International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 02 | February - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

A Survey on Multi objective Clustering Classification Algorithms

Nasir Khan[1], Ritesh Kumar Yadav2

1 M.TECH Scholar, SRK University, Bhopal

2 Associate Professor, SRK University, Bhopal

E Mail Nasirkhan.ind@gmail.com , er.ritesh1987@gmail.com

ABSTRACT

Clustering is a popular data mining technique which can be applied to a given dataset to identify the data objects that belong to a single class, such that data objects in different clusters are distinct while similarity exists for data objects belonging to the same cluster. Usually, clustering techniques are based on optimizing single objective function criteria, which may not be capable of performing well in many real time scenarios. Motivated by this many multi-objective based optimization techniques are discussed in this paper. Multi-objective based optimization techniques are capable of optimizing several conflicting objective functions simultaneously. Under this context, evolutionary based approach and simulated annealing based techniques are adopted in various MOO techniques and proven well in case of noise, non-spherical and high dimensional feature space. The paper further discusses various validity measures to evaluate the goodness of clustering techniques.

Keywords: Clustering, K-means, Intra-cluster homogeneity, Inter-cluster separability.

Introduction

Clustering is a 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. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification.

It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods surveyed below.

For example, a company those sales a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Clustering techniques fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. There is no target variable to be predicted, thus no distinction is being made between independent and dependent variables.

Clustering techniques are used for combining observed examples into clusters (groups) which satisfy two main criteria:
1. Each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
2. Each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters. Depending on the clustering technique, clusters can be expressed in different ways:
1. Identified clusters may be exclusive, so that any example belongs to only one cluster.
2. They may be overlapping; an example may belong to several clusters.
3. They may be probabilistic, whereby an example belongs to each cluster with a certain probability.

II Clustering Methods

In this section we describe the most well-known clustering algorithms. The main reason for having many clustering methods is the fact that the notion of "cluster" is not precisely defined [1]. Consequently many clustering methods have been developed, each of which uses a different induction principle. [2] Suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. [3] suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods. An alternative categorization based on the induction principle of the various clustering methods is presented in [1].

A. Hierarchical Methods

These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These method scan be sub-divided as following:
• Agglomerative hierarchical clustering- Each object initially represents a cluster of its own.Then clusters are successively merged until the desired cluster structure is obtained.
• Divisive hierarchical clustering — All objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained.

The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.
The merging or division of clusters is performed according to some similarity measure, chosen so as to optimize some criterion (such as a sum of squares). The hierarchical clustering methods could be further divided according to the manner that the similarity measure is calculated [4]:
• Single-link clustering (also called the connectedness, the minimum method or then earest neighbor method)—methods that consider the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, the similarity between a pair of clusters is considered to be equal to the greatest similarity from any member of one cluster to any member of the other cluster [5]
• Complete-link clustering (also called the diameter, the maximum method or the furthest neighbor method) - methods that consider the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster [6].
• Average-link clustering (also called minimum variance method) - methods that consider the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster. Such clustering algorithms may be found in [7].
The disadvantages of the single-link clustering and the average-link clustering can be summarized as follows [8].
• Single-link clustering has a drawback known as the "chaining effect":A few points that form a bridge between two clusters cause the single-link clustering to unify these two clusters into one.
• Average-link clustering may cause elongated clusters to split and for portions of neighboring elongated clusters to merge.
The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single-link methods are more versatile.

Generally, hierarchical methods are characterized with the following strengths:

• Versatility—The single-link methods, for example, maintain good performance on data sets containing non-isotropic clusters, including well-separated, chainlike and concentric clusters.

• Multiple partitions — hierarchical methods produce not one partition, but multiple nested partitions, which allow different users to choose different partitions, according to the desired similarity level. The hierarchical partition is presented using the dendrogram.

