山西财经大学

硕士学位论文

聚类分析在数据挖掘中的应用

姓名：许存兴

申请学位级别：硕士

专业：统计学

指导教师：雷钦礼

20040510

摘　要

聚类分析是数据挖掘方法中的一个重要的方法。本文首先对数据挖掘进行了简要的描述；其次、着重对数据挖掘中的聚类分析法进行讨论；最后、以一个超市的商品销售为例，用数据挖掘中的聚类分析法进行了挖掘。因此，本文从研究数据挖掘的算法角度出发，从三个方面对数据挖掘进行了论述：

一、数据挖掘的概述

通过对数据挖掘的概念、方法、过程、特点、作用及其与统计学关系的描述，使我们对数据挖掘有一个整体的了解。

二、聚类分析在数据挖掘中的应用

在这部分首先介绍了统计学中的聚类分析基础知识，即距离与相似系数和聚类的特征与聚类间的距离。其次，介绍了具体的聚类分析方法，包括分层聚类法（最短距离法、最长距离法和中间距离法）、分割聚类算法（PAM 算法、CLARA 算法）、基于密度的方法、基于网格的方法和基于模型的方法。

三、数据挖掘在超市中的应用

在这部分以某一超市为例，以数据挖掘的过程为线索，对这个超市的销售数据用聚类分析法中的层次法进行了数据挖掘；其次，对数据挖掘的结果进行了描述；最后，分析了数据挖掘的结果。

关键词：数据挖掘　聚类分析　数据仓库　分层聚类法　分割聚类法　数据

Abstract

Cluster analysis is a fundamental method of data mining. The issue includes three parts, a brief description of data mining, a discussion about cluster analysis in data mining, and a case study of it.

1.Introduction of data mining

In order to get a gestalt view of data mining, its conception, methods, procession, character, roles, and the relation with statistics are introduced.

2.The application of cluster analysis in data mining

In this section, the basic knowledge of cluster analysis is introduced, namely distance and relative coefficient, character of clustering analysis and distance between clusters. Furthermore, the specific methods of clustering analysis is presented, consists of hierarchical agglomerative methods (nearest neighbor, furthest neighbor and median clustering), partition cluster approach (PAM method, CLARA method), methods concerned of density, nets and models.

3.Application of data mining in supermarkets

In this section, through the process of data mining of the sale data in the supermarket, the cluster analysis methods are presented. Second, the results are showed, the last, the results are analyzed.

Key words: data mining, clustering analysis, data storage, hierarchical agglomerative methods, partition methods, data

前　言

数据挖掘技术是一门交叉性、边缘性学科，它涉及到数据库、统计学、人工智能与机器学习等多个领域。计算机的应用普及产生了大量的数据，数据挖掘就是利用这些学科的技术进行大数据量的处理。从数据挖掘的应用领域来看，其应用非常广泛，从农业生产的预测到基因分类，从化学分子结构的识别到医疗疾病的分析，从信用卡欺诈到税务稽查，从产品的销售到顾客特征的分析。因此，数据挖掘技术对未来社会的各个领域将起到越来越重要的作用。

从数据挖掘技术来看，我国的数据挖掘技术还停留在科研机构的学术研究方面，真正利用数据挖掘技术解决实际工作的问题还有一定的难度；另外，从软件方面来看，我们在数据挖掘方面的软件还很博弱，仅有个别公司开发此相关软件，解决实际问题的能力还不是十分的理想。

本文以数据挖掘的理论理体系为基础，以数据挖掘的具体应用为目的，分别从三个方面对其进行了介绍。

第一部分、对数据挖掘的总体进行了描述，包括数据挖掘的概念、数据挖掘的数据来源、数据挖掘的分类、数据挖掘的发展阶段、数据挖掘的过程、数据挖掘的特点、数据挖掘的作用及数据挖掘与统计的关系等方面进行了描述。

第二部分、描述了统计学中的聚类分析法在数据挖掘中的应用，其内容包括对聚类分析基础知识的介绍和聚类分析方法。聚类分析方法包括分层聚集法、分割聚类法、基于密度的方法、基于网格的方法和基于模型的方法。

第三部分、以列举了某超市近一周的销售数据为例，进行数据挖掘分析。本例采用的是聚类分析法中的分层聚集法，希望从该数据中能得到商品之间的关联性，以便于管理人员进行商品摆放的决策。

一、数据挖掘的概述

（一）、什么是数据挖掘

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在的有用信息和知识的过程。一般而言，人们把数据挖掘比作采矿，即将原始的数据看作是形成知识的源泉，就像采矿一样，通过一定的方法和手段从这些数据里来寻找知识。相对来说，原始的数据又可以分为结构化的和半结构化的，前者如一个二维的数据库的数据，后者如文本、图像等。挖掘知识的方法可以是数学的，也可以是非数学的。挖掘出来的知识可以被用于信息的管理，也可以用于辅助决策等等。因此，相对而言，数据挖掘是一门具有广义的交叉和边缘性的学科，它汇总了不同领域的研究者，特别是数据库、计算机、通信、人工智能、数理统计、可视化等方面的学者和工作人员。

当今社会，我们正处于数据如山，而知识贫乏的社会。如何从这些大量的数据中提取隐藏在其背后的知识是我们数据挖掘要解决的问题。我们可以利用计算机作为工具，统计学作为方法，从这个“数据矿山”中找到所蕴藏的“知识金块”，用其可以帮助企业进行决策，减少不必要的投资同时提高资金回报。因此，数据挖掘是一个利用各种分析工具在海量的数据中发现模型和数据间关系的过程，这些模型和关系可以用来做出预测。但是我们应该确信，数据挖掘不会替代有经验的商业分析师或管理人员所起的作用，它只是提供一个强大的工具。

（二）、数据挖掘的数据源泉

数据挖掘的数据来源可以说是多种多样，可以是关系数据库、事务数据库、文本数据库、多媒体数据库等。最终使用哪一种取决于用户的目的及所处的领域。目前，数据挖掘的数据主要取自关系数据库和数据仓库。

（1）关系数据库

企业日常运行的业务系统拥有大量的数据库，如客户的记录、商品的记录，但是随着时间的推移，这些数据库的格式会发生变化，需要对这些数据先进行整理及清洗。

（2）数据仓库

通常情况下，数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市

中。数据挖掘库可能是数据仓库的一个逻辑的子集，而不一定非得是物理上单独的数据库。

（3）事务数据库

数据仓库不是必需的，因为建立一个数据仓库往往要花巨额资金和大量的时间。因此，若只是为了数据挖掘，可以将一个或几个事务数据库集中到一个只读的数据挖掘库中，把它当作数据集市，然后在它上面进行数据挖掘。

（三）、数据挖掘的分类

数据挖掘可以根据不同的分类方法进行分类。根据任务划分，可以分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等。根据挖掘对象划分，有关系数据库、面向对象的数据库、空间数据库、时态数据库、文本数据库、多媒体数据库、异构数据库、遗产数据库以及 Web。根据挖掘方法，可分为机器学习方法、统计方法、神经网络方法和数据库方法。机器学习包含归纳学习方法、基于案例学习、遗传算法等。统计方法包含回归分析、判别分析、聚类分析、探索性分析等。神经网络方法包含前向神经网络、自组织神经网络等。数据库方法主要是多维数据分析方法，另外还有面向属性的归纳方法。本文将着重讨论统计方法中的聚类分析法在数据挖掘中的应用。

（1）分类分析

预言模型以通过数据库中的某些数据得到另外的数据为目标。若预测的变量是离散的，这类问题称之为分类；如果预测的变量的连续的，这种问题称之为回归。在数据挖掘中使用分类的方法有决策树法、神经网络法。

（2）聚类分析

聚类分析用于从数据集中找出相似的数据并组成不同的组。与预测模型不同，聚类中没有明显的目标变量作为数据的属性存在。聚类算法通过检测数据判断“隐藏属性”。

（3）关联分析

关联分析的目的在于生成部分数据的概要，寻找数据子集间的关联关系或者一些数据与其数据之间的派生关系，最常见的技术是利用关联规则。一旦关联数据被推导出，即可用于生成关联规则。关联规则生成后，选定关联数据中的某一类为预测目标，

给其他类赋值作为预测规则的条件。

（4）序列分析及时间序列

序列分析和时间序列说明数据中的序列信息和与时间相关的序列分析。

（四）、数据挖掘的发展阶段

数据挖掘在其发展的过程中，大致经历了以下阶段：

（1）第一代数据挖掘系统

这一代的数据挖掘系统支持一个或少数几个数据挖掘算法，这些算法设计用来挖掘向量数据，这些数据模型在挖掘时候，一般一次性调入内存进行处理。许多这样的系统已经商业化。第一代数据挖掘系统，直接将需要挖掘的数据一次性调入内存，这些系统的成功依赖于数据的质量。

（2）第二代数据挖掘系统

第二代数据挖掘系统支持数据库和数据仓库，和它们具有高性能的接口，具有高的可扩展性。这一代系统支持数据挖掘模式和数据挖掘查询语言增加系统的灵活性。第二代数据挖掘系统提供数据仓库和数据挖掘系统之间的有效接口。在实施策略方面，如果数据足够大，并且频繁的变化，这就需要利用数据库或者数据仓库技术进行管理，因此，第二代数据挖掘系统是必须的。

（3）第三代数据挖掘系统

第三代数据挖掘的特征是能够挖掘 Internet/Extranet 的分布式和高度异质的数据，并且能够有效地和操作系统集成。这一代数据挖掘系统关键的技术之一是提供对建立在异质系统上的多个预言模型以及管理这些预言模型的元数据提供第一级别的支持。第三代系统另外还提供数据挖掘系统和预言模型系统之间的有效接口。在实施策略方面，如果使用多个预言模型，或者预言模型需要经常修改，那么应该选择正在出现的第三代数据挖掘系统，以支持这些功能，当然第三代系统也能与数据库或者数据仓库集成。第三代数据挖掘系统和预言模型系统的一个重要的优点是由数据挖掘系统产生的预言模型能够自动地被操作型系统吸收，从而与操作型系统中的预言模块相联合提供决策支持的功能。

（4）第四代数据挖掘系统

第四代数据挖掘系统能够挖掘嵌入式系统、移动系统和普遍存在计算设备产生的

各种类型的数据。

就现状而言，第一代数据挖掘系统仍未发展完全，第二代、第三代数据挖掘系统已出现。第二代、第三代数据挖掘和预言模型系统将与数据仓库合并，以提供一个集成的系统来管理日常的商业过程。

（五）、数据挖掘的过程

数据挖掘的过程可粗略地分为：问题定义、数据收集和预处理、数据挖掘算法执行，以及结果的解释和评估。（如图 1-1 所示）。

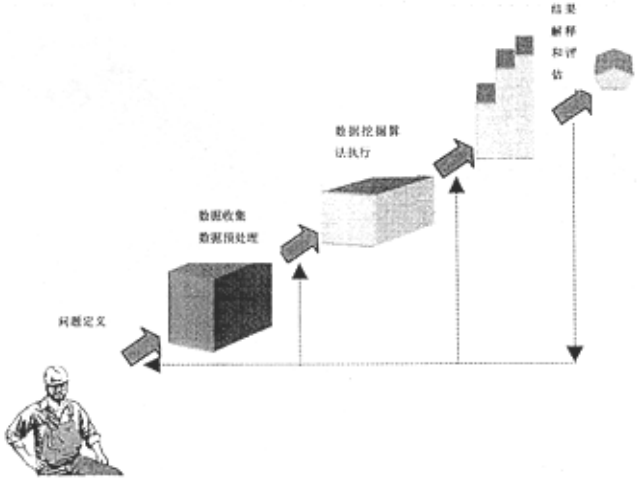


图 1-1

（1）、问题定义

数据挖掘是为了在大量数据中发现有用的令人感兴趣的信息，因此发现何种知识就成为整个过程中第一个也是最重要的一个阶段。在问题定义过程中，数据挖掘人员必须和领域专家以及最终用户紧密协作，一方面明确实际工作对数据挖掘的要求；另一方面通过对各种学习算法的对比进而确定可用的学习算法。

（2）、数据收集和数据预处理

数据准备又可分为三个子步骤：数据选取、数据预处理和数据变换。

数据选取的目的是确定发现任务的操作对象，即目标数据，是根据用户的需要从原始数据库中抽取的一组数据。

数据预处理一般可能包括消除噪声、推导计算缺值数据、消除重复记录、完成数

据类型转换等。当数据挖掘的对象是数据仓库时，一般来说，数据预处理已经在生成数据仓库时完成了。

数据变换的主要目的是消减数据维数或降维，即从初始特征中找出真正有用的特征，以减少数据挖掘时要考虑的特征或变量个数。

（3）、数据挖掘

数据挖掘算法执行阶段首先根据对问题的定义明确挖掘的任务或目的。确定了挖掘任务后，就要决定使用什么样的算法。选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；二是用户或实际运行系统的要求，有的用户可能希望获取描述型的容易理解的知识，而有的用户只是希望获取预测准确度尽可能高的预测型知识，并不在意获取的知识是否易于理解。

