分类预测方法及其在经济管理决策中的应用

王昱 著

科学出版社

分类预测方法及其在经济管理决策中的应用

王昱 著

国家自然科学基金项目（编号：71001112）
重庆大学现代物流重庆市重点实验室 资助出版

科学出版社
北京

内 容 简 介

本书是关于分类预测方法及其在经济管理决策中应用的研究专著。全书总结了国内外最新的研究资料和作者及所在的研究团体多年的科研成果，涉及面较广，内容新颖，反映了当前该领域的研究水平。

全书理论联系实际，使读者能很快地将最新的分类预测方法应用到经济管理实践中。全书共分为八章，内容包括分类预测基本概念及其与经济管理决策的关系、数据预处理、常用的分类预测方法、基于组合和集成的新分类预测方法及其在企业财务困境预测中的应用、信用评估、数据库营销中的应用、分类预测方法的研究进展等。

本书既可作为从事数据挖掘、商务智能、管理决策等方面研究的科技人员的参考资料，也可以作为高等院校管理科学与工程专业研究生和高年级本科生的教学用书和参考用书。

图书在版编目（CIP）数据

分类预测方法及其在经济管理决策中的应用 / 王昱著. —北京：科学出版社，2012

ISBN 978-7-03-034616-2

Ⅰ. ①分… Ⅱ. ①王… Ⅲ. ①统计预测－应用－经济－管理－研究 Ⅳ. ①F2

中国版本图书馆 CIP 数据核字（2012）第 115785 号

责任编辑：徐 倩 / 责任校对：黄江霞
责任印制：阎 磊 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

铭浩彩色印装有限公司 印刷

科学出版社发行 各地新华书店经销

*

2012 年 6 月第 一 版 开本：B5（720×1000）

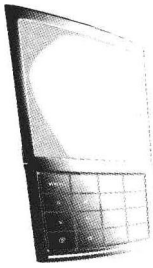
2012 年 6 月第一次印刷 印张：10 1/2

字数：208 000

定价：42.00 元

（如有印装质量问题，我社负责调换）

目 录



第一章

绪　论

第一节　数据挖掘中的分类预测

随着数据库技术发展和应用的普及，企业和组织存储数据的速度和数量飞速增长。据粗略估计，一个中等规模的企业每天要产生 100MB 以上来自生产经营和销售等各方面的商业数据。美国宇航局每天从卫星下载的数据量达 3TB 以上，并且为了科学研究的需要，这些数据要保存七年之久。虽然传统的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，更无法根据现有的数据预测未来的发展趋势，从而导致出现“数据爆炸但知识贫乏”的现象。另外，大量数据背后隐藏着许多重要的信息。面对如此巨大的数据资源，管理者和决策者需要新技术和新工具，以便能够应用这些技术和工具对数据进行更高层次的分析，将巨大的数据资源转化为有用的知识和信息，进而进行科学有效的管理决策。

数据挖掘(data mining，DM)又称为数据库中的知识发现(knowledge discovery from database，KDD)，是 20 世纪末从信息技术领域迅速兴起的计算机技术。数据

挖掘是一个从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中抽取挖掘出隐含其中的、事先未知的、有价值的模式或规律等知识的复杂过程。它是一个多学科交叉的研究与应用领域，所涉及的内容包括数据库技术、统计学、模式识别、信息检索与获取、机器学习、人工智能、高性能计算以及可视化计算等。由于数据挖掘所具有的广阔应用前景，对其进行的理论方法和应用的研究得到了广泛关注。例如，高德纳咨询公司(Gartner Group Inc.)的一次高级技术调查结果将数据挖掘和人工智能列为"将对未来三到五年内工业产生深远影响的五大关键技术"之首；世界500强企业中80%都涉足数据挖掘的前瞻性研究。

