# 聚类应用综述

2023263041牛远卓

## 摘要

数据挖掘是一个从海量数据收集并分析有用模式的过程，聚类是其中一个功能。在聚类中，对相同数据制作簇。同一个簇中的数据是相似的，然而与其他簇集合的数据不同。在图像分割中，聚类应用于制作图像的簇并确定疾病，比如Fuzzy C-Mean（FCM）算法。在社交网络分析中，聚类被用作协同过滤推荐，其中包括基于记忆与基于模型的算法。K-means算法与隐蔽马尔可夫（HMM）模型一起用来辨识信用卡欺诈。在医药的疾病分析领域，一种名叫3SW-FCM的进化的FCM算法帮助医生诊疗。贪心聚类技术是一种聚类，它基于相似性将物体结合一起，这关于高维记录的参照匹配。

## 1.介绍

聚类是将一组个体进行合并的过程，其方式是使相同集合中的个体比其他集合中的个体更相似，如图1所示。它是研究数据挖掘的核心工作——统计记录研究的常用方法。它的一些应用包括植物和动物生态学、序列分析、人类基因聚类、医学成像、市场研究、购物个体分组、社会网络分析、软件进化、图像分割、进化算法、犯罪分析、推荐系统、石油地质学和自然地理学。根据不同的研究，聚类有不同的表示方法：连接性表示法、中心点表示法、分布表示法、密度表示法、子空间表示法和基于图形的表示法。

我总结了聚类的五大应用，分别是：图像分割中的聚类、使用推荐系统进行网络分析的聚类、疾病分析中的聚类、具有参考相似性的海量记录聚类以及信用卡欺诈检测中的聚类。我将简要介绍上述论文中使用的所有技术和算法。

图表, 散点图

描述已自动生成

图 1聚类的定义

## 2.文献综述

随着当今时代科技的飞速发展，海量数据不断涌现。如何控制和管理数据，将其转化为有价值的资源，已变得十分重要。不同的技术被用来以有用的方式维护数据，数据挖掘就是其中之一。数据挖掘是我从不同角度检查数据并将其转化为有用信息的过程，这些信息可进一步用于增加利润和降低成本。数据挖掘工具可以预测未来的发展和行动，使商人能够做出积极的判断，而这些判断是由知识决定的。数据挖掘工具可以节省时间，并能对那些耗时过长的业务查询做出回应。它允许用户检查来自不同数据库的记录，对其进行分类，并精确识别其中的关系。数据挖掘功能和算法可用于不同的业务领域，预测和发现趋势。它有五大功能。聚类分析、泛化、关联和相关分析、分类、离群值分析。聚类分析是最主要也是使用最广泛的规则之一，几乎在所有申请中都会用到。

聚类是一种程序：将那些彼此较为相似的个体归为一组，而将它们与其他组中的个体区分开来。它用于探索性数据挖掘技术，是一种检查数据的集体程序。聚类可用于多种领域，包括机器知识、模式学习、图像检查、信息检索和生物信息学。根据不同的研究，有不同类型的聚类模型，它们是连接性表示法、中心点表示法、分布表示法、密度表示法、子空间表示法、基于图的表示法。常用的聚类算法有：K-均值聚类、模糊C-均值聚类、层次聚类和高斯混合聚类。

聚类分析已广泛应用于生活的各个领域，并有不同的应用：

聚类方法用于使用推荐系统的协同过滤。协同过滤是买家和商品的集体记录。推荐系统是根据评价来推荐商品。推荐系统可分为三类：协作过滤、基于内容的推荐和混合方法。协作过滤法是推荐系统中广泛使用的一种方法。协同过滤有两种方法，一种是基于记忆的方法，另一种是基于模型的方法。

聚类被用来根据公司盈亏的相似性对股票市场上的公司进行分类，对股票市场上的公司进行分组对投资者、管理者和政策制定者都有帮助。根据公司在股票市场上的相似性，采用了三阶段聚类方法对公司进行分类。首先，使用低确定时间序列事实对公司进行分类。然后，在第二阶段，将预先分组的公司划分为一些纯粹的子组。然后，在第三阶段将子群合并。