（4）、结果解释和评估

数据挖掘阶段发现出来的模式，经过评估，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要整个发现过程回到前一阶段，如重新选取数据、采用新的数据变换方法、设定新的参数值，甚至换一种算法等。因此，整个挖掘过程是一个不断反馈的过程。

（六）、数据挖掘的特点

数据挖掘具有以下特点：

（1）、模型复杂性

在建模上数据挖掘的重点大多放在“学习”上，对模型的复杂性和需要的计算量较为关注，而很少放在大样本的渐进推论上。数据挖掘技术有能力对复杂的数据关系进行建模，更适应于解决复杂的问题。

（2）、问题大型性

数据挖掘所涉及到的数据集合远远大于统计分析研究的数据对象。相对于古典统计学而言，数据挖掘则是从实际的海量数据源中抽取知识，这些海量数据源通常是一些大型数据库。由于数据挖掘使用的数据直接来自数据库，数据的组织形式、数据规模都具有依赖数据库的特点，数据挖掘处理的数据量非常巨大，数据的完整性、一致性和正确性都难以保证。所以，数据挖掘算法的效率、有效性和可扩充性都显得至关重要。

（3）、变量的离散性

在实践中，涉及到连续和离散变量的数据集是非常普遍的，统计学中的大多数变量分析方法是设计为连续变量模型的，但许多数据挖掘方法适合离散变量的分析。实际中，一些基于规则的方法只能使用离散变量，需要将连续变量离散化。

（4）、评价标准的有效性

在传统的数据分析方法中，评判一个方法的好坏标准是优良性，在什么范围内，按什么标准，可以证明它是最优的，在一些情况下，最优解还有明确的表达式。面对数据挖掘算法，要论证什么算法是最优的，困难是非常大的，在此评价的标准从优良性转向有效性。

（七）、统计学与数据挖掘

1、统计学在数据挖掘技术创新中的贡献

统计学在数据挖掘技术创新中的贡献主要体现在两个方面：

第一、统计学在数据挖掘方法创新中的贡献

数据挖掘方法主要包括聚类分析、决策树分析、关联分析、人工神经网络、遗传算法、机器学习和可视化方法等。在这些方法中有许多都是从统计方法中衍生而来。

首先、统计理论在人工神经网络技术中的应用——概率分析网(PLN)

人工神经网络是由一系列称为节点的处理单元组成，通过调整节点、输入和输出的权-域值来实现非线性模式识别。PLN 网络是基于概率逻辑的神经网络，它是在传统权-阈值神经网络(典型的一类是 BackpropagationHopfield，简称 BP 学习算法)的基础上提出的。它的学习速度比相同问题的 BP 算法的学习速度快百倍(两个数量级)，这说明基于统计逻辑的 PLN 网络在某些性能上比权-阈值网络强。

由于神经网络节点构造的特殊性，人们早已通过随机过程，比如马尔科夫链等工具，对 PLN 网络进行定量分析，研究神经网络各状态之间转移的概率和收敛情况。甚至在未完全知道网络对应的转移矩阵的情况下，借用统计模拟计算工具，给出平均收敛步长的变异结果。

其次、统计思想在数据挖掘学习方法上的贡献——贝叶斯网络

贝叶斯网络是一个带有概率注释的有向无环图。这种概率图模型能表示变量之间

的联合概率分布，分析变量之间的相互关系，利用贝叶斯定理提示的学习和统计推断功能，可以实现预测、分类、聚类、因果分析等数据挖掘任务。学习贝叶斯网络指的是利用样本数据更新网络原有参数或结构的先验分布。比较简单的问题是：给定贝叶斯网络的结构，利用给定样本数据学习网络的参数(概率分布)。更为复杂的问题是：网络的结构也没有确定，利用给定样本数据学习网络的结构和参数。

最后、统计在遗传算法中的应用——概率进化算法

遗传算法(Genetic Analysis,简称 GA)，是基于人工选择和交叉、变异、重组等操作构成的一种优化方法，GA 通过对大量的构造块进行选择和重组操作，再生和混合更多好的构造块，最后逼近解，但由于实际的重组操作常导致构造块破坏，导致算法或者逼近局部最优或者早熟，构造块破坏问题一般称为连锁问题。为了克服 GA 因交叉重组导致的连锁问题,人们通过从优选的解集合中提取信息的方式代替重组操作，然后利用这种信息的分布概率产生新的解，由此实现算法的连锁学习，这种将构造性概率模型引入进化算法的思想形成概率分析进化算法(PMEA)的理论依据。目前，概率分析进化算法已成为并行计算中的重要和流行的研究方向。PMEA 的特点是把自然进化算法和构造性统计分析方法结合，以指导对问题空间的有效搜索。

第二、统计对数据挖掘过程的贡献

数据挖掘是一个过程，它从大量数据中抽取出有价值的信息或知识。由于不同数据挖掘技术特点和实现步骤各不相同，成功应用数据挖掘技术、达到目标的过程就是一件很复杂的系统工程。一般，数据挖掘项目要经历的过程包括：问题的理解，数据的理解、收集和准备、建立数据挖掘模型、评价所建的模型、应用所建的模型等一系列任务。数据挖掘过程的系统化、结构化和支持系统(软件或工具)对解决问题起着至关重要的作用。统计思想在数据挖掘整个系统中的各个阶段都担负着不可忽视的重任，用统计学方法开发的工具可用于数据的抽取、清洗、转换、整合等方面，统计逻辑推理还可以让数据分析员站在更高层次上进行数据的模式识别。

2、数据挖掘为统计学的发展带来了机遇

需求是任何学科发展的动力，包括统计学在内的许多科学中，很多方法和思想都来源于现实的需求。现今，当处理的数据单位已经以 GB 或 TB 字节来计算时，仅能应付数据集的统计分析方法，已经不能满足数据挖掘的要求。这种挑战不仅体现在统计方法的计算方面，同样也体现在统计理论方面，具体表现为：统计推断的基础“总体”和“样本”的概念是否还继续适用？理由是面对如此大量的数据很难定义总体和

样本；大样本渐近性质是否满足？理由是由于数据量太大，传统的统计量无论真实情况如何都会变得“显著”；统计假设检验使用的小概率原理是否还适用？因为假定小概率事件在一次实验中不会发生是合理的，而数据量大到一定程度之后，小概率事件一定会发生。无论如何，这些问题都将带给统计学再次发展自己的机遇。具体体现在三个方面的问题：

第一、强调需求，重视过程和结果。

虽然统计学和数据挖掘一样，都是在寻求实际数据解决方案的过程中成长起来的，然而统计学家更关注模型，运用数据仅仅是为了发现新的模型，而数据挖掘则更强调知识的价值，模型是用来发现知识的工具。强调需求，重视过程和结果才能实现统计创新。

第二、借鉴机器学习的特点，提炼方法，以算法的形式体现方法。

由于统计方法的复杂性，造成实际可以被直接使用的成果太少，这不仅阻碍了人们对统计方法的运用，甚至造成对先进统计方法的不甚了解。数据挖掘的兴起，为统计学与信息技术的结合带来了发展的契机。计算机技术将成为继数学之后，又一推动统计学发展的强大工具。

第三、发挥统计软件的优势。

许多“傻瓜”统计软件的设计，更适合统计学家研究使用，任何一个初通统计的数据分析员要想通过软件来进行数据分析，都极有可能由于对数据涵义的不求甚解，导致脱离实际的统计模型的滥用，数据挖掘软件也是如此；另外，统计软件为统计研究的目的，在图形和可视化方面的互动操作，应该在数据挖掘的软件中体现这一思想，使统计软件真正成为傻瓜软件，因为它可以帮助数据分析员理解高维数据复杂的结构。

（八）、数据挖掘与聚类分析

数据挖掘的目的是从大量的数据中找出潜在、有用的信息。面对海量的资料，首要的任务是将它合理的归类。而聚类分析就是将数据合理归类的一种方法，它把分类对象按一定的规则分组或类，这些组或类不是事先给定的，而是根据数据特征而定的。在一个给定的类里，这些对象在某种意义上是倾向于彼此相似，而在不同的类里的对象差别较大。聚类分析是多元统计分析的重要组成部分，在传统的统计分析中已有多种算法，随着数据挖掘技术的兴起，又有许多新的算法被提出。目前，聚类分析已经

被广泛地用在许多领域中，包括模式识别、数据分析、图像处理以及市场研究等。

1、聚类分析的作用

聚类分析是数据挖掘中一门非常有用的技术，可以用于从大量数据中寻找隐含的数据分布和模式。在商务上，聚类能帮助市场分析人员从客户基本库中发现不同的客户群，并用购买模式来刻画不同的客户群的特征。在生物学上，聚类能用于推导植物和动物的分类，对基因进行分类，获得对种群中固有结构的认识。聚类在地球观测数据库中相似地区的确定，汽车保险单持有者的分组，以及根据房子的类型、价值和地理位置对一个城市中房屋的分组上也可以发挥作用。聚类也能用于对 Web 上的文档进行分类，以发现信息。在实际应用中，聚类分析的结果并不是最终目的。人们通过聚类分析，将数据划分为若干类，然后才在每一类中寻找模式或各种潜在的有用信息。这时，数据挖掘的其它技术将用于聚类分析的结果之上。此外，聚类分析还可用于对孤立点的监测。有时进行聚类不是为了将对象聚集在一起而是为了更容易地使某个对象从其他对象中分离出来。

2、聚类分析算法

目前文献中存在大量的聚类算法。算法的选择取决于数据的类型、聚类的目的和应用。如果聚类分析被用作描述或探查的工具，可以对同样的数据尝试多种算法，以发现数据可能揭示的结果。常用的聚类分析的算法包括：层次法、分裂法、基于密度的方法、基于网格的方法和基于模型的方法。

（1）、层次法

这种方法对给定的数据集进行层次的分解，直到某种条件满足为止。具体又可分为自底向上和自顶向下两种方案。例如在自底向上方案中，初始时每一个数据记录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。代表算法有：BIRCH 算法、CURE 算法、CHAMELEON 算法等。

（2）、分裂法

给定一个有 N 个元组或记录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，$K < N$。而且这 K 个分组满足下列条件：

第一、每一个分组至少包含一个数据记录。

第二、每一个数据记录属于且仅属于一个分组；对于给定的K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案较前一次好，而所谓好的标准就是：组与组之间的记录越远越好。使用这个基本思想的算法有：K－MENAS 算法、K－MEDOIDS 算法、CLARANS 算法。

（3）、基于密度的方法

基于密度的方法与其他方法的一个根本区别是：它不是基于各种各样的距离的，而是基于密度的，这样就能克服基于距离的算法只能发现类圆形的聚类的缺点。这个方法的指导思想就是，只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。代表算法有：DBSCA 算法、OPTICS 算法、DENCLUE 算法等。

（4）、基于网格的方法

这种方法首先将数据空间划分成为有限个单元的网格结构，所有的处理都是以单个单元为对象的。这样处理的一个突出的优点就是处理速度很快，通常与目标数据库中记录的个数无关的，它只与把数据空间分为多少个单元有关。代表算法有：STING 算法、CLIQUE 算法、WAVE－CLUSTER 算法。

（五）、基于模型的方法

基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好的满足这个模型的数据集。这样一个模型可能是数据点在空间中的密度分布函数或其他。它的一个潜在的假定就是：目标数据集是由一系列的概率分布所决定的。通常有两种尝试方向：统计的方案和神经网络的方案。

二、聚类分析在数据挖掘中的应用

聚类，顾名思义是要将相近似的对象聚成一类。聚类分析是指研究如何将研究对象按照多个方面的特征进行综合分类的一种统计方法。

（一）聚类分析基础

1、距离与相似系数

为了度量分类对象之间的接近与相似程度，需要定义一些分类统计量，常用的分类统计量有距离和相似系数。

距离是聚类分析常用的分类统计量。对于有 p 个变量的样品来说，n 个样品可以视为 p 维空间中的 n 个点，自然可以设想用点间距离度量样品间的接近程度。常用 d_{ij} 表示第 i 个样品与第 j 个样品间的距离。作为点间距离应满足以下条件：

（1）非负性，即对所有的 i 和 j，恒有 $d_{ij} \geq 0$。同时，当且仅当两个样品的 p 个变量对应相等时，其等式才成立。

（2）对称性，即对所有的 i，j 恒有 $d_{ij} = d_{ji}$。

（3）满足三角不等式，即对所有的 i、j、k 恒有 $d_{ij} \leq d_{ik} + d_{kj}$。

按上述条件可见，两个样品的距离在 $0 \to \infty$ 之间，距离越小，两个样品越接近。在聚类分析中，最常用的距离定义如下。

（1）明氏距离

$$d_{ij}(q) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{q}\right)^{1/q} \qquad q > 0$$

