维普资讯 http://www.cqvip.com



# 聚类分析方法及工具应用研究

王 鑫 王洪国 张建喜 胡宝芳
(山东师范大学信息管理学院 济南 250014)

**摘 要** 聚类是数据挖掘领域的一个重要的研究方向。本文介绍了聚类的基本概念及主要方法,通过具体实例对当今国际上先进的数据挖掘工具(SPSS和DBMiner)聚类的性能进行了对比,最后得出了结论。

**关键词** 数据挖掘,聚类,聚类工具,SPSS,DBMiner

## Research of Clustering Methods and Tools in Application

WANG Xin WANG Hong-Guo ZHANG Jian-xi GU Jian-Jun HU Bao-Fang
(Information Management School of Shandong Normal University, Jinan 250014)

**Abstract** Clustering is an important research direction in the field of Data Mining. This paper introduces the concept and main methods of clustering, analyzes and compares the clustering functions of overseas leading data mining tools (SPSS and DBMiner) based on the real examples. The conclusion is given in the last of paper.

**Keywords** Data mining, Clustering, Clustering tool, SPSS, DBMiner

## 1 引言

数据挖掘[1](Data Mining),也称知识发现(KDD),是从数据库中便捷地抽取出以前未知的、隐含的、有用的信息,所挖掘出来的知识可应用于信息管理、决策支持、过程控制等等。聚类分析是数据挖掘研究和应用中一个重要的部分。简单地讲,聚类分析就是将数据对象分组成多个类或簇(cluster),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。

### 1.1 聚类的应用

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十分重要的作用。聚类还可以用于对Web上的文档进行分类,以发现信息。此外,聚类可以发现孤立点,在金融领域,可以用于发现欺诈和其他金融犯罪行为等等。

### 1.2 数据挖掘领域中常用的聚类算法

聚类分析方法主要有以下几种:划分方法,层次方法,基于密度的方法,基于网格的方法和基于模型的方法等。聚类分析的典型算法如下[2~4]:

(1)划分方法的代表算法有:K-MEANS算法,K-MEDOIDS算法,CLARANS算法等;

(2)层次方法的代表算法有:BIRCH算法,CURE算法,CHAMELEON算法等;

(3)基于密度的方法的代表算法有:DBSCAN算法,OPTICS算法,DENCLUE算法等;

(4)基于网格的方法的代表算法有:STING算法,CLIQUE算法,WAVE-CLUSTER算法;

(5)基于模型的方法通常有两种方案:统计的方案和神经网络的方案。

### 1.3 国外先进的数据挖掘工具

国外比较有影响的典型数据挖掘系统有:SAS公司的Enterprise Miner、IBM公司的Intelligent Miner、SGI公司的MinerSet、SPSS公司的Clementine、加拿大Simon Fraser大学开发的DBMiner、RuleQuest Research公司的See5,还有CoverStory、EXPLORA 、KnowledgeDiscovery Workbench、Quit等[5]。

## 2 用SPSS Clementine做聚类分析

### 2.1 SPSS Clementine

SPSS[6](Statistics Package for Social Science)是世界著名的统计分析软件之一,SPSS Clementine是一个开放式数据挖掘工具,它不但支持整个数据挖掘流程,从数据获取、转化、建模、评估到最终部署的全部过程,还支持数据挖掘的行业标准——CRISP-DM。Clementine的可视化数据挖掘使得"思路"分析成为可能,即将精力集中在要解决的问题本身,而不是局限于完成一些技术性工作(比如编写代码)。它还提供了多种图形化技术,有助于理解数据间的关键性联系,指导用户以最便捷的途径找到问题的最终解决办法。

SPSS的应用领域非常广泛,在宏观经济管理、企业经营管理、行业管理与特征分析、社会科学领域等的诸多领域都有用武之地。SPSS for Windows V11.0是模块化的统计分析软件,由基本模块、分类模块、趋势模块、回归分析模块、高级模块等十余个模块组成。

