

# 数据挖掘中聚类分析的研究

陈学进

(合肥工业大学 计算机与信息学院, 安徽 合肥 230009;  
安徽工业大学 计算机学院, 安徽 马鞍山 243002)

**摘要:**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。

**关键词:**数据挖掘;聚类分析;客户行为

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2006)09-0044-02

## Research of Cluster Analysis in Data Mining

CHEN Xue-jin

(Computer and Information College of Hefei University of Technology, Hefei 230009, China;  
Computer College, Anhui University of Technology, Maanshan 243002, China)

**Abstract:** Cluster analysis is made up of patterns, and becoming increasingly essential in data mining field. This paper briefly introduces the basic concept, means and application of cluster analysis discussing about cluster analysis by using a case of customer transaction. In order to know about much important information of running, funds, profits and customers. And analyze state of client, bargaining action, natural attribute and other information, subdivide customer groups and fix on core client. By using various methods of cluster analysis, it is effective project to develop pertinence service.

**Key words:** data mining; cluster analysis; customer action

### 0 引言

自20世纪60年代数据库系统诞生以来,数据库技术已经得到了飞速的发展,并且已经深入到社会生活的各个方面。现在,数据无处不在,可以存放在不同类型的数据仓库中,数据仓库技术可以将异构的数据库集成起来进行综合管理,从而提供更好的服务。

但是,随着科学技术的进步,新的数据采集和获取技术不断发展,使得数据库中所存储的数据量也随之急剧增长。另一方面,数据处理技术的发展却相对落后,数据库技术仍然停留在相对简单的录入、查询、统计、检索阶段,对数据库中的数据之间存在的关系和规则、数据的群体特征、数据集内部蕴涵的规律和趋势等,却缺少有效的技术手段将其提取出来,从而出现所谓的“被数据淹没,却饥渴于知识”(John Naisbett, 1997)的现象<sup>[1]</sup>。为了解决这种现象,科学家们于20世纪80年代末期创立了一个新的研究

领域,即数据挖掘(Data Mining),或称数据挖掘和知识发现(Data Mining and Knowledge Discovery, DMKD)。这是在数据库技术、机器学习、人工智能、统计分析等基础上发展起来的一个交叉性的学科。区别于简单地从数据库管理系统检索和查询信息。数据挖掘是指“从数据中发现隐含的、先前不知道的、潜在有用的信息的非平凡过程”(Frawley, 1991),其目的是把大量的原始数据转换成有价值的、便于利用的知识。

自从数据挖掘和知识发现的概念于1989年8月首次出现在第11届国际联合人工智能学术会议以来,数据挖掘和知识发现领域的研究和应用均得到了长足的发展,形成了一些行之有效的理论和方法,并逐渐成为计算机信息处理领域的研究热点。

数据挖掘(Data Mining)是一个多学科交叉研究领域,它融合了数据库(Database)技术、人工智能(Artificial Intelligence)、机器学习(Machine Learning)、统计学(Statistics)、知识工程(Knowledge Engineering)、面向对象方法(Object-Oriented Method)、信息检索(Information Retrieval)、高性能计算(High-Performance Computing)以及数据可视化(Data Visualization)等最新技术的研究成果<sup>[2,3]</sup>。

收稿日期:2005-11-09

**作者简介:**陈学进(1972-),男,安徽六安人,讲师,硕士研究生,研究方向为计算机软件理论及数据挖掘;导师:胡学钢,博士,教授,研究方向为知识工程、数据挖掘、数据结构。

## 1 数据挖掘中聚类分析

利用数据挖掘技术可以分析各种类型的数据,例如结构化数据、半结构化数据以及非结构化数据、静态的历史数据和动态数据流数据等。关系数据库 (relational database) 中通常存储和管理的是结构化的数据,它将一个实体的各方面信息通过离散的属性进行描述;而文本数据库 (text database) 或文档数据库 (document database) 则通常存储和管理的是半结构化的数据,例如新闻稿件、研究论文、电子邮件、书籍以及 WEB 页面等属于半结构化数据;空间数据库、多媒体数据库中存放的是非结构化数据,例如地图、图片、音频、视频等都属于非结构化数据。相对半结构化和非结构化数据来说,针对结构化数据的数据挖掘技术比较成熟,市场有很多的商品软件可以使用,其中用得较多的包括 IBM Intelligent Miner, SAS Enterprise Miner, SGI MineSet, Clementine SPSS 及 Microsoft SQL Server 2000 等。

聚类分析方法即根据实体的特征对其进行聚类或分类,进而发现数据集的整个空间分布规律和典型模式的方法。常用的聚类方法有 K-mean, K-medoids 方法;Ester 等提出的基于 R-树的数据聚焦法及发现聚合亲近关系和公共特征的算法;周成虎等提出的基于信息熵的时空数据分割聚类模型等。

