计算机研究与发展
Journal of Computer Research and Development

ISSN 1000-1239/CN 11-1777/TP
43(Suppl.): 171～175, 2006

缺失数据处理方法的比较研究

乔珠峰　田凤占　黄厚宽　陈景年
(北京交通大学计算机与信息技术学院　北京　100044)
(qiaozhufeng@126.com)

A Comparison Study of Missing Value Datasets Processing Methods

Qiao Zhufeng, Tian Fengzhan, Huang Houkuan, and Chen Jingnian
(*School of Computer and Information Technology. Beijing Jiaotong University, Beijing* 100044)

Abstract Data mining approaches have been applied widely in many fields now, but most datasets are missing values. Therefore it is very important to study data mining methods based on missing data. Four methods of treating missing attribute values are introduced in this paper. They are directdeletion, special values completer, mean completer and robust method. Based on the four methods above, four models of treating missing attribute values and corresponding four Naïve Bayesian classifiers are built. These models are conducted on five datasets. Five-folds cross-validation is used to estimate the performances of each model, which shows these naive Bayesian classifiers are effective.

Key words data mining; missing value; naive Bayesian classifier; robust; cross-validation

摘　要　由于数据挖掘技术日益广泛地应用于各个领域,而大多数领域中数据都存在缺失值,因此基于缺失数据的数据挖掘方法的研究具有重要意义.利用直接删除、特殊值填充、平均值填充、Robust方法4种处理缺失值的方法建立4个缺失值处理模型以及相应的朴素贝叶斯分类器模型.通过在5个实际数据集上进行实验比较,并采用五重交叉验证来检验这些模型的性能.结果表明,用这些模型处理缺失值构建的朴素贝叶斯分类器是有效的.

关键词　数据挖掘;缺失值;朴素贝叶斯分类器;Robust;交叉验证

中图法分类号　TP18

近年来,数据挖掘技术被广泛地应用到各个行业领域.数据挖掘的过程包括问题理解、数据采集和理解、预处理、数据挖掘工具、模型评估和知识应用.根据研究,在数据挖掘过程中20%的时间用于目标识别,60%的时间用于数据准备,数据挖掘和知识分析的时间只占10%.为什么人们要将超过50%的精力放在数据预处理上呢?在现实世界的数据库中存在着严重的质量问题:

① 数据不完整;

② 数据冗余;

③ 数据不一致;

④ 噪音数据.

这些严重的质量问题会降低数据挖掘算法的性能,因此,人们不得不将大量的时间和精力花在数据预处理上.在保证不减少数据所含信息的前提下,合理有效的数据预处理可以压缩数据量,改善数据质量,提高数据挖掘算法的性能,减少学习时间.

缺失数据的处理问题是数据挖掘过程中的一个严重问题.本文介绍数据预处理过程中的几种常见缺失数据处理技术以及Ramoni和Sebastiani提到

收稿日期:2006-06-05
基金项目:国家自然科学基金项目(60503017)

万方数据

的 Robust 方法[1],并根据这些方法建立 4 个缺失处理的模型进行实验分析比较.

1 研究意义

1.1 数据缺失处理的重要性和复杂性

数据挖掘算法本身更致力于避免数据过分适合所建的模型,这一特性使得它难以通过自身的算法去很好地处理不完整数据. 因此,空缺的数据需要通过专门的方法进行推导、填充等,以减少数据挖掘算法与实际应用之间的差距.

数据缺失在许多研究领域都是一个复杂的问题. 对数据挖掘来说,缺值的存在造成了以下影响:①系统丢失了大量的有用信息;②系统中所表现出的不确定性更加显著,系统中蕴涵的确定性成分更难把握;③包含缺值的数据会使挖掘过程陷入混乱,导致不可靠的输出.

1.2 数据缺失的原因

在各种实用的数据库中,属性值缺失的情况经常发生甚至是不可避免的. 因此,在大多数情况下,信息系统是不完备的,或者说存在某种程度的不完备;造成数据缺失的原因是多方面的,主要有以下几种:

1) 有些信息暂时无法获取. 例如在医疗数据库中,并非所有病人的所有临床检验结果都能在给定的时间内得到,致使一部分属性值空缺出来;又如在申请表数据中,对某些问题的反映依赖于对其他问题的回答.

2) 有些信息是被遗漏的. 可能是因为输入时认为不重要、忘记填写或对数据理解错误而遗漏,也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等原因而丢失.

3) 有些对象的某个或某些属性是不可用的. 也就是说,对于这个对象来说该属性值是不存在的,如一个未婚者的配偶姓名、一个儿童的固定收入状况等.

4) 有些信息(被认为)是不重要的. 如一个属性的取值与给定语境是无关的,或训练数据库的设计者并不在乎某个属性的取值.

5) 获取这些信息的代价太大.

6) 系统实时性能要求较高,即要求得到这些信息前迅速做出判断或决策.

1.3 数据缺失的类型[2-4]

在对缺失数据进行处理前,了解数据缺失的类型和形式是十分必要的. 将数据集中不含缺失值的变量(属性)称为完全变量,数据集中含有缺失值的变量称为不完全变量,Little 和 Rubin 定义了以下 3 种不同的数据缺失类型:

1) 完全随机缺失(missing completely at random, MCAR). 数据的缺失与不完全变量以及完全变量都是无关的.

$$P(Y \text{ Missing} \mid X, Y) = P(Y \text{ Missing}).$$

2) 随机缺失(missing at random, MAR). 数据的缺失仅仅依赖于完全变量.

$$P(Y \text{ Missing} \mid X, Y) = P(Y \text{ Missing} \mid X).$$

3) 非随机、不可忽略缺失(not missing at random, NMAR)(nonignorable, NI). 不完全变量中数据的缺失依赖于不完全变量本身,这种缺失是不可忽略的.