数据挖掘包含以下六个关键步骤：

(1)数据清洗(data cleaning)：清除数据噪声和与挖掘主题明显无关的数据。

(2)数据集成(data integration)：将来自多数据源的相关数据组合在一起。

(3)数据转化(data transformation)：将数据转换为易于进行数据挖掘的数据存储形式。

(4)数据挖掘：利用智能方法发掘潜在的数据模式或数据知识。

(5)模式评估(pattern evaluation)：根据一定的评估标准从挖掘结果中筛选出有意义的模式知识。

(6)知识表示(knowledge presentation)：利用可视化和知识表达技术，向用户展示所挖掘的相关知识。

概括而言，数据挖掘的主要流程如图1.1所示。

数据挖掘的数据一般来源于数据库或者数据仓库，这些数据首先经过清洗和集成，如清除噪声数据及与挖掘主题无关的数据，将多个数据源中的相关数据组合在一起等，然后转换为易于进行挖掘的数据存储形式，形成与挖掘任务相关的数据集。在此基础上，应用各种数据挖掘技术对任务相关数据集进行分析处理，并根据一定的标准评估和筛选出数据中潜在的有意义的模式和知识。与传统的数据库技术相比，数据挖掘具有若干新特征，如表1.1所示。

表1.1 数据挖掘与传统数据库技术的比较

比较的角度	传统数据库技术	数据挖掘
工具特点	回顾型、验证型	发现型、预测型
分析重点	已经发生了什么	解释发生的原因，预测未来的情况

续表

比较的角度	传统数据库技术	数据挖掘
分析目的	从最近的销售文件中列出最大客户	锁定未来的可能客户，以减少未来的销售成本
数据大小	数据量和数据维度均是少量的	数据量和数据维度均是庞大的
控制方式	企业管理人员、系统分析员、管理顾问启动与控制	数据与系统启动，少量的控制人员
发展状况	成熟	发展中

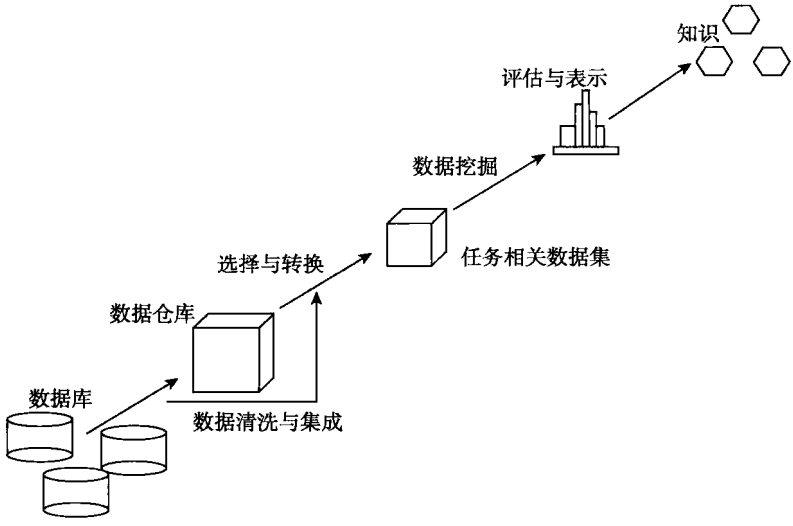


图 1.1 数据挖掘的主要流程

数据挖掘任务一般分为两类：描述式数据挖掘和预测式数据挖掘。描述式数据挖掘刻画数据库或数据仓库中数据的一般特性和性质，主要包括概念描述，即利用数据属性中更广义的(属性)内容对数据进行归纳和总结；预测式数据挖掘通过对数据应用特定方法进行分析，获得一个或一组模型，并将模型用于预测未来新数据的有关性质，其分析方法主要包括关联分析、分类预测、聚类分析、孤立点分析、演化分析等。

分类预测是数据挖掘中的一个重要领域，它可以用于提取描述重要数据类别的模型。分类是指从一组训练样本数据(其类别归属已知)中，通过学习获得一个

能够描述数据集合典型特征的模型(函数或规则)，以便能够分类识别未知数据的归属或类别(class)，即将未知事例映射到某种离散类别之一，如银行在信用评估中判断某个客户的信用等级是属于 A 级、B 级还是 C 级。分类挖掘所得到的模型可以采用多种形式加以描述输出，其中主要的方法有 If-Then 分类规则、决策树(decision tree，DT)、人工神经网络(artificial neural network，ANN)、支持向量机(support vector machine，SVM)等。在很多问题中，需要预测某一数值属性的值(连续数值)，这种情况被称为预测。例如，可以建立预测模型，给定潜在顾客的收入和职业，预测这些顾客在计算机设备上的花费。尽管预测既包括连续数值的预测，也包括有限离散值的预测，但一般使用预测这一术语表示对连续数值的预测，使用分类这一术语表示对有限离散值的预测。