当 q 分别为 1，2，∞ 时，明氏距离即为绝对值距离：

$$d_{ij}(1)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$$

欧氏距离：

$$d_{ij}(2)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\right]^{1/2}$$

和切比雪夫距离：

$$d_{ij}(\infty)=\max_{1\le k\le p}\left|x_{ik}-x_{jk}\right|$$

（2）马氏距离：

上面定义的明氏距离适用于一般的欧氏空间。考虑到样品各变量的观测值往往为随机变量，因为第i个样品的p个变量的观测值$x_i=(x_{i1},x_{i2},\cdots,x_{ip})'$应是$p$维随机向量。由于随机向量有一定的分布律，各个分量之间又可能相关，因此两个样品作为两个随机向量的个体，将第i个与第j个样品间马氏距离的平方记为：

$$d_{ij}^2(M)=(x_i-x_j)'\sum{}^{-1}(x_i-x_j)$$

式中，$\sum$是随机变量的协方差矩阵。若它未知，则可用它的估计值。

对于p维空间中的两个向量，可以用相似系数度量它们之间的相似程度。设a_{ij}表示第i个与第j个向量间的相似系数，则a_{ij}应满足以下条件：

（1）绝对值不大于 1，即对所有的i，j恒有$|a_{ij}|\le 1$。同时，当且仅当两个向量存在线性关系，即$x_i=cx_j$，C 为不等于 0 的任一常数时，才成立。

（2）对称性，即对所有的i，j恒有$a_{ij}=a_{ji}$。

两个对象间的相似系数可有多种定义形式：

（1）夹角余弦

在聚类分析中，样品作为p维空间中的向量，它们的相似系数可用两个向量的夹角余弦表示，于是第i个与第j个样品间的相似系数可记为

$$q_{ij}=\cos(\theta_{ij})=\frac{\sum_{k=1}^{p}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{p}x_{ik}^2\sum_{k=1}^{p}x_{jk}^2}}$$

在聚类分析中，分类对象为指标，每个指标可以看作n维空间中的向量，第i个指标可表示为：

$$x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T$$

即为样本数据矩阵的第j列。指标间的接近程度常用相似系数表示，它可用夹角余弦表示，也常用相关系数表示。

（2）相关系数

第i个指标与第j个指标间的相关系数可记为：

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{ki}-\bar{x}_j)^2}}$$

其中$\bar{x}_i$、$\bar{x}_j$为均值，$\bar{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ki}$，$\bar{x}_j=\frac{1}{n}\sum_{k=1}^{n}x_{kj}$。

若观测值视为随机变量，则每个指标为n维随机向量。上式实际上就是两个n维随机向量之间的相关系数的估计值。

对于样品也可以定义相关系数，同样对于指标也可以定义距离。各分类对象两两之间的距离或相似系数可能构成p维或n维的系数矩阵。系数矩阵比较全面地反映了各分类对象间的接近与相似程度，是进行聚类分析所依据的基础。由距离与相似系数的对称性可知，这些系数矩阵是对称的。

2、聚类的特征与聚类间的距离

聚类是相似事物的集合。从数学的角度，难以给出一种通用的严格定义。常用的有以下几种定义，可以适合于不同的场合。

设 G 为元素的集合，它共m有个元素，记为g_i，$i=1,2,\cdots,m$　另外给定一个阈值$T>0$，则有以下几种类定义。

（1）若G中任意两个元素g_i,g_j之们的距离不大于阈值，即有$d_{ij} \le T$，则称G为类。

（2）若G中任意两个元素g_i，它与其他元素间的距离均值不大于阈值，即有$\frac{1}{k-1}\sum_{1\le j\le k} d_{ij} \le T$，则称$G$为类。

（3）对G中的任意一个元素g_i，总存在另一个元素g_j，它们的距离不大于阈值，即有$d_{ij} \le T$，则称G为类。

由此可见，它们均通过限制元素间的距离来定义类，只是限制的方法有所不同，其中第一个定义的要求最高，凡满足它的条件，一定也满足其他定义的条件。

若将类G的元素g_i视为随机向量x_i，则可用以下几种特征来描述类。

（1）类的重心

即为各元素的均向量：

$$\overline{X}_G = \frac{1}{m}\sum_{i=1}^{m} X_i$$

（2）类的样本离差矩阵与样本协差矩阵

这与随机向量样本的这两种矩阵相同，它们的定义分别为：

$$A_G = \sum_{i=1}^{m}(x_i - \overline{x}_G)(x_i - \overline{x}_G)^{\mathrm{T}}$$

$$S_G = \frac{1}{m-1}A_G$$

（3）类的直径

类的直径也有多种定义，比较简单的有两种：

$$D_G = \max_{1\le i,j\le m}(d_{ij})\text{或}$$

$$D_G = \sum_{i=1}^{m}(X_i - \overline{X}_G)^{\mathrm{T}}(X_i - \overline{X}_G) = tr(A_G)$$

前者将类中元素间的最长距离定义为类直径。后者为类中各元素至类重心的欧氏距离之和，也可理解为类中各元素指标的离差平方和的总和。前者定义的类直径与距离的量纲相同，后者类直径的量纲则为距离的平方。类的直径反映了类中各元素间的差异。

另外，还可定义类间距离以描述两类间的关系，其前提是取自不同类的两个元素间的距离是可定义的。类间距离也有多种定义形式。

设有两个类G_a与G_b，它们分别有m和n个元素，它们的重心分别为x_a与x_b。又设元素$g_i \in G_a$，元素$g_j \in G_b$，这两个元素间的距离记为d_{ij}。又将类间距离记为$D(a,b)$。它们的定义方法如下：

（1）最短距离法。它定义两类中最靠近的两个元素间的距离为类间距离，即为：

$$D_s = (a,b) = \min\{d_{ij} \mid g_i \in G_a, g_j \in G_b\}$$

（2）最长距离法。它定义两类中最远的两个元素间的距离为类间距离，即为：

$$D_L = (a,b) = \max\{d_{ij} \mid g_i \in G_a, g_j \in G_b\}$$

（3）重心法。它定义两类的两个重心间的距离为类间距离，即为：

$$D_c = (a,b)$$

（4）类平均法。它将两类中任意两个元素间距离的平均值定义为类间距离，即为：

$$D_G(a,b) = \frac{1}{mn}\sum_{g_i \in G_a}\sum_{g_j \in G_b} d_{ij}$$

（5）离差平方和法。用类中各元素指标的离差平方和的总和得到两类G_a与G_b的直径分别为D_a与D_b，类$G_{a+b} = G_a \cup G_b$，则可定义类间距离的平方为：

$$D_w^2(a,b) = D_{a+b} - D_a - D_b$$

（二）、聚类分析方法

1、分层聚集法

分层聚类法基上有两种：聚集法和分割法。聚集法是先将所有研究对象都各自算作一类，将最靠近的首先进行聚类，再将这个类和其他类中最靠近的结合，这样继续

合并直至所有对象都综合成一类或满足一个阈值条件为止。分割法正好相反，先将所有对象看成一大类，然后割成两类，使一类中的对象都自成一类或满足一个阈值条件为止。聚集或分割的过程可以用树形图直观地表示出来。分层聚集法包括以几种：

（1）、最短距离法

最短距离法又称单一连接或最近邻连接。两个类之间的距离如果定义为两类中元素之间距离最小者，并依此逐次选择最靠近的类聚集的方法叫最短距离法。见图 2-1：

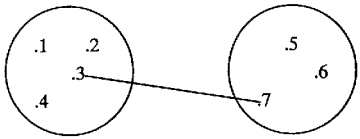


图 2-1

类间距离为：$$d\{1,2,3,4\}\{5,7,7\} = \min\{d_{15},d_{16},d_{17},d_{25},d_{26},d_{27},d_{35},d_{37},d_{45},d_{46},d_{47}\} = d_{37}$$

（2）、最长距离法

最长距离法又叫完全连接或最远紧邻连接。最长距离法与最短距离法聚类方式相同，所不同的是类与类之间的距离定义为是两类元素之间距离最大者。见图 2-2：

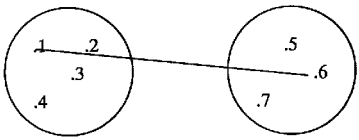


图 2-2

类间距离为：$$d\{1,2,3,4\}\{5,5,7\} = \max\{d_{15},d_{16},d_{17},d_{25},d_{26},d_{27},d_{35},d_{37},d_{45},d_{46},d_{47}\} = d_{16}$$

（3）、中间距离法

类与类之间距离如果不取两类元素间的最短距离，也不取最长距离，而是取某个中间的距离，就称为中间距离法。例如，假定在聚类的过程中两个类G_1和G_2合并成一个新类$G_N = (G_1, G_2)$。那么G_N和其他任意一类G_3的距离就定义为如下图的三角形的中线的平方。见图 2-3

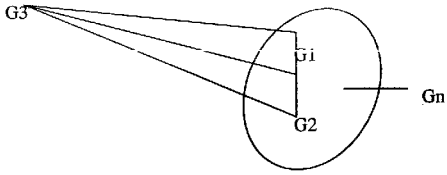


图 2-3

类间距离 $d^2_{G_3G_N} = d^2 = \frac{1}{2}\left(d^2_{G_3G_1} + d^2_{G_3G_2} - \frac{1}{2}d^2_{G_1G_2}\right)$

(4)、其他方法

a、重心法

两个类之间的距离定义为这两个类的重心间的距离。

b、类平均法

两个类之间的距离定义为是这两个类中的元素两两之间的平均距离。

c、变差平方和法

其分类的思想和方差分析的思想类似。在分类过程中，使类内元素间的变差平方和尽可能小，而类与类之间的变差平方和尽可能大。

以上六种方法我们已经了解到了聚类分析的思路。在这六种方法中，最短距离法和最长距离法的聚类过程都是单调，即每一步聚类时的距离都大于前一步，因此，若用图示聚类树将一目了然。而中间距离法是非单调的，即有时聚类的距离可能反而小于前一步聚类时的距离，因此树形图有时不易让人理解。但是中间距离法是属于空间守恒的，即两个类之间的距离基本上是取中间的，不取最短，也不取最长，这是一个优点。后三种聚类方法与前面三种方法相同。重心法是空间守恒的，但也是非单调的。类平均法即是空间守恒又具有单调的性能，比较常用，效果也较好。变差平方和法也是单调的，但并不是空间守恒而是空间扩张的，也是比较常用效果较好的。

比较这六种方法，我们发现，分层聚类法虽然比较客观，但距离的选择以及各种聚集法的选择仍带有一定的主观性。因此，在进行聚类分析时，不妨多用几种距离，多试几种方法，最后根据实际问题的性质确定一个比较合适的聚类结果。

这六种聚类方法的比较可以归纳如下（表 2-1）

表 2-1 六种聚类方法比较

方法	空间性质	单调性	对距离的要求	结果的要求	适用形	备注
最短距离法	压缩	单调		条形，S 形	惟一	太压缩、不够灵敏
最长距离法	扩张	单调		椭球形	不惟一	太扩张、样本大时易失真
中间距离法	守恒	非单调	欧氏距离的平方	椭球形	不惟一	
重心法	守恒	非单调	欧氏距离的平方	椭球形	不惟一	
类平均法	守恒	单调		椭球形	不惟一	不太压缩也不太扩张，效果较好，较常用
变差平方和法	扩张	单调	欧氏距离的平方	椭球形	不惟一	效果较好，较常用

2、分割聚类算法

给定一个有 N 个元组或记录的数据集，分割法将构造 K 个分组，每一个分组就代表一个聚类，$K < N$。而且这 K 个分组满足下列条件：

第一、每一个分组至少包含一个数据记录。

第二、每一个数据记录属于且仅属于一个分组；对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案较前一次好，而所谓好的标准就是：组与组之间中的记录越远越好。

（1）、PAM 算法

PAM 算法，即围绕中心点的分割算法。为了发现 k 个聚类，PAM 方法为每一个聚类确定了一个代表对象，称为中心点，即是聚类的最中心的位置的对象。一旦选定了中心点，每一个未被选中的对象与该中心点分在一组应该是最相似的。更进一步的说，若 o_j 是一个未被选中作为中心点的对象，而 o_i 是选中的中心点，若 $d(o_j,o_i)=\min_{o_e} d(o_j,o_e)$，我们说 o_j 属于由 o_i 代表的聚类，其中 $d(o_a,o_b)$ 定义为对象 o_a 与 o_b 之间的距离或者非相似性，$\min_{o_e}$ 表示 o_j 与所有的中心点之间的距离是最小的。所有给定的距离的值都作为 PAM 的输入。最后，聚类的质量（即与中心点的紧密程度）是由对象与它所属的中心点之间的平均距离决定的。