2 缺值数据的处理方法及模型

2.1 数据缺失的处理方法[5-6]

1) 直接删除元组

也就是将存在遗漏信息属性值的对象(元组、记录)删除,从而得到一个完备的信息表.

2) 特殊值填充(treating missing attribute values as special values)

将缺值作为一种特殊的属性值来处理,它不同于其他的任何属性值. 如所有的缺值都用"unknown"填充,这样将可能导致严重的数据偏离,一般不推荐使用.

3) 平均值填充(mean/mode completer)

将信息表中的属性分为数值属性和非数值属性来分别进行处理. 如果缺值是数值型的,就根据该属性在其他所有对象取值的平均值来填充该缺失的属性值;如果空值是非数值型的,就根据统计学中的众数原理,用该属性在其他所有对象的取值次数最多的值(即出现频率最高的值)来补齐该缺失的属性值.

4) 使用最有可能的值来填充缺值

可以用回归、基于推导的使用贝叶斯形式化方法的工具或判定树归纳确定,这些方法直接处理的是模型参数的估计而不是空缺值预测本身. 与前面的方法相比,它使用现存数据的多数信息来推测空缺值.

万方数据

5) 保留缺失数据不予处理

不对缺失数据做任何处理,直接在含大量缺失数据的数据集上进行数据挖掘,比如在文献[1]中提到的 Robust 模型上进行.

2.2 缺失数据的处理模型

1) 直接删除缺值模型

直接删除数据集中所有含缺失数据的记录,仅用剩余的无缺失数据的记录组成新的完整的数据集.

2) 特殊值填充模型

所有的缺值都用"unknown"补齐,如果类变量有缺值则删除该元组.

3) 平均值填充模型

我们仅考虑离散情况下的分类,所以我们对缺值的数据用出现频率最高的值来补齐.

4) 直接在缺失数据上建立分类模型

对缺值数据不给予处理,直接利用文献[1]中提到的 Robust 方法,这种方法不是通过对缺失数据进行假设填充来处理缺失数据,它通过一个特殊方法对不完整数据进行处理,算出一个条件概率和先验概率区间,我们在这些区间内取满足概率条件的条件概率值和先验概率值构建朴素贝叶斯分类器[7-8].

3 模型实现过程

3.1 模型 1、模型 2、模型 3 构建朴素贝叶斯分类器过程

考虑到模型 1、模型 2、模型 3 都是把有缺失的不完整数据集转化为完整数据集,故这 3 个模型构建朴素贝叶斯分类器的过程基本一致:

1) 对有缺失的数据集进行相应缺失数据处理;

2) 在步骤 1)完成的完整数据集上算出先验概率: $P(C_i)=\dfrac{a_i+N(C_i)}{a_i+\sum_i N(C_i)}$,

和条件概率: $P(Aik/C_j)=\dfrac{a_{ikj}+N(Aik\mid C_j)}{a_{ikj}+N(C_j)}$,

其中, $N(C_i)$ 是训练实例中类 C_i 的记录个数; $N(Aik\mid C_j)$ 是在训练实例中属于类 C_j 的实例中属性 A_i 取第 k 个值的个数;分子分母加 a_i 和 a_{ikj} 是为了防止出现零的情况,实验中我们取值 1.

3.2 模型 4 的构建朴素贝叶斯分类器过程

1) 根据文献[1]的办法对每一个类、每一个属性以及相应属性的每一个取值求出先验概率区间和条件概率区间.

先验概率区间: $[\underline{p}(c_j),\bar{p}(c_j)]$,

其中:

$$\underline{p}(c_j)=\frac{a_j+n(c_j)}{\sum_l[a_l+n(c_l)]+n(?)},$$

$$\bar{p}(c_j)=\frac{a_j+n(c_j)+n(?)}{\sum_l[a_l+n(c_l)]+n(?)}.$$

条件概率区间: $[\underline{p}(a_{ik}\mid c_j),\bar{p}(a_{ik}\mid c_j)]$,

其中:

$$\underline{p}(a_{ik}\mid c_j)=\frac{a_{ijk}+n(a_{ik},c_j)}{\sum_h[a_{ijh}+n(a_{ih},c_j)]+\underline{n}(a_{ik},c_j)},$$

$$\bar{p}(a_{ik}\mid c_j)=\frac{a_{ijk}+n(a_{ik},c_j)+\bar{n}(a_{ik},c_j)}{\sum_h[a_{ijh}+n(a_{ih},c_j)]+\bar{n}(a_{ik},c_j)},$$

其中: a_i 和 a_{ikj} 是为了防止出现零的情况,实验中我们取值 1.

2) 我们在这两个区间内取满足概率条件的值作为构建朴素贝叶斯分类器的先验概率和条件概率表,我们令:

$p(a_{ik}\mid c_j)=\bar{p}(a_{ik}\mid c_j)-\lambda(\bar{p}(a_{ik}\mid c_j)-\underline{p}(a_{ik}\mid c_j))$,

以及

$$p(c_j)=\bar{p}(c_j)-\mu(\bar{p}(c_j)-\underline{p}(c_j)),$$

其中, λ 和 μ 被选择,以使得概率条件 $\sum_k p(a_{ik}\mid c_j)=1$ 和 $\sum_j p(c_j)=1$ 成立,即,令:

$$\lambda=\frac{\sum_k\bar{p}(a_{ik}\mid c_j)-1}{\sum_k(\bar{p}(a_{ik}\mid c_j)-\underline{p}(a_{ik}\mid c_j))},$$

$$\mu=\frac{\sum_j\bar{p}(c_j)-1}{\sum_j(\bar{p}(c_j)-\underline{p}(c_j))}.$$

3.3 模型及其检验指标

本文采用的模型是在上面缺失数据处理的基础上算出类先验概率和条件概率,构建朴素贝叶斯分类器,再使用所构建出朴素贝叶斯分类器进行数据测试,采用五重交叉验证法检验分类效果.