### 2.2 采用的聚类方法

在SPSS中,有两类聚类分析方法:层次聚类法(Hierarchical Cluster Procedure)和迭代聚类法(Iterative Partitioning

王 鑫 硕士研究生,研究方向:数据挖掘,知识发现。王洪国 博士后,教授,硕士生导师。张建喜 硕士研究生。胡宝芳 硕士研究生。

维普资讯 http://www.cqvip.com

Procedure)。迭代聚类,在算法上需要有聚类中心,或由使用者输入各类的中心,或者机器自己确定一个初始中心。层次聚类方法[7]是聚类分析的一个重要方法。层次聚类方法可以分成凝聚(agglomerative)法和分裂(divisive)法两种。

凝聚法也称为自底向上的方法,先将每个对象作为一个单独的组,然后合并相近的对象或组,直到达到一个终止条件为止。分裂法也称为自顶向下的方法,开始时所有的对象都在一个组,然后每一步将每个组被分成更小的组,直到达到一个终止条件为止。凝聚法的主要步骤:

(1)初始化:把每个点都当作是一个类;

(2)计算所有类间的距离;

(3)把最近的两个类合并为一个类;

(4)循环执行第 2 步和第 3 步,直到最后所有的样本都合并到一个类中,或者直到最后得到的类别数目小于或等于事先给定的类别数。

分裂法采用的步骤与凝聚法正好相反,它先把所有的点都归入一个类,每一步都把类进行二分,直到每个点都属于仅包括自己的类。

### 2.3 SPSS 中的聚类分析

2.3.1 *将数据标准化* 为了消除不同变量的单位对聚类结果的影响,首先对所有的数据标准化。

2.3.2 *计算对象之间的"距离"* "距离"有多种表达,如:(1)欧氏距离(的平方);(2)偏差距离;(3)相关系数;(4)明考夫斯基距离(的 q 次方);(5)马氏距离(的平方)等等。应用者可根据实际问题,选其一种。

2.3.3 *选择类与类之间的距离定义*

(1)类的定义:由 1 个以上(含 1 个)对象组成的集合。

(2)类与类之间的距离,可由类的"代表点"之间的距离表示。类的代表点,有如下几种选择方法:a)用两个类之间的距离最近(或最远)的点,分别作为这两类的代表点。b)用两个类(类 S 与类 T)中所有点之间的距离(或距离的平方)平均值,作为两个类之间的距离。显然,后者是一种比较好的方法,类间的距离不再依赖于类内的个别点。

(3)按照某一规则,选择类中的某一点,代表该类。

2.3.4 *聚类*

(1)把每个点(对象)作为一类(称为第一层的类)。

(2)找出距离最小(或最大)的 $d_{ij}$,从而得出距离最近(或最远)的两类 $i$ 与 $j$,把它们合并成为层次最高的一类。如果同时有两个距离 $d_{ij}$ 和 $d_{st}$ 一样最小(或最大),则同时把 $i,j$ 作为一类,把 $s,t$ 作为一类;若 $i,j$ 和 $s,t$ 中有一个是相同的,则把这三个小类合并成为一个大类。如果有更多的两类之间的距离一样小(或大),可类似处理。

(3)重复做(2),直至所有的点(对象)都并成一个大类。

如果每做一次(2)之前的类是第 $k$ 层的类,那么每做一次(2)之后的类,就是第 $k+1$ 层的类。

2.3.5 *分类* 依据实际需要,确定以第几层的类为最终的分类标准。通常采用如下准则:

