



经管学术文库

数据预处理

——数据归约的统计方法研究及应用

刘云霞 / 著



厦门大学出版社
XIAMEN UNIVERSITY PRESS

国家一级出版社
全国百佳图书出版单位

E c o n o m i c
Management

经管学术文库

数据预处理

——数据归约的统计方法研究及应用

刘云霞 / 著



厦门大学出版社
XIAMEN UNIVERSITY PRESS

国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

数据预处理:数据归约的统计方法研究及应用/刘云霞著. —厦门:
厦门大学出版社, 2011. 3

(经管学术文库)

ISBN 978-7-5615-3825-8

I. ①数… II. ①刘… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字(2011)第 028464 号

厦门大学出版社出版发行

(地址:厦门市软件园二期望海路 39 号 邮编:361008)

<http://www.xmupress.com>

xmup@public.xm.fj.cn

沙县方圆印刷有限公司印刷

(地址:沙县长安路金沙园区 邮编:365500)

2011 年 3 月第 1 版 2011 年 3 月第 1 次印刷

开本:889×1240 1/32 印张:4.625 插页:1

字数:120 千字 印数:1~1 000 册

定价:20.00 元

本书如有印装质量问题请寄承印厂调换

摘要

数据归约是数据挖掘过程的关键环节,因此对数据归约技术的研究具有重要的意义。当前已有的数据归约方法多偏重于有监督学习,而无监督情形下归约方法的研究还相对不够丰富。鉴于这种情况,本书的重点内容是尝试对无监督数据归约的统计方法及其应用进行研究。

在本书第一章中,首先阐述了选题的研究背景和研究意义。之后,在概述相关背景知识和总结国内外数据归约研究方法现状的基础上,明确了本书的研究内容及创新之处。

数据归约两项重要的基础工作——缺失值填补和异常值探测是第二章探讨的内容。在本章,根据对统计学中常用的各种缺失值填补和异常值探测方法的分析,总结出了一些适合数据挖掘使用的方法。此外,通过将几种异常值探测方法应用在某地区移动通信用户缴费数据库上,对手机用户的消费行为进行了实证分析。

数据归约包括元组的归约和属性的归约。本书在第三章探讨了元组归约的两种主要方法——连续属性离散化和概念分层。在对当前的离散化方法和概念分层中面向属性归纳方法综述的基础上,提出了两种从独立性角度考虑的连续属性离散化方法,分别是基于可辨识矩阵的离散化方法和基于似然比假设检验的离散化方法。并通过在 Iris 样本集上对这两种方法进行模拟,验证了它们的有效性。

属性重要性排序以及属性的提取和属性子集的选择是属性归约

的两类方法。本书在第四章探讨了属性重要性的排序问题。数据挖掘中目前常见的排序问题是有监督属性的排序,本章首先对它们作了介绍和比较。然后在无监督属性重要性的排序方面,提出了单向有序列联资料的属性排序方法——改进秩和法和基于因子分析的无监督属性排序方法,这两种方法分别在一份调查问卷的列联资料 and 全国居民人均消费支出样本集的模拟中,取得了较为满意的结果。

第五章探讨的是属性的提取和属性子集的选择问题。首先对目前在数据挖掘中用于属性线性提取的几种统计学和其他学科的方法作了介绍和评价。然后是本章的重点内容——属性子集的选择,在对属性子集选择的基本知识及目前已有的研究成果详细阐述和分析之后,提出了逐步向前的无监督属性选择方法,并通过实例验证了该方法的有效性。

第六章对全书的主要工作进行了总结,并指出了有待进一步改进和完善的地方。

本书的创新之处主要有以下四个方面:

- (1) 提出了分别基于可辨识矩阵和基于似然比假设检验的两种连续属性离散化方法。
- (2) 提出了单向有序列联资料属性排序的方法——改进秩和法。
- (3) 提出了基于因子分析的无监督属性重要性的排序方法。
- (4) 提出了逐步向前的无监督属性选择方法。

关键词: 数据归约;数据挖掘;统计学

Abstract

Data reduction is the key step of Data Mining and it is important to study the methods of data reduction. Majority of existing methods pay more attention to supervised learning currently. However the study of the unsupervised data reduction wasn't abundant relatively. Therefore this dissertation focuses on the study to the statistical methods and application of the unsupervised data reduction.

In Chapter one, the backgrounds and significance of the selected topic were illustrated firstly. Afterwards, on the bases of summarizing relevant backgrounds and study methods of the data reduction from both home and abroad, we pointed out the contents and the innovative places of this paper.

In Chapter two, it was discussed the missing value imputation and the outliers detection which are the base work of data reduction. In this chapter, we summarized some methods which can be applied in Data Mining on the basis of the analysis to those statistical methods. In addition, we analyzed consumers' consumptive behavior by the methods of the outliers detection applied in the database of the some consumptive mobile telecommunication.