为了发现k个中心点，PAM 在开始任意选择了k个对象。然后在每一步都进行一个选择的中心点o_i与非中心点o_h之间的交换，这样的交换导致聚类质量的改进。为了评价o_i与o_h之间交换的效果，PAM 为所有的非中心点的对象o_j准备了一个代价变量C_{jih}，根据o_j属于下面的那种情况，定义等式中的一种。

（1）设o_j当前属于由o_i代表的聚类，更进一步说，设$Q_{j,2}$与o_h相比，o_j更相似于$Q_{j,2}$，其中$Q_{j,2}$是o_j第二个最相似的中心点，则有$d(o_j,o_h) \geq d(o_j,Q_{j,2})$。于是，若由$o_h$替换$o_i$作为中心点，则$o_j$就属于由$Q_{j,2}$代表的聚类。故在涉及$o_j$交换的代价就为：

$$C_{jih} = d(o_j,Q_{j,2}) - d(o_j,o_i)$$

该等式的结果总是非负的，表明在用o_h替换o_i时总是出现非负的代价。

（2）o_j当前属于由o_i所代表的聚类中，但在这时，o_j相似于o_h而不是$Q_{j,2}$，即$d(o_j,o_h) < d(o_j,Q_{j,2})$。于是，若由$o_h$替换$o_i$作为中心点，则$o_j$就属于由$o_h$代表的聚类，则$o_j$就属于由$o_h$代表的聚类，则$o_j$的代价就由下式给出：

$$C_{jih} = d(o_j,o_h) - d(o_j,o_i)$$

该等式的结果不同于上面的等式，其结果可能是正的，也可能是负的，依赖于o_j是更相似o_i还是o_h。

（3）假设o_j当前属于不是由o_i所代表的一个聚类中，设$Q_{j,2}$是这个聚类的中心点。设$Q_{j,2}$与o_h相比，o_j更相似于$Q_{j,2}$，则即使o_h替换o_i，o_j仍属于由$Q_{j,2}$所代表的聚类中。于是，代价为：

$$C_{jih} = 0$$

（4）o_j当前属于由$Q_{j,2}$所代表的聚类中，但$Q_{j,2}$与o_h相比，o_j不相似$Q_{j,2}$而相似于o_h。由替换o_i作为中心点，则引起o_j从$Q_{j,2}$所代表的聚类跳到o_h的聚类。于是，代价为：

$$C_{jih} = d(o_j,o_h) - d(o_j,Q_{j,2})$$

其代价总是负的。综合上述四种情况，用o_h替换o_i作为中心点的总代价为：

$$TC_{ih} = C_{jih}$$

PAM 的算法表述如下：

（1）任意选择 k 个代表对象。

（2）计算所有的对象对 o_i, o_h 的 TC_{ih}，其中 o_i 是当前聚类选择的中心点，而 o_h 是非中心点。

（3）选择对应 $\min o_i, o_h TC_{ih}$ 的对 o_i, o_h。若最小的 TC_{ih} 是负的，用 o_h 替换 o_i，返回到步聚（2）。

（4）否则，对每一个非中心点，找出相似的代表对象。

实践表明，PAM 对于小数据集合是非常合适的，但对于中数据集及大数据集其效率不高。通过分析 PAM 的计算复杂性就可以验证这个结论。在步骤（2）及步骤（3）中，共有 $k(n-k)$ 个 o_i，o_h 数据，对每一对数据，计算 TC_{ih}，需要检查 $(n-k)$ 个非中心对象，于是，步骤（2）及步骤（3）合起来计算复杂性为 $o\left(k(n-k)^2\right)$。而且，这只是一个循环的计算复杂性。因此，在 n 及 k 的值非常大时，PAM 的开销代价就非常大。

（2）、CLARA 算法

CLARA 是为了处理大数据量而开发的。与在整个数据集中发现代表对象不同，CLARA 是从数据集的样本中发现代表对象，即将 PAM 用于样本上而不是整个数据集，并且找出样本的中心点。若从数据集中抽取样本的方法是真正随机的，样本的中心点可近似看作是整个数据集的中心点。为了更好地达到近似，CLARA 抽取多个样本并将最好的聚类作为输出。但是，在精度方面，度量聚类的质量是基于在整个数据集上所有对象的平均非相似性，而不只是样本上这些对象的平均非相似性。实践证明，CLARA 在数据集大小为 $40+2k$ 时抽取 5 个样本结果最好。

CLARA 算法描述如下：

（1）for i=1 to 5，重复执行下列步骤：

（2）随机地从整个数据集中抽取一个 $40+2k$ 个对象的样本，调用算法 PAM 找出样本的 k 个中心点。

（3）对于整个数据集中的每一个对象 Q_j，判断 k 个中心点中的哪一个中心点与 Q_j 最相似的。

（4）计算前一步中得到的聚类的平均非相似性。若该值小于当前的最小值，用该值替换当前的最小值，保留在步骤（2）中得到的 k 个中心点作为到目前为止得到的最好中心点的集合。

（5）返回到步骤（1），开始下一个循环。

CLARA 弥补了 PAM 的不足，对于大数据集其性能较好。因为，PAM 的每一步循环的计算复杂性为 $o\left(k(n-k)^2\right)$，对于 CLARA 来说，由于只是将 PAM 算法应用于样本上，

每一步循环的计算复杂性为$o(k(40+k)2+k(n-k))$。这就说明了为什么对于n的值较大时 CLARA 的效率比 PAM 更高。

（3）、CLARANS 算法

给定n个对象，描述发现k个中心点的过程可以抽象地看作搜索一个图，在该图中，由$G_{n,k}$定义的节点可以用一个k个对象的集合$\{o_{m1},o_{m2},\ldots,o_{mk}\}$来表示，即$o_{m1},o_{m2},\ldots,o_{mk}$实际是选择的中心点。在图中节点的集合是$\{\{o_{m1},o_{m2},\ldots,o_{mk}\}|o_{m1},o_{m2},\ldots,o_{mk}\}$是数据集中的对象。

若两个节点的集合之间只有一个对象不同，则这两个节点是相邻的。正规的表达方式为：两个节点$S_1=\{o_{m1},o_{m2},\ldots,o_{mk}\}$及$S_2=\{o_{w1},o_{w2},\ldots,o_{wk}\}$是邻居，当且仅当$S_1$与$S_2$交集有$k-1$个元素，即$|S_1\bigcap S_2|=k-1$。可以容易地看出，每一个节点有$k(n-k)$个邻居。因为一个节点是由$k$个中心点组成的，则每一个节点对应着一个聚类，于是，每一个节点可赋予一个代价，其定义为每一个对象与它的聚类之间非相似性的总和。

很明显，可以将 PAM 看作是在图$G_{n,k}$上搜索一个最小代价。在每一个步骤，对所有当前节点的邻居都进行检查，当前的节点然后被代价下降最远的邻居替代，搜索一直到得到最小值才结束。在n及k的值较大时，检查一个节点的所有$k(n-k)$个邻居耗费时间很多，这也说明 PAM 在大数据集时效率不高。

另一方面，CLARA 试图检查较少的邻居并限制在子图上搜索，子图在大小上要远小于初始图$G_{n,k}$。但是，问题是所检查的子图完全是由样本中的对象定义的，设S_a是在样本中对象的集合，子图$G_{sa,k}$是由所有节点组成的，而这些节点是S_a的子集。虽然 CLARA 通过 PAM 完全检查了$G_{sa,k}$，但问题就是搜索全部封闭在$G_{sa,k}$内部。

类似 CLARA，算法 CLARANS 不对一个节点的每一个邻居都进行检查。与 CLARA 不同的是，它不能限制它对一个特殊的子图的搜索，事实上，它搜索了初始图$G_{n,k}$。在 PAM 与 CLARANS 之间的一个关键的差别是后者只检查一个节点的邻居的样本。与 CLARA 不同的是，CLARANS 动态地抽取每一个样本，在某种意义上讲不限制对应特殊对象的节点。换句话说，CLARA 抽取一个节点的样本是在搜索开始时进行的，而 CLARANS 抽取邻居的样本是在搜索过程中的每一个步骤进行的，其优点是不用限制在局部区域搜索。CLARANS 的搜索产生的聚类比 CLARA 的质量高，而且 CLARANS 需要更少的搜索次数。

CLARANS 的算法如下：

（1）输入参数 numlocal 及 maxneighbor。将 i 初始化为 1，而 mincost 设置为一个较大的数。

（2）将 current 设置为 $G_{n,k}$ 中一个任意节点上。

（3）将 j 设置为 1。

（4）根据 current 的一个随机邻居 S 及等式 5，计算两个节点的代价差。

（5）若 S 具有较低的代价，将 S 赋予 current，转移到步骤（3）。

（6）否则，对 j 增加 1，若 $j \le \max neighbor$，转移到步骤（4）。

（7）否则，当 $j > numlocal$ 时，比较 current 与 mincost。若前者小于 mincost，则有 $current \rightarrow \min cost$ 及 $bestnode \rightarrow current$。

（8）i 的值增 1。若 $i > numlocal$，输出 *bestnode* 并结束程序的运行。否则，转移到步骤（2）。步骤（3）到步骤（6）搜索节点的代价逐步降低。但若当前的节点已与节点的邻居的最大数比较过而且仍然是最低的代价，则就认为当前的节点是局部最小的。然后在步骤（7）中，将这个局部最小与目前得到的最小代价比较，两者值小的存入 mincost。CLARANS 算法然后再重复搜索其他的局部最小者，直到发现它们的 numlocal 为止。

CLARANS 有两个参数：检查到的最大邻居数（maxneighbor）及得到的局部最小（numlocal）。Maxneighbor 的值越高，CLARANS 越接近 PAM，而且每一个搜索局部最小就越长。但这样的一个局部最小质量越高，需得到的局部最小就越低。

3、基于密度的方法

基于密度的方法上依据密度的概念对分类对象进行聚类。它或者根据领域对象的密度或者根据某种密度函数来生成聚类。DBSCAN 就是一个有代表性的基于密度的方法。此外还有 OPTICS，DENCLUE 等算法。

DBSCAN 是一个基于密度的聚类算法。这个方法将密度足够大的那部分记录组成类，其基本思想涉及一些新的定义：

1、对于给定的对象，我们称在其半径 ε 范围内的一个记录为这个记录的 ε – 邻域；

2、如果一个对象的 ε – 邻域个数超过一个最小值，*MinPts*，那么我们就将这个记录称作核心对象；

3、一个对象的集合 D，我们说一个对象 p 在 q 的 ε – 邻域内，且 q 是一个核心对象，我们说对象 p 是从对象 q 出发直接密度可达的；

4、一个对象链 $p_1,p_2,\cdots p_n$，如果 $p_1=q,p_n=p$，对 $p_i\in D,(1\le i\le n)$。p_{i+1} 是从 p_i 出发的关于 ε 和 *MinPts* 直接密度可达的，则对象 p 是从对象 q 关于 ε 和 *MinPts* 密度可达的；

5、如果对象集合中存在一个对象 o，使得对象 p 和对象 q 是从 o 关于 ε 和 *MinPts* 密度可达的，那么对象 p 和对象 q 是关于 ε 和 *MinPts* 密度相连的。

DBSCAN 通过检查数据库中每个点的 ε－邻域来寻找聚类。如果一个点 p 的 ε－邻域包含多于 *MinPts* 个点，则创建一个以 p 为核心对象的新类。然后，DBSCAN 反复地寻找从这些核心对象直接密度可达的对象，这个过程可能涉及一些密度可达类的合并。当没有新的点可以被添加到任何类时，该过程结束。

4、基于网格的方法

基于网格的方法把对象空间量化为有限数目的单元，形成一个网格结构。所有的聚类操作都在这个网格结构(即量化的空间)上进行。代表算法有：STING 算法、CLIQUE 算法、WAVE－CLUSTER 算法。

STING 是一种基于网格的多分辨聚类技术，它将空间区域划分为矩形单元。针对不同的分辨率，通常存在多个级别的矩形单元，这些矩形单元形成了一个层次结构：高层的每个单元被划分为多个低一层的单元。关于每个网格单元属性的统计信息被预先计算和存储。

首先，在层次结构中选定一层作为查询处理的开始点，通常该层包含少量的单元。对当前层次的每个单元，我们计算置信区间(或估算其概率范围)，用以反映该单元与给定查询的关联程度。不相关的单元就不再考虑。低一层的处理就只检查剩余的相关单元。这个处理过程反复进行，直到到达底层。此时，如果查询要求被满足，那么返回相关单元区域。否则，检索和进一步的处理落在相关单元中的数据，直到它们满足查询要求。

5、基于模型的方法

基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好的满足这个模型的数据集。COBWEB 就属于这种方法。