分类预测过程一般包括以下两个步骤：

第一步，建立一个描述已知数据样本集合的类别或概念的模型，该模型通过对数据集合中各行记录(数据样本)内容的分析而获得。为建立模型而被分析的数据样本形成训练样本集合，其中的每一数据样本都属于一个确定的数据类别，其类别值由一个属性描述。该属性通常被称为目标变量(target variable)或类标记(class label)。除目标变量这一属性外，训练样本集合中每一个属性称为一个特征(feature)。由于分类是在已知训练样本类别的情况下，通过学习建立相应模型，因此分类又可以称为有监督学习(supervised learning)。与之相对应的无监督学习(unsupervised learning)则在训练样本的类别和类别个数均未知的情况下进行。

分类学习所获得的模型通常可以表示为分类规则形式、决策树形式或数学公式形式。例如，给定一个顾客的信用信息数据库，通过分类学习所获得的分类规则可用于识别顾客是否具有良好的信用等级或一般的信用等级。分类规则可用于对(今后)未知(所属类别)的数据样本进行识别判断，同时也可以帮助用户更好地了解数据库中的内容。

第二步，利用所获得的模型进行分类操作，即利用该模型对类别未知的对象进行分类。这一步中的一个关键问题是对所得到模型的分类准确率进行估计。其常用的方法是交叉验证(cross validation)方法，即从训练数据集中随机选择一部分作为测试样本，然后用剩余的训练样本进行学习得到分类模型，最后考察该分类模型在测试样本上的分类准确率。为了避免在选择训练样本和测试样本时的随

机性，同时也使得分类结果具有统计意义，在分类预测准确率评估中一般采用 k 层交叉验证(k-fold cross validation)，即将样本分为 k 个子集合，分别将每一个子集合作为测试样本，其余 $k-1$ 个子集合作为训练样本，最终得到 k 个分类结果。如果一个通过学习所获得模型的准确率经测试后被认为可以接受，则可以将其应用于对类别未知的对象进行分类。

分类的流程可描述如下[不失一般性，仅考虑二分类(binary classification)，即只有两个数据类集的情况]：设 S 为 p 维向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_p)'$ 的集合，即 $S=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，其中样本 $\boldsymbol{X}_i=(x_1^i,x_2^i,\cdots,x_p^i)'$ $(i=1,2,\cdots,N)$ 的取值为已知。S 中的样本来自两个不同的数据类集 $A_k(k=1,2)$，即每个样本具有一个标记(label)y_i，表明该样本来自哪一个类集，例如，$y_i=+1$ 代表样本 $\boldsymbol{X}_i$ 来自于类集 A_1，$y_i=-1$ 代表样本 $\boldsymbol{X}_i$ 来自于类集 A_2。在分类模型中，一般将样本中的属性 $x_j(j=1,2,\cdots,p)$ 称为特征。分类要解决的问题是给定若干特征和类标号已知，但两者之间依赖关系[线性或非线性映射 $f:(x_1,x_2,\cdots,x_p)'\rightarrow y$]未知的训练样本，对于特征已知但类标记 y^u 未知的待分类样本 $\boldsymbol{X}^u=(x_1^u,x_2^u,\cdots,x_p^u)'$，如何将 $\boldsymbol{X}^u$ 正确划分到某个类集中，即预测类标记 y^u 的取值。假设每个样本 $\boldsymbol{X}_i$ 对应的类标号 y_i 为二分变量(-1 或 $+1$)，且可以表示为 $\boldsymbol{X}_i$ 的函数：

$$y_i=f(\boldsymbol{X}_i) \tag{1.1}$$

分类的目的是建立模型对以上函数进行估计，即估计以下函数：

$$y_i=\hat{f}(\boldsymbol{X}_i) \tag{1.2}$$

使得 $\sum_{i=1}^{N}|f(\boldsymbol{X}_i)-\hat{f}(\boldsymbol{X}_i)|$ 最小的 $\hat{f}^*(\boldsymbol{X}_i)$即为最终得到的分类模型。可以看出，分类的关键是在特征 $(x_1,x_2,\cdots,x_p)$ 和类标记 y 之间建立准确合理的依赖关系，并根据该依赖关系预测 $\boldsymbol{X}^u$ 的类标记 y^u。