聚类通过使用模糊C-means技术用于图像分割。图像分割是将图像实际分割成一定数量的非重叠区域，这些区域具有相似的特征，如灰度、色调、颜色和纹理，用于计算机视觉和理解。Szilagyi提出了增强型FCM算法（EnFCM）和Caifast提出了广义FCM算法（FGFCM），这些算法的计算时间较短，但无法直接与独特的图像建立联系。为了克服和解决这些问题，Stelios提出了模糊有限信息c-means聚类算法（FLICM），这种聚类算法是在kerne技术的基础上发展起来的，被应用于许多图像分割领域。

聚类被用于信用卡欺诈检测。由于电子商务方式的迅速发展，信用卡的使用量也在急剧增加。在发达国家，大多数在线和离线交易都是通过信用卡支付的。随着信用卡的使用在全球范围内急剧增加，攻击者入侵信用卡信息的机会也随之增加，他们很容易进行虚假交易。攻击者只需要与持卡人相关的少量信息就能入侵信用卡。通过使用隐蔽马尔可夫模型对信用卡操作处理过程中使用的程序安排进行聚类建模，并展示了如何将其用于检测信用卡欺诈行为。该模型（HMM）对信用卡欺诈识别非常有用，它能在欺诈交易正在进行时提供即时报告。

聚类用于参考相似性的高维记录。许多问题都包括对大型数据集进行聚类。本文给出了一种对这些巨大的高维记录进行聚类的新方法。该方法的理念是使用一种廉价的估计距离量，将记录专业地分成名为"树冠"的上覆子群。聚类是通过测量共同覆盖中出现的点之间的精确空间来实现的。使用树冠方法可以解决大型聚类问题。树冠方法适用于许多领域，并与多种聚类方法一起使用，包括贪婪聚类分组法、K-均值法和期望最大化法。

聚类方法用于医学分析。随着数据挖掘方法的发展，聚类检查方法成为医学研究的活跃工具。聚类技术在检查疾病特征方面发挥了重要作用，为医生诊断提供了准确的依据。FCM算法被广泛应用于疾病检测中。但过时的FCM算法在准确性和回复时间方面存在一些缺陷。因此，本文提出了一种增强算法，该算法对点浓度进行加权组合，并根据配平方法调节理想的组数。其结果优于过时的FCM算法。

## 3.应用

### 3.1针对图像聚类的Fuzzy C-Means算法

图像分割是一种可以分割图像（将图像分割成小部分）以确认疾病和识别疾病创建部分的程序。它将图像分割成具有相同特征（如灰度、色调和纹理）的明确的非重叠部分。许多基于聚类程序的算法被推荐用于图像分割。在所有这些程序中，最流行的程序是模糊分组，它比硬聚类更可行，能保留更多的图像信息。最有名、使用最广泛的聚类技术是模糊C-means技术，该技术最早由Dunn[4]提出，后来由Bezdek[5]加以改进。传统的FCM方法用于无噪声图片时效果很好，但却无法对受声音和异常值影响的图片进行分类。为了克服这个问题，人们提出了一种改进的FCM\_S方法，但它有一个缺点，即在每个迭代阶段都必须计算空间邻域方法，因此这是一种非常耗时的方法。为了减少计算阶段和成本，Chen Zhang[6]提出了两种变体，即（FCM-S1和FCM-S2）。随后，Szilagyi[7]提出了改进的FCM（EnFCM），以加快图像分割过程，因此EnFCM方法的计算时间比其他分组方法非常短。随后，Cai[12]提出的快速生成FCM（FGFCM）技术是一种更好的FCM方法，其计算时间非常短。另一方面，这些方法不能直接作用于唯一的图像，它们需要一些指令，而这绝非易事。针对参数选择困难的问题，Setlios[1]提出了一种局部信息均值法FLICM，它通过描述模糊因子来保持图像分割的性能。随后又引入了内核距离度量方法来扩展FLICM的性能，这种方法在机器学习领域越来越受到重视。因此，建立在核方案基础上的技术在图像分割的各个领域都很有用。Zhan[12]在FCM（KCM）的主要功能中引入了一种新的核方法，将距离度量引入其中，从而提高了可预测性。Chen[6]推荐了KCM的两个变体：KFCM-S1和KFCM-S2，以重塑图像分割的计算成本。FLICM使用的是欧几里得距离，但这种距离对图像的分割效果并不理想。因此，一些研究人员采用了另一种方法，即所谓的鲁棒距离测量法，以减少离群值对分组结果的影响。受所有这些聚类方法的启发，有人提出了一种改进的FLICM算法，即KWFLICM算法，它是核方法和加权模糊方面的结合。KWFLICM算法与其他聚类算法（NNcut法、FLICM、RFLICM、WFLICM）进行了测试和比较。结果表明，与其他四种技术相比，所建议的技术可以消除噪声，同时保持图像事实，实现图像分割的精彩呈现。

