关于聚类分析的调研报告

软件学院 2020010108 徐浩博

**摘 要** 信息时代数据膨胀。从大量数据中，如何提取有价值信息成为焦点问题，从而促进出数据挖掘领域的繁荣——而聚类分析恰是其重要手段之一。本文将从聚类的基础概念出发，对聚类分析的实现方法和应用进行简单概括和总结。

**关键词** 数据挖掘 聚类分析 相似性衡量 划分聚类 层次聚类

1. 引言

在“大数据概论”的讲座中，王建民教授为我们深入介绍了大数据的内涵及其工业应用。教授提及，在数据膨胀的当代，如何发掘、利用数据集，推动相关技术和产业的持续发展，成为了必须要解决的问题。而这一问题的研究，也正是知识发现（KDD）领域的主要方向。

数据挖掘作为KDD的重要研究部分，主要承担数据清洗和选择后的分析任务。传统的数据挖掘主要可以分为回归Regression、分类Classification和聚类Clustering[1]。本文将主要围绕聚类的相关概念和常见的传统算法作一总结，同时还将介绍一些最新的研究进展和应用成果。

2. 聚类基础概念

**2.1 聚类的定义和特点**

聚类是将数据划分为不同类别的一种方法，可以理解为将采集到的数据按照其属性特征，划分为若干不同组的过程。作为data mining的一种重要方法，聚类区别于其他手段，如分类等，最重要的一点在于它是基于无监督学习而进行的，即预先不设定任何预期的分类，在机器学习和分析的过程中，依靠数据本身属性的相似性确定分类数量并进行分类划分。

**2.2 为什么要聚类？**

聚类是无监督学习过程，这就意味着通过聚类，人类可以发现并识别数据特征，并识别不同数据之间的相似程度。这意味着在众多数据之中，有意义的特征可以被发现并提取出来，衡量数据彼此之间的相似性确定特征，同时也可以对数据进行压缩。这样的基础性方法几乎在计算机的各个领域乃至于医学、气象、生物等交叉领域中均有广泛应用[2]。

**2.2 相似性衡量**

完成聚类的首要任务是确定相似性衡量（similarity measurement）的标准，即如何衡量各组数据之间的相似性。一般我们将相似性衡量定义为一种距离度量，L1 norm和L2 norm均是常用的度量函数，除此之外还有相关系数r、核函数等。相关系数r一定程度上可以排除线性变换的干扰；核函数使得数据被映射到高维空间并进行内积，升维操作有时会使得数据线性可分[12]。

**2.4 评判依据**

## 从大体上讲，只要组内数据点彼此相似性好，并且组与组间数据相似性低，就可以认为分组的质量较好。评判时可以用一些评价因子来量化。在没有参考模型时，Dunn指数(DI)利用所有不同类中距离最小的两组数据的最小值比所有类内元素最大距离的最大值之比值来衡量；Davies-Bouldin指数(DBI)利用任意两类的类内平均距离之和比类中心点的欧氏距离的比值最大值衡量[3]。

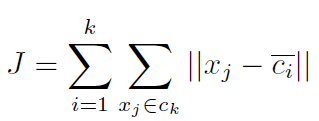
## 

3. 常用聚类方法

**3.1 划分聚类（partition clustering）**

划分聚类是指，将n个数据分配到无层次结构的k个类C1……Ck类中，同时类间两两交集为空：也即每个数据落于某一固定类内。最常用的算法为k-means算法。

k-means的流程是：随机选择k个点，每个均作为一个类的中心，并把n个数据中的每一个都划分进离它最近的聚类中心对应的簇里[13]。定义损失函数：



可以从损失函数的定义中看出，它显示了每一簇的点聚合的紧密程度，J越低则聚合越紧密。下面我们采用迭代算法，平均每一簇的所有数据，计算每一簇的中心点，将得到的新中心点作为簇的中心点，在这一步骤中，J递降。然后，再对所有n个数据操作，将其划分进离它最近的聚类中心所在的簇里，J也递降。迭代过程的流程图如下所示[12]：

