浅议聚类分析方法

伍育红
(移通学院计算机系 重庆 401539)

摘 要 在简要介绍传统聚类算法的基础上,归纳了目前出现的新的聚类算法,并指出了聚类算法今后的发展方向。

关键词 群,粒度,模糊,量子聚类
中图法分类号 TP301 **文献标识码** A

Discussion on the Method of Cluster Analysis

WU Yu-hong
(Department of Computer Science of Mobile College, Chongqing 401539, China)

Abstract On the basis of brief introduction of the traditional clustering algorithm, this paper summarized the new clustering algorithm, and pointed out the future development direction of clustering algorithm.

Keywords Swarm, Particle size, Fuzzy, Quantum clustering

聚类分析是数据挖掘技术中重要的组成部分,它能够在潜在的数据中发现令人感兴趣的数据分布模式。聚类分析被广泛应用在金融数据的分类、空间数据处理、卫星图片分析和医学图像的自动检测中。聚类分析就是把数据集分成簇,使簇内数据尽量相似,簇间数据尽量不同。

聚类已经被广泛地研究了许多年,迄今为止,研究人员已经提出了很多聚类算法,大体上这些算法可以分为基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法。

经典聚类分析方法在很多领域已经得到了成功的应用。例如在商业上,聚类可以帮助市场分析人员从消费者数据库中区分出不同的消费群体,并且概括出每一类消费者的消费模式或习惯;在生物学中,它可以被用来辅助研究动物、植物的分类,可以用来分类具有相似功能的基因,还可以用来发现人群中一些潜在的结构等;另外它在空间数据处理、金融数据、卫星图像等领域都得到非常成功的应用。但是由于每一种方法都有缺陷,再加上实际问题的复杂性和数据的多样性,使得无论哪一种方法都只能解决某一类问题。近年来,随着人工智能、机器学习、模式识别和数据挖掘等领域中传统方法的不断发展以及各种新方法和新技术的涌现,数据挖掘中的聚类分析方法得到了长足的发展。

1 基于群的聚类方法

这种方法可看作进化计算的一个分支。它模拟了生物界中蚁群、鱼群和鸟群在觅食或逃避敌人时的行为。纵观文献中对于群的分类方法的研究,将这种方法分为两类:一类是蚁群算法或蚁群优化(ant colony optimization ACO);另一类称为PSO(particle swarm optimization)。

用蚁群算法或蚁群优化来进行分类规则挖掘的算法称为Ant-miner。Ant-miner是将数据挖掘概念和原理与生物界中蚁群行为结合起来形成的新算法。受生物进化机理的启发,1991年意大利学者A. Dorigo等人提出了蚁群算法,它是一种新型的优化方法。该算法不依赖于具体问题的数字描述,具有全局优化能力。后来其他科学家根据自然界真实蚂蚁堆积尸体及分工行为,提出蚂蚁的聚类算法;2002年,Labroche等人提出基于蚂蚁化学识别系统的聚类方法。总的说来,基于蚁群算法的聚类方法从原理上可以分为4种:运用蚂蚁觅食的原理,利用信息素来实现聚类;利用蚂蚁自我聚集行为来聚类;基于蚂蚁堆的形成原理实现数据聚类;运用蚁巢分类模型,利用蚂蚁化学识别系统进行聚类。

蚁群聚类算法的许多特性,如灵活性、健壮性、分布性和自组织性等,使其非常适合本质上是分布、动态及又要交错的问题求解中,能解决无人监督的聚类问题,具有广阔的前景。后来将K-means算法跟Ant-miner算法相结合,提出了Kmant-miner算法,它的预测精度高于Ant-miner。

PSO是进化计算的一个新的分支,它模拟了鱼群或鸟群的行为。PSO将群中的个体称为particles,整个群称为swarm。在优化领域,PSO可以与遗传算法相媲美。文献将PSO用于分类,对discrete PSO(DPSO) linear decreasing weight PSO(LDWPSO)和constricted PSO(DPSO)进行了比较,并选取CPSO作为挖掘分类规则的工具。对CPSO进行了改进,并与遗传算法进行了比较。实验结果表明,在预测精度和运行速度方面,PSO方法都占优势。