准则 1:各类重心之间的距离必须较大。

准则 2:各类所包含的元素都不要过分地多。

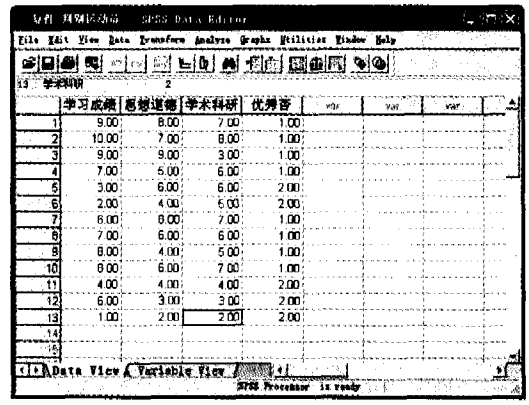
准则 3:分类的数目应该符合使用的目的。

准则 4:若采用几种不同聚类方法处理,则在各自的聚类图上应发现相同的类。

当然,不能把第一层的类,作为分类的准则,因为这时每一个初始对象都是一类,这等于没有分类。也不能把最后一层的类,作为分类的标准,因为此时所有的对象都被合并为一个类了,这也等于没有分类。在 SPSS 中我们对于这个问题一般分为三到四类即可。

### 2.3 用 SPSS 做聚类分析举例

例 某专业评选优秀研究生,主要从学习成绩、思想道德和学术科研三个方面进行考核。请依据这些分值对研究生分类。数据见图 1。



| | 学习成绩 | 思想道德 | 学术科研 | 优秀否 |
|---|---|---|---|---|
| 1 | 9.00 | 8.00 | 7.00 | 1.00 |
| 2 | 10.00 | 7.00 | 8.00 | 1.00 |
| 3 | 9.00 | 9.00 | 3.00 | 1.00 |
| 4 | 7.00 | 6.00 | 6.00 | 1.00 |
| 5 | 3.00 | 6.00 | 6.00 | 2.00 |
| 6 | 2.00 | 4.00 | 6.00 | 2.00 |
| 7 | 6.00 | 8.00 | 7.00 | 1.00 |
| 8 | 7.00 | 6.00 | 6.00 | 1.00 |
| 9 | 8.00 | 4.00 | 5.00 | 1.00 |
| 10 | 6.00 | 6.00 | 7.00 | 1.00 |
| 11 | 4.00 | 4.00 | 4.00 | 2.00 |
| 12 | 6.00 | 3.00 | 3.00 | 2.00 |
| 13 | 1.00 | 2.00 | 2.00 | 2.00 |

图 1 数据

本例中,参与聚类分析的变量为"学习成绩"、"思想道德"、"学术科研",聚类对象为 cases(样本个体),聚类方法采用组间连接法 Between-groups linkage(依据不同类之间个体的距离的平均值,进行聚类合并),相似性测度选择 Square Euclidean distance(欧几里得空间距离的平方)。因为本例的数据都是分值,没有不同变量的不同计量单位的影响,所以不用再选择标准化数据的方法。

利用 SPSS 对我们上面的统计数据进行分析的结果用图表表示如下:

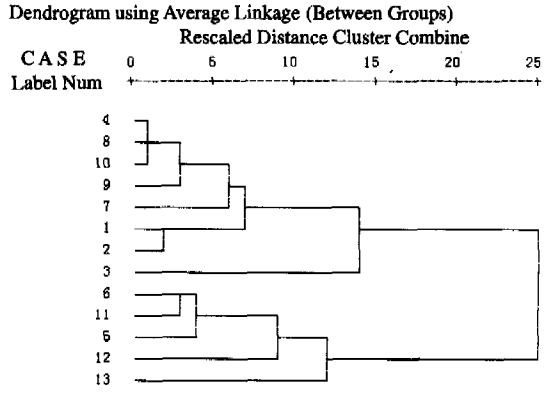



图 2 聚类树形图

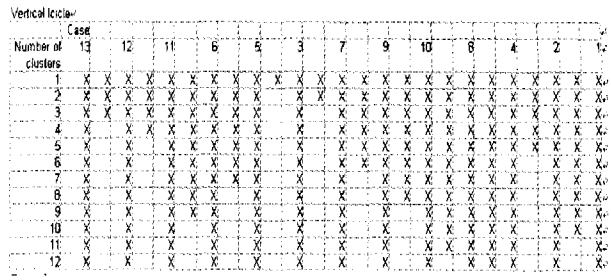



图 3 聚类冰状图

聚类树形图(图 2):显示了在什么尺度(例如欧几里得空间距离)上,哪些个体被聚为一类。

维普资讯 http://www.cqvip.com

聚类冰柱图(图 3):横坐标表示被聚类的个体。每个个体占一列,都用"×"符号占满,成柱状。每个相邻的两个个体之间,由一个没有个体号的列(柱)隔开。"×"号在哪个高度出现,表示从那个高度的聚类步数开始,它左右两个个体被聚为一类了。

