第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

浅议网状聚类算法

伍育红 王伟峰

(重庆邮电大学移通学院 重庆 400065)

摘 要 数据挖掘是信息产业界近年来非常热门的研究方向,聚类分析是数据挖掘中的核心技术,而层次聚类算法是众多聚类算法中应用最为广泛的一种,但该方法经常会遇到合并或分裂点选择的问题,一旦一组对象被合并或者分裂,下一步的处理将在新生成的簇上进行,已做的处理不能被撤销,聚类之间也不能交换对象。若在某一步没有很好地选择合并或分裂,可能导致低质量的聚类结果。因此提出了一种改进的层次聚类算法——网状聚类算法,实验结果表明该算法具有更高的准确率。

关键词 层次聚类,实验,网状聚类

中图法分类号 TP311 文献标识码 A

Discussion on Network Clustering Algorithm

WU Yu-hong WANG Wei-feng

(College of Mobile Telecommunications, Chongqing University of Posts and Telecom, Chongqing 400065, China)

Abstract Data mining is a very popular research direction in the IT industry recently, clustering analysis is the core technology of data mining, and hierarchical clustering algorithm is the most widely used one. But the method often encounter merging or split point selection problem. Once a set of objects is merged or split, the next processing will be performed on the newly created cluster, the processing has been done can not be undone, and the object can not be exchanged between the clusters. If a step does not choose to merge or split very well, that may lead to lower quality clustering results. The author presented an improved hierarchical clustering algorithm, called mesh clustering algorithm, and the experimental results show that the algorithm has a higher accuracy rate.

Keywords Hierarchical clustering, Experiment, Mesh clustering

1 引言

随着经济社会和科学技术的高速发展,各行各业积累的数据量急剧增长,如何从海量的数据中提取有用的信息成为当务之急。聚类是将数据划分成群组的过程,即把数据对象分成多个类或簇,在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大。它对未知数据的划分和分析起着非常有效的作用。通过聚类,能够识别密集和稀疏的区域,发现全局的分布模式,以及数据属性之间的相互关系等。常用的聚类算法有层次聚类算法、密度聚类算法、划分聚类算法、网格聚类算法以及模糊聚类算法等。

在这些聚类算法中,层次聚类算法由于其简单,得到了广泛的应用。但该方法经常会遇到合并或分裂点选择的问题,一旦一组对象被合并或者分裂,下一步的处理将在新生成的簇上进行,已做的处理不能被撤销,聚类之间也不能交换对象。若在某一步没有很好地选择合并或分裂,可能导致低质量的聚类结果。

为了解决上述问题,笔者提出了一种改进的 Newman 快速算法,即网状聚类算法(net-agglom erative clustering algorithm),并利用该算法在 Iris 数据集上做了测试。实验结果表明,该算法比传统层次聚类算法有更高的准确率。

2 基于层次的聚类算法

对给定数据对象进行层次上的分解,可分为凝聚算法和分裂算法。

(1)自底向上的凝聚聚类方法。这种策略是以数据对象作为原子类,然后将这些原子类进行聚合,逐步聚合成越来越大的类,直到满足终止条件。凝聚算法的过程为:在初始时,每一个成员都组成一个单独的簇,在以后的迭代过程中,再把那些相互邻近的簇合并成一个簇,直到所有的成员组成一个簇为止。其时间和空间复杂性均为 $O(n^2)$。通过凝聚式的方法将两簇合并后,无法再将其分离到之前的状态。在凝聚聚类时,选择合适的类的个数和画出原始数据的图像很重要[1]。

(2)自顶向下分裂聚类方法。与凝聚法相反,该方法先将所有对象置于一个簇中,然后逐渐细分为越来越小的簇,直到每个对象自成一簇,或者达到了某个终结条件。其主要思想是将那些成员之间不是非常紧密的簇进行分裂。跟凝聚式方法的方向相反,它从一个簇出发,一步一步细化。它的优点在于研究者可以把注意力集中在数据的结构上面。一般情况下

伍育红(1981—),女,硕士,讲师,主要研究方向为数据挖掘、信息安全、物流网等;王伟峰(1978—),男,博士,讲师,主要研究方向为网络安全、数据挖掘、现代物流等。