对ACO或PSO在数据挖掘中应用的研究仍处于早期阶段,要将这些方法用到实际的大规模数据挖掘的聚类分析中还需要做大量的研究工作。

2 基于粒度的聚类方法

从表面上看,聚类和分类有很大的差异——聚类是无导师的学习,而分类是有导师的学习。更进一步地说,聚类的目

伍育红(1981—),硕士生,讲师,主要研究方向为数据挖掘、网络控制与信息安全、电子商务与现代物流。

的是发现样本点之间最本质的抱团性质的一种客观反映；分类在这一点上却不大相同。分类需要一个训练样本集，由领域专家指明哪些样本属于一类，哪些样本数据属于另一类，但是分类的这种先验知识却常常是纯粹主观的。如果从信息粒度的角度来看的话，就会发现聚类和分类有很大的相通之处：聚类操作实际上是在一个统一粒度下进行计算的；分类操作是在不同粒度下进行计算的。所以说在粒度原理下，聚类和分类是相通的，很多分类的方法也可以用在聚类方法中。作为一个新的研究方向，虽然目前粒度计算还不成熟，尤其是对粒度计算语义的研究还相当少，但是相信随着粒度计算理论本身的不断完善和发展，今后几年它必将在数据挖掘中的聚类算法及其相关领域得到广泛的应用。

3 基于模糊的聚类方法

在实践中大多数对象没有严格的属性，它们的类属和形态存在着中介性、适合软划分。模糊聚类分析具有描述样本类属中介性的优点，能客观地反映现实世界，因而成为当今聚类分析研究的主流。最早系统表达和研究模糊聚类问题的著名学者 Ruspini 率先提出了模糊划分的概念。利用这一概念，人们相继提出了多种模糊聚类分析方法。比较典型的有基于相似性关系和模糊关系的方法、基于模糊等价关系的传递闭包方法、基于模糊凸轮的最大树方法以及基于数据集的凸分解、动态规划和难以辨识关系等方法。然而上述方法均不适于大数据的情况，难以满足实时性较高的场合。基于目标函数的模糊聚类方法把聚类归结成一个带约束的非线性规划，通过优化求解获得数据集的模糊划分和聚类。基于目标函数的模糊聚类算法成为新的研究热点。FCM(基于目标的模糊聚类方法)的原理为：

设集合 $X=(x_1,x_2,\cdots,x_n)$ 中元素有 m 个特征，即 $x_i=(x_{i1},x_{i2},\cdots,x_{im})$，要把 X 分为 c 类($2\leqslant c\leqslant n$)。设有 c 个聚类中心 $V=(v_1,v_2,\cdots,v_c)$。其中 $v_i\in\{v\mid v=\sum_{i=1}^{n}(a_ix_i)/\sum_{i=1}^{n}a_i,a_i\in R,x_i\in x\}$。取 $d_{ik}=\|v_k-v_i\|=[\sum_{j=1}^{n}(x_{kj}-x_{ij})^2]^{1/2}$ 为样本 x_k 与聚类中心 v_i 的欧式距离，那么理想的分类显然是目标 $J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{n}u_{ik}(d_{ik})^2$ 函数为最小的 U。其中：u_{ik} 表示样本 x_k 对于聚类中心 v_i 的隶属度。

由于梯度法的搜索方向总是沿着能量减小的方向，使得算法存在易陷入局部极小值和对初始化敏感的缺点。为了克服上述缺点，近几年来人们提出了各种算法来对目标函数进行优化。采取的主要措施是在 FCM 算法中引入全局寻优法。例如 1989 年徐雷提出用模拟退火对硬分类矩阵 U 进行退火处理的硬 C-均值算法；1993 年 Selim 和 Asultan 等人提出模拟退火＋模糊聚类算法；1995 年刘健庄、谢维新等人提出用遗传算法进行硬聚类和模糊聚类的分析方法；1999 年杨广文等人利用确定性退火技术提出一种聚类模型及聚类算法，然而由于模拟退火算法只有当温度下降足够慢时才能收敛于全局最优点，极长的运行时间限制了其实用性；1994 年 Babu 和 Murty 提出利用进化策略对目标函数进行聚类的方法；2002 年陈金山、韦岗提出遗传＋模糊 C-均值混合聚类算法。这些算法利用遗传算法的全局搜索能力来摆脱 FCM 聚类运算时可能陷入的局部极小点，优化了聚类的性能。众所周知，传统的进化算法是一种具有“生成＋检测”迭代过程的搜索算法。这种算法多是由体现群体搜索和群体中个体之间信息交换的两大策略的交叉和变异算子组成，为每个个体提供了优化机会，即进化的趋势。进化算法在进化过程中不可避免地产生退化现象的固有缺点，导致了进化后期的波动现象，并会出现迭代次数过多和聚类准确率不太高的现象。在某些情况下，这种退化现象还比较明显。