COBWEB 是 D. Fisher 于 1987 年提出的一种聚类方法，是一种简单实用的概念增量聚类算法。其算法思想是：

1、把第一个数据项分配到第一个类里；

2、将下一个数据项分配到目前某个类中或一个新类中。其分配是基于一些准则，例如采用新数据项到目前类的重心的距离法，在这种情况下，每次添加一个新数据项到一个目前的类中时，需要重新计算重心的值。

3、重复步骤（2），直到所有的数据样本都被聚类完毕。

例如：我们要对一个平面上的 5 个点$\{x_1,x_2,x_3,x_4,x_5\}$进行聚类分析。其中，各点的坐标如下：

$$x_1=(0,2),x_2=(0,0),x_3=(1.5,0),x_4=(5,0),x_5=(5,2)$$

假定样本的顺序是x_1,x_2,x_3,x_4,x_5，则类间相似的阈值水平是$\delta=3$。

聚类过程如下：

1、第一个样本x_1将变成第一个类$C_1=\{x_1\}$。x_1的坐标就是重心坐标$M_1=\{0,2\}$。

2、开始分析其他样本。

（1）把第 2 个样本x_2和M_1比较，距离d为：

$$d(x_2,M_1)=(0^2+2^2)^{1/2}=2.0<3$$

因此，x_2属于类C_1，新的重心是：

$$M_1=\{0,1\}$$

（2）第 3 个样本x_3和重心M_1比较：

$$d(x_3,M_1)=(1.5^2+1^2)^{1/2}=1.8<3$$
$$x_3\in C_1\Rightarrow C_1=\{x_1,x_2,x_3\}\Rightarrow M_1=\{0.5,0.66\}$$

（3）第 4 个样本x_4和重心M_1比较：

$$d(x_4,M_1)=(4.5^2+0.66^2)^{1/2}=4.55>3$$

因为样本到重心M_1的距离比阈值δ大，因此该样本将生成一个自己的类$C_2=\{x_4\}$，其相应的重心为$M_2=\{5,0\}$。

(4)第 5 个样本和这两个类的重心相比较：

$$d(x_5, M_1) = \left(4.5^2 + 1.44^2\right)^{1/2} = 4.72 > 3$$
$$d(x_5, M_2) = \left(0^2 + 2^2\right)^{1/2} = 2 < 3$$

这个样本更靠近重心M_2。它的距离比阈值δ小，因此，样本x_5被添加到第 2 个类C_2中。

$$C_2 = \{x_4, x_5\} \Rightarrow M_2 = \{5,1\}$$

3、分析完所有的样本，最终的聚类解决方案是获得两个类：

$$C_1 = \{x_1, x_2, x_3\} \text{和} C_2 = \{x_4, x_5\}$$

在这中方法中，如果样本的排序不同，增量聚类过程的结果也不同。通常这个算法不是迭代的。一次迭代中分析完所有的样本生成的类便是最终类。

三、数据挖掘在超市中的应用

假设一家超市的高级管理人员想知每天超市的销售情况，顾客的购买模式，通过分析顾客购买特征，采取相应的优惠政策和商品摆设、陈列方法，增加顾客满意度和商品的销售额。如果只靠以前的传统人工技术，从巨大的购买信息中找出相应的答案就像大海里捞针，非常困难。数据挖掘技术可以帮助解决这一难题。数据挖掘的主要工作流程如下:

(一)、数据挖掘所要解决的问题

在超市中，数据挖掘所能解决的问题包括商品的销售情况，顾客的购买行为和习惯，商品摆放，货架安排，购买特定商品的顾客的特征、类型等，本例针对某一问题展开讨论。

某一超市高级管理人员想通过分析顾客商品篮子，分析商品与商品的关联性，从而找出商品与商品之间的关系，以便于商品的摆放，货架的安排等。

(二)、相关数据的收集

根据上述问题收集有关数据。需要收集的数据包括顾客信息(识别号码、姓名、性别、生日、婚姻状况、收入、教育程度、会员卡类别、经济状况、联系电话、住址等)，产品信息(识别号码、名称、品牌名称、规格、单位数量、重量、单价、库存量、销售量等)，产品分类信息(分类号码、分类名称、所属产品族名称等)，商店信息(商店代号、名称、类型、地址、负责人名称、电话、传真等)，员工信息(识别码、姓名、职位、生日、雇用时间、住址、相关描述等)等等，这些数据通过整合、集成，存入数据仓库形成分析型数据。

(三)、数据的预处理

在确定数据仓库的信息需求后，首先进行数据建模，然后确定从源数据到数据仓

库的数据提取、清理、转换、汇总和加载。

数据提取：从综合数据库中取出当前主题需要的数据，即顾客和销售情况信息。在本例中，为了分析商品与商品之间的关联程度，特提取了 20 位顾客的购买情况，即：产品名称、森通、国皓、祥通、广通、同恒、万海、世邦、中通、嘉叶、文成、康浦、东旗、建资、业兴、实翼、阳林、悦海、华科、千固和东帝望。在这些顾客中，他们总计购买了 71 种产品。其数据库形式见下表（表 3-1）(部分)：

数据清洗：数据清洗是整个数据仓库的数据入口，通过数据清洗将获取有效的数据。典型的数据清洗任务有：数据验证、数据映射。从上表中，我们可以发现大部分的单元是 0，这是因为一个顾客在一次或几次购物中不可能购买超市中所有的商品。我们对其未购买的商品均视购买量为 0。

数据转换：由于数据仓库中各个主题中的数据是按照前端应用需求存放的，因此在数据清洗后必然存在一个数据整理和转换的过程，这一过程需要对数据进行变形，使之适应前端应用需要。

数据汇总：数据仓库的一个重要应用是针对所有数据的多维查询，要进行多维查询和分析，就必须根据不同的维度对公司所有数据进行汇总，并将结果保留在数据仓库中。

表 3-1　数据库

	产品名称	森通	国皓	祥通	广通	同恒	万海	世邦	中通	嘉业	文成	康浦	东旗	建资	业兴	实翼	阳林	悦海	华科	千固	东帝望
1	肉　　肉	20	2	39	0	0	3	0	0	6	0	20	0	0	0	0	0	0	0	6	0
2	烤　肉　酱	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
3	小　　米	4	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	16
4	柠　檬　汁	28	1	20	0	0	0	12	0	0	0	10	5	0	20	0	0	6	0	0	0
5	龙　　生	23	0	0	0	0	0	0	0	40	0	0	0	6	0	0	0	40	40	0	0
6	鱼　　鱼	10	1	50	0	0	0	0	0	60	8	0	0	0	0	20	0	0	40	0	0
7	蛋　　糕	14	0	30	26	0	0	0	0	30	5	0	5	0	15	0	0	0	16	0	0
8	猪　肉　干	8	0	40	24	0	0	0	0	15	10	0	0	40	28	35	0	0	0	4	40
9	海　鲜　酱	4	0	0	4	0	0	0	0	0	0	0	0	10	0	10	0	0	5	30	0
10	猪　　肉	0	3	16	56	0	0	0	0	0	0	0	0	20	0	15	0	0	70	0	0
11	苏　打　水	0	2	0	30	0	0	0	0	0	0	0	0	4	0	0	0	0	0	30	0
12	绿　　茶	0	1	0	0	0	0	0	0	0	0	0	0	0	0	64	0	0	49	0	0
13	薯　　条	0	3	0	15	0	0	0	30	0	0	0	0	0	0	40	30	0	0	0	40
14	锅	0	5	21	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0
15	白　　米	0	4	30	48	0	0	0	0	0	0	42	5	0	20	20	0	16	80	0	28
16	黄　　豆	0	3	28	0	0	0	0	0	0	0	0	7	24	0	0	0	0	24	0	0
17	浓缩　咖啡	0	2	20	10	0	0	12	0	2	0	0	0	0	20	0	0	0	0	0	0
18	蟹	0	1	36	46	0	0	0	0	0	0	24	5	10	20	0	0	0	16	40	0
19	三合一麦片	0	2	0	0	0	0	0	0	0	0	0	0	0	0	70	0	0	0	8	3
20	山　渣　片	0	3	14	107	0	0	0	0	36	0	0	0	20	0	0	0	0	31	40	0

数据加载：在经历了数据提取、清洗、整理和汇总后，需要将所获得的结果载入数据仓库中，这个过程应定时进行，并且不同主题的数据加载有各自不同的执行任务。

上面的各种处理并非一个应用中全部用到，这要根据任务的要求和数据的质量来确定。应该注意的是，尽管上面的各种处理都有相应的工具来实现，但是这些工

具毕竟不具备人类智能，有时反而将正确的数据当作错误的数据来处理，增加了数据的噪声。因此人的参与不可缺少。

（四）、数据仓库的逻辑模型

数据仓库的数据是面向主题的，数据仓库的设计应围绕主题展开，最常用的是“星型模型”，也可以采用“雪花模型”，“雪花模型”是“星型模型”的变种，它把一些数据进一步分解到附加表中，以实现多维分析。“雪花模型”中包括“事实表”和“维表”。“事实表”存储事实的度量值和各个维的码值；“维表”存储维的描述信息，包括维的层次，成员类别和码值等。针对前面的问题，最重要的主题是商品销售。超市的高级管理人员关心的是商品销量，销售额和利润；也关心顾客购买行为和习惯、特征。从这个主题出发，设计以下一个雪花结构的模型,见下图(图 3-1)：

本例的数据仓库的逻辑模型为（表 3-2）

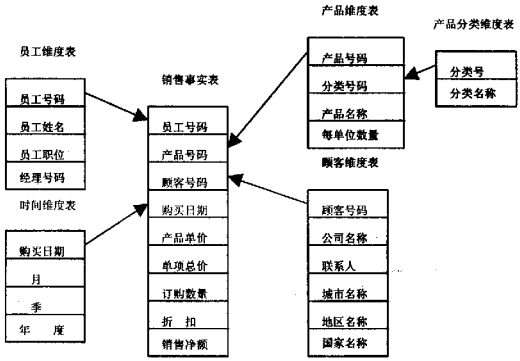


图 3-1

（五）、数据挖掘算法的选择

在数据仓库的基础上，从产品这个主题出发，将上述问题转化为一系列数据挖掘任务，主要有分类、估值、预测、篮子分析、聚集、描述。主要分析的是在顾客购买产品中，产品与产品之们的关联性。这里重点运用聚类分析法中的聚集法。

聚集法的基本思想是先将所有研究对象都各自算作一类，将最靠近的首先进行聚类，再将这个类和其他类中最靠近的结合，这样继续合并直至所有对象都综合成

一类或满足一个阈值条件为止。最终，可以用树形图直观地表示出来。本例将采用不同的聚集法对其进行分析。

表 3-2 数据库逻辑模型

	Name	Type	Width
1	产品名称	String	12
2	森通	Numeric	8
3	国酷	Numeric	8
4	祥通	Numeric	8
5	广通	Numeric	8
6	同恒	Numeric	8
7	万海	Numeric	8
8	世邦	Numeric	8
9	中通	Numeric	8
10	嘉业	Numeric	8
11	文成	Numeric	8
12	康浦	Numeric	8
13	东旗	Numeric	8
14	建资	Numeric	8
15	业兴	Numeric	8
16	实翼	Numeric	8
17	阳林	Numeric	8
18	悦海	Numeric	8
19	华科	Numeric	8
20	千固	Numeric	8
21	东希望	Numeric	8

1、最短距离法

最短距离法是把一个类的所有样品与另一个类的所有样品的两两样品之间的最短距离找出来，并将其定义两个类之间的距离。

用 d_{ij} 表示样品 i 和样品 j 的距离，G_1、G_2、$\cdots$ 表示类，则最短距离法定义类 G_p 与类 G_q 之间的距离为两类最近样品间的距离，用 D_{pq} 表示 Gp 与类 Gq 的距离，则：

$$D_{pq} = \min_{j\in G_p, j\in G_q} d_{ij}$$

用最短距离的法聚类的步骤如下：

（1）、规定样品这间的距离，计算样品两两距离的对称矩阵，记作 $D_{(0)}$。开始每个样品自成一类，这时显然 $D_{pq} = d_{pq}$。

（2）、找出 $D_{(0)}$ 的非对角线上的最小元素，设为 D_{pq}，将 G_p 与 G_q 合并成一新类，记为 $G_r = \{G_p, G_q\}$。

（3）、计算新类与其它类的距离：

$$D_{rk} = \min_{i\in G_r, j\in G_K} d_{ij}$$
$$= \min\left\{\min_{i\in G_r, j\in G_K} d_{ij}, \min_{j\in G_r, j\in G_K} d_{ij}\right\} = \min\{D_{pk}, D_{qk}\}$$