朴素贝叶斯分类器是一个简单、有效而且在实际使用中很成功的分类器. 假设给定类变量 C,所有的属性变量都是相互独立的,而且每一个属性变量都以类变量作为惟一的父节点.

所谓五重交叉验证就是将所有的已知数据均分为 5 份,每一份都独立地建立一个模型,并用其余的

万方数据

数据进行验证. 交叉验证的一个优点是即使数据量很小也能达到很好的效果.

模型检验指标为模型预测准确率和类预测准确率. 分别定义如下:

$$\text{模型预测准确率} = \frac{\text{整个数据集中预测正确的记录个数}}{\text{整个数据集中的记录总数}} \times 100\%,$$

$$\text{类预测准确率} = \left(\sum_{c_i} \frac{\text{某类中预测正确的记录个数}}{\text{该类中的记录总数}}\right) / (\text{类数量}) \times 100\%,$$

其中, c_i 代表一个类"整个数据集中预测正确的记录个数"和"某个类中预测正确的记录个数"是利用朴素贝叶斯分类器预测模型在补缺模型产生的数据集上计算得到的.

4 实验结果及分析

本次实验我们采用了 5 组来自 UCI 的数据集作为实验数据来验证我们的不同缺失数据处理的模型,这些数据集的属性都是离散的,表 1 列出了每个数据集的实例数量、属性数量、类数量以及各个模型的模型预测准确率和类预测准确率,每个检验指标都是通过五重交叉验证后得到的.

表 1 四个模型在 5 个典型数据集上分类情况的比较

数据集	实例数量	属性数量	类数量	缺值情况	模型 1		模型 2		模型 3		模型 4	
					R_{Model}	R_{Class}	R_{Model}	R_{Class}	R_{Model}	R_{Class}	R_{Model}	R_{Class}
Agaricuslepiota	8124	22	2	一般	0.9701	0.9624	0.9331	0.9145	0.9338	0.9162	0.9405	0.9244
Anneal	798	38	6	严重	0.0101	0.2000	0.9586	0.9077	0.9146	0.8429	0.9360	0.7949
Audiologystandardized	199	70	24	严重	0.0101	0.0167	0.67227	0.2416	0.6822	0.2505	0.6724	0.2433
CreditApproval	690	15	2	很少	0.8464	0.8393	0.84493	0.8378	0.8478	0.8407	0.8449	0.8378
Housevotes	434	16	2	一般	0.9515	0.9564	0.9007	0.9052	0.9030	0.9083	0.9053	0.9101

注: R_{Model} 代表模型预测准确率, R_{Class} 代表类预测准确率. 缺值情况:很少(小于 10%),一般(30%~60%之间),严重(大于 60%)

从表 1 中的实验数据来看,在数据缺值严重的情况下模型 1 的效果很差,说明了对于缺值严重的数据简单删除缺值做法抛弃了很多起决定作用的信息. 而在缺值很少的情况下,模型 1 表现出的分类效果要比其他模型好;说明这种方法简单易行,在对象有多个属性缺失值、被删除的含缺失值的对象与信息表中的数据量相比非常小的情况下是非常有效的,类标号缺少时通常使用. 然而,这种方法却有很大的局限性. 它是以减少历史数据来换取信息的完备,会造成资源的大量浪费,丢弃了大量隐藏在这些对象中的信息. 在信息表中本来包含的对象很少的情况下,删除少量对象就足以严重影响到信息表信息的客观性和结果的正确性;从这些数据我们可以看出,模型 2、模型 3 在处理缺失数据上表现比较稳定,它们在缺失数据很少时分类效果不比模型 1,说明它们的缺失数据填充处理导致了数据倾斜,最终导致分类错误的增加;从实验中也可以看出,模型 4 和模型 2 以及模型 3 的效果差不多,但是它没有对缺失数据进行处理,而是直接在有缺失的数据集上进行的数据训练,另外这些在有些数据集上的类预测准确率很低,通过对这些数据集的分析,导致类预测准确率低的原因是训练集有些类的训练实例比较少.

5 总 结

本文通过建立 4 个处理缺失数据的模型,并通过实际的实验分析对 4 种处理缺失数据的技术进行了比较研究,通过实验我们得到了有效的朴素贝叶斯分类器;也验证说明了各个模型的各自有效性和不足;不存在一种处理缺值的方法可以适合于任何问题. 另外,这些模型还可以结合别的分类技术更进一步地改进分类器的分类效果,而且除了上述提到的几个模型以外还有其他的一些缺失数据处理方法,这将是我们下一步的研究工作.

参 考 文 献

[1] Marco Ramoni, Paola Sebastiani. Robust Bayes classifiers [J]. Artificial Intelligence, 2001, 125(1-2): 209-226

[2] Sameer Agarwal. Learning from incomplete data [OL]. http://www. cs. ucsd. edu/user/elkan/254springol/sagarwalrep. pdf, 2006

万方数据

[3] Zoubin Ghahramani, Michael I Jordan. Learning from incomplete data [R]. MIT Center for Biological and Computational Learning, Tech Rep: AIM-1509, 1994

[4] R J A Little, D B Rubin. Statistical Analysis with Missing Data [M]. Wiley Series in Probability and Mathematical Statistics. New York: Wiley and Sons, 1987

[5] Jiawei Han, Micheline Kamber. 数据挖掘概念与技术[M]. 范明, 孟小峰, 等译. 北京: 机械工业出版社, 2001

[6] J W Grzymala-Busse, M Fu. A comparison of several approaches to missing attribute values in data mining [C]. In: Proc of the 2nd Int'l Conf on Rough Sets and Current Trends in Computing. Berlin: Springer-Verlag, 2000. 378-385

[7] David Heckerman. Bayesian networks for data mining [G]. In: Data Mining and Knowledge Discovery. Berlin: Springer, 1997. 79-119

[8] Nir Friedman, Dan Geiger, Moises Goldszmidt. Bayesian network classifiers [J]. Machine Learning, 1997, 29(2-3): 131-163

乔珠峰 男,1979 年生,硕士研究生,主要研究方向为贝叶斯网络分类器等.