不使用分裂型方法，因为在较高的层很难进行正确的拆分。

3 网状聚类算法

3.1 算法基本思想

笔者提出的网状聚类算法的基本思想是：依次合并有边相连的社区并计算合并后的模块性增量。不只是计算新合并形成的社区与各个节点进行合并的模块性的增量，还要计算新合并形成的社区内部各个节点与新合并社区之外节点进行合并的模块性的增量，取其中模块性增大最多或减少最小的方向进行下一次合并[2]。每次合并以后，对相应的元素进行更新。这样，就可能存在一个数据对象与其他数据对象进行多次合并的情况，而不像一般层次聚类，每个数据对象可能只合并一次。通过不断合并社区，直到模块性增量矩阵中最大的元素由正变到负以后，即可停止合并，并认为此时的社区结构就是网络的社区结构。

3.2 算法描述

若有 n 个数据对象，网状聚类算法过程如下：

(1)初始化同 Newman 快速算法，将得到的模块性增量矩阵称为 $M_{init}(n*n)$。

(2)从 $M_{init}(n*n)$ 中选择最大的 ΔQ_{ij}，合并相应的社区 i 和 j，合并后的社区的簇号为 c_1。

(3)计算并更新模块性增量矩阵 M、最大堆 H 和最大向量 a。

M 的更新。计算 c_1 与剩余社区(除 c_1 所包含数据对象之外)的 ΔQ，进而更新 ΔQ。可能出现两种情况：①剩余的社区是单个数据对象，即需要合并的是 c_1 和另外一个数据对象 k。这时，除了要计算出 ΔQ_{kc_1} 外，还需要进一步比较 ΔQ_{kc_1} 与 $M_{init}(n*n)$ 中的第 k 行或列，选择最大值所对应的两个社区进行合并，并更新 M。如图 1 所示，当需要合并 c_1 和数据对象 5 时，要计算出 ΔQ_{5c_1}，并与 $M_{init}(n*n)$ 中 ΔQ_{52}，ΔQ_{56}，ΔQ_{59}，ΔQ_{54} 进行比较。假如 $\Delta Q_{56}>\Delta Q_{5c_1}$，则合并数据对象 5 和数据对象 6，如图 1 虚线 a 所示[3]。②剩余的簇是另一个簇，即需要合并的是 c_1 和另外一个簇 c_2。这时，除了要计算出 $\Delta Q_{c_1c_2}$ 外，还需要进一步比较 c_1、c_2 中所有数据对象之间的模块性增量，并找出最大值对应的两个数据对象进行合并，并更新 M。如图 1 所示，当需要合并 c_1 和 c_2 时，要计算出 $\Delta Q_{c_1c_2}$，并与 $M_{init}(n*n)$ 中 ΔQ_{28}，ΔQ_{21}，ΔQ_{68}，ΔQ_{61}，ΔQ_{98}，ΔQ_{91}，ΔQ_{48}，ΔQ_{41} 进行比较。假如 ΔQ_{48} 最大，则合并数据对象 4 和数据对象 8，如图 1 虚线 b 所示。

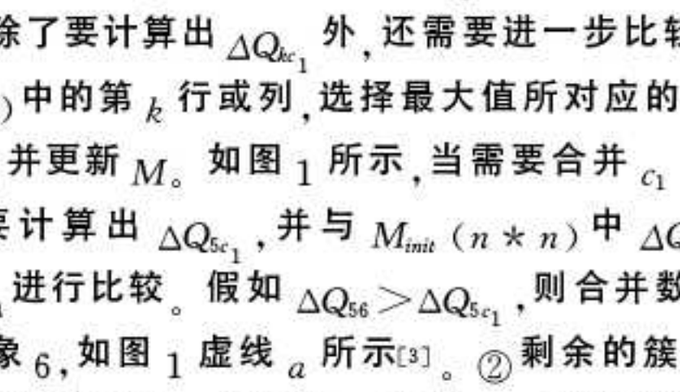

图 1 网状聚类合并点选择示意图

最大堆 H 的更新。每一次更新 ΔQ_{ij} 后，就要更新最大堆 H 中相应的行和列的最大元素。

重复步骤(2)直到模块性增量矩阵中最大的元素由正变到负，即可停止合并。

4 实验

4.1 实验数据

为了测试以上算法的聚类精度，笔者选取 UCI 的 Iris 数据集作为测试数据。Iris 数据集包含手工标注的类标识，常用于测试和评价空间聚类算法的聚类性能。Iris 包含 150 个样本，分别取自 3 种不同类型的 setosa、versicolor 和 virgi-nica 的花朵样本，每一类各 50 条记录，其中每条记录有 4 个属性：萼片长度、萼片宽度、花瓣长度和花瓣宽度[4]。