Data reduction includes tuples reduction and attributes reduc-

tion. In Chapter three, we discussed the discretization of continuous attributes and the concept hierarchy which are two main methods of teples reduction. On the bases of the summary of the current methods of the discretization and attribute oriented induction, we put forward two methods which were the discretization of continuous attributes based on discernibility matrix and the discretization of continuous attributes based on likelihood ratio hypothesis testing. The simulation to these methods in the Iris database validated their validation.

The methods of attributes reduction include the importance order, the extraction and the selection of attributes. In Chapter four, we discussed the importance order of attributes. The supervised importance order of attributes is familiar in Data Mining. We firstly, made an introduction to it. And then on the aspect of the unsupervised order, two methods were put forward which were the improved rank sum applied in the single ordinal contingency data and the unsupervised order of attributes based on factor analysis. The simulation to the methods of the contingency data of the survey questionnaire and national inhabitant average per person consumptive expend in the databases gained satisfying results.

Attributes extraction and attribute subset selection were discussed in Chapter five. We firstly introduced and evaluated the several methods of statistics and other disciplines applied in attributes linear extraction and followed by the main contents of this paper—attributes subset selection. After introducing and evaluating the basic knowledge and existed study productions, we put forward the method of the unsupervised stepwise forward selection. Then we

validated their validation by examples.

In Chapter six, we made a summary of this paper and raised some questions need to be improved and perfected in the future study.

The main innovation ideas in this paper are as follows:

We put forward (1) the method about the discretization of continuous attributes based on discernibility matrix and the discretization of continuous attributes based on likelihood ratio hypothesis testing.

(2) The method about the improved rank sum that applied in the single ordinal contingency data.

(3) The method about the unsupervised order of attributes based on factor analysis.

(4) The method about the unsupervised stepwise forward attributes selection.

Keywords: Data reduction; Data Mining; Statistics

目 录

摘要

第一章 绪论/1

第一节 选题的研究背景和研究意义/1

第二节 相关背景知识/2

一、数据挖掘的发展概况/3

二、数据归约的主要内容/5

三、数据归约的重要作用/6

第三节 国内外研究现状/7

一、属性离散化方法研究的现状/7

二、属性排序和属性子集选择方法研究的现状/9

第四节 本书的研究内容及创新点/11

一、本书的研究内容和结构/11

二、本书的创新点/12

第二章 缺失值的填补与异常值的探测/14

第一节 缺失值的填补/14

一、单一填补法和多重填补法/15

二、基于距离的填补方法/17

三、贝叶斯填补方法/18

第二节 异常值的探测/20

一、异常值及形成原因/20

二、异常值的探测方法/21

三、异常值探测的步骤及应用/29

第三节 移动通讯用户消费行为的分析/31

一、单个属性异常值探测的应用及分析/32

二、多个属性的异常值探测方法的应用及分析/35

第三章 数据挖掘中元组的归约/39

第一节 面向属性归纳/39

一、面向属性归纳/39

二、面向属性归纳的步骤/41

三、面向属性归纳的算法/42

第二节 连续属性离散化方法及分类/43

一、离散化方法的分类/43

二、典型离散化的过程及结果评价/45

三、相关的离散化方法/46

第三节 基于可辨识矩阵的连续属性离散化方法/52

一、基于可辨识矩阵离散化方法的基本思路/52

二、基于可辨识矩阵离散化方法的框架/55

三、基于可辨识矩阵离散化方法的统计模拟/56

第四节 基于似然比假设检验的连续属性离散化方法/58

一、基于似然比假设检验的离散化方法/59

二、基于似然比假设检验离散化方法的步骤/60

三、基于似然比假设检验离散化方法的验证/61

四、两种离散化方法结果的比较/63

第四章 属性重要性的排序/64

第一节 有监督属性重要性的排序/64

- 一、粗糙集理论中属性重要性的排序方法/64
- 二、信息论和决策树中属性重要性的排序方法/66
- 三、神经网络中属性重要性的排序方法/69
- 四、三种方法的比较/71

第二节 单向有序列联资料的属性重要性的排序/72

- 一、单向有序列联表/73
- 二、以秩效应为标准的方法/75
- 三、改进秩和法/75
- 四、改进秩和法对一份调查问卷的分析/77
- 五、以秩效应为标准的方法和改进秩和法的比较/80

第三节 基于因子分析的无监督属性重要性的排序/82

- 一、基于因子分析的属性重要性排序方法/82
- 二、基于因子分析的属性重要性排序方法的步骤/84
- 三、基于因子分析排序方法的框架/85
- 四、基于因子分析属性重要性排序方法的验证/86
- 五、值得注意的问题和局限性/88