聚类分析是根据“物以类聚”的原理,将本身没有类别的样本聚集成不同的组,并且对每一个这样的组进行描述的过程。其主要依据(即目的)是聚到同一个组中的样本应该彼此相似,而属于不同组的样本应该足够不相似。与分类分析不同,进行聚类前并不知道将要划分成几个组和什么样的组,也不知道根据哪些空间区分规则来定义组。其目的旨在发现空间实体的属性间的函数关系,挖掘的知识用以属性名为变量的数学方程来表示。聚类方法包括统计方法、机器学习方法、神经网络方法和面向数据库的方法。基于聚类分析的空间数据挖掘算法包括均值近似算法,CLARANS, BIRCH, DBSCAN 等算法。目前,对空间数据聚类分析方法的研究是一个热点。

以客户关系管理为例,利用聚类技术,根据客户的个人特征以及消费数据,可以将客户群体进行细分。例如,可以得到这样的一个消费群体:女性占 91%,全部无子女、年龄在 31 到 40 岁占 70%,高消费级别的占 64%,买过针织品的占 91%,买过厨房用品的占 89%,买过园艺用品的占 79%。针对不同的客户群,可以实施不同的营销和服务方式,从而提高客户的满意度。

对于空间数据,利用聚类分析方法,根据地理位置以及障碍物的存在情况可以自动进行区域划分。例如,根据分布在不同地理位置的 ATM 机的情况将居民进行区域划分,根据这一信息,可以有效地进行 ATM 机的设置规划,避免浪费,同时也避免失掉每一个商机。

对于文本数据,利用聚类技术可以根据文档的内容自动划分类别,从而便于文本的检索<sup>[4,5]</sup>。

## 2 利用聚类做客户行为分析

### 2.1 分析目标

用客户交易数据统计出每个客户的交易情况,根据客户行为进行聚类。通过对客户数据进行聚类,将客户进行分群,考察每类客户的对证券公司的贡献情况,这样可以客户产生类别的交易行为等其他特点知道该类用户是否对公司最有价值,并且证券公司根据客户行为的特点对贡献度大的客户类采取相应的政策照顾,并且还能吸引某些行为类似的贡献度较低类的客户发展为较高贡献的客户。

### 2.2 数据解释

对交易数据(变量描述见表 1)进行数据总结,生成客户股票交易行为数据表。变量描述见表 2。使用的数据包括股票代码,买卖股票的最大、最小数量/金额,平均价格,总金额等。

表 1 聚类总结

类别	类中元素数目	均方偏差	到聚类种子的最大距离	最近类	到最近类的距离
1	1379	0.302471	6.210544	4	1.998038
2	371	2.015032	105.2028	1	2.255088
3	5	2.861125	10.78603	4	12.33607
4	560	0.540016	14.99704	1	1.998038

表 2 每类的特征(各个变量的平均值)

类别	交易次数	股票种类	卖股票数	卖平均价格	卖总金额	买股票数	买平均价格	买总金额
1	19.09292	7.915535	73159.4	9.749066	480011	70831	11.6	70831
2	179.3968	34.51056	1421653	12.83477	9453650	1408731	13.6	1408731
3	4.601296	2.80039	446.0216	87.66555	34432	2919	76.8	2919
4	25.50872	10.26739	33911.36	19.89891	1739863	27917	21.3	27917

### 2.3 步骤

文中使用数据挖掘的聚类算法,聚类数为 4,聚类准则采用 Newton,其流程如图 1 所示。

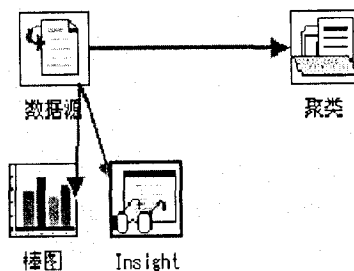


图 1 客户聚类数据挖掘流程图

### 2.4 模型结果

聚类后给数据集增加了新的类别变量,标志客户所属的类别。

## 3 结束语

通过数据挖掘,可以及时了解营业状况、资金情况、利

(下转第 49 页)

可以很容易地直接映射到这种关系模型,并且应用程序能够权衡 SQL 针对特殊查询和数据访问的能力。这里重点是数据。

#### 4.2 何时在应用程序中使用 JDO

对于以对象为中心的应用程序,它的主要目的是处理相互关联的数据的复杂图或者层次结构,JDO是最适合开发这类应用程序的工具。对象模型可以用于表示相互关联的复杂图或者层次结构,这里应用程序可以通过访问这些数据来完成处理。此处的重点是对象模型。

#### 4.3 何时在应用程序中使用 JDO 和 JDBC

有时候,你可能希望把 JDO 代码和一些 JDBC 结合起来使用。虽然应该避免这种混合的体系结构(除非有很好的理由使用它,而且这种体系结构严重限制了其他数据仓库的可移植性),但它在技术上是可行的。一个方法是使用实现特有的 API 扩展来使用 SQL 而不是 JDO SQL(或者两者混合使用)作为查询语句。另一个方法是根据应用程序需要使用混合体系结构的原因来使用 JDO 实现特有的方法获得 Connection。Connection 应该通常来自于某个 JDO 的 PersistenceManager。