将$D_{(0)}$中p行与q行和p列与q列合并成一个新行新列，新行新列对应着G_r，所得到的距离矩阵记作$D_{(1)}$。

（4）、对$D_{(1)}$重复施行对于$D_{(0)}$的步骤得$D_{(2)}$，由$D_{(2)}$按同样的步骤计算得$D_{(3)}$，……,这样一直进行下去，直到所有样品都成一类为止。

（5）、将聚类过程做出聚类谱系图，根据谱系图进行分类。

2、最长距离法

最长距离法对类之间距离的定义与最短距离正好相反，此聚类法定义类与类之间的距离为它们之间两个最远样品之间的距离。即：

$$D_{pq} = \max_{j\in G_p, j\in G_q} d_{ij}$$

最长距离法和最短距离法的并类步骤完全一样，也是各样品先自成一类，然后将距离最小的两类合并，设某一步将类G_p和G_q合并成为G_r，则类G_r与其他任意一类G_k的距离为：

$$D_{rk} = \max\{D_{pk}, D_{qk}\}$$

然后，再找距离最小的两类合并，直至将所的样品都合并为一类为止。

3、中间距离法

类与类之间距离如果不取两类元素间的最短距离，也不取最长距离，而是取某个中间的距离，就称为中间距离法。例如，假定在聚类的过程中两个类G_1和G_2合并成一个新类$G_N = (G_1, G_2)$。那么G_N和其他任意一类G_3的距离就定义为如下图的三角形的中的线的平方。(见图 3-2)，类间距离

$$d^2_{G_3G_N} = d^2 = \frac{1}{2}\left(d^2_{G_3G_1} + d^2_{G_3G_2} - \frac{1}{2}d^2_{G_1G_2}\right)$$

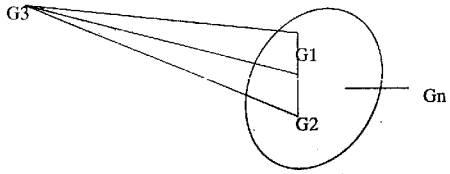


图 3-2

4、重心法

一个类用它的重心即该类样品的均值做代表比较合理，类与类之间的距离就用重心之间的距离来代表。设G_p和G_q的重心分别是$\bar{x}_p$和$\bar{x}_q$，则G_p和G_q之间距离为：

$$D_{pq}=d_{\bar{x}_p\bar{x}_q}$$

对应于这种距离定义的系统聚类法就叫重心法。

设某一步设G_p、G_q的重心为$\bar{x}_p$、$\bar{x}_q$，它们分别有样品n_p、n_q个，将G_p与G_q合并为G_r，则G_r内有样品$n_r=n_p+n_q$个，它的重心是$\bar{x}_r$，显然：

$$\bar{x}_r=\frac{1}{n_r}(n_p\bar{x}_p+n_q\bar{x}_q)$$

对于其他某一类G_k，它一重心是$\bar{x}_k$，若采用欧氏距离，则类G_k与新合并成的新类G_r之间的距离是：

$$\begin{aligned}D_{kr}^2&=d_{\bar{x}_k\bar{x}_r}^2=(\bar{x}_k-\bar{x}_r)'(\bar{x}_k-\bar{x}_r)\\&=\left[\bar{x}_k-\frac{1}{n_r}(n_p\bar{x}_p+n_q\bar{x}_q)\right]'\left[\bar{x}_k-\frac{1}{n_r}(n_p\bar{x}_p+n_q\bar{x}_q)\right]\\&=\bar{x}_k'\bar{x}_k-2\frac{n_p}{n_r}\bar{x}_k'\bar{x}_p-2\frac{n_q}{n_r}\bar{x}_k\bar{x}_q+\frac{1}{n_r^2}\left[n_p^2\bar{x}_p'\bar{x}_p+2n_pn_q\bar{x}_p'\bar{x}_p+n_q^2\bar{x}_q'\bar{x}_q\right]\end{aligned}$$

利用：$\bar{x}_k'\bar{x}_k=\frac{1}{n_r}(n_p\bar{x}_k'\bar{x}_k+n_q\bar{x}_k'\bar{x}_k)$

则有：

$$\begin{aligned}D_{kr}^2&=\frac{n_p}{n_r}\left(\bar{x}_k'\bar{x}_k-2\bar{x}_k'\bar{x}_p+\bar{x}_p'\bar{x}_p\right)+\frac{n_q}{n_r}\left(\bar{x}_k'\bar{x}_k-2\bar{x}_k'\bar{x}_q+\bar{x}_q'\bar{x}_q\right)\\&-\frac{n_pn_q}{n_r^2}\left(\bar{x}_p'\bar{x}_p-2\bar{x}_p'\bar{x}_q+\bar{x}_q'\bar{x}_q\right)=\frac{n_p}{n_r}D_{kp}^2+\frac{n_q}{n_r}D_{kq}^2-\frac{n_p}{n_r}\frac{n_q}{n_r}D_{pq}^2\end{aligned}$$

这就是重心法的递推公式。

（六）运行数据挖掘算法

根据选定的数据挖掘算法对经过处理的数据仓库中的数据进行模式提取，即数据挖掘。本例将采用 SPSS 软件进行数据挖掘。数据仓库见前表。

1、最短距离法

具体操作步骤如下：

(1)、通过 File 菜单中的 Open 中的 Date...项，弹出 OpenFile 对话框。从对话框中，选择超市文件，点击打开（图 3-3）。

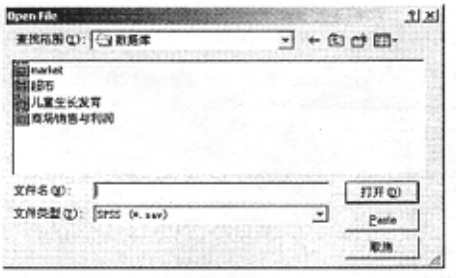


图 3-3

(2)、激活 Analyze 菜单选 Classify 中的 Hierarchical Cluster...项，弹出 Hierarchical Cluster Analysis 对话框。（图 3-4）

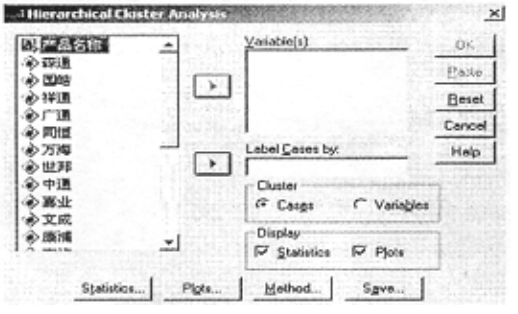


图 3-4

从对话框左侧的变量列表中选中除产品名称外的全部变量，点击⇒钮使之进入 Variable[s]框；同时，将产品名称变量，点击⇒钮使之进入 Label Cases by:框;在 Cluster 处选择聚类类型的 Case 项。

(3)、点击钮 Statistics...，弹出 Hierarchical Cluster Analysis：Statistics 对话框，选择 Proximity matrix，要求显示距离矩阵，同时在 Cluster Membership 选项框中

选择 Range of solutions，在 Minimum number of clusters:输入 4，在 Maximum number of clusters:输入 12，点击 Continue 钮返回 Hierarchical Cluster Analysis 上面的对话框。（图 3-5）

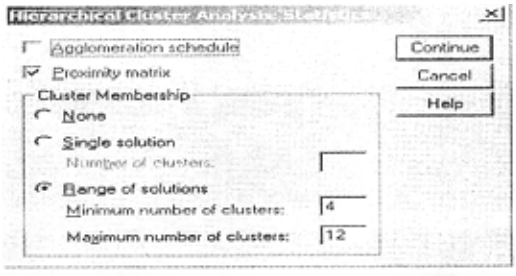


图 3-5

（4）、点击 Plots...钮弹出 Hierarchical Cluster Analysis:Plots 对话框(见图 3-6)，选择 Dendrogram 项，点击 Continue 钮返回 Hierarchical Cluster Analysis 对话框。

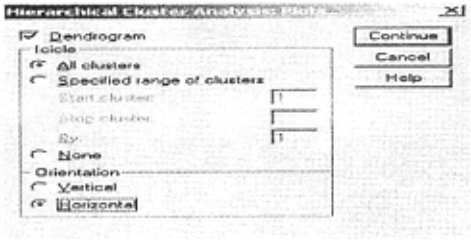


图 3-6

（5）、点击 Method...钮弹出 Hierarchical Cluster Analysis: Method 对话框，选择 Nearest neighbor。点击 Continue 钮返回 Hierarchical Cluster Analysis 对话框。（图 3-7）

图 3-7

（6）、点击 OK 钮，即完成分析。

2、最长距离法

最长距离法与最短距离法方法相似，唯一不同的是在第五步，选择 Furthest Neighbor 方法。

3、中间距离法

中间距离法与最短距离法方法相似，唯一不同的是在第五步，选择 Median clustering 方法。

4、重心法

重心法与最短距离法方法相似，唯一不同的是在第五步，选择 Centroid clustering 方法。

（七）、结果输出

1、样品量统计

见（表 3-3）

表 3-3　样品量统计

Case Processing Summary(a)

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
71	100.0	0	.0	71	100.0

a　Single Linkage

2、样品之间的距离矩阵（局部）

见（表 3-4）

3、全部样品聚类结果：

见（表3-5）

4、对结果进行分析

本例选择了 20 位顾客对 71 种商品的采购情况，用最短距离法进行了从 4 类到 12 类的聚类过程。从聚类的结果中，我们可以发现如下知识：

1、从聚类结果我们得知，在鸡肉、烤肉酱、小米、柠檬汁、花生、蛋糕、海鲜酱、猪肉、苏打水、绿茶、薯条、蚵、白米、黄豆、浓缩咖啡、蟹三合一麦片、虾子、苹果汁、苏澳奶酪、柳橙汁、巧克力、燕麦、盐、胡椒粉、鱿鱼、虾米、义大利奶酪、饼干、黑奶酪、运动饮料酸、龙虾、海参、沙茶、鸭肉、浪花、奶酪、蕃茄酱、民众奶酪、绿豆糕、麻油、辣椒粉、牛奶、糯米、棉花糖、　盐水鸭、蜜桃汁、黄鱼、肉松、干贝、　甜辣酱、雪鱼、鸡精、玉米片、矿泉水、糙米、德国奶酪、白奶酪、牛肉干、啤酒这 58 类商品中，我们从聚 4 类到聚 12 类中，它们一直属于第 1 类，因此，我们可以得出结论，这 58 类商品中有相当高的关联性，在陈列商品时，应将这 58 类商品放在一起。这样一方面，有利于顾客的购物，减小顾客的购物成本；另一方面，有利于提高商品的销售量，顾客在购买一种商品时，极有可能也购买另一种商品。

表 3-4　样品之间的距离矩阵

Case	1:鸡肉	2:烤肉酱	3:小米	4:柠檬汁	5:花生	6:墨鱼	7:蛋糕	8:猪肉干
1:鸡肉	.000	2695.000	2416.000	1535.000	7151.000	8306.000	3078.000	7308.000
2:烤肉酱	2695.000	.000	1947.000	2290.000	4936.000	11557.000	4377.000	9051.000
3:小米	2416.000	1947.000	.000	1867.000	5843.000	10570.000	2862.000	5948.000
4:柠檬汁	1535.000	2290.000	1867.000	.000	5586.000	10457.000	2507.000	5811.000
5:花生	7151.000	4936.000	5843.000	5586.000	.000	6693.000	4193.000	11107.000
6:墨鱼	8306.000	11557.000	10570.000	10457.000	6693.000	.000	5100.000	11056.000
7:蛋糕	3078.000	4377.000	2862.000	2507.000	4193.000	5100.000	.000	5278.000
8:猪肉干	7308.000	9051.000	5948.000	5811.000	11107.000	11056.000	5278.000	.000
9:海鲜酱	3823.000	2442.000	887.000	2922.000	5818.000	11665.000	3837.000	6951.000
10:猪肉	10075.000	9326.000	10031.000	10586.000	9342.000	12325.000	6969.000	10989.000
11:苏打水	3922.000	2325.000	1978.000	3805.000	7533.000	12884.000	3866.000	8006.000
12:绿茶	8223.000	7306.000	7259.000	8386.000	8042.000	11145.000	6181.000	9907.000
13:薯条	7676.000	5659.000	5912.000	7739.000	11815.000	14978.000	8852.000	7520.000
14:蚵	2183.000	2164.000	629.000	2204.000	6520.000	10263.000	2197.000	6289.000
15:白米	11759.000	13102.000	13815.000	12940.000	12518.000	14581.000	10575.000	13159.000
16:黄豆	2267.000	2394.000	2763.000	3084.000	6042.000	8473.000	3519.000	6515.000
17:浓缩咖啡	1882.000	1493.000	1726.000	1245.000	6829.000	9828.000	2356.000	6154.000
18:蟹	4719.000	7034.000	5019.000	5504.000	10530.000	12566.000	4095.000	7027.000
19:三合一麦片	7343.000	5482.000	5705.000	6962.000	10738.000	13241.000	8525.000	7959.000
20:山楂片	16290.000	16265.000	16126.000	17765.000	16901.000	17704.000	10926.000	15620.000
21:虾子	3444.000	2695.000	1412.000	2999.000	7471.000	11462.000	3168.000	5532.000