在过去的几十年中，统计学、人工智能、生物学等领域已经出现了许多分类预测方法。作为数据挖掘的重要方法之一，分类预测广泛应用于各个学科领域，如银行的信用评估、文献的整理归档、客户的重要级别认定、疾病的严重程度划分、网络浏览板块的隶属等。

第二节 分类预测与经济管理决策

随着信息技术日新月异的发展，企业和组织能够更加容易地获取和存储大量

数据。因此，如何应用数据挖掘技术对海量数据进行分析以发现潜在的数据模式，使得管理者能够更加科学有效地进行决策，减小行动风险，提高收益和市场竞争力，已经成为经济管理决策中的重点研究问题。在这些问题中广泛存在着具有如下特征的一类问题，即管理决策者需要先建立历史数据样本与自然状态之间的依赖关系，然后根据该依赖关系估计新的数据样本所对应自然状态的出现概率，最后建立风险决策模型，以最大化收益函数(或最小化风险损失函数)为目标选择最优行动方案。

从估计历史数据样本与自然状态出现概率的依赖关系这一意义上，可以将上述风险决策问题归纳为数据挖掘中的分类预测问题，即从一组训练样本数据(其类别归属已知)中通过学习获得一个能够描述数据集合典型特征的模型(函数)，以便能够分类识别未知数据的归属或类别，即将未知事例映射到某种离散类别之一。这类问题的典型例子有海关查验走私(Hua et al.，2006)、信用评估(West，2000；Wang et al.，2005)、企业财务困境预测(Min and Lee，2005；Hua et al.，2007；Xu and Wang，2009)、市场营销(Smith et al.，2000；Kim and Street，2004；Kim et al.，2005)等。例如，在海关查验走私问题中，海关监管人员需要决策是否对一单报关货物进行走私查验。在该风险决策问题中存在两种自然状态：一种自然状态是报关货物符合海关的报关货物法规，即未走私；另一种自然状态是报关货物不符合海关的报关货物法规，即走私。由于海关监管人员无法预先知道报关货物是否走私，并且报关货物数量巨大且海关的人力和资源有限，监管人员不可能对每一单报关货物进行查验。因此，监管人员需要建立一个风险决策模型来进行查验决策(Hua et al.，2006)。为了建立有效的风险决策模型，海关监管人员需要先根据一单报关货物的各项属性(如货物种类、单价、原产国、进出口岸、毛净重比等)判断该报关货物是否存在走私风险，即将其划分为走私和不走私两类情况，然后再选择最优的行动方案。在这一问题中，用于学习的训练样本为海关数据库中历史报关商品的各项属性指标值(如商品进出口岸、单价、毛净重比、商品进出口贸易单位等)及商品类别(表示是否走私，如类别0表示不走私，类别1表示走私)，建立分类模型的目的是准确地对当前报关货物进行分类，再根据分类结果决策是否需要进行走私查验。

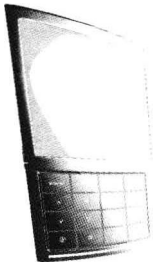
在应用分类预测技术的经济管理决策问题中，另一个典型例子是消费者个人

信用评估。随着我国近年来消费信贷的飞速发展，各个商业银行消费信贷业务纷纷加入竞争行列，希望在新一轮的市场竞争中获得优势。但是，随着消费信用市场的飞速发展，在消费者债务突飞猛进的同时，消费者破产及拖欠债务的现象也与日俱增。在商业银行确定信用发放对象的过程中，一个关键因素是对申请贷款消费者的信用风险进行科学有效的分析和评估。消费者信用评估的主要目的就是对可能引起信用风险的因素进行定性分析和定量计算，以测量消费者的违约风险，为授信方(银行)决策提供依据。目前信用评估中最常用的方法为分类预测技术，即根据消费者的一些原始资料，通过建立一种易于应用的分类预测模型，得到消费者的个人信用评级。授信方根据该评级，对提出信用贷款或消费申请的个人进行评估，并对符合要求的申请人授予相应的信用额度。例如，West(2000)使用人工神经网络这一在各个领域得到广泛应用的分类预测方法对消费者信用进行评估，Wang 等(2005)提出了一种新的模糊支持向量机进行消费者信用评估。在信用评估中应用到的其他分类方法还包括 K-近邻(k-nearest neighbor，K-NN)(Henley and Hand，1996)、遗传算法(genetic algorithms，GAs)(Ong et al.，2005)等。