田凤占 男,1972 年生,副教授,硕士生导师,主要研究方向为机器学习、数据挖掘、贝叶斯网络等.

黄厚宽 男,1940 年生,教授,博士生导师,主要研究方向为数据挖掘、分布式人工智能等.

陈景年 男,1970 年生,博士研究生,主要研究方向为数据挖掘、贝叶斯网络等.

万方数据

缺失数据处理方法的比较研究

万方数据 WANFANG DATA 文献链接

作者： 乔珠峰， 田凤占， 黄厚宽， 陈景年， Qiao Zhufeng， Tian Fengzhan， Huang Houkuan， Chen Jingnian

作者单位： 北京交通大学计算机与信息技术学院,北京,100044

刊名： 计算机研究与发展 ISTIC EI PKU

英文刊名： JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT

年，卷(期)： 2006，43(z1)

引用次数： 0次

参考文献(8条)

1. Marco Ramoni. Paola Sebastiani Robust Bayes classifiers 2001(1-2)

2. Sameer Agarwal Learning from incomplete data 2006

3. Zoubin Ghahramani. Michael I Jordan Learning from incomplete data[,Tech Rep:AIM-1509] 1994

4. R J A Little. D B Rubin Statistical Analysis with Missing Data 1987

5. Jiawei Han. Micheline Kamber. 范明. 孟小峰 数据挖掘概念与技术 2001

6. J W Grzymala-Busse. M Fu A comparison of several approaches to missing attribute values in data mining 2000

7. David Heckerman Bayesian networks for data mining 1997

8. Nir Friedman. Dan Geiger. Moises Goldszmidt Bayesian network classifiers 1997(2-3)

相似文献(10条)

1. 学位论文 刘健 不可能即排除准则在缺失值情况下的分类研究及其在数据挖掘中的应用 2003

该文在研究了国内外在缺失值情况下分类的最新进展后,提出了一种全新的分类思想.文章主要内容包括以下几个部分:首先,介绍了现有的主流的分类模型——贝叶斯网络和决策树(C4.5算法),以及这些模型针对缺失值情况的改进算法——EM算法、SASC、SACMB,并讨论了它们对缺失值的不同的处理方式.其次,提出了一种新的分类思想——"不可能即排除准则"(Impossibility-Excluded Criterion,简称IEC).与以往的分类思想不同,IEC不是试图直接找到最合适的类标,而是通过排除可能性较小的候选类标来缩小选择的范围.为了表述的方便和保证IEC在逻辑上的正确性,给出相关概念的形式化定义和定理证明.再次,提出了以IEC思想为核心的新的分类算法CLassification with Impossibility-Excluded Criterion(CLIEC),给出算法描述并对算法的各个部分进行了说明.随后,提出了称之为Discretization with Universal Gravitation(DUG)的离散化算法.该算法借鉴了万有引力的模型,解决了单纯的CLIEC算法不能处理连续量的问题,并结合实例对DUG加以介绍.最后,将CLIEC应用于实际的数据挖掘项目,其良好的效果表明算法具有相当的实际应用价值.

2. 学位论文 李晓菲 数据预处理算法的研究与应用 2006

随着信息时代的来临，人类在各种领域中面临着越来越多的数据信息，与此同时，这些数据的规模还在以惊人的速度不断增长。因此，为了提高工作效率和生活质量，人们必须获取蕴藏在这些数据中的有价值信息。为了达到这个目的，人们开始致力于从数据库中挖掘知识的研究。然而，众所周知，数据库中往往存在冗余数据、缺失数据、不确定数据和不一致数据等诸多情况，这些数据成了发现知识的一大障碍。因此，在从数据库中挖掘知识之前必须对数据进行预处理。 本论文着重研究数据挖掘中的数据预处理技术，尤其是数据清洗技术，并实现了数据挖掘试验平台(DataMiningLaboratory,DMLab)的数据预处理模块的功能。 首先对数据预处理知识做了全面和详细的描述，介绍了数据预处理的研究背景、定义和主要的预处理技术研究现状等。然后对现有的数据预处理技术进行了深入的分析，涉及到数据清洗、数据选择、数据变换和数据归约等技术。之后重点对缺失值填充技术及各种填充算法进行了深入地研究和探讨，并提出了基于聚类技术的缺失值填充法。最后，在前面讨论的各种技术的基础上，实现了数据挖掘试验平台的数据预处理模块功能，主要包括数据清洗、数据选择、数据转换、数据归约等功能。 在对数据预处理技术进行的研究中，着重介绍了缺失值清洗的基本知识和方法，并探讨了当前缺失值清洗技术，客观地评价了它们的优缺点。本文对目前广泛应用的各种数据预处理技术进行了深入的研究，并在此基础上完成了DMLab系统中数据预处理模块的设计和实现，既根据系统需要实现了部分基础的预处理算法，又提出了如何应用聚类算法进行缺失值填充的新方法，并给出了在数据集上的试验结果及结论。 本文的主要创新点在于提出的基于聚类技术的缺失值填充算法。