第五章 属性的提取与属性子集的选择/90

第一节 属性的提取/90

- 一、小波变换/91
- 二、投影寻踪/92
- 三、多维标度/94
- 四、多元统计分析方法/97
- 五、几种属性提取方法的比较/98

第二节 属性子集的选择/99

一、属性子集选择方法的两个组成部分/99

二、属性子集选择方法的两种模式/102

三、基于各学科知识的属性子集选择方法/102

四、关于模式识别中基于距离的评价函数的思考/105

第三节 逐步向前无监督属性子集的选择方法/108

一、逐步选择方法的不足/108

二、逐步向前无监督属性子集选择方法的思路/109

三、逐步向前无监督属性子集选择方法的基本框架/110

四、统计模拟及方法验证/111

五、逐步向前无监督属性子集选择方法的合理性和局限性/113

第六章 全书的总结/117

第一节 全书的主要工作/117

第二节 尚需研究的问题/118

参考文献/119

后 记/128

攻读博士学位期间发表的论文/130

Catalogue

Abstract

Chapter 1 Exordium/1

Section 1 Backgrounds and significance of the selected topic/1

Section 2 Knowledge about relevant backgrounds/2

1 The developmental overview of Data Mining/3

2 The main contents of data reduction/5

3 The important effect of data reduction/6

Section 3 Domestic and overseas study actuality/7

1 The study actuality of the attributes discretization/7

2 The study actuality of the order and subset selection about the attributes/9

Section 4 Contents and innovative places of this paper/11

1 The contents and configuration of this paper/11

2 The innovative places of this paper/12

Chapter 2 The missing value imputation and the outliers detection/14

Section 1 The missing value imputation/14

1 Single imputation and multiple imputation/15

2 Distance-based imputation/17

3 Bayes imputation/18

Section 2 The outliers detection/20

1 outliers and it's cause of formation/20

2 The methods of the outliers detection/21

3 The approach and application on the outliers detection/29

Section 3 The analysis about consumers' consumptive behavior in
the mobile telecommunication/31

1 The application and analysis of the outliers detection on single attribute/32

2 The application and analysis of the outliers detection on multiple attributes/35

Chapter 3 Tuples reduction in Data Mining/39

Section 1 Attribute oriented induction/39

1 The introduction/39

2 The approach/41

3 The arithmetic/42

Section 2 The discretization of continuous attributes/43

1 The sort of the discretization methods/43

2 The course and evaluation of the model discretization methods/45

3 The relevant discretization methods/46

Section 3 The discretization of continuous attributes based on
discernibility matrix/52

1 The basic idea/52

2 The framework/55

3 The statistical simulation/56

Section 4 The discretization of continuous attributes based on
likelihood ratio hypothesis testing/58

- 1 The introduction/59
- 2 The approach/60
- 3 The validation/61
- 4 The consequential comparison of two discretization methods/63

Chapter 4 The importance order of attributes/64

Section 1 The supervised importance order of attributes/64

- 1 The importance order of attributes in Rough Sets/64
- 2 The importance order of attributes in Information theory and decision tree/66
- 3 The importance order of attributes in neural network/69
- 4 The comparison of three methods/71

Section 2 The method of the improved rank sum applied in the single ordinal contingency data/72

- 1 The single ordinal contingency data/73
- 2 The method of the rank effect/75
- 3 The method of the improved rank sum/75
- 4 The analysis of the research questionnaire with the method of the improved rank sum/77
- 5 The comparison of the rank effect and the improved rank sum/80

Section 3 The unsupervised order of attributes based on factor analysis/82

- 1 The introduction/82
- 2 The approach of the unsupervised order of attributes based on factor analysis/84
- 3 The framework/85

- 4 The validation/86
- 5 Watchful problem/88

Chapter 5 Attributes extraction and attribute subset selection/90

—————→

Section 1 Attributes extraction/90

- 1 Wavelet/91
- 2 Projection pursuit/92
- 3 Multidimensional scaling/94
- 4 Multivariable statistics analysis methods/97
- 5 The comparison of some methods about attributes selection/98

Section 2 Attributes subset selection/99

- 1 Two makeup part of attributes subset selection/99
- 2 Two pattern of attributes subset selection/102
- 3 The methods of attributes subset selection based on several subjects/102
- 4 The considering about the distance-based evaluate function in pattern recognition/105

Section 3 The method of the unsupervised stepwise forward selection/108

- 1 The shortage of the stepwise selection/108
- 2 The elements of the method of the unsupervised stepwise forward selection/109
- 3 The approach of the method of the unsupervised stepwise forward selection/110
- 4 The statistical simulation and methodological validation/111
- 5 The rationality and limitation about the method of the unsupervised stepwise forward selection/113

Chapter 6 Summarization of this paper/117

Section 1 Summary of this paper/117

Section 2 Questions need to be improved and perfected in the future study/118

References/119

Acknowledgements/128

Publication articles/130