分类预测技术还被广泛应用于市场营销这一领域。例如，Kim 和 Street(2004) 提出了一个针对市场营销中顾客确定(customer targeting)问题的风险决策支持系统。该系统首先应用遗传算法进行特征选择，其次利用人工神经网络进行顾客分类，判断各个顾客积极参与企业营销活动的可能性大小，最后根据上述结果确定对哪些潜在客户以邮件形式发送活动内容和声明，以最大化企业的收益。特征选择的第一个目的是去除与问题无关的顾客特征，提高人工神经网络的准确度和推广性能；第二个目的是使决策者了解与市场营销效果相关的顾客特征信息。Kim 等(2005)对上述工作进行了一些修改，用局部选择进化算法(evolutionary local selection algorithm，ELSA)替代一般的遗传算法。该系统比仅使用人工神经网络能够取得更好的市场营销收益。

综上所述，在经济管理决策中各种分类预测方法及其应用已经引起了学术界和企业界的持续关注。在瞬息万变的社会环境和激烈的市场竞争中，有效的风险决策可以减小企业或公共组织的行动风险，提高收益和市场竞争能力，使企业保持有利地位。其中，分类预测方法的准确性和高效性对风险决策效果具有至关重要的影响。例如，在海关查验走私这一风险决策问题中，准确高效的分类预测方

法能够缓解海关查验资源与日益增长的进出口贸易量之间的矛盾，提高海关的通关效率和监管服务效率，有利于海关监管工作的顺利开展。在消费者信用评分这一风险决策问题中，准确高效的分类预测方法可以使信贷机构科学高效地确定信用发放对象，提高自身的市场竞争力与效益，从而增大市场份额，降低运营成本，获得比行业平均水平更高的超额利润。因此，数据挖掘中分类预测方法及其在经济管理决策中的应用研究具有重要的理论和现实意义。



第二章

数据预处理

第一节　数据预处理的重要性

在现实世界中，由于数据库或数据仓库所获取的数据量迅速膨胀，使得数据中极易受噪声数据、遗漏数据和不一致性数据的影响。所谓噪声数据是指数据中存在着错误或异常(偏离期望值)的数据；遗漏数据是指与数据挖掘任务相关的某些属性没有值；不一致性数据是指数据内涵出现不一致的情况(例如，作为关键字的某一属性在同一数据库中出现不同值)。导致以上问题产生的主要原因如下：①在数据采集时，某些属性的内容有时并未记录，这可能是记录疏忽，也可能是这些属性在当时被认为不必要；②由于误解或设备问题导致相关数据未得到记录；③与其他记录内容不一致而被删除；④历史记录或数据的修改被忽略；⑤数据传输过程中发生错误；⑥由于命名规则或数据代码不同而引起不一致。

对现实世界中大规模的数据库或数据仓库而言，噪声数据、遗漏数据和不一致性数据是非常普遍的情况。例如，一个负责公司销售数据分析的人员会仔细检

查公司数据库或数据仓库中的数据内容，挑选与挖掘任务相关的数据对象的描述性特征，这些特征可能包括商品类型、价格、销售量或者顾客类型等。但是，他(她)或许会发现数据中有些记录的一些特征值并未记录，甚至存在一些错误或不一致的情况。对于这样的数据对象进行数据挖掘，必须首先进行数据预处理。

概括而言，现实世界的数据一般是存在噪声的、不完整的和不一致的。数据预处理技术可以改进数据的质量，从而有助于提高其后的挖掘过程的精度和性能(朱明，2002)。由于高质量的决策必然依赖于高质量的数据，因此数据预处理是知识发现过程的重要步骤。数据预处理在数据挖掘之前使用，可以大大提高数据挖掘模式的质量，降低实际挖掘所需要的时间。数据预处理包含以下几方面的内容：①数据清洗。该过程通常包括填补遗漏值，识别并消除噪声和异常值，纠正数据中的不一致性等。②数据集成，将来自多个数据源的数据合并在一起。由于描述同一个概念的属性在不同数据源中可能有不同的名字，在进行数据集成时常常会引起数据的不一致性或冗余。如在一个数据源中顾客的身份编码为 customer _ id，而在另一个数据源中则可能为 cust _ id。因此，在数据集成中除了进行数据清洗外，还要消除数据的冗余。③数据规范化。该过程可以改进涉及距离度量的挖掘算法的精度和有效性。④数据约简(data reduction)。该步骤的目的是缩小所挖掘数据的规模，同时不影响(或基本不影响)最终的挖掘结果。例如，可以通过聚集、删除冗余特征或聚类等方法来压缩数据。

