We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

6,500
Open access books available

177,000
International authors and editors

195M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Chapter

Introductory Chapter: Development of Data Clustering

Niansheng Tang and Ying Wu

1. Introduction

Data clustering is a popular method in statistics and machine learning and is widely used to make decisions and predictions in various fields such as life science (e.g., biology, botany, zoology), medical sciences (e.g., psychiatry, pathology), behavioral and social sciences (e.g., psychology, sociology, education), earth sciences (e.g., geology, geography), engineering sciences (e.g., pattern recognition, artificial intelligence, cybernetics, electrical engineering), and information and decision sciences (e.g., information retrieval, political science, economics, marketing research, operational research) [1]. Clustering analysis aims to group individuals into a number of classes or clusters using some measure such that the individuals within classes or clusters are similar in some characteristics, and the individuals in different classes or clusters are quite distinct in some features.

2. Measures of similarity or dissimilarity

There are a lot of measures of similarity or dissimilarity for data clustering. Generally, assessing the similarity of individuals in terms of the number of characteristics, which can be regarded as the points in space (e.g., a plane, the surface of a sphere, three-dimensional space, or higher-dimensional space) directly relates to the concept of distance from a geometrical viewpoint [1]. The widely used measures include Euclidean distance, Manhattan distance (also called city-block distance), and Mahalanobis distance for measuring the similarity of two data points. Euclidean distance depends on the rectangular coordinate system, Manhattan distance depends on the rotation of the coordinate system, but Euclidean and Manhattan distances do not consider the correlation between data variables and data dimensions. Mahalanobis distance can be regarded as a correction of Euclidean distance, the dependence of the data points is described by covariance matrix, which can be used to deal with the problem of non-independent and identically distributed data. In addition, there are other distances such as chebychev distance, power distance, and sup distance.

In many applications, different types of data are related to different distances. For example, the simple matching distance is used to measure the similarity of two categorical data points; a general similarity coefficient is adopted to measure the distance of two mixed-type data points or the means of two clusters; probabilistic model, landmark models, and time series transformation distance are used to measure the similarity of two time-series data points. In particular, Wu et al. [2] considered
Data Clustering


Although the aforementioned similarity and dissimilarity measures can be applied to various types of data, other types of similarity and dissimilarity measures such as cosine similarity measure and a link-based similarity measure have also been developed for specific types of data. Also, one may require computing the distance between an individual and a cluster, and the distance between two clusters based on various central data points or representative data points. In these cases, the widely used distances include the mean-based distance, the nearest neighbor distance, the farthest neighbor distance, the average neighbor distance, which are extensions of data point distances. Particularly, the Lance-Williams formula can be used to compute the distances between the old clusters and a new cluster formed by two clusters. Again, to assess the similarity among probability density distributions of random variables, one can use Kullback-Leibler (K-L) distance (relative entropy) and Wasserstein distance. K-L distance does not satisfy three properties of the distance and is asymmetric, while Wasserstein distance possesses three properties of the distance and is symmetric. More importantly, Wasserstein distance can be used to deal with the mixture of discrete and continuous data. To this end, data clustering based on the Wasserstein distance has received a lot of attention over the past years. For example, see [6–9] for dynamic clustering of interval data, complex data clustering, variational clustering, geometric clustering, respectively.

3. Data clustering algorithms

Many useful data clustering algorithms have been developed to cluster individuals into different clusters over the past years. For example, hierarchical clustering algorithm, partitioning algorithm, fuzzy clustering algorithm [10], center-based clustering algorithm, search-based clustering algorithm, graph-based clustering algorithm, grid-based clustering algorithm, density-based clustering algorithm, model-based clustering algorithm, and subspace clustering [11]. Hierarchical clustering algorithm, which divides individuals into a sequence of nested partitions has two key algorithms: agglomerative algorithm and divisive algorithm, and partitioning algorithms are two important clustering algorithms. Fuzzy clustering algorithm allows an individual to belong to two or more clusters with different probabilities, has three major algorithms: fuzzy k-means, fuzzy k-modes, and c-means. Center-based clustering algorithm is more used to cluster large scales and high-dimensional data sets has two major algorithms: k-means and k-modes in which k-means is the most widely used clustering algorithm, and is a non-hierarchical clustering method. Search-based clustering algorithm is usually used to find the globally optimal clustering for fitting the data set in a solution space, its main algorithms include genetic algorithm, tabu search algorithm, and simulated annealing algorithm. Graph-based clustering algorithm is suitable for clustering graphs or hypergraphs via the dissimilarity matrix of the data set. Grid-based clustering algorithm is sequentially implemented by partitioning the data space into a finite number of cells, estimating the cell density for each cell, sorting the cells with their densities, determining cluster centers, and traversal of neighbor cells, it can largely reduce the computational complexity. Density-based clustering
algorithm is clustered according to dense regions separated by low-density regions, can be used to cluster any shaped clusters but is not suitable for high-dimensional data sets. The commonly used density-based clustering algorithms include DBSCAN (Density-based spatial clustering of application with noise), which cannot deal with clustering for data sets with different densities, OPTICS (Ordering points to identify the clustering structure) which can solve the clustering problem for data sets with different densities and outliers, BRIDGE, DBCLASD, DENCLUE, and CUBN algorithms. Recently Ma et al. [12] developed a density-based radar scanning clustering algorithm that can discover and accurately extract individual clusters by employing the radar scanning strategy.