2、从聚类结果中我们也可以得知，一些商品属于某一类，随着聚类数目的增多，它有可能成为另一类。例如，像墨鱼这种商品，在聚 4 类时，其属于第 1 类，而在聚 5 到 12 类时，其属于第 2 类。由此，我们可以说这种商品相对于上述 58 类的商品关联性要低一些。

3、我们还可以发现一些奇怪的现象。例如，猪肉干和糖果，当划分 4 至 7 类时，它们均属于第 1 类商品，当划分到 8 至 12 类时，它们又都属于第 3 类商品，由此我们可以推理出这种商品在距离上具有很高的相似性，存在一定的关联性，可以考虑将这两种商品集中放在一起，是否有利于商品的销售。

4、另外，还有一些商品在我们主观认为其应为同类商品，应该放在一起，有利商品销售。但是，通过聚类分析后发现，事实情况并不是如此。例如：温馨奶酪、花奶酪和其它奶酪之间，主观认为这三种商品之间应有很高的相近性，应将奶酪产品放在一起。事实恰好相反，从聚类的结果来看，这两种奶酪和其它奶酪从始到终都不是同

一类。但是，我们不能因此就得出这种奶酪与其它奶酪不同。相反，是由于奶酪之间本身具有相当高的替代性，是由于顾客消费了其它品牌的奶酪而没有消费此品牌。因此，我们可以认为是否由于这种奶酪宣传不够，或者是摆放位置而影响销售。

5、通过聚类后，我们还发现大部商品属于第 1 类，但是，我们不能由此推出这些商品之间存在相同的关联度。从树状图，我们可以发现，干贝、玉米片、德国奶酪、雪鱼、甜辣酱、鸡精、盐水鸭、蕃茄酱、棉花糖、肉松和白奶酪它们具有几乎相等的距离，因此，我们可以认为在第 1 类中，这几种产品之间具有更高的相关性。其次，从树状图我们还可以发现小米、沙茶、海参这几种产品具有相等的距离，并且与上面所列的产品有相等的距离，但在这两小类之间似乎要比它们内部之间的关联程度要弱一些。同时，我们也发现山渣花、海苔酱、温馨奶酪、墨鱼、酱油、花奶酪、猪肉干、海鲜粉、耗油等产品它们的距离相当的大，我们也发现这些产品基本上不属于第 1 类，或者在聚类数较小的情况下，它们属于第 1 类，但随着聚类数目的增多，它们成为其它类，甚至自成一类。因此，我们可以推出这些产品与第 1 类产品有较大的距离，关联程度较弱。

6、最后，我们还发现，当划分 4 至 12 类时，大部分产品属于第 1 类，随着聚类数目的增多，第 1 类产品的数目将变的越来越小。

表 3-5　聚类结果

Cluster Membership

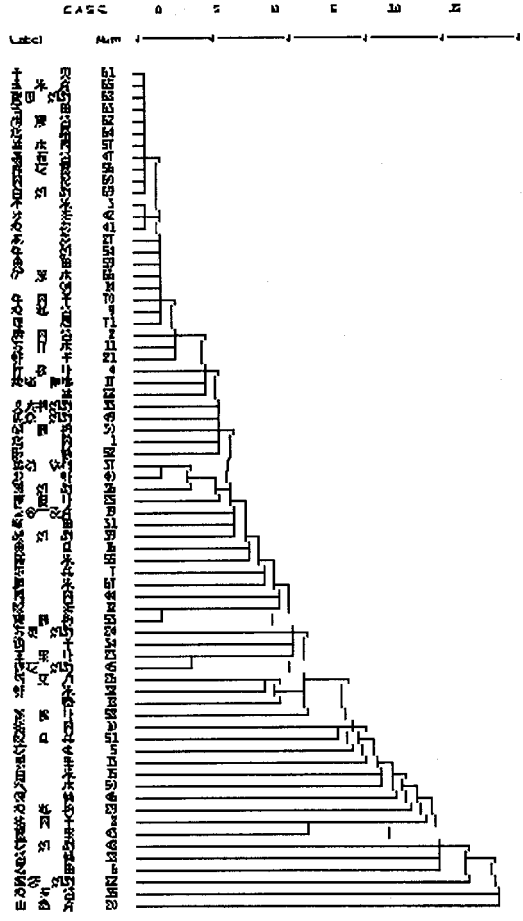
Case	12类	11类	10类	9 类	8 类	7 类	6 类	5 类	4 类
1: 鸡　　肉	1	1	1	1	1	1	1	1	1
2: 烤 肉 酱	1	1	1	1	1	1	1	1	1
3: 小　　米	1	1	1	1	1	1	1	1	1
4: 柠 檬 汁	1	1	1	1	1	1	1	1	1
5: 花　　生	1	1	1	1	1	1	1	1	1
6: 鳕　　鱼	2	2	2	2	2	2	2	2	1
7: 蛋　　挞	1	1	1	1	1	1	1	1	1
8: 猪 肉 干	3	3	3	3	3	1	1	1	1
9: 海 鲜 酱	1	1	1	1	1	1	1	1	1
10: 猪　　肉	1	1	1	1	1	1	1	1	1
11: 苏 打 水	1	1	1	1	1	1	1	1	1
12: 绿　　茶	1	1	1	1	1	1	1	1	1
13: 酱　　条	1	1	1	1	1	1	1	1	1
14:　　柯	1	1	1	1	1	1	1	1	1
15: 白　　米	4	1	1	1	1	1	1	1	1
16: 黄　　豆	1	1	1	1	1	1	1	1	1
17: 浓缩 咖啡	1	1	1	1	1	1	1	1	1
18:　　盐	1	1	1	1	1	1	1	1	1
19: 三合一麦片	1	1	1	1	1	1	1	1	1
20: 山 渣 片	5	4	4	4	4	3	3	3	2
21: 虾　　子	1	1	1	1	1	1	1	1	1
22: 蓝莓 奶酪	6	5	5	5	5	4	4	4	3
23: 苹 果 汁	1	1	1	1	1	1	1	1	1
24: 布康 奶酪	1	1	1	1	1	1	1	1	1
25: 柳 橙 汁	1	1	1	1	1	1	1	1	1
26: 巧 克 力	1	1	1	1	1	1	1	1	1
27: 燕　　麦	1	1	1	1	1	1	1	1	1
28:　　盐	1	1	1	1	1	1	1	1	1
29: 海 鲜 酱	7	6	6	6	1	1	1	1	1
30: 胡 椒 粉	1	1	1	1	1	1	1	1	1
31: 鱿　　鱼	1	1	1	1	1	1	1	1	1
32: 虾　　米	1	1	1	1	1	1	1	1	1
33: 艾人利奶酪	1	1	1	1	1	1	1	1	1
34: 饼　　干	1	1	1	1	1	1	1	1	1
35: 海 苔 酱	8	7	7	7	6	5	5	5	4
36: 里 奶 酪	1	1	1	1	1	1	1	1	1
37: 运动　饮料	1	1	1	1	1	1	1	1	1
38: 酱　　油	9	8	8	8	7	6	6	1	1
39: 酸 奶 酪	1	1	1	1	1	1	1	1	1
40: 花　　生	1	1	1	1	1	1	1	1	1
41: 海　　参	1	1	1	1	1	1	1	1	1
42: 沙　　茶	1	1	1	1	1	1	1	1	1
43: 腰　　果	3	3	3	3	3	1	1	1	1
44: 鸭　　肉	1	1	1	1	1	1	1	1	1
45: 浪花 奶酪	1	1	1	1	1	1	1	1	1
46: 鲍　　鱼	10	9	9	1	1	1	1	1	1
47: 番 茄 酱	1	1	1	1	1	1	1	1	1
48: [illegible]	11	10	10	9	8	7	1	1	1
49: 民众　奶酪	1	1	1	1	1	1	1	1	1
50: 汽　　水	12	11	1	1	1	1	1	1	1

5、树状图

******HIERARCHICAL CLUSTER ANALYSIS******

Dendrogram using Single Linkage

Rescaled Distance Cluster Combine



结　论

本文通过分析聚类分析在数据挖掘中的应用，将统计理论与现代的计算机技术进行了有机的结合。从而可以利用统计理论作为指导，计算机技术作为工具，能够对大容量的数据进行处理，从而克服了统计学对大容量数据处理的局限性。但是，需要说明的是，要完全将统计理论与计算机技术进行结合还有一定的距离，这并不是说计算机技术或统计理论有何问题。问题在于统计软件在处理数据时对数据格式有一定的要求，只有符合这种格式的数据，才有可能运用统计软件进行分析，进行数据挖掘。然而，我们目前的数据格式大都不符合这一要求，仅仅是将数据予以存储，能满足查询、汇总等简单的需要，而对其进一步的分析尚不能满足。因此，这就要求我们在收集及存储初始数据时，一定要考虑到数据的进一步利用，为数据的进一步利用作好准备。真正做到，将数据变成信息，以便于管理者进行决策所有；而不是只见数据，不见信息和知识，数据如山而知识贫乏的状况。

参考文献

[1]雷钦礼：《经济管理多元统计元析》，中国统计出版社，2002。

[2]米子川：《统计软件方法》，中国统计出版社，2002。

[3]贺铿：《经济计量学》，中国统计出版社，1999。

[4]Mehmed Kantardzic 著，《数据挖掘——概念、模型、方法和算法》，闪四清、陈茵、程雁等译，清华大学出版社，2003。

[5]Richard J.Roiger、Michael W.Geatz 著，《数据挖掘教程》，翁敬农译，清华大学出版社，2003。

[6] 李松辉、戚昌文、周祖德："企业的客户数据挖掘系统设计"，《武汉理工大学学报》，2003.9

[7] 黄永锋、刘同明："聚集式聚类分析方法及其应用"，《华东船舶工业学院学报》，2002.8。

[8] MichelineKambe:《数据挖掘概念与技术》，机械工业出版社，2001。

[9]罗晓沛，"数据挖掘在科学数据库中的应用探索"，《中国科技大学》，2002。

[10]陈佳，《信息系统开发方法教程》，清华大学出版社，1998。

[11]黄梯云，《管理信息系统》，高等教育出版社，1999。

聚类分析在数据挖掘中的应用

作者：许存兴
学位授予单位：山西财经大学

相似文献(10条)

1.学位论文 席景科 数据挖掘中聚类分析的研究与实现 2003

该文首先概述了数据挖掘的概念、分类和数据挖掘过程;其次,介绍了聚类分析的定义,对聚类分析算法进行了系统地归纳和总结,并简要介绍了每一种代表性算法的实现思想及其优点和不足;然后,重点讨论了k-prototypes算法——一种能对数值型和分类型混合属性数据集进行聚类的算法,在此基础上,提出了基于k-prototypes算法的改进算法,并使用实验室数据集对改进算法进行了测试,证明新算法是有效的;最后,根据目前的研究状况,提出了聚类分析技术需要进一步的研究方向.