免疫进化算法(immune evolutionary algorithm，IEA)借鉴生命科学中的免疫概念和理论，在保留原算法优良特性的前提下，力图有选择、有目的地利用待求问题中的一些特征或知识来抑制其优化过程中出现的退化现象。免疫算法的核心在于免疫算子的构造。免疫算子通过接种疫苗或免疫选择两个步骤来完成。免疫进化算法能提高个体的适应度和防止群体的退化，从而达到减轻原有进化算法后期的波动现象和提高收敛速度。文献[3]提出了基于免疫进化的模糊聚类算法(IFCM FCSS IFCL)和基于免疫进化的硬聚类算法。这种算法既较大地提高了获取全局最优的概率，又减轻了基于遗传聚类算法在遗传后期的波动现象。进一步的工作是参数的适当选取和减小运行时间等。文献[5]提出了一种基于有限资源的模糊网络结构聚类算法。由于该算法引入模糊识别球，大大提高了运算效率，使得该算法更加适合于大数据集聚类分析；同时，因为采用了有限资源网络，克服了标准基于网络聚类算法对噪声点敏感的缺点，使得到的网络具有清晰的结构；通过分析网络神经元的最小树，能够快速准确地获得类别数以及相关的分类信息，从而实现了聚类分析的自动化。该算法不依赖于初始原型的选择，也无须类数的先验知识，可以真正做到无监督自学习。该算法中只需要预先设定最大资源数一个参数，而初始的网络规模并不影响最终的结果，所以该算法在现实生活中是非常方便的。

人们对于客观事物的认识往往带有模糊性。人类大多用一些模糊的词语来交流思想、互通信息，然后进行推理分析、综合判断，最后作出决策。客观事物是有确定性的，而反映在人的认识上却带有模糊性。人对于客观事物的识别往往只通过一些模糊信息的综合，便可以获得足够精确的定论。实质上，上述模糊聚类算法就是利用了人认识事物的规律，使计算机接近人类的智能。模糊聚类分析仍然是今后研究的重要课题之一。

4 综合其他领域的聚类方法

4.1 量子聚类

目前常用的聚类算法是基于距离的分割聚类算法，它仅仅根据数据间的几何相似性进行分类，是一种无监督的学习方法。一般来说，它并不加入数据间的几何相似性进行分类，是一种无监督的学习方法，其效果并不尽如人意；而且在现有的聚类算法中，聚类数目一般需要事先指定，如 Kohenon 自组织算法、K-means 算法和模糊 K-means 聚类算法。然而，在很多情况下类别数是不可知的，而且绝大多数聚类算法的结果一般都要依赖于初值，即使类别数目保持不变，聚类的结果也可能相差很大。

受到物理学中量子机理和特性的启发，可以用量子理论解决此类问题。一个很好的例子就是基于相关点的 Pott 自旋和统计机理提出的量子聚类模型。它把聚类问题看作一个物理系统。许多算例表明：对于传统聚类算法无能为力的几

种聚类问题，该算法都得到了比较满意的结果。

Horn 等人提出了一种新的量子聚类算法。该方法是对尺度空间向量聚类和支撑矢量机聚类固有思想的一种扩充。类似于支撑机聚类算法，该方法也与 Hilbert 空间中向量的每个点相关联；同时，还强调了它们的总和，这等于尺度空间概率函数。在这一点上与尺度空间聚类算法类似。新方法是研究 Hilbert 空间的一个算子，由 Schrodinger 等式表示，其概率函数是一个解。这个 Schrodinger 等式包括一个从概率函数中解析导出的势函数。本文将聚类中心与势能最小值联系在一起，最后验证了新方法在已知数据集合上的可行性，并通过限定 Schrodinger 势能对数据点位置的估价，将此方法应用到高维空间中的聚类问题。

4.2 核聚类算法

目前比较经典的聚类算法，如 K-means、模糊 K-means 聚类算法和 Kohonen 自组织神经网络等，只能对一些经典分布的样本奏效。它们没有对样本的特征进行优化，而是直接利用样本的特征进行聚类。因此这些方法的有效性在很大程度上取决于样本的分布情况。例如在一类样本散布较大，而另一类散布较小的情况下，这些方法的聚类效果就比较差。如果样本分布更加混乱，则聚类的结果反而会面目全非。