The main disadvantages of the hierarchical methods are:

• Inability to scale well — The time complexity of hierarchical algorithms is atleast O(m2)(where m is the total number of instances), which is non-linear with the number of objects. Clustering a large number of objects using a hierarchical algorithm is also characterized by huge I/O costs.

• Hierarchical methods can never undo what was done previously. Namely there isno back-tracking capability.

B. PartitioningMethods

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. To achieve global optimality in partitioned based clustering, an exhaustive enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization. Namely, a relocation method iteratively relocates points between the k clusters. The following subsections present various types of partitioning methods.

Error Minimization Algorithms

These algorithms, which tend to work well with isolated and compact clusters, are the most intuitive and frequently used methods. The basic idea is to find a clustering structure that minimizes a certain error criterion which measures the “distance” of each instance to its representative value. The most well-known criterion is the Sum of Squared Error (SSE), which measures the total squared Euclidian distance of instances to their representative values. SSE may be globally optimized by exhaustively enumerating all partitions, which is very time-consuming, or by giving an approximate solution (not necessarily leading to a global minimum)using heuristics.The latter option is the most common alternative.

III Related Work

A number of different types of clustering techniques are used based on the type of given data objects. A.K. Jain et. al [9] has discussed the various types of clustering techniques such as; Partitional Clustering, Hierarchical Clustering, Density-based Clustering and Grid-based Clustering techniques. Partitional Clustering technique decomposes data set directly into an integer number of disjoint cluster sets.Hierarchical Clustering algorithm finds successive cluster using the previously established clusters. These algorithms can be either agglomerative or divisive type. The key point of density based clustering algorithms is to create clusters based on density functions. Thus it is able to deal with non-spherical data objects. DBSCAN is one such density based clustering algorithm. While, Grid based Clustering technique is used in spatial data mining.

Ramandeep Kaur and Gurjit Singh Bhathaln [10] have described various clustering techniques. The authors have compared K-mean algorithm and FuzzyC-meanalgorithms.FuzzyC-meanprovides superior results to K-mean because of membership property and also gives better flexibility than K-mean.

The paper[11] shows that evaluating the quality of clustering results is essential before absorbing them in real world problems. The authors have discussed various internal clustering validation measures. Further, the goodness of clusters is evaluated in different important aspects such as, presence of outliers or noise, high density or low density of data points, monotonic or lack of variety in features, presence of sub-clusters and skewed distributions. As a result of experiment, S_Dbwhas proven

superior to other validity measures in all different aspects. The other validity measures show lower performance mainly in presence of outliers and sub-clusters.

Ferenc Kovacs et al [12] have discussed the various techniques for evaluating the significance of clustering solutions- External criteria, Internal criteria and relative criteria. External criteria are evaluated based on external information like existing class labels. Internal criteria is based on intrinsic property of data set such as, the best solution is selected with maximum value of dissimilarity measure and minimum value of similarity measure. Further, relative criteria is based on the comparison of different clustering schema. There are large variety of features, the input parameters which can change the behavior of the same algorithm for such parameters and thus the clustering results may vary for the same data sets.

Erendira Rendon et. al.[13] have defined that the evaluation of clustering result is one of the major challenges in machine learning. Validity indexes are used to evaluate the quality of clustering algorithms. Depending on the availability of external information, the validity indexes varies for the evaluation in different types of problems. The paper compares the two main indexes-external and internal indexes. Here the data sets are grouped using K-means and Bissection K- means clustering techniques. Here, internal indexes show superior performance to external indexes.
Sriparna Saha et. al.[14] have proposed a multi-objective clustering technique - GenClustMOO algorithm, that is capable to handle non-spherical and overlapped data points and able to partition them appropriately. The three objective functions based on total compactness of the partitioning, the other reflecting the total symmetry of the clusters and the last reflecting the connectedness, are optimized here. AMOSA algorithm outputs a set of non dominating solutions; then an approach to select the best solution is also discussed.