3. 学位论文 殷杰 数据挖掘在医疗信息分析中的研究与应用 2007

数据挖掘技术在商业方面应用较早，目前已经成为电子商务中的关键技术。由于数据挖掘在开发信息资源方面的优越性，数据挖掘已逐步推广到保险、医疗、制造业和电信等各个行业。 国家军字一号医院信息系统在近7年时间里，已在军队、武警、地方的近500所医院推广使用。随着时间的推移，医院的业务数据正通过不同的途径源源不断的汇入服务器数据库中，其数据量以每日成百上千万条记录的速度快速增长。如何有效地利用这些海量的医疗信息，让“信息”变成“知识”，较好的办法是借助数据挖掘技术对医疗数据进行分析。本文选取了新桥医院最近三年内的冠心病病人的基本信息和费用信息进行数据挖掘，以建立医疗费用的分类模型。 由于各种原因，数据中存在各种程度的缺失。为了提高数据挖掘的效率和精确度，需要采取数据填补技术对缺失数据进行填补。本文在介绍现有的缺失值处理技术和对比各种算法的优劣的基础上，通过实验证实了多重填补法有较好的填补性能，故采用多重填补法对缺失数据进行填补。 本文介绍了多种数据挖掘算法。因为决策树是以实例为基础的归纳学习算法，它着眼于从一组无次序、无规则的事例中推理出决策树表示形式的分类规则，在对比适用性以后，决定采用决策树算法作为核心的数据挖掘算法进行医疗数据挖掘。 在决策树的生成过程中，测试属性的选择对决策树的优劣起着重要的作用。在测试属性的选择方面，本文首先分析了利用条件属性对样本集进行划分，得到正确划分的赞同度。利用赞同度可以找到对正确决策贡献最大的属性。以该度量为启发式信息，提出了一种基于赞同度的决策树生成算法。 赞同度决策树采用阈值预剪枝作为剪枝方法。当叶子结点的样本数量达不到某个设定阈值时，对该叶子结点所在的最小子树进行剪枝，保留该叶子结点的父结点为新的叶子结点。阈值预剪枝虽然可能得不到样本量较小的事例规则，但采用阈值预剪枝不必生成整棵决策树，且算法相对简单，效率很高。 对比CHAID，CART和QuEST算法生成的决策树，利用赞同度算法得到的决策树规模适中，分类精度和预测精度有了一定的提高。

4. 期刊论文 覃泽. QIN ZE 基于信息增益的数据库缺失值填充算法 -微计算机信息2007,23(12)

在数据挖掘以及机器学习等领域,都需要涉及一个数据预处理过程.其中,缺失值的填充是一个非常具有挑战性的任务,因为填充效果的好坏会极大的

影响学习算法及挖掘算法的后续处理过程.目前已有的一些填充算法在一定程度上能够处理缺失值问题.与已有的方法不同,提出了一种扩展的基于信息增益的缺失值填充算法,它充分利用数据集中各属性之间隐含的关系对缺失的数据进行填充.大量的实验表明,提出的扩展的基于信息增益的缺失值填充算法是有效的.

5.学位论文 韩冰 股票投资行为模式研究——基于交易数据库的数据挖掘 2007

经典金融理论认为，证券投资者是理性的，他们依据效用最大化原则进行投资。然而近年来，随着种种金融异象的发现，金融学界开始对投资者的具体决策过程和投资行为进行实验研究和实证分析，发现人的情绪、性格及心理感觉等主观因素在金融投资中起着不可忽视的作用。投资者并不总是理性的，其决策行为不仅受到自身固有的认知偏差的影响，同时还受到外界环境的干扰，于是金融投资研究中开始考虑行为科学对决策的影响，逐渐形成了行为金融学。　　中国理论界对于行为金融理论的研究时间还比较短，在金融实际投资和监管领域的应用更是少见。事实上，行为金融所发现的很多现象在中国证券市场也存在，如过度自信、羊群效应、反转效应和动量效应等等。本文以行为金融理论为指导，以我国股票投资者的交易数据作为研究对象，通过对投资者在二级市场的交易数据的挖掘，探索存在于我国股票投资者中的行为规律和交易模式。在研究我国股票投资者交易行为模式时，采用了SAS所提倡的SEMMA过程，即抽样(sample)、探测(Explore)、修正(Modify)、建模(Model)、评估(Assess)、打分(Scoring)的数据挖掘过程。