通过把核方法引入到聚类算法中，本文提出了一种核聚类方法。该方法增加了对样本特征的优化过程，通过利用 Mercer 核把输入空间的样本映射到高维特征空间，并在特征空间中进行聚类。核聚类方法是普适的，并在性能上优于经典的聚类算法，它通过非线性映射能够较好地分辨、提取并放大有用的特征，从而实现更为准确的聚类；同时，算法的收敛速度也较快。在经典聚类算法失效的情况下，核聚类算法仍能够得到正确的聚类。

4.3 谱聚法

最近一类有效的聚类方法开始受到广泛关注。该类方法建立在谱图理论基础之上，并利用数据的相似矩阵的特征向量进行聚类，因而统称为谱聚类方法。谱聚类算法是一种基于两点间相似关系的方法，这使得该方法适用于非测度空间。算法与数据点的维数无关，而仅与数据点的个数有关，可以避免由特征向量的过高维数所造成的奇异性问题。谱聚类算法是一个判别式算法，不用对数据的全局结构作假设，而是首先收集局部信息来表示两点属于同一类的可能性；然后根据某一聚类判据作全局决策，将所有数据点划分到不同的数据集合中。通常这样的判据可以在一个嵌入空间中得到解释，该嵌入空间是由数据矩阵的某几个特征向量张成的。谱方法成功的原因在于：通过特征分解，可以获得聚类判据在放松了的连续域中的全局最优解。

与其他方法相比，谱聚类方法具有明显的优势。该方法不仅思想简单、易于实现、不易陷入局部最优解，而且具有识别非凸分布的聚类能力，非常适合于许多实际应用问题。目前，谱聚类方法已应用于语音识别、视频分割、图像分割、VLSI 设计、网页划分、文本挖掘等领域。

谱聚类方法尽管取得了很好的效果，但目前仍处在发展的初期。算法本身仍存在许多值得深入研究的问题。

结束语 聚类分析作为数据挖掘中的重要组成部分，已经广泛应用于各个领域。在实际应用中，应根据具体问题具体分析，选择使用最佳的聚类方法。纵观数据挖掘中聚类分析方法的发展，可以看出聚类分析的新趋势：a)新方法不断涌现，如基于群的分类方法和基于粒度计算的分类方法。b)根据实际问题的需要，可以有针对性地综合众多领域的技术，以提高分类的性能。总之，数据挖掘中的聚类算法综合了机器学习、数据挖掘、模式识别、物理等领域的研究成果。相信随着这些领域中相关理论的发展、完善和相互渗透，聚类方法也将得到更进一步的发展。

参 考 文 献

[1] 马刚，李志刚. 数据仓库与数据挖掘的原理及应用[M]. 北京：高等教育出版社，2008：20-42

[2] 陈志泊. 数据仓库与数据挖掘[M]. 北京：清华大学出版社，2009：8-37

[3] 郭子龙，等. 免疫进化模糊聚类算法在边缘检测中的应用[M]. 西安：西安交通大学学报，2004：372-373

[4] Chen Y，Tu L. Density-Based Clustering for Real-Time Stream Data[J]. ACMKDD，San Jose，California，USA，2007：133-142

[5] 曲福恒，等. 浅议模糊网络结构聚类算法[J]. 吉林大学学报，2008：18-96

[6] 孙玉芬. 基于网格方法的聚类算法研究[J]. 华中科技大学，2006：62-104

[7] Han J，Kamber M. Data Mining：Concepts and Techniques [J]. Morgan Kaufmann Publishers，2001：33-82

[8] Chenm S，Han Jia-wei，Yup S. Datamining：an overview from a database perspective [J]. IEEE Trans on Knowledge and Data Eng，1996：886-883

[9] Han J，Kamber M. Data Mining：Concepts and Techniques [J]. Morgan Kaufmann Publishers，2001

[10] Wei Yong-qing，Yang Ren-hua，Liu Pei-yu. An improved Apriori algorithm for association rules of mining[C]//Proc of IEEE International Symposium on IT in Medicine &Education. Beijing：IEEE Press，2009：942-946

[11] 黄名选，严小卫，张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报，2009，20(7)：1854-1865

(上接第 321 页)

性、降低维修成本，优化维修策略方面起到重要作用[4]，对设备全寿命管理有着很好的辅助支撑作用。

参 考 文 献

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

[2] 蒋盛益. 基于聚类的入侵检测算法研究[M]. 北京：科学出版社

[3] ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/espinosa/datasets/powplant.dat

[4] 刘天安. 浅谈电力设备状态监测技术[Z]. China Science and Technology Review，115