Sanghamitra Bandyoupadhyay et. al.[15] have discussed various multi-objective optimization techniques in addition to several single objective based approach. Numerical methods (e.g. hill climbing) belong to calculus based techniques. Simulated annealing belongs to single point search, Genetic algorithm follows multiple point searchtechnique.Simulated annealing and Genetic algorithmare the most popular approaches adopted in multiple Multi-objective optimization techniques.MOGA,NSGA-II,PAES,AMOSA
are some of the popular multi-objective optimization techniques.

Ujjwal Maulik et. al. [16] have proposed genetic algorithm based clustering technique called GA-clustering algorithm. Here, genetic algorithm has been used to search for cluster centre which minimize the clustering metric that is given by the sum of the absolute Euclidean distance of each point from their respective cluster centers. Further, the proposed GA-clustering algorithm is compared to K-means algorithm for four artificial and three real-life data sets with the number of features ranging from two to ten and number of clusters ranging from two to nine. The results show that GA-clustering algorithm provides superior performance over K-means algorithm.

In several situations, data points have belongingness to more than one cluster which may lead to overlapped clustering.Two-stage clustering algorithm referred to SiMM-TS [17] is proposed which utilizes the concept of points having significant multi-class membership(SiMM).The results are compared with average linkage method and SOM algorithm with respect to three real life gene data sets. The proposed approach has been proven superior after evaluating the results using severalvalidityindices-SilhouetteindexandRand
index.

Ujjwal Maulik et.al. [18] have addressed a multi-objective optimization problem associated with fuzzy clustering approach to identify the co-expressed genes. Here, the pareto-optimal front is obtained after optimizing various fuzzy internal validity indices. The clustering information from allsuchfinalsolutionsarefurthercombinedby the proposed technique known as, novel fuzzymajority voting approach. This basically utilizes SVM to identify the training examples that comprises all

such genes having high membership degree for a particular cluster. These training examples further classify the remaining genes.

IV Conclusions

Clustering lies at the heart of data analysis anddata mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change. At the same time, it is notable that any clustering "is a division of the objects into groups based on a set of rules – it is neither true nor false"

References

[1] Estivill-Castro, V. and Yang, J. A Fast and robust general purpose clustering algorithm. Pacific Rim International Conference on Artificial Intelligence, pp. 208-218, 2000.

[2] Fraley C. and Raftery A.E., "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", Technical Report No. 329. Department of Statistics University of Washington, 1998.

[3] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[4] Jain, A.K. Murty, M.N. and Flynn, P.J. Data Clustering: A Survey. ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[5] Sneath, P., and Sokal, R. Numerical Taxonomy.W.H.FreemanCo.,SanFrancisco,CA, 1973.

[6] King, B. Step-wise Clustering Procedures, J. Am. Stat. Assoc. 69, pp. 86-101, 1967.

[7] Murtagh, F. A survey of recent advances in hierarchical clustering algorithms which usecluster centers. Comput. J. 26 354-359, 1984.

[8] Guha, S., Rastogi, R. and Shim, K. CURE: An efficient clustering algorithm for large databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 73-84, New York, 1998.

[9] A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, ACM Computing Surveys, Vol. 31, No. 3 (1999) 264-323.

[10] Ramandeep Kaur, Gurjit Singh Bhathal, A Survey of Clustering Techniques, IJARCSSE, Volume 3, Issue 5 (2013) 153-157.

[11] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao1, Junjie Wu, Understanding of Internal Clustering Validation Measures, IEEE International Conference on Data Mining (2010) 911-916.

[12] Ferenc Kovács, Csaba Legány, Attila Babos, Cluster Validity Measurement Techniques source: https://pdfs.semanticscholar.org/c4f9/df3c66105382d05e58ec35faa8d435f55c91.pdf.

[13] Erendira Rendón, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz, Internal versus External cluster validation indexes, International Journal of Computers and communications, Issue 1, vol-5 (2011) 27-34.