论文主要分为以下五个部分：　　第一部分回顾了行为金融学在我国的发展，对我国股票市场投资者的行为偏差进行总结。首先，我国的股票市场在市场交易、政府监管、上市公司、投资者构成等方面同国外的成熟证券市场存在极大的差异。一些上市公司将股权融资视为“免费贷款”，机构投资者可以一定程度操纵股票价格。个人股票交易者投资意识差，投机意识浓。其次，简要概述了行为金融学的基本理论，对个人投资偏差和处置效应的国内外研究成果进行了综述。　　第二部分阐述数据挖掘的方法。数据挖掘的理论基础来源于统计学，实现机制则依赖于数据库管理系统。但是数据挖掘模型同统计学原型有一定的差距，数据挖掘面对的是海量数据，要考虑数据分析的执行效率，而统计学中的种种分析方法在数学上太过完美，如果用于分析海量数据将导致低效。数据挖掘中探索模式使用的主要模型是决策树。决策树之所以有强大的吸引力，主要是因为它的结果容易解释。将枝干节点合取形成条件，叶节点就形成了结果，这就是一个规则。本文研究就是利用决策树模型探索个人交易数据中的规则，即发现交易模式。　　第三部分数据预处理。数据挖掘是针对次级数据的挖掘，其数据收集是在分析目的确定前，这种途径收集来的数据可能并不存在对挖掘有益的直接信息，需要对数据进行加工汇总。首先是数据准备过程，将原始的FoxPro格式数据转化为了SAS格式；对不同的数据库数据进行了整合，重点插补了不完整数据。本文通过自创的插补方法，充分利用最原始数据信息，对不完整的文本信息和数字信息进行插补，为后续的分析打下了良好的基础。　　第四部分原始数据的探测。按照交易者背景信息和交易记录信息两个方面进行了数据描述统计，发现数据库中存在于隐性冗余和数据异常。隐性冗余主要表现在交易者背景数据库中，交易者背景数据库是截止2006年6月为止开户的所有交易者信息，但从交易流水库看，2002年8月后开户的投资者是没有交易记录的，这就形成了隐性冗余；数据异常主要表现在本次资金余额这个字段上，理论上本次资金余额应大于0，由于2000年前证券公司存在对大客户的融资，导致了本次资金余额出现为负的情况。　　第五部分构建数据挖掘模型。首先研究了中国股票市场是否存在处置效应，实证表明在损失还未超过30%的情况下，处置效应并不明显；在损失超过了30%后处置效应显著。其次，分别对买入和卖出决策、盈亏分析、不同类型交易者差异分析三个部分分别采用决策树模型，发现人们更加注重20个交易日内的低点；交易期限低于5天的情况下，亏损的可能性增加，这个结论对没有完成交易的交易者尤为显著；盈利的能力同大盘呈现了明显的相关，由于个人的预测能力有限，不能探测何时到底，在大盘下降过程中买入股票，亏损的可能性大大增加；投资额在1-20万的投资者交易次数最多，资金量较小的投资者(20万以下)在一支股票上的总买入金额有集中趋势，而资金量超过50万的投资者没有显著差异。结合上述的研究结果，提出了相应的政策建议。

6.学位论文 朱晓峰 缺失值填充的若干问题研究 2007

数据缺失在实际应用中是经常发生的，甚至是不可避免的。造成数据缺失可能是信息(暂时)无法获取或者在操作过程中被遗漏等。数据缺失对数据挖掘的过程和结果都有十分严重的影响。数据缺失可能直接影响到模式发现的准确性和运行性能，甚至导致错误的挖掘模型。处理有缺失数据的数据集是极端困难的，因为，现有的模式发现算法通常假设输入的数据是无缺失的。于是，这些可用的模式发现算法和实际数据之间存在一条不可逾越的鸿沟。　　缺失数据的处理方法可分为删除元组、缺失数据填充和不处理三大类。Han 和Zhang等认为，从使用的频率和研究的程度等各方面来看，填充方法是最常用的一种处理缺失值的方法，因此，本论文研究如何用填充的方法处理缺失数据。填充缺失数据的方法无论是在技术上还是理论上都得到了空前的重视，国际上有很多专门机构研究这个问题，例如：　　美国宾州大学和佛蒙特大学都成立有专门的研究小组。但是，无论在统计方面还是数据挖掘领域的缺失填充方法仍然存在许多致命的缺陷。首先，现实数据集通常缺失十分严重，常见的填充方法仅仅利用没有缺失值的完全事例去填充缺失的数据。这类处理方式一方面可能要面对可用信息不足；另一方面忽略了含有缺失值的事例中的有效信息，这样不仅造成了资源浪费，而且填充效果也会出现偏差。其次，用户对所处理的数据集通常没有任何先验知识，常用的参数填充方法经常可能由于参数的错误估计而导致填充的结果严重失实，存在的非参数填充方法在技术上和理论上都很粗糙，并且只在本应用范围内十分有效，一旦被应用到其他应用领域或者一些交叉应用领域，这些在某领域内十分优秀的方法可能会导致极差的填充效果。　　上述表明，缺失数据填充是一个实际且具有挑战性的研究课题。本论文研究缺失数据填充的如下三方面问题。　　填充决策属性的缺失问题：提出的DAIM 算法能处理混合类型的条件属性，算法首次使用基于混和核的非参重复填充方法填充离散型或者连续型缺失决策属性，并且提出了一种新颖的发现最优窗宽（bandwidth）的网格搜索（gridsearch）方法,能在有限的空间内穷举式地搜索最优窗宽，大大地减少搜索空间和时间。　　在缺失值填充过程中研究了填充代价和填充代价约束的问题：本论文首次提出建立代价敏感的填充器必须考虑构造填充器的有效信息问题，算法折中考虑了经济因素和构造填充器所需有效信息来对缺失数据进行排序，提出了一个考虑填充顺序的条件属性缺失的增量式填充算法CAIM。　　条件属性和决策属性同时有缺失的问题：分析了kNN 算法中Minkowski距离公式正确选择Minkowski 参数的复杂性，提出了用灰色分析的方法代替Minkowski 距离的思想，然后分析了填充缺失值充分利用所有有效信息的必要性，并且提出非参重复填充方法来充分利用所有有效信息的理论，最后的填充算法CDAIM 能处理条件属性和决策属性同时缺失的情况。　　本论文的每种算法都用模拟数据和真实数据进行评估和分析，在各个评价指标的比较中，本论文的算法都优于存在的一些经典算法。　　本论文的主要创新点如下：　　（1）在对所处理的数据集的分布没有任何先验知识的情况下，参数填充方法经常由于错误的参数估计导致填充的结果严重失实，此时非参方法是一个很好的替换，但是存在的非参方法在技术上和理论上都很粗糙，本论文的三个算法都对传统的非参方法进行了改进。为了充分利用所有有效的信息，本论文的三个算法都采用重复填充技术。区别于存在的参数重复填充算法（例如EM 算法），本论文提出的非参数重复填充算法收敛速度要比现有的参数重复填充算法EM 算法快，且填充效果上优于一次填充或者多重填充的效果。本论文的三个非参重复填充方法既丰富了重复填充算法理论，也是对非参理论无重复算法的填补。　　（2）在核填充方法中首次引入混合核，在填充过程中能加强核函数的内插能力和外延能力；在最近邻算法中使用灰色分析代替Minkowski 距离的方法，弥补了由于选择Minkowski 参数造成填充效果不稳定的缺陷。这些研究建立了新的缺失值填充的理论、方法和技术。　　（3）首次把填充代价和构造填充器所需的有效信息综合考虑，把填充理论和代价理论有机地融合在一起进行研究。