### 3.2.协同过滤参考的聚类方法

在过去几年里，大量数据被上传至全球网络和社交网络。管理数据并将其存储为有用的模式是非常必要的。在推荐系统的帮助下，使用协同过滤的聚类方法可以根据以往的记录搜索个体。推荐系统会根据用户之前对商品的评价来推荐商品。评分高的个体比评分低的个体更受推荐。推荐系统中使用的是协作过滤。集体过滤有两种：1)基于记忆的2)基于模型的。在基于记忆的系统中，它以消费者评分的整个矩阵为基础，通过识别个体来生成推荐。基于模型的技术则根据以往的评分构建一个模型，并利用该模型进行推荐。在基于记忆的协同过滤中，用户以前的评分被用来制作商品的整个矩阵，因为新用户有可能根据以前的建议和评分来购买商品。在协同过滤算法中，相似性计算是最重要的一步。比较计算的主要指标是个体I和个体j的共同评分。如果是以消费者为中心的相似性分析，则使用两个消费者都排序的个体来计算两个消费者之间的关系；如果是以个体为基础的相似性分析，则通过对两个个体都排序的消费者来计算两个物质之间的比较。分析相似性的技术有很多，但使用最广泛的是皮尔逊相关和矢量余弦。发现相似性计算后，CF技术必须为有活力的客户找出最相似的客户。这是最重要的阶段，因为参照物是利用邻居的得分来确定的，因此邻居会对高分产品的认可度产生影响。邻域选择策略是通过关注相似度测量和应用领域来选择的。然后，对评分极高的产品进行极力推荐，对评分较低的产品进行较少推荐。

### 3.3信用卡欺诈检测中的聚类算法

Ghosh和Reilly[8]发明了（信用卡故障）CCF传感机，该传感机对标签信用卡的巨大模型进行了熟练（检查）。Cho和Park[14]推荐了一种隐蔽马尔可夫模型（Hidden Markov Model）构建的干扰查找方案，该方案提高了显示周期和性能。Ourston等人[15]给出了隐蔽马尔可夫模型的应用，该模型可检测多阶段的网络攻击。Hoang等人[16]也给出了使用HMM进行系统排列的新方法。HMM有两个分级层次。HMM是一组与可能性分布相关的状态。一个可能的结果或陈述可以与可能性分布的观测符号联系起来。因此，状态从外部被隐藏起来，这就是所谓的隐蔽马尔可夫模型。CCFD系统是在HMM的基础上创建的，它不需要伪造符号，只需通过持卡人的通知行为和信用卡用户不知道的特定交易中获得的物品就能检测出欺诈行为。连续在银行中使用的FDS会对持卡人的识别卡产生争议。每项操作都会提交给FDS，然后用于确认目的。FDS接受信用卡信息，如信用卡号。隐蔽马尔可夫模型会彻底检测信用卡的虚假交易，创建一组训练集，然后对持卡人的轮廓进行分类。它以集群的形式存储不同金额的交易数据，这些数据可以是低值、中值和高值范围。出于安全考虑，SI被存储在数据库中，安全表单有不同的查询，如账号、出生日期、母亲姓名，用户必须给出正确的答案以验证操作单元。这些信息只有持卡人知道。该方法分为两个阶段1)训练阶段。2)检测阶段。训练阶段：这是FDS的重要组成部分。在这一阶段，隐蔽马尔可夫模型将被熟练使用。在教学过程中，隐蔽马尔可夫模型将持卡人的交易转换为报表符号和表格结构。教学结束后，我将HMM转换为与每个持卡人一致的HMM。这一步是离线完成的，不会干扰要求在线回答的信用卡性能。虽然可以使用不同类型的聚类方法，但我使用K-均值聚类过程来定义聚类。K-means聚类技术的基本步骤如下：我定义当前群组的数量，并将其定为K，同时采用这些群组的中间值。