第二节 数据清洗

数据清洗是数据预处理的一个重要方面，主要包括填补遗漏值，识别并消除噪声，纠正数据中的不一致性等。

一、填补遗漏值

对于数据中某些记录(元组)的某些属性出现的遗漏情况，可以采用以下几种方法进行填补。

1. 忽略该记录

若一条记录中某个或者某些属性被遗漏，则在数据挖掘中直接排除此记录，尤其在进行分类或者概念描述时，如果类别属性缺失，则可以忽略该记录。但是，该方法在属性缺少值所占百分比较高时性能较差。

2. 人工填补遗漏值

该方法一般很费时，并且对于有较多属性缺失值的大规模数据集而言，该方法缺乏可行性。

3. 使用常量填补遗漏值

将一个属性的所有遗漏值均用同一个常数(如“unknown”)填补。但是当某一属性的遗漏值较多时，这种方法可能会误导挖掘过程。例如，如果遗漏值都用“unknown”替换，挖掘程序可能误以为它们形成了一个有趣的概念，因为它们的该属性都具有相同的取值。

4. 使用属性的平均值填补遗漏值

计算某个属性的均值，然后利用该均值填补该属性中所有的遗漏值。例如，假定数据集中某个属性的均值为 x，则用 x 替换该属性中所有的遗漏值。

5. 使用同类别记录的属性平均值填补遗漏值

例如，如果将记录按某一属性分类，则用同一类记录的平均值填补该类记录中相应的属性遗漏值。

6. 使用最可能的值填补遗漏值

可以用回归、贝叶斯(Bayes)形式化方法或决策树归纳等工具推导出记录缺失属性值的最大可能取值。例如，利用数据集中的属性构造一棵决策树来预测某个属性的遗漏值(决策树将在第三章中讨论)。

二、识别并消除噪声

噪声是指被测变量的一个随机错误和变化。如果用于数据挖掘的数据记录中存在噪声，可能会误导挖掘过程，产生错误的挖掘结果。因此，在进行数据挖掘前，需要首先识别并消除噪声。概括而言，识别并消除噪声可以采用以下几类方法。

1. 分箱法

分箱(Bin)法通过利用被平滑记录的“邻居”(即周围的记录)，对一组排序后的属性值进行平滑。排序后的属性值被分布到若干箱中。分箱法的一个简单例子如下所示。

假定数据库中某个属性为工资(salary)。首先，分箱法对 salary 属性值进行排序，假定排序后的 salary 属性值为 1000，1200，2000，2400，2600，3400，3800，4500，5200。然后，以上的 salary 被划分并存入等深的箱中(在该例中设定深度 3)，则得到如下三个 Bins，即 Bin 1：1000，1200，2000；Bin 2：2400，2600，3400；Bin 3：3800，4500，5200。最后，可以用不同的方法对箱内的数值进行平滑。例如，根据 Bin 均值进行平滑，可得如下结果：

Bin1：1400，1400，1400；

Bin2：2800，2800，2800；

Bin3：4500，4500，4500。

根据 Bin 边界进行平滑，可得如下结果：

Bin1：1000，1000，2000；

Bin2：2400，2400，3400；

Bin3：3800，3800，5200。

一般来说，深度越大，平滑效果越大。除上述等深度箱划分外，也可以用等宽的箱进行划分，即每个箱值的区间范围是个常量。

2. 聚类方法

聚类的基本思想是根据“各聚集(cluster)之内数据对象的相似度最大化和各聚集之间数据对象的相似度最小化”这一原则将数据对象划分为若干组。由于在聚类分析中，相似或相近邻的数据对象聚合在一起形成了各个聚集，而位于这些聚集之外的数据对象则可以认为是异常数据。

3. 人机结合检查方法

通过人与计算机相结合的方法，可以检查数据中存在的异常数据。例如，使用信息理论度量可以帮助识别手写体字符数据库中的异常数据。这些异常数据可能提供信息(如字符“0”或“7”的不同版本)，也可能是“垃圾”(如错误的字符)，其差异程度大于某个阈值的模式被输出到一个表中。此时，人可以审查表中的模