### 5 结束语

JDO 是纯 Java 的 API<sup>[8]</sup>,并且是为透明对象的持久性而设计的高级 API。程序员通常会编写 Java 代码,并像往常一样通过 Java 对象进行工作。对象中的数据总是被自动持久化,可以在应用程序的不同运行中重复使用,可被查询,甚至其他应用程序也可以访问这些数据。所有这些都可以通过一个相对透明的方式来实现<sup>[9]</sup>。它综合了以前所有持久存储 APIs 的最好的特征,而没有相关的缺陷,它使得应用程序开发人员可以做到“只写一次,处处持久”。

JDBC 是专门为访问关系数据库而设计的低级 API,因为在任何地方都不可能把关系数据库以任何方式“隐藏”起来<sup>[9]</sup>,而且它通过一个基于 SQL 的接口提供了可靠性和可测量性。JDBC 被应用于关系模型而使得它很难和

应用程序中的对象模型相结合,这是因为 JDBC 不支持 Java 类,所以 JDBC 必须直接和 SQL 的数据模型一同工作。

通过 JDBC 你可以具体说明哪些内容应该被持久化和这一切是怎么发生的。通过 JDO 你仅仅定义了你的类中的持久部分,至于是怎么被持久化的取决于 JDO 的供应商。两个 API 显然是根据编程者心中的不同目标而设计的,虽然两个 API 都是要在高层次上为“持久数据”服务的,但实际上,它们只是通过不同的方法来达到这个目的。而且它们是互补的 API,两者都具有独特的功能,可以被不同层次的程序员使用,并可以根据不同的开发目的而应用到项目或项目的模块中。

#### 参考文献:

- [1] 刘柯,杨贯中. J2EE 工程中持久性存储技术的比较[J]. 株洲师范高等专科学校学报,2004,2(9):43-45.
- [2] Jordan D, Russell C. Java Data Objects[M]. USA: O'Reilly, 2003. 1-2.
- [3] 何成万,余秋惠. JDO 初探[J]. 计算机工程,2002,28(6):282-283.
- [4] 翟鸿鸣,张惠娟. JDO 技术研究[J]. 微机发展,2004,14(6):78-81.
- [5] Jordan D. A Comparison Between JDO, Serialization and JDBC for Java Persistence[EB/OL]. <http://www.jdocentral.com/pdf/DavidJordan-JDOversion-12Mar02pdf>, 2002-01.
- [6] 张伟燕,夏涛,席传裕. 在 Java 企业应用中选择正确的对象持久技术[A]. 中国计算机科学与技术-2004. 合肥:中国科学技术大学出版社,2004. 945-946.
- [7] 夏科军,徐良贤. JDO 实例的状态管理研究[J]. 计算机仿真,2005,22(4):269-272.
- [8] Jones B L. Data Persistence and Java Data Objects—JDO[EB/OL]. <http://www.developer.com/java/article.php/918111>, 2002.
- [9] Tyagi S, Vorburger M, McCammon K, et al. JDO 核心技术[M]. 北京:清华大学出版社,2005. 258-259.

(上接第 45 页)

润情况、客户群分布等重要信息。并结合大盘走势,提供不同行情条件下的最大收益经营方式。同时,通过对各营业部经营情况的横向比较,以及对本营业部历史数据的纵向比较,对营业部的经营状况作出分析,提出经营建议。通过对客户状态、交易行为、自然属性和其他信息的综合分析,细分客户群,确定核心客户。同时通过对其进行关联分析,可为协助制定各种有效的营销方案,开展针对性的个性化服务。今后还应继续采用预测、相关、关联等技术,挖掘更深层次的规律<sup>[6]</sup>。

#### 参考文献:

- [1] Berson A. 构建面向 CRM 的数据挖掘应用[M]. 北京:人

民邮电出版社, 2001.

- [2] Hand D, 张银奎. 数据挖掘原理[M]. 北京:机械工业出版社,2003.
- [3] 陈京民. 数据仓库与数据挖掘技术[M]. 北京:电子工业出版社,2002.
- [4] Han Jiawei, Kamber M. Data Mining Concepts and Techniques(影印版)[M]. 北京:高教出版社,2001.
- [5] 邹涛,戚广智,蔡丽娟,等. 网络信息挖掘系统 IIGS 的实现[J]. 南京大学学报(自然科学版),2000(2):183-188.
- [6] 胡乐群. 数据挖掘在证券行业中的应用[EB/OL]. [http://industry.ccidnet.com/art/465/20020904/24238\\_1.html](http://industry.ccidnet.com/art/465/20020904/24238_1.html), 2002-09-04.