Model-based clustering algorithm becomes an increasingly popular tool and is conducted by assuming that data sets under consideration come from a finite mixture of probability distributions, and each component of the mixture represents a different cluster, which indicates that it requires knowing the number of components in the mixture including finite mixture model (a parametric method) and infinite mixture model (a nonparametric method), the clustering kernel including multivariate Gaussian mixture models, the hidden Markov mixture models, Dirichlet mixture models, and non-Gaussian distributions-based mixture models. Also, model-based clustering algorithms can be divided into non-Bayesian and Bayesian methods, its implementation is challenging [13]. Recently Goren and Maitra [14] developed a clustering methodology using the marginal density for the observed values assuming a finite mixture model of multivariate t distributions for partially recorded data. For clustering problems with missing data, the most common treatment is deletion or imputation. Deletion methods may lead to poor clustering performance when the missing data mechanism is not missing completely at random. In contrast, imputation method using a predicted value to impute each missing value may lead to a better clustering performance when the missing data mechanism is missing at random. But it is rather difficult to impute a suitable value for each missing value for missing not at random. The defects of deletion and imputation do not consider the missing data structure. Model-based clustering via the finite mixture of the multivariate Gaussian or t distributions has been applied to many fields, for example, see [15, 16].

Subspace clustering is conducted by identifying different clusters embedded in different subspaces of the high-dimensional data, whose clustering has several difficulties: distinguishing similar data points from dissimilar ones due to the same distance between any two data points, and different clusters lying in different subspaces. In this case, dimension reduction techniques such as principal component analysis or feature selection techniques [17, 18]. It is rather difficult to tell readers which algorithm should be used for some settings considered and how to compare novel ideas with the existing results because of its unsupervised learning process. But Gan, Ma and Wu [11] gave a comprehensive review of the applications of the aforementioned clustering algorithms.

4. Assessment of clustering algorithm

Since data clustering is a non-supervised method, the assessment of clustering algorithm is rather important. In the data clustering, there are no pre-specified clusters, it is rather challenging to find an appropriate index for measuring whether the obtained cluster result is acceptable or not. The process of assessing the results of a clustering algorithm is usually referred to as clustering validity evaluation.
Generally, clustering validity assessment includes judging the quality of clustering results, the degree to which the clustering algorithm is suitable for a special data set, and finding the best number of clusters. There are two criteria for clustering validity, for example, compactness that the individual within each cluster should be as close to each other as possible and the common measure of compactness is the variance, and separation that the clusters themselves should be separated and the commonly used methods for measuring the distance between two different clusters are the distance between the closest individual of the clusters, distance between the most distant individuals and distance between the centers of the clusters. There are three indices for assessing the results of the clustering algorithm, for example, internal indices measuring the inter-cluster validity, external indices measuring the intra-cluster validity, and relative indices.

Both internal and external indices are based on statistical methods and involve intensive computation. The comprehensive review can refer to [19]. With the increase in the dimension of data points and variables, the cluster analysis method needs to be combined with the corresponding dimension reduction technology. Extracting features through dimension reduction technology and using features to realize clustering is a method of cluster analysis of high-dimensional data.

5. Future interesting topics

Some interesting research fields in the future include model-based clustering with missing not at random data and skew-normal or skew-t distribution, model-based tensor clustering, which is a challenging topic due to the correlation structure, ultrahigh-dimension and sparsity of tensor data, and the dimension of each mode of the tensors growing at an exponential rate of the sample size, and high-dimensional and ultrahigh-dimensional data clustering that is also challenging due to sparsity of data. In these cases, data clustering needs to incorporate the dimension reduction technique and imputation technique of missing data. Also, variational and distributed techniques for data clustering may be important and challenging research with the development of computing techniques.

Acknowledgements

We would like to cordially thank Maja Bozicevic for polishing this chapter and fruitful remarks regarding chapter structure. This work was funded by NSFC grant 11731011.
Author details

Niansheng Tang* and Ying Wu
Department of Statistics, Yunnan University, Kunming, P.R. China

*Address all correspondence to: nstang@ynu.edu.cn

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
References


Introductory Chapter: Development of Data Clustering
DOI: http://dx.doi.org/10.5772/intechopen.104905

for feature screening. Computational Statistics and Data Analysis. 2018;122:18-32