由于J递减，故在经历若干次迭代后，J必然收敛，到达极小值，此时聚类划分完成[12]。由于J函数non-convex，故极小值不一定为全局最优。而求解J函数最小值又是NP-Hard问题，故选取不同的初始中心，进行k-means多次，选取多次结果中的最好结果，以此作为最优聚类的近似解。

k-means复杂度较低，每次迭代的复杂度均约为O(N)，但存在易被噪声和离异点干扰、特殊不规则形状聚类效果不好等缺点。

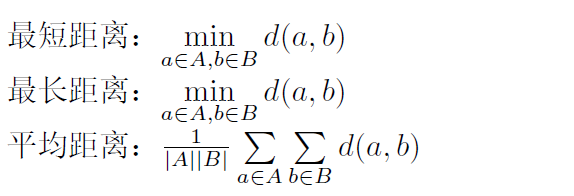
Kaufman和Rousseeuw在k-means算法的基础上加以改进，将中心点强制选为已有数据点，提出k-medoid方法，避免了k-means对于噪声和离异点过于敏感的缺陷[4]。

Vassilvitskii和Arthur针对k-means设置的初始化条件加以改进。考虑到k-means选择初始化中心点是随机选取的，若选取的点过于集中，必然获得较差的收敛结果，因此应该在随机的基础上保证选取的初始点分散[5]。由此，初始时选择任一数据点作为聚类中心；在选择好i个点之后，计算剩余未选择的点，它们每一个距离所有已知中心点的最小值，将这个最小值的倒数作为该点此次选作中心点的概率权重，再从剩余(n-i)个数据点中按照概率分布随机选择中心点。由此，初始中心点的选择会更加分散，使得k-means的结果更加接近最优解。

除此之外，Dunn还提出了Fuzzy c-means clustering(FCM)算法，他求解的是聚类的软分配(soft assignment，也即一个数据可以被多个聚类包含)问题。与k-means不同，他定义了一个ui向量表示数据点i与各中心点的相似程度（可以理解为k-means的ui为除一个方向为1外，其他方向均为0的向量），如前所述，相似程度可以用任意的度量函数度量。在一个聚类中选取中心点时，也需要与uij相关的值作为权重[6]。

**3.2 层次聚类（****hierarchical clustering）**

层次聚类是指将聚类划分层次，从较少数据点构成的小聚类到包含众多小聚类的大聚类，聚类之间有层次的包含关系。传统做法有两种观点，一种是自上向下分割聚类，一种是自下向上聚合聚类。无论是自顶向下还是自底向上，连接（linkage）比较两个聚类时，必然要考虑如何定义聚类之间的距离这一问题。常见的三种定义方式如下：