4.2 实验方法

由于所提出的算法是基于 Newman 快速算法改进而来，而 Newman 算法是基于图的，因此实验对数据进行图构造预处理，具体步骤如下：

(1)根据原始数据集计算相似度(欧式距离的倒数)得到相似度矩阵。

(2)根据相似度矩阵建立加权复杂网络。首先以数据为节点，相似度作为数据之间连边的权值，表示数据之间的连接强度，然后将权值小于阈值 β 的弱连接边删去。这样在保证网络性能的前提下，大大简化了计算。

(3)在构造的 Iris 图上运行笔者提出的算法[5]。

4.3 实验结果分析

通过以上实验，聚类的结果如表 1—表 3 所列。

表 1 直接 K-means 聚类结果

类	植物品种			数量
	setosa	versicolor	virginica	
0	0	3	36	39
1	0	47	14	61
2	50	0	0	50

表 2 加权欧氏距离 K-means 聚类结果

类	植物品种			数量
	setosa	versicolor	virginica	
0	0	2	46	48
1	0	48	4	52
2	50	0	0	50

表 3 网状聚类结果

类	植物品种			数量
	setosa	versicolor	virginica	
0	0	1	48	49
1	0	49	2	51
2	50	0	0	50

由表 1—表 3 可以看出，网状聚类结果只有 3 个样本被分错类，错分率为 2.0%。该结果与采用直接 K-means 算法和加权欧氏距离 K-means 算法相比，准确率分别提高了 9.3%和 2.0%，可见笔者所提出的网状聚类算法能够提高聚类质量，优化聚类性能。笔者提出的改进算法主要是在进行合并点选择时，不再只选取堆中最大值直接合并，而是还要充分比较剩余数据对象所有节点，以避免合并点选择错误带来的误差放大。但也不必对每个数据对象和全部其他数据对象进行比较，因此该算法的时间复杂度为 $O(n^{3/2})$。

结束语 传统的凝聚式层次聚类固有的迭代算法，使得在合并时，合并点的选择不能存在错误，一旦出错，这种错误将会被累积。基于这种情况，笔者提出了一种基于 Newman

快速算法的网状聚类改进算法。该算法由于在每个合并点都计算了模块性，这就保证了聚类结果的正确性。但是该算法的时间复杂度较高是一个致命的弱点，今后的研究工作将主要围绕这一点展开。

参 考 文 献

[1] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

[2] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev E，2009，69(6)：66-69

[3] Acqueen J. Some methods for classification and analysis of multivariate abservations [C] // Proceedings of the fifth Berkeley Sympium of Math and Statist. [S. l]：s[n.]，2012：281-297

[4] 李东琦. 聚类算法的研究[D]. 成都：西南交通大学，2010

[5] 李明华，刘全，刘忠，等. 数据挖掘中聚类算法的新发展[J]. 计算机应用研究，2010，25(1)：60-62

(上接第 490 页)

图 4 为图 3 所示算法流程的图形方式。其中 A、B、C、D、E 5 个点表示 5 个已经被 TFIDF 算法向量化的文档，两个实心灰点表示两个聚类的质心。由于 TFIDF 算法和相似度计算公式的存在，因此可以计算出任意两个点之间的距离，K-means 算法就是根据这个距离计算每个文档距离聚类质心的距离，并将每个文档分配给离它最近的聚类(第二幅图)，分配之后将每个聚类的质心按它所拥有的文档坐标更新(第三幅图)，并重复第一步(第四幅图)，直到所有文档所属的聚类都不再变化为止(第五幅图)。