## 3 用 DBMiner 做聚类分析

### 3.1 DBMiner

DBMiner 是加拿大 Simon Fraser 大学(简称 SFU)智能数据库研究所开发的商品化数据仓库与知识发现集成系统,其前身是 DBLearn。该系统设计的目的是把关系数据库和数据开采集成在一起,以面向属性的多级概念为基础发现各种知识。

该系统采用了层次化开发,系统划分为数据存储层、数据获取层、数据挖掘层、应用层四个层次(如图 4 所示)。数据获取层从原始数据中提取数据,构建数据挖掘层的挖掘平台;数据表示层运用挖掘层的提供的各种挖掘模式的 API,创建新的模型。整个挖掘过程及其相关信息由元数据管理模块所控制;主动式信息处理模块实现了系统的主动式服务。DBMiner 系统具有如下特色:

(1)能完成多种知识的发现:泛化规则、特性规则、关联规则、分类规则、演化知识、偏离知识等;

(2)综合了多种技术开采技术:面向属性的归纳、统计分析、逐级深化发现多级规则、元规则引导发现等方法;

(3)提出了一种交互式的类 SQL 语言——数据开采查询语言 DMQL;

(4)能与关系数据库平滑集成;实现了基于客户/服务器体系结构的 Unix 和 Pc(Windows/NT)版本的系统。

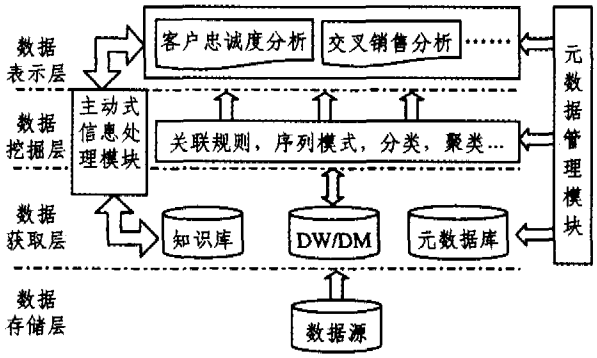



图 4 DBMimer 系统架构

### 3.2 相关的聚类算法

当前该系统实现了 K-Means[8],针对丢失数据等现实情况进行了优化。K-Means 算法是一种分割的而非分层的聚类方法,在数据挖掘领域中得到了最广泛的应用。这种聚类方法通常使用的聚类准则函数是聚类集中的每个样本点(数据或者对象)到该类中心的距离之和,并使它最小化。其基本步骤为:

(1)选择 $K$ 个初始聚类中心:$Z_1(1),Z_2(1),\cdots,Z_K(1)$,括号内的序号为寻找聚类中心的迭代运算的次序号,聚类中心的向量值可以任意设定,比如可以用开始 $k$ 个样本点作为初始聚类中心。

(2)逐个将需分类的样本$\{x\}$按最小距离原则分配给聚类中心的某一个 $Z_j(t)$,假如当 $i=j$ 时,即 $D_j(t)=\min\{\|x-Z_i(t),j=1,2,\cdots,k\|\}$,则 $x\in S_j(t)$,其中,$t$ 为迭代运算的次序号,第一次迭代 $t=1$,$S_j$ 表示第 $j$ 个聚类,其聚类中心为 $Z_j$。