定义好linkage的度量，我们便可以进行层次划分。

自顶向下分割法的大致流程如下：

1.将所有数据点作为最大聚类

2.分割某一聚类成为两个，使二者能够产生最大linkage度量距离

3.重复步骤2直至聚类数量满足要求

自底向上凝聚法的大致流程如下：

1.将所有的数据点均作为独立的聚类

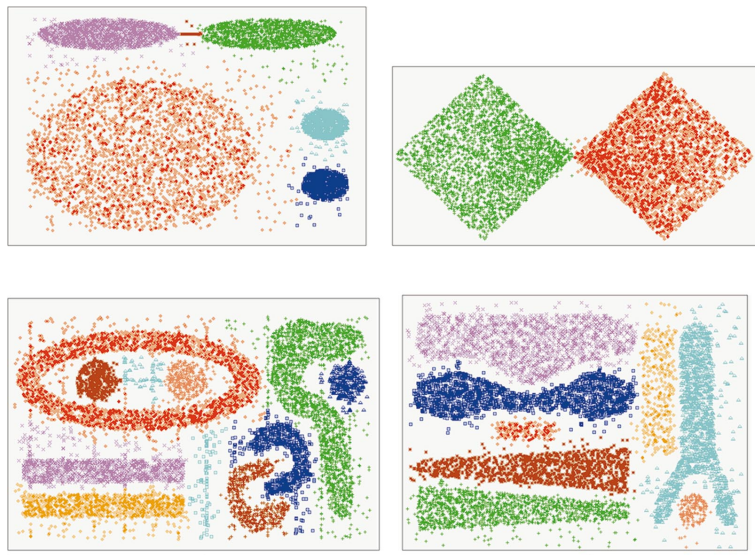
2.聚合具有linkage的最小度量距离的两个聚类

3.重复步骤2直至聚类数量满足要求[7]

我们能够通过两个流程看出，两种方法本质上是相似的。然而由于这种方法是贪心法，每一次分割或聚合后，后续步骤均不能更改以前步骤的操作，对于离异点和噪声的处理能力较弱，这造成的可能结果之一是聚类质量较低；而且该种方法复杂度较高，处理大数据集开销过大。

层次聚类的优化版本也有许多。BIRCH优化了处理离异值的鲁棒性，并且将复杂度优化为了线性[8]；CURE也在处理离异值方面进行了优化，且可以较好地处理处理各种形状和大小的集群[9]。

较为成熟的改进方案还包括Chameleon算法[10]，它由Karypis提出。Chameleon算法基于凝聚法，在合并两个聚类时，利用KNN算法充分比较两个聚类之间的相似度和聚类内部的相似度，从而使得聚类形成效果更好，结果更优。下图是原论文中展示的效果图，每个颜色代表一种聚类。可以看到，Chameleon处理较为复杂的数据时也具有较好的效果。



**3.3 基于密度的聚类**

基于密度的聚类最基础算法是Ester提出的DBSCAN算法。该算法类似于定义好“邻接”数据点的含义后再做宽度优先搜索DFS。首先我们定义一个核心点core point为在以其为中心，𝜀为半径的闭球内数据点超过minPts个的点，定义邻接可达为从核心点到半径𝜀内的其他任何点。DFS时从一个核心点开始，将所有邻接可达的点均做某聚类标记，并将可达的核心点加入队列，队列为空时则该聚类划分完毕。以此可以划分所有聚类。

该种方法具有处理离异点较优、不用确定聚类数量等特性，但由于一是需要确定𝜀和minPts参数，二是针对密度差异较大的情况无法给出较优解。后续多项研究针对DBSCAN提出优化算法，如CLIQUE等，在此不进行赘述。

**3.4 其他传统聚类方法**

基于概率模型的聚类方式如EM算法。它通过迭代，使得似然函数最大化，从而推导出高斯混合模型中的参数获得问题的解。Latent Dirichlet Allocation(LDA)等算法利用贝叶斯方法，对混合模型加以改进。[12]

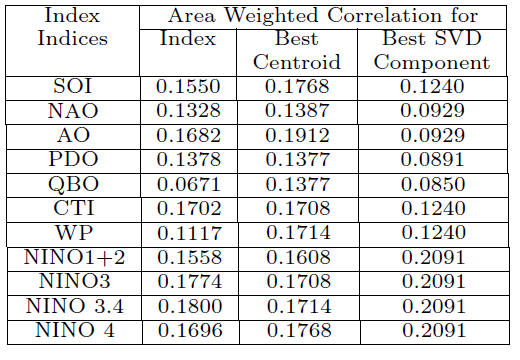
图论聚类，也作谱聚类，将数据点视作某加权图的节点，核心思想是将图的点集划分为两个子集，使其满足一定的条件。Shi和Malik提出归一化切割(Normalized Cut), Yu和Shi(2003)在此基础上进行了多级版本的改良。[2]

**3.5 聚类前沿研究领域**

深度学习方法得到深入研究并推广之后，涌现出大量基于深度学习方法的聚类方法。自动编码器AutoEncoding可以为数据特征降维，从而获得clustering的分类。此外，还有基于CDNN、基于VAE、基于GAN等多种深度学习聚类研究领域。[13]