2.学位论文 周东华 数据挖掘中聚类分析的研究与应用 2006

数据挖掘是目前信息领域和数据库技术的前沿研究课题，被公认为是最具发展前景的关键技术之一。数据挖掘涉及到统计学、人工智能(特别是机器学习)、模糊理论和数据库技术等多种技术，它强调的是大量数据和算法的可伸缩性，是一门很接近实用的技术，其技术含量比较高，实现难度也较大。　　聚类分析是数据挖掘的重要功能之一，近年来在该领域的研究取得了长足的发展，出现了许多聚类分析方法，如划分聚类方法、层次聚类方法、基于密度的聚类方法、基于网格的聚类方法、基于模型的聚类方法等。这些方法所涉及的领域几乎遍及人工智能科学的方方面面，而且在特定的领域中，特定的情形下取得了良好的效果。但是当处理数大量数据、具有复杂数据类型的数据集时，仍存在若干尚未解决的问题。
本文系统地研究了数据挖掘的概念、功能、处理过程及技术算法，数据挖掘的核心技术是数据挖掘的算法，本文就数据挖掘的算法做了分析和比较，选取了K—平均算法和IDBSCAN算法做了深入的研究，并给出了一种基于距离的异常数据挖掘算法。本文以山西省一所高职院校的学生成绩数据为背景，通过数据预处理工作，应用以上几种算法对上述数据进行了聚类分析，实现了可视化，最终挖掘到一定价值的信息。

3.学位论文 陈志强 数据挖掘中聚类分析技术的研究及应用 2004

随着计算机技术的发展,数据库得到了广泛应用.在数据库中积累了大量可用的数据,但是数据库管理系统却没有提供有效的工具和方法来分析和利用这些数据,如何充分利用这些数据,进行决策支持成为当今需要深入研究的课题.数据库的知识发现或数据挖掘随之出现,成为有效利用数据,进行数据分析的有力武器.聚类分析是数据挖掘的重要组成部分,应用范围非常广泛,其常规应用包括:模式识别、空间数据分析、图像处理、经济学(尤其是市场研究方面)、www文档分类等等,因而成为各界研究的对象.本文首先阐述了数据挖掘技术,是本文的研究背景和基石;接着讨论了聚类分析技术,是本文研究的核心和关键;最后在以上两部分基础之上开展应用研究,是本文理论结合实际、多种技术和知识综合应用的体现.由此本文通过对以上各部分的研究,提供了一个实用的、有效的数据挖掘中聚类分析技术应用的参考模型,可供准备开展数据挖掘项目的广大企业用户参考使用,这也正是本文研究的创新和目的所在.本文分为五章,内容结构如下:第1章绪论介绍了本文的研究背景,所做的主要工作和本文的内容结构.第2章数据挖掘技术概述详细介绍了数据挖掘理论和技术,包括:数据挖掘和数据库知识发现定义,数据挖掘起源,数据挖掘处理过程模型,数据挖掘技术,数据挖掘分类,数据挖掘与相关学科的区别与联系,数据挖掘研究现状及存在的问题.第3章聚类分析技术概述详细介绍了聚类分析技术,包括:聚类分析综述,聚类分析中常用的两种数据结构,聚类分析中常用的几种数据类型,聚类分析中的常用方法.第4章聚类分析应用研究在以上两章的基础上,结合一个具体的数据挖掘任务实例,研究了数据挖掘应用的解决方案,包括:数据挖掘技术应用方案研究,聚类分析应用实例,实例系统的技术重点与难点及不足,实例系统的使用介绍.第5章结束语总结全文,给出本文研究中存在的问题和今后研究工作的方向.

4.学位论文 吴晓彬 数据挖掘中金融时间序列的粗糙聚类分析 2008

传统统计分析与现代金融计量经济方法研究时间序列的主要思路是建立基于严格数学推导下的统计模型并对其进行参数估计与数据检验，目前已建立起一套较为成熟的理论体系。但该方法既依赖于苛刻的假设条件，又要求所有数据都符合一个固定的数学模型，显得过于牵强。数据挖掘研究时间序列的思路则不同，它由数据直接驱动建立模型，克服了上述的缺陷。　　时间序列数据挖掘已是当前的研究热点之一，人们也取得不少的研究成果，但对于时间序列相似性度量这一关键难题一直未能得到较好的解决，而很多时序挖掘方法都是建立在相似性的基础上，显然时间序列相似性度量直接影响着

果，为此本文首先就该关键的基础性问题展开研究，进一步讨论了该度量方法在序列挖掘中的应用。由于数据挖掘方法众多，本文不可能一一涉及，所以只针对聚类分析进行深入的探讨。聚类分析不仅是数据挖掘的重要组成部分，同时也是多元统计分析的重要方法，在实际中有广泛的运用。本文绕开了已有较多成熟方法的硬聚类，而深入地研究了一种软聚类——粗糙聚类的方法及其在时间序列挖掘中的应用，同时从侧面反映了本文度量序列相似性方法的实用性。全文的主要工作及创新可归纳为以下几点。　　首先，结合小波分析的思想方法，提出一种基于小波多尺度变换的时间序列相似性度量方法，并通过金融时间序列的实例研究，说明该方法全面考虑了影响序列相似性度量的各种因素，很好地克服了已往方法无法兼顾序列整体形状轮廓与细节差异的缺陷。　　其次，在相似性度量方法的基础上，研究了序列粗糙聚类方法，通过金融实证研究表明粗糙聚类方法的优点。并深入研究了以下三个问题：　　(1)建立粗糙聚类质量指标，并研究不同阈值参数对聚类结果的影响；　　(2)将粗糙聚类法与层次聚类法进行整合，各取所长；　　(3)将软聚类转化为硬聚类，通过迭代剔除法对粗糙聚类结果精简化，并与之前聚类结果进行比较，说明其可行性。　　最后，本文模型方法尚无现成的软件模块实现，故本文还给出Matlab软件上具体实现的参考程序，结合实证研究取得较好的效果。

5.学位论文 李浪波 聚类分析在科学数据挖掘中的应用研究 2006

如何让各种数据挖掘技术更好地为实际工程所服务，一直是数据挖掘领域的一个挑战。一方面是人们对快速、准确而全面获取信息的渴望，而另一方面却是各种信息的纷繁芜杂，在这两者之间架设一座桥梁的确是一个巨大的挑战。聚类分析在数据挖掘技术中占有重要的位置。所谓聚类，是将一个数据单位的集合(数据源)分割成几个称为类或类别的子集，每个类内的对象之间是相似的，但不同类的对象间区别相对较大。聚类分析是在没有先验知识支持的前提下，根据事物本身的特性研究被聚类对象的类别划分，实现满足这种要求的类的聚合，它所依据的原则是使同一类中的对象具有尽可能大的相似性，而不同类中的对象具有尽可能大的差异性。　　论文基于大规模核物理科学数据挖掘的背景，全面介绍了数据挖掘的关键技术和主要任务，从理论、算法和应用三个层次，结合科学数据的特点来分析预处理技术和聚类方法，提出了很多实用的预处理方法：对HDF5科学数据进行分块、除噪、集成、变换等，同时对它使用“截断法”和“逐层求差法”进行规约，并对数据进行信息提取。在聚类方面，经过比较各种聚类算法和分析科学数据的特点，提出了结合k一平均思想的改进型系统聚类算法。此聚类算法有如下特点：能生成具有代表性的数据簇中心；使用相似系数计算距离，避免了距离受量纲影响的缺点；不需要多次迭代计算，减少了计算量；不需要指定初始中心；改进了聚类图，更容易得出聚类阀值。实验结果表明这种改进的系统聚类算法非常适合科学数据的处理。　　本文最后简单介绍了我们开发的科学数据挖掘系统。其中重点介绍了聚类分析模块的设计和功能。

6.学位论文 张国云 数据挖掘中的聚类分析及其在控制中的应用研究 2002

该文对数据挖掘中的聚类分析方法及其在工业过程控制中的应用研究作了偿试,重点研究了基于统计理论的聚类分析理论和方法,模糊聚类分析理论和方法及模糊Kohonen网络(FKN)的结构与学习算法,即模糊C——均值算法与自组织特征映射神经网络(Kohonen网络)的有机融合,并根据硬分类思想及软分类思想提出了改进的模糊Kohonen网络(IFKN),通过Matlab编程对人工合成控制时序图数据集进行聚类分析,其聚类效果与当今广泛使用的数据挖掘软件平台,德国MIT公司著名的DataEngine智能数据分析和数据挖掘软件的聚类效果相当,最后,论述了聚类分析在控制中的应用,它可以用于过程控制中的参数变化趋势的模式识别及图象分割处理等具体应用中,并以贵州铝厂熟料烧结回转窑自动控制系统为工程背景,利用IFKN识别其热工参量变化趋势,取得了较理想的效果.把数据挖掘技术用于工业过程控制,是自动化领域中一项崭新的研究内容,而且也是自动化领域很值得研究的新课题.

7.学位论文 李婷 聚类分析在交通流时序数据挖掘中的应用 2007

在中国智能交通系统快速发展的大背景下，如何利用丰富的交通检测数据进行交通规划及控制优化是交通控制领域中需要解决的重要问题之一，而数据挖掘技术的发展正好给海量交通数据的知识发现提供了理论基础和技术实现手段。　　数据挖掘指的是从大量数据中提取出有效的、新颖的、潜在有用的，以及最终可理解的模式的高级过程。聚类分析是数据挖掘研究的一个十分重要的方面。在智能交通系统中，时间序列是最常见的数据形式。采用聚类分析方法对海量交通流时间序列进行研究分析具有很大的应用价值，一方面可以发现典型的交通流变化趋势规律，另一方面可以对检测点具有不同交通流特性的时段进行合理分组，进而针对各时段制定出相应的交通控制策略，同时还可进一步结合空间信息发现一些有意义的交通流时空分布规律。　　本文详细介绍了聚类算法在时间序列数据挖掘中的研究和应用，借助某市网络化智能交通系统数据库中的历史车流量数据，结合聚类分析理论，对快速路段上两个月的单点车流量数据进行了聚类分析。作者将传统聚类算法进行改进，提出了一种阶梯型系统聚类方案，在得到交通流量分类初始模式的基础上进一步挖掘具有相应交通流特征的时段模式，最终得到了较为满意的实验结果，发现了有意义的交通流时间分布模式并以C++、MATLAB程序实现了阶梯型系统聚类挖

了此次聚类结果，并提出了进一步的研究方向。

8.期刊论文 陈学进.CHEN Xue-jin 数据挖掘中聚类分析的研究 -计算机技术与发展2006,16(9)

聚类分析是由若干个模式组成的,它在数据挖掘中的地位越来越重要.文中阐述了数据挖掘中聚类分析的概念、方法及应用,并通过引用一个用客户交易数据统计出每个客户的交易情况的例子,根据客户行为进行聚类.通过数据挖掘聚类分析,可以及时了解经营状况、资金情况、利润情况、客户群分布等重要的信息.对客户状态、交易行为、自然属性和其他信息进行综合分析,细分客户群,确定核心客户.采用不同的聚类方法,对于相同的记录集合可能有不同的划分结果对其进行关联分析,可为协助各种有效的方案,开展针对性的服务.

9.学位论文 尹波 聚类分析及其在移动通信企业数据挖掘分析中的应用研究 2008

数据挖掘是指从数据库中发现隐含的、新颖的、有用的信息的过程，已经在许多领域得到了广泛的应用。聚类分析是数据挖掘的主要技术手段之一，至今已在理论和方法上取得了丰硕的研究成果。随着近年来数据密集型企业数据仓库等决策支持系统的建设以及企业对商业智能的需求，数据挖掘面临新的应用，聚类分析研究也面临更多新的内容和挑战。移动通信企业是典型的数据密集型企业，随着电信市场竞争的不断加剧，如何对客户进行细分和分类、并针对不同的客户群实施差异化营销和服务，已成为当前电信企业的迫切需求。本文针对移动通信企业的客户细分需求以及数据特性，研究和提出一种针对混合属性数据的聚类算法，并将其应用于移动通信企业的客户细分，在此基础上提出了基于客户细分的市场营销方法。所做工作归纳如下：　　1.介绍了数据挖掘技术，详细论述了数据挖掘中的聚类分析，总结了聚类分析的方法、特点和分类，重点讨论了混合属性数据聚类，具体研究了模糊K-Prototypes(FKP)算法，并指出了它的优缺点。　　2.针对模糊K-Prototypes算法对初始值敏感、容易陷入局部极小值的问题，提出了一种基于粒子群优化(Particle Swarm Optimization，PSO)算法和FKP算法有机结合的混合聚类算法。该算法首先利用PSO算法确定FKP的初始聚类中心，再将PSO聚类结果作为后续FKP算法的初始值。实验结果显示，新算法具有良好的收敛性和稳定性，聚类效果优于单一使用FKP算法。另外考虑到样本矢量中各维特征对模式分类的不同影响，采用了ReliefF算法对特征进行加权选择。　　3.研究了聚类技术在移动通信企业客户细分领域的应用。论述了客户细分的基本理论、方法和步骤，建立了基于客户行为特征/消费心理的细分模型，在对湖南移动经营分析系统的客户信息原始数据进行商业理解之后，聚类技术实现了客户群的一种细分，并将该细分模型成功地应用于市场营销过程中的决策支撑。

10.学位论文 邓冰 宝钢实用数据挖掘系统的设计及聚类分析在质量控制中的应用研究 2003

二十世纪九十年代起,数据挖掘的应用与研究成为国际上的热点,数据挖掘的主要目的就是从大量的数据中获取知识,为决策支持服务.值得注意的是,数据挖掘是应用驱动的,这就意味着,对具体的应用领域,应该度身定制自己的数据挖掘方案.针对冶金行业的质量控制领域,我们从项目实践中得出适合宝钢特点的数据挖掘方案,并把它应用与质量控制,取得了一定的成功. 该文主要工作在于:1.针对质量控制领域,将通常的数据挖掘方案具体化,给出它的框架、特点和实施步骤;2.研究了适合宝钢数据特点的数据预处理方法;3.研究了各种聚类方法和特性,找出了K均值法、类平均法和WARD方法作为适合宝钢数据特点的聚类方法;4.针对实际问题,提出定性调优和定量调优,前者能指导质量工程师在进行质量控制时调整哪些变量,后者具体给出可调变量的范围;5.开发基于SAS的数据挖掘软件.