(3)计算各个聚类中心新的向量值 $Z_j(t+1),j=1,2,\cdots,k$,即求各聚类域中所包含样本的均值向量,即

$$Z_j(t+1)=(1/N_j)\sum x \quad x\in S_j(t),j=1,2,\cdots,k$$

式中 $N_j$ 是第 $j$ 个聚类域 $S_j$ 包含的样本个数。以均值向量为新的聚类中心,可以使聚类准则函数

$$J_j=\sum x\{\|x-Z_j(t+1),j=1,2,\cdots,k\|\} \quad x\in S_j(t),j=1,2,\cdots,k$$

最小。在这一步中,要分别计算 $k$ 个聚类中的样本均值向量,$k$ 均值算法由此得名。

(4)如果 $Z_i(t+1)\neq Z_j(t),j=1,2,\cdots,k$,则 $t=t+1$,返回(2),将样本逐个重新分类,重复迭代计算;如果 $Z_j(t+1)=Z_j(t),j=1,2,\cdots,k$,则算法收敛,计算完毕。

### 3.3 用 DBMiner 做聚类分析举例

图 5 中的源数据库是一个以超市为原型的模拟数据库,由 7 个表组成,在交易记录表中保存着客户在某个时间购买的商品情况,其它表通过外键与交易记录表相连。我们用 DBMiner 提供的 K-Means 算法对顾客购买商品的情况进行聚类,结果如图 6 所示。

## 4 结论

通过以上实例的对比,我们可以得出以下结论:

### 4.1 SPSS 具有如下特点

(1)不单支持整个数据挖掘流程,从数据获取、转化、建模、评估到最终部署的全部过程,还支持数据挖掘的行业标准—CRISP≠DM。

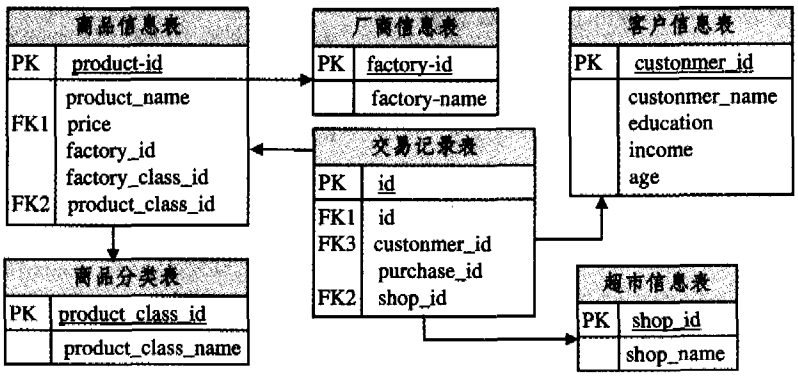



图 5 模拟数据库结构

维普资讯 http://www.cqvip.com

图 6　聚类结果

(2)支持平台、数据展现:适用于多种操作系统和多种数据源。通过将操作推向数据库,提升数据库和数据仓库的投资回报率。

(3)用户界面:通过连接节点的表示形式,模型在可视编程环境中被确定。

(4)数据准备:设置了全部的数据挖掘过程,包括大量的数据准备功能,不需要通过 SQL Sever 来处理数据库。

(5)模型发布:Clementine Solution Publisher 使分析人员能够抽出全部的数据挖掘过程。发布模型和升级模型既容易也经济。Clementine 也可将模型输出到 C、SQL 语言,通过编程来实现应用。

(6)聚类方法采用层次聚类方法。层次方法的特点在于,一旦一个步骤(合并或分裂)完成,它就不能撤消。其优点是不用担心组合数目的不同选择,计算代价会较小。其缺点是它不能更正错误的决定。