4. 聚类的应用——以气象学的一个应用为例

气象学一个可关注的领域是分析海洋和大气对于陆地气候的影响，传统方法是通过卫星和各种观测手段得到的数据进行奇异值分解SVD和主成分分析PCA，但这种分析方法存在诸多缺陷：在奇异值分解中，只有前几个SVD向量是可被相信的，如季节变化；而后面的向量则无法分辨是噪音还是数据真的具有某种特性，这就导致了传统方法必然会导致某些数据特征的忽略。而这些特征，如容易被忽略的厄尔尼诺现象，在某些场景中就会格外突出，甚至大于主导的季节变化因素。聚类就为发掘数据特征提供了一种良好方案。在[14]这篇文章中，Steinbach就采取了k-means、DBSCAN等算法。



上表中，每一行代表一种异常天气状况的气象指数。传统方法是通过测量气象指数获得异常天气出现的可能性，但指数的参数确定，这意味着大量细节特征可能被忽略。第一列展示了通过气象指数预测天气的空间加权相关指数（越大越精准），第二、三列分别展示用clustering方法获得的空间加权相关性。可以看出，SVD大多情况下均较差，而clustering与用气象指数预测的结果差不多，甚至在某些情况下更优。

5. 总结

本篇报告首先从数据挖掘中的聚类入手，总结了与聚类相关的基础概念，归纳了几类聚类传统方法，同时还提及了聚类前沿的研究。最后，通过一个气象学相关的聚类应用，我们再次回顾了聚类的性质、算法，看到了聚类研究与应用的现实意义。我们从聚类算法的整体脉络可以看出，每一个时代的研究都具有鲜明的时代特色；从基础算法，到数学工具优化，再到近年来深度学习工具的发展，为聚类分析再添新的活力。相信随着越来越多的新技术新工具的发展，相关的研究还将继续持续而深入地进行下去，并展现出更多迷人的色彩。

参考文献

1. Chen, F., Deng, P. et al. Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks[J], 2015, 11(8): 431047.
2. Jain, A. K. Data clustering: 50 years beyond K-means. Pattern recognition letters[J], 2010, 31(8): 651-666.
3. Arbelaitz, O., Gurrutxaga, I. et al. An extensive comparative study of cluster validity indices[J]. Pattern Recognition, 2013, 46(1): 243-256.
4. Kaufman, L., & Rousseeuw, P. J. Finding groups in data: an introduction to cluster analysis (Vol. 344) [M], 2009.
5. Vassilvitskii, S., & Arthur, D. k-means++: The advantages of careful seeding[C]. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2006: 1027-1035.
6. Dunn, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[M], 1973.
7. Jain, A. K., Murty, M. N., & Flynn, P. J. Data clustering: a review[J]. ACM computing surveys (CSUR), 1999, 31(3), 264-323.
8. Zhang, T., Ramakrishnan, R., & Livny, M. BIRCH: an efficient data clustering method for very large databases[C]. ACM sigmod record, 1996, 25(2), 103-114.
9. Guha, S., Rastogi, R., & Shim, K. CURE: An efficient clustering algorithm for large databases[C]. ACM sigmod record, 1998, 27(2), 73-84.
10. Karpis, G., Han, E. H., & Kumar, V. Chameleon: Hierarchical clustering using dynamic modeling[J]. Computer, 1999, 32(8): 68-75.
11. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of ACMKDD-96, 1996, 96(34): 226-231.
12. Saxena, A., Prasad, M. et al. A review of clustering techniques and developments[J]. Neurocomputing, 2017, 267, 664-681.
13. Min, Erxue, et al. A survey of clustering with deep learning: From the perspective of network architecture[J]. IEEE Access, 2018, 6: 39501-39514.
14. Steinbach, M., Tan, P. N., Kumar et al. Discovery of climate indices using clustering[C]. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003: 446-455.