# 高维数据的聚类分析技术

聚类技术随着数据挖掘这些年的发展已经成为一个热点研究的领域。聚类分析的目的是将数据划分成有意义或有用的类。当目标是划分成有意义的类，则类能发现数据的自然结构。随着计算机科学的发展，聚类分析在心理学、社会科学、生物学、统计学、模式识别、信息检索、机器学习等广泛领域扮演着越来越重要的角色。目前聚类算法主要分为层次化聚类方法、分为式聚类方法、基于密度的聚类方法、基于网格的聚类方法、基于核的聚类算法、基于谱的聚类方法、基于模型的聚类方法、基于遗传算法的聚类方法、基于SVM的聚类方法、基于神经网络的聚类方法等。

聚类问题已出现于很多不同应用中，如数据挖掘和知识发现[[[1]](#endnote-1)]、数据压缩和向量量化[[[2]](#endnote-2)]、模式识别和模式分类[[[3]](#endnote-3)]。发现合适的类要根据不同的应用来进行，并且根据不同的标准，可以有多种不同的方法来发现类，其中包括了基于分裂和合并的方法。G. H. Ball提出的ISO Data高维数据类型的聚类分析方法[[[4]](#endnote-4)]。R. T. Ng提出的基于选择的方法CLARA[[[5]](#endnote-5)]。Gabrys B提出的模糊神经网络聚类方法[[[6]](#endnote-6)]。

基于最小化目标函数的聚类方法中，研究与使用得最为广泛的方法就是K-Means聚类。在K-Means聚类中，给定d维空间中的N个数据对象（即）和一个整数k。要解决的问题是将N个d维数据划分到K个聚簇中，使同一个聚簇中的数据对象具有相似的属性。同时，每个聚簇都和一个“中心”值相关联，且这个值是数据所在类的代表。评判划分质量的手段之一是类内差异[[[7]](#endnote-7)]，它是指每个数据对象到聚簇中心的平方距离之和（欧几里德）。对于固定划分，每个中心的最优（某种程度上是指最小的类内差异）位置是每个类中所有数据对象的中心。而且对于固定的中心选择，最优的划分是将数据对象划分到距离最近的中心所在聚簇。K-means聚类算法（其中有几个变量）使用迭代的方法在固定数目K的聚簇上进行操作，其目标是同时最优化中心的位置及数据对象的分配。

大部分聚类算法都是针对低维度的数据而设计，也能在低维数据上取得较好的聚类效果，然而很多算法在高维数据上的性能并不好。这种现象称为“维数灾难”(Curse of Dimensionality)，它泛指在分析数据时遇到的由于特征（变量或属性）过多而引起的一系列问题。点云模型本身只存在三维空间里一系列的点坐标，维数并不高，但是在虚拟装配过程中只有模型的模型的点坐标并不足够完成装配，还需要有点云模型的自由度信息、位姿信息、干涉距离信息等辅助完成。10维以上的数据就可以认为是高维数据，如果点云模型包含上述全部信息则已经属于高维数据聚类分析的范畴，而且许多聚类算法在处理高维数据时就会遇到困难[[[8]](#endnote-8)]。

目前，一般适用两种方法解决高维数据上的聚类问题：一是特征转换；二是特征选择或者子空间聚类。特征转换一般通过主成分分析或奇异值分解等策略，把原始高维数据线性合并至一个低维的空间，然后使用传统的聚类算法（如K-Means）在该低维空间上进行聚类，从而达到降低维数的目的。特征选择则是选取有效的特征，然后将其组成相关的子空间，并在该子空间上执行聚类任务。特征选择一般适用贪心策略搜索不同的特征子空间，使用一些准则来评价各个子空间并选取最优子空间，从而找到相应的类[[[9]](#endnote-9)]。

# 参考文献

1. [] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. [↑](#endnote-ref-1)
2. [] A.Gersho and R.M. Gray, Vector Quantization and Signal Compression. Boston: Kluwer Academic, 1992. [↑](#endnote-ref-2)
3. [] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973. [↑](#endnote-ref-3)
4. [] G.H. Ball and D.J. Hall, Some Fundamental Concepts and Synthesis Procedures for Pattern Recognition Preprocessors[J], Proc. Int'l Conf. Microwaves, Circuit Theory, and Information Theory, Sept. 1964. [↑](#endnote-ref-4)
5. [] R.T. Ng and J. Han, Efficient and Effective Clustering Methodsfor Spatial Data Mining, Proc. 20th Int'l Conf. Very Large Databases, pp.144-155, Sept. 1994. [↑](#endnote-ref-5)
6. [] Gabrys B, Bargiela A. General fuzzy min-max neural network for clustering and classification. IEEE Trans. on Neural Networks, 2000, 11(3):769−783. [↑](#endnote-ref-6)
7. [] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, N.J. Prentice Hall, 1988. [↑](#endnote-ref-7)
8. [] Kriegel H P, Kroger P, Sander J, et al. Density-based clustering[J]. wires. [↑](#endnote-ref-8)
9. [] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review[J]. ACMSIGKDD Explorations Newsletter, 2004, 6(1): 90–105. [↑](#endnote-ref-9)