Economic 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

Economic
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

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和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 Ming 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

Catalogue