检测阶段：训练阶段通常是离线进行的，而检测则是在线进行的。当隐蔽马尔可夫模型的限制条件很宽泛时，我会利用持卡人提供的信号来设计初始信号排列。阈值可根据经验确定，新的排列被用作恶意排列，并用于定义后续操作的权限。计算非恶意信号的目的是检测持卡人的不同表现。

### 3.4针对疾病分析的聚类方法

聚类研究可以观察疾病的特征，为医生的分析提供准确的依据。过时的FCM算法被用于此目的，但它的研究范围有限，对大量的聚类效果不佳。为了减少这一问题，科学家们需要将过时的算法转换为基于平滑知识的模糊c-means算法。在平滑知识中，测量一个平滑函数，以精确处理事实并得出全局理想解决方案。改进后的FCM算法结合了两个特征：点密度和确定最佳聚类的过程，从而获得良好的结果。聚类算法基于聚类方法。

聚类方法有基于密度的方法、基于网格的方法和模型方法、分区方法和分层方法。聚类算法中的分区法是将数据分成K个分区，每个分区必须只属于一组数据。使用这种方法的典型算法有K-MEDIODS算法、CALARNS算法和K-means算法。层次法是一种对给定数据集进行层次设置的方法，基于这种方法的算法有CHAMELEON算法、BIRCH算法和CURE算法。基于密度法的算法有DENCLUE算法、DBSCAN算法和OPTICS算法。基于网格的方法是一种快速处理方法，基于这种方法的算法有WAVECLUSTER算法、CLIQUE算法和STING算法。基于模型的算法，我认为模型包含数据集、数据集处理、特征提取、模型和算法的选择与设计、聚类分析、结果分析和新知识。Alizadeh[9][10][11]在观察肿瘤数据时使用了基于层次的算法[13]。受上述所有算法的启发，提出了基于平滑技术的新算法3SW-FCM算法。这种新算法将点密度作为权重因子、确定最优聚类数的方法和模糊c-means聚类算法结合在一起。FCM是一种聚类算法，最早由Dumm[2]提出，后来由Bezdek[3]加以改进。通过实验证明，所提出的3SW-FCM算法比传统的FCM算法更有效，而传统的FCM算法在准确性和响应时间方面存在一些缺点。

### 3.5.在对高维记录进行分组时，参考相似性

聚类方法被广泛应用于许多重要领域。通过对病人病历进行分组，可以揭示医疗保健的情况。通过对地址进行分组，聚类可用于创建新的类别。通过对文件进行分组，可以得出排名靠前的组织。树冠技术用于对个体进行分组。树冠的主要用途是减少将记录分组到上层组所需的空间。不同的信息源都会用到树冠技术。搜索引擎使用的文本空间度量基于倒排索引。倒排索引，是一种我可以直接进入列表的矩阵演示形式。倒排索引可用于高维实值数据[17]。每个数据都能有效地转换成文档。贪婪聚合法（Greedy Agglomerative method）是一种分组类型，用于根据相似性将对象定位在一起。期望最大化聚类法也使用了天棚。这是天棚使用的另一种聚类方法。由于无法指定使用多少个聚类，所以通常情况下，Canopies不会使用这种方法。

聚类可以正式将计算节省用于顶棚技术。这种方法有两个机制：首先是设计顶篷，然后是低分组程序。如果我通过倒序目录创建檐篷，我就不需要完成整个成对空间关系。如果使用贪婪聚合法生成分组，则必须提供书目引文的距离度量[18]。许多相关工作都是为了对大型数据记录进行聚类，而树冠法则是用于对相似数据集进行分组。树冠聚类会带来一些问题，但它能减少数据记录的重复。

