, 2nd International Symposium, 513-524, London, UK. M. 1997.

[35] SCHIKUTA, "Grid-clustering: a fast hierarchical clustering method for very large ". E. 1996.data sets. In Proceedings 13th International Conference on Pattern Recognition, 2, 101-105.

[36] SHEIKHOLESLAMI, G. , CHATTERJEE, S., and ZHANG,. WaveCluster: "A multiresolution clustering approach for very large spatial databases". In Proceedings of the 24th Conference on VLDB, 428-439, New York, NY. A. 1998

[37] ERTOZ, L., STEINBACH, M., and KUMAR, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, Technical Report". V. 2002

[38] Rui Xu, *"Survey of Clustering Algorithms", VOL. 16, NO. 3, MAY 2005*