相关工作

数据聚类是将无标签的数据集 $O = o_1, o_2, \ldots, o_n$ 分成 k 个相似集合的过程，其中 $1 < k < n$。聚类方法有很多种 [1], 总体来讲主要可以分为四类：基于划分聚类、基于分层聚类、基于分布聚类、基于密度聚类等，另外还有一些少数的方法：基于有向图的聚类方法等。基于划分聚类的方法以 k-means、k-medoids 为代表，先将数据集划分成不同的 k 簇，然后迭代交换不同类之间的元素从而提升分类效果。k-means 算法是将簇内元素的平均值作为簇的中心。而 k-medoids 算法，例如 PAM，CLARA[],CLARANS[] 则是采用簇内的某一个点作为簇的中心，相比 k-means，更不容易受到异常点的影响。但划分聚类需要根据先验知识提供簇的数目 k，k 值的设定会直接影响聚类结果。分层聚类法通常有从下至上凝聚型聚类和从上至下分裂型聚类，凝聚型聚类初始将每个数据点作为一个类别，如 single-link 算法就是将距离最近的两个类别聚成新的一类。分裂型聚类是将初始所有数据看成一个类别，然后不断细分成更小的类别。层次聚类方法的时间复杂度比较高，为 $O(N^2)$, 其中 N 为数据的个数，此外对离异点也比较敏感。BIRCH[2] CURE[3] 和 C2P[4] 等算法对传统的分层聚类算法做出了相应的改进。基于分布的聚类假设数据满足某种分布如高斯分布，聚类的目标是寻找最合适的参数，基于分布聚类的最大问题在于过拟合，并且如何选择合适的模型也是比较困难的。基于密度的聚类可以发现空间中的高密度区域并将其与低密度区域分开，DBSCAN[5] 和 OPTICS[6] 算法是密度聚类的典型代表。这种方法不需要知道任何先验知识，可以通过密度分布寻找到任何形状的聚类类别，并且对异常点不敏感，且不需要先验知识定义聚簇的个数。

[1] HAN J, PEI J, KAMBER M. Data mining: Concepts and techniques[M]. Elsevier, 2011.

[2] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: An introduction to cluster analysis[M]. John Wiley & Sons, 2009, 344.

[3] LIVNY M. BIRCH: An efficient data clustering method for very large databases[J].

[4] NG R T, HAN J. Efficient and effective clustering methods for spatial data mining[C]//Proceedings of vLDB. Citeseer, 1994: 144–155.

[5] ESTER M, KRIEGEL H-P, SANDER J, 等. A density-based algorithm for discovering clusters in large spatial databases with noise.[C]//Kdd. 1996, 96: 226–231.

[6] ANKERST M, BREUNIG M M, KRIEGEL H-P, 等. OPTICS: Ordering points to identify the clustering structure[C]//ACM sigmod record. ACM, 1999, 28: 49–60.