**4.2　DBMiner 具有如下特点**

(1)融入了多种知识发现思想;系统功能较齐全;系统提供了一种较完整的知识发现语言 DMQL。

(2)支持平台、数据展现:适用于多种操作系统。使用 excel 作数据展现。

(3)用户界面:提供了直观的图形用户界面,可视化的数据浏览工具。

(4)数据准备:直接通过 Microsoft SQL Sever OLAP Manager 实现访问多种关系数据库系统。

(5)模型发布:提供一个开放式体系结构,能够轻松实现与流行数据库和前端工具的集成。

(6)能处理千兆级的大型数据库。

(7)聚类算法采用 K-Means 算法,K-Means 算法的不足之处在于它要多次扫描数据库,此外,它只能找出球形的类,而不能发现任意形状的类。还有,初始质心的选择对聚类结果有较大的影响。

总之,在选择数据聚类工具时需要考虑很多因素,应根据特定的应用需求加以选择。国外许多行业已经大量利用数据挖掘工具来协助其业务活动,国内在这方面的应用还处于起步阶段,研究和学习国外先进的数据挖掘工具是很有必要的。

## 参 考 文 献


1　Han Jiawei, Kamber M. Data Mining : Concepts and Techniques [C]. Mongan Kaufmann publishers, 2000. 225～278

2　Alsabti K, Ranka S, Singh V. An eficient k-means clustering algorithm[A]. In: IPPS-98, Proc. of the First Workshop on High Performance Date Mining[C]. Orlando, Florida, USA, 1998

3　Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm forDiscovering Clustersin Large Spatial Databaseswith Noise. In: Proc. 2nd International Conference on Knowledge Discovery an dData Mining[C]. Portland, OR, 1996. 226～231

4　Wang I-IX, Zaniolo C. Database System Extensions for Decision Support: the AxL Approach[A]. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery[C]. 2000. 11～20

5　张雪英. 国外先进数据挖掘工具比较分析. 计算机工程, 2003(9): 1～3

6　马庆国. 管理统计:数据获取、统计原理、SPSS 工具与应用研究[M]. 北京:科学出版社, 2002

7　Zhang T, et al. BIRCH: An efficient data clustering method for very large databases. In: Proc. of the ACM SIGMOD Int' l Conf. on Management of Data. Montreal: ACM Press, 1996. 73～84

8　王实, 高文. 数据挖掘中的聚类方法. 计算机科学, 2000(4): 42～45


(上接第 190 页)

理。基于这些理论研究,提出了一个基于 Max-Code 寻找极大完全子图的算法——FMCSG (Finding Maximal Complete Subgraph)。该算法在搜索极大完全子图的过程中,通过剪掉非 Max-Code 码对应的子矩阵,有效地减少对子矩阵集的遍历次数,提高了算法的搜索效率。文中给出了 FMCSG 算法复杂性的基本估计,并证明了算法的正确性和完备性,即 FMCSG 算法能够找出全部的极大完全子图。

## 参 考 文 献


1　Washio T, Kunio N, et al. Complete Mining of Frequent patterns from Graphs: Mining Graph Data. Maching Learning, 2003, 50 (3): 321～354

2　Symth P, Goodman R M. An Information Theoretic Approach to Rule Induction from Database. IEEE Trans. Knowledge Data Eng., 1992, 4(4): 301～316

3　Chen A L, Tang C J, et al. An improved algorithm based on maximum clique and FP-tree for mining association rules. Journal of Software, 2004, 15(8): 1198～1207

4　Loukakis E, Tsouros C. A depth first search algorithm to generate the family of maximal independent sets of a graph lexicographically. Computing, 1981, 27: 249～266

5　Kopf R, Ruhe G. A computational study of the weighted independent set problem for general graphs. Foundations of Control Engineering, 1987, 12(4): 167～180

6　Bomze I M, Pelillo M, Stix V. Approximating the maximum weight clique using replicator dynamics. IEEE Trans. Neural Networks, 2000, 11(6): 1228～1241

7　Biggs N. Algebraic graph theory. Cambridge University Press, Cambridge, England, 1974

8　Adleman L M. Molecular computation of solutions to combinatorial optimization. Science, 1994, 226: 1021～1024

9　Sun S L. Algebra Structure. China University of Science and Technology Press, Hefei, China, 1990