7.学位论文 金义富 高维稀疏离群数据集延伸知识发现研究 2007

数据是当今信息社会最宝贵的一种资源，发现隐藏在那些复杂数据集中的有用知识并利用这些知识已经成为科学决策的前提。数据挖掘就是运用基于计算机的智能技术从大量甚至海量数据集中获取知识的过程，它通过关联规则、分类与聚类等方法实现从数据集中挖掘出潜在的有用知识。　　离群数据是那些与众不同的远离常规数据对象的数据，它们表现为与多数常规对象有明显差异，以至于被怀疑可能是由另外一种完全不同的机制产生的。离群数据不等同于错误数据，有的离群数据中可能蕴含着极重要的信息，如在信用卡欺诈检测、疾病诊断、网络入侵检测、通信欺诈分析、故障检测、灾害预测等诸多领域中离群点是数据分析的主要对象，在所有的科学研究领域，离群数据可能给予我们新的视角，从而导致新理论或新应用的出现，因此，对离群数据进行研究具有十分重要的意义。已有离群数据研究主要集中于离群数据挖掘，而且其挖掘的目的也仅仅是为了通过去除被发现的离群对象获得更好质量的数据集，力图为常规数据挖掘与分析提供更稳定可靠的结果，较少涉及对已发现的离群数据的进一步分析。　　本文认为对离群数据的研究包括离群挖掘与离群分析两个方面。论文的主要贡献是：以现有的离群挖掘算法为基础，重点对高维稀疏离群数据集的分类、产生来源、含义、特征以及离群趋势等进行分析，结合粗糙集(Rough Set)理论定义了离群数据关键域子空间(Key Attribute Subspace，KAs)等一系列概念，提出了相应的离群约简及关键域子空间搜索算法、离群聚类算法、缺失值处理及离群趋势分析方法等，建立了高维稀疏离群数据集特征描述及延伸知识发现的整体框架。作为一项具有创新性意义的工作，论文在研究方法与思路上力求有所突破，其主要研究成果包括如下几个方面。　　①对离群挖掘技术进行了较为全面的分析与总结，设计了一种基于k-最近邻的离群检测算法，介绍了基于分区的离群挖掘算法，详细分析与设计了基于似然的一元离群检测算法以及多元回归分析离群检测法等多种基于统计的离群检测方法，并从离群挖掘的角度探讨了聚类算法中对离群对象的处理技术，分析了不平衡分类及非频繁模式关联规则挖掘与离群检测的相似性。　　②结合粗糙集理论以离群划分的观点去揭示离群对象子空间特性，提出了离群划分相似度、离群约简等概念，其目的是寻找一个范围较小的属性子集，从这个子集中去探索离群数据集的出现原因和概率。提出的基于遗传算法(Genetic Algorithm)的离群约简技术可以较好地解决离群约简搜索问题。　　③对提出的离群对象关键域子空间KAS的意义、作用及搜索方法进行了深入地研究。基于KAS将

缺失值、普通离群点与噪声统一为离群对象，认为具有非空KAS的离群点均蕴含了一定的知识，是普通离群点，而不存在对应KAS的离群点是噪声。提出了离群包络与离群核、属性值离群状态矩阵等概念及相应的一系列KAS搜索算法，包括基于统计的、基于显著域子空间的单个离群对象KAS搜索算法，以及基于离群核、基于离群属性频度、基于统计的离群集KAS搜索算法，并对算法性能进行了分析与测试。　　④根据离群共享属性定义了离群簇，提出了簇数量、簇对象数以及相似度等离群聚类三原则，并在此原则基础上提出了基于KAS和基于离群邻接图的离群聚类算法，对算法的分类能力与性能进行了测试与比较。在离群簇分析方面，提出了离群数据的内、外及单关键域子空间分析方法以及基于离群K-最近邻的离群分析技术，并可从离群最近邻与离群簇的相互关系中获取知识。　　⑤含缺失值的对象作为一种特殊离群对象进行研究，提出了一种基于灰预测模型GM(1,1)的序列缺失数据灰插值推理方法，该算法在估计每一个缺失值时均会充分利用其时区窗口内全部信息，并建立对插补值的误差修正模型，从而可以获得性能较好的插补效果。
⑥对序列数据离群趋势进行了分析，提出了原子离群类及离群变异类等概念，研究了这两种离群类数据一般特性，给出了对象离群概率估计方法，并结合关键域子空间对属性离群频度进行了预测。