## 4.结论

我介绍了不同的聚类应用。在社交网络研究中，我使用修改后的CF算法来生成基于用户标记的推荐，并在记录上应用复杂的网络分组方法来寻找相似用户组。HMM应用于信用卡故障检测，在不同的交易金额范围内检测HMM的有效性。对大量记录进行聚类是一项通用任务，可以通过制作天篷来完成。树冠技术具有广泛的适用性。聚类方法可以是贪婪聚类法、K-近邻法或K-均值法，这些方法可用于创建用于测量大型数据集的天篷。树冠具有将组内所有对象归入相同树冠的能力，而且这些对象没有精确性。在图像分割中，传统FCM算法的增强算法KWFLICM被用来将图像分割成小部分，为医生诊断提供正确的依据。基于平滑技术的过时FCM算法的改进形式称为3SW-FCM算法，用于医学分析。

## 参考文献

[1] S. Krinidis and V. Chatzis, “A robust fuzzy local information C-means clustering algorithm, ”IEEE Trans. Image Process., vol. 19, no. 5, pp.1328–1337, May 2010.

[2] Sun Yang. “Fuzzy clustering in the research and application of intelligent medical diagnosis system”. MS thesis. Zhejiang University, 2006.

[3] Yang Cuiqiong, Jiang Hong, and Yu Xiaolei. ”An improved FCM clustering research”. Computer and Digital Engineering. vol. 38.pp.1-3, 2010

[4] J. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” J. Cybern., vol. 3, no. 3,pp. 32–57, 1974.

[5] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[6] S. Chen and D. Zhang, “Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure, ”IEEE Trans. Syst., Man, Cybern., B, Cybern., vol. 34, no. 4, pp. 1907–1916,Aug. 2004.

[7] L. Szilagyi, Z. Benyo, S. Szilagyii, and H. Adam, “MR brain image segmentation using an enhanced Fuzzy C-Means algorithm,” in Proc. 25th Annu. Int. Conf. IEEE EMBS, Nov. 2003, pp. 17–21

[8] Ghosh, S., and Reilly, D.L., 1994. Credit Card Fraud Detection with a Neural-Network, 27th Hawaii International Conference on Information Systems, vol. 3 (2003), pp. 621- 630.

[9] Chae, Young Moon, et al. "Data mining approach to policy analysis in a health insurance domain." International journal of medical informatics. vol. 62, no. 2, pp. 103-111, 2001.

[10] Shi Yifang et al. “Data mining and knowledge discovery technology in patient flow analysis”. Journal of Preventive Medicine. Vol. 33, no. 2, pp. 237-238, 2006 .

[11] Liu Mingxia, Ren Shiquan. “Self-organizing data mining in total health expenditure forecast”. Health Economics Research. Vol.12, pp. 10-12. 2003.

[12] W. Cai, S. Chen, and D. Zhang, “Fast and robust Fuzzy C-Means clustering algorithms incorporating local information for image segmentation, ”Pattern Recognition., vol. 40, no. 3, pp. 825–838, Mar.2007

[13] Bouchachia, Abdelhamid, and Witold Pedrycz. "Data clustering with partial supervision." Data Mining and Knowledge Discoveryvol. 12, no. 1, pp. 47-48, 2006.

[14] S.B. Cho and H.J. Park, “Efficient Anomaly Detection by Modeling Privilege Flows Using Hidden Markov Model,” Computer and Security, vol. 22, no. 1, pp. 45-55, 2003.

[15] D. Ourston, S. Matzner, W. Stump, and B. Hopkins, “Applications of Hidden Markov Models to Detecting MultiStage Network Attacks,” Proc. 36th Ann. Hawaii Int’l Conf. System Sciences, vol. 9, pp. 334-344, 2003.

[16] X.D. Hoang, J. Hu, and P. Bertok, “A Multi-Layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls,” Proc. 11th IEEE Int’l Conf. Networks, pp. 531-536, 2003.

[17] H. Hirsh. Integrating multiple sources of information in text classification using whril. In Snowbird Learning Conference, April 2000.

[18] D. Sankoff and J. B. Kruskal. Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley, 1983