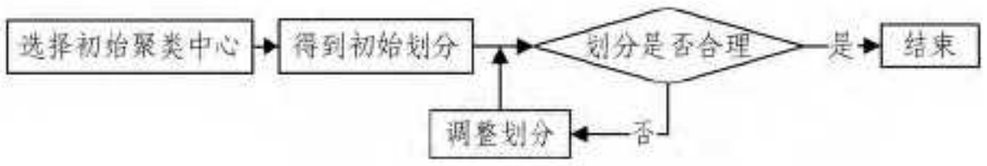

图 3 K-means 算法流程

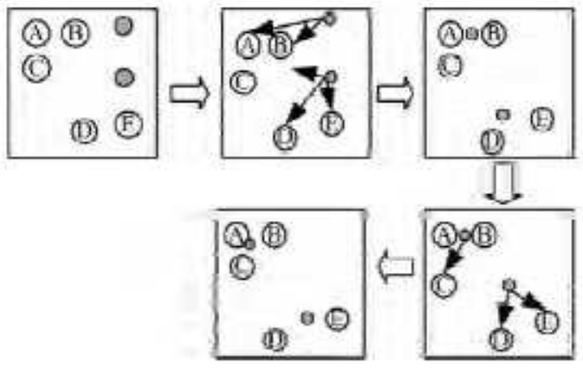

图 4 K-means 算法的模型演示

3.2.3 词库搜索

词库搜索是语义搜索引擎的扩展功能之一。它是根据用户提交给系统的关键词，连接 JWordNet[10-12] 开源接口，并向 WordNet[13] 词库发起请求，WordNet 会根据关键词和词性检索符合语义的相关词，该结果即为基于词典搜索引擎的搜索结果。

3.2.4 偏好搜索

该模块同目前的基于统计学的搜索引擎原理相似，它的实现建立在这样一个事实基础之上：同一个用户在相近一段时间内的搜索关键词同该用户的需求是呈正相关的。其实现原理是将用户提交的关键词进行分词，然后到以前的日志记录中进行比对，如果发现有符合要求的结果，则寻找那次搜索中用户在短时间内还搜索过的其它关键词，然后把这些关键词返回给用户。

结束语 本文使用了 Lucene 搜索引擎作为索引搜索的核心类库，使用 K-means 算法和 TFIDF 算法作为聚类搜索引擎的核心算法，并且引入了偏好搜索和基于词典的语义搜索，可以完成高价值的关键词推介，并具备了一定程度上的“语义搜索引擎”[14] 特征。但它的使用体验依然很大程度上依赖于训练文本库的选取。在本系统中，使用了训练文本库与索引文本库相分离的做法，用以提高聚类效率，但这都必须建立在训练文本库足够“精选”的基础上，即聚类搜索结果完全取决于训练文本库是否足够精炼，这其中的差距很大。另外，更新聚类需要耗费大量的计算机资源和时间，所以训练文本库并不能频繁刷新，这也就导致了如果在一段时间内出现某个相当风靡的词汇，聚类搜索引擎是无法将这些词及时列入聚类数据库的，时效性较差。

总之，聚类搜索是一个值得研究和期待的技术，但同大多数技术一样，它也有自己的缺陷和不足，需要不断地去弥补和完善才能满足用户的需求。

参 考 文 献

[1] 谷照升. RIA 技术解析[J]. 长春工程学院学报：自然科学版，2010，11(1)：85-88

[2] 侯丽. Web2.0 的特性及对信息服务的创新性思考[J]. 图书馆建设，2008(1)：66-69

[3] 熊回香，陈姗，许颖颖. 基于 Web 3.0 的个性化信息聚合技术研究[J]. 情报理论与实践，2011，34(8)：95-99

[4] 刘兴宇. 基于倒排索引的全文检索技术研究[D]. 武汉：华中科技大学，2004

[5] 吴洁明，冀单单，韩云辉. 基于 Web 的 DCI 垂直搜索引擎的研究与设计[J]. 计算机工程与设计，2013，34(4)：1481-1487

[6] Tan Pang-ning，Steinbach M，Kumar V. 数据挖掘导论[M]. 北京：人民邮电出版社，2011

[7] 施聪莺，徐朝军，杨晓江. TFIDF 算法研究综述[J]. 计算机应用，2009，29(1)：167-170

[8] 吴夙慧，成颖，郑彦宁，潘云涛. K-means 算法研究综述[J]. 现代图书情报技术，2011，205(5)：28-35

[9] 张睿. 基于 k-means 的中文文本聚类算法的研究与实现[D]. 西安：西北大学，2009

[10] 郑廷，郑诚. 基于 Lucene 的语义检索系统[J]. 计算机工程，2008，34(16)：92-94

[11] 王学松. Lucene+nutch 搜索引擎开发[M]. 北京：人民邮电出版社，2008

[12] 徐会生，康爱媛，何启伟. 深入浅出 Ext JS[M]. 北京：人民邮电出版社，2009

[13] 翟延冬. 基于 WordNet 的短文本语义网挖掘算法研究[D]. 长春：吉林大学，2012

[14] 张体首，蔡明. 语义搜索引擎概念模型[J]. 微电子学与计算机，2007，42(3)：171-174