8.学位论文 唐合文 基于国家作物种质资源数据库的知识发现研究 2007

国家作物种质资源数据库拥有180种作物、39万份种质信息、135万条记录,数据量达40GB,是世界上最大的植物种质资源数据库之一.利用知识发现(KDD)的原理、方法和技术发掘这些海量数据中蕴藏的信息,已成为当前作物信息科学研究的重要内容,对于充分发挥国家作物种质资源数据库的作用,更好地保护和利用我国丰富的作物种质资源具有十分重要的意义.本研究主要进行了以下两方面的研究.　　在分析国家作物种质资源数据库数据特点的基础上,提出了基于正态模拟的连续型数据缺失值处理方法以及基于随机数的离散型数据缺失值处理方法,并结合基于语义的离散化方法对国家作物种质资源数据进行了缺失值处理和离散化处理.研究分析了统计分析、决策树、关联规则、神经网络、遗传算法、模糊集、粗糙集等知识发现方法,结合国家作物种质资源数据库的特点,提出了基于关联规则的国家作物种质资源数据库知识发现方法.在此基础上,综合分析了现有的关联规则挖掘算法,重点分析了事务数据库中关联规则挖掘的经典算法-Apriori及其改进算法的特点,根据国家作物种质资源数据库中的关联规则具有多维性的特征,改进了Apriori算法,使其适用于多维关联规则挖掘,并提出了基于sQL的国家作物种质资源数据库Apriori关联规则挖掘方法.　　研究分析了国内外典型的知识发现系统,完成了国家作物种质资源数据库知识发现系统的总体设计,研制了国家作物种质资源数据库知识发现系统的原型.该系统接口简洁直观、易操作、挖掘结果易懂.在系统中设计了支持度过滤、置信度过滤、规则前件过滤及规则后件过滤等四种方法来精减规则数量.利用该系统,开展了大豆种质资源数据库的知识发现,初步获得了有关大豆种质资源农艺性状、品质、抗逆、抗病虫等特征特性的关联知识.

9.学位论文 雷蕾 缺失数据处理技术与NBI模型 2006

数据挖掘致力于从大型数据库中挖掘有价值的信息。然而，现实世界中的数据集往往不可避免地含有一些缺失数据。这使得数据挖掘算法的性能下降，甚至影响到知识发现的有效性。本文主要研究缺失数据的处理技术，并提出一个有效可行的缺失数据处理模型，朴素贝叶斯归因模型(NBI)。
通过缺失数据灵敏度分析发现，数据集中的缺失数据对分类器的预测准确率有明显的不利影响。在各种分类器中，朴素贝叶斯分类器对缺失数据最不敏感，适用于建立缺失数据归因模型。　　在简单阐述了目前流行的缺失数据处理方法之后，本文提出了一种基于朴素贝叶斯分类器的新缺失数据处理方法——NBI模型。首先，确定需要进行归因的属性，然后将归因属性作为目标属性，利用数据集中的其他属性建立NBC分类模型，将归因问题转换为分类问题。最后，利用已建立的NBC模型预测归因属性中的缺失值，并用预测值替换缺失值，完成归因过程。　　选取归因属性需要考虑两个方面：属性所含缺失数据的比例和属性对数据挖掘任务的重要程度。属性对数据挖掘任务的重要程度可以由基于信息增益的属性重要因子或基于决策树结构的属性重要因子确定。根据归因过程中的归因顺序相关性，NBI 模型可以分为三大类策略：顺序无关策略、顺序相关策略和混合策略。NBI 模型根据属性缺失数据率和属性重要因子相互加权来确定归因顺序。本文还通过基于决策树结构的属性选择策略来改进贝叶斯分类器预测准确率，从而提高了 NBI 模型的性能。本文在多个数据集上对 NBI 模型的不同策略进行了测试。实验发现，在所有缺失数据都需要处理的情况下，顺序无关策略是一个很有竞争力的策略。与其他缺失数据处理方法相比，NBI 模型的性能优于流行的均值/众数归因法和C4．5内置模型。而且随着缺失比例的上升，NBI 模型的优势更为明显。　　最后，本文将研究成果应用于医疗数据集Clinics，并取得了良好的效果。NBI模型对提高病人住院持续时间(LOS)的预测准确率有显著作用，尤其是中期和长期的预测准确率有明显的提高。在 NBI 模型的众多策略组合中，仅对重要归因属性运用 NBI 模型进行归因，其效果要优于对全部归因属性进行 NBI 归因处理。

10.学位论文 杨涛 基因表达缺失数据填充算法研究 2005

DNA微阵列技术使人们可以同时观测成千上万个基因的表达水平，对其数据的分析已成为生物信息学研究的焦点。但是，在基因表达数据产生过程中存在一些因素导致获得的数据中包含有大量的缺失值，为后续的数据分析工作带来了极大的困难，甚至使分析结果出现严重错误。因此，基因表达缺失数据的填充是生物数据挖掘过程中的重要预处理步骤，也是研究重点之一。　　基于K个最近邻居的填充算法是基因表达数据中经典的缺失值填充算法。但算法没有考虑基因表达数据间的相关性，本文提出一种基于马氏距离的缺失值填充算法。该算法使用考虑了数据间相关性的马氏距离选择邻居基因，并利用Shannon信息熵确定更为合理的邻居基因权重系数，有效地提高了对缺失数据的填充准确度。　　模糊C-均值算法是聚类分析中广泛使用的聚类方法，在基因表达数据分析中也有较多的应用。本文利用模糊C-均值算法能很好地处理数据间的重叠性和相关性的特点，将它应用到基因表达数据的缺失问题处理中，提出了基于模糊C-均值的填充算法。算法针对不同的数据集，给出了动态确定聚类参数的方法，然后对经过初始填充的非完整基因表达数据进行聚类分析，利用聚类结果对缺失数据进行估计和填充。该算法自适应地确定聚类参数，增强了聚类的有效性，从而提高了填充结果的正确率。　　模糊C-均值算法受初始条件影响较大，在迭代过程中容易陷入局部极小。因此，论文在上述算法的基础上，利用迭代局部搜索策略来解决局部最优问题，并且使用新的聚类有效性指标优化聚类结果，较大程度上改善了聚类结果，提高了缺失值估计的准确度。实验结果表明填充准确度较原算法有较大的提高。

本文链接：http://d.g.wanfangdata.com.cn/Periodical_jsjyjyfz2006z1035.aspx

下载时间：2010年4月15日