学校编码：10384　　分类号_____密级_____
学　　号：200231046　　UDC _____

厦 门 大 学
硕 士 学 位 论 文

基于模糊聚类与分类规则的数据挖掘技术研究

Research on Data Mining Techniques Based on Fuzzy Clustering and Classification

杨 剑 雄

指导教师姓名：李 茂 青 教授
专 业 名 称：系 统 工 程
论文提交日期：2005 年 6 月
论文答辩时间：2005 年 月
学位授予日期：2005 年 月

答辩委员会主席：_____
评　　阅　　人：_____

2005 年 6 月

基于模糊聚类与分类规则的数据挖掘技术研究
杨剑雄
指导教师 李茂青教授
厦

厦门大学学位论文原创性声明

兹呈交的学位论文是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权力和责任。

声明人（签名）：

年　月　日

摘　要

数据挖掘是利用一种或多种计算机学习技术，从数据库的数据中自动分析并提取有用知识的处理过程。作为一门新兴的交叉性学科，它以数据库技术作为基础，将统计学、逻辑学、机器学习、神经网络、模糊数学、可视化计算等相关领域的成果综合在一起．数据挖掘技术已经成为目前的研究热点．而聚类分析和模式分类作为数据挖掘技术的两个重要课题，在模式识别、图像分割、图形识别等许多方面应用十分广泛．

本文认真研究了数据挖掘技术的基本理论和一般方法，聚类分析和模式分类的一般过程、主要方法以及目前国内外的研究现状，并对此进行认真的归纳和总结．在此基础上，对模糊 *c*-均值算法和基于超盒表示的分类方法进行深入的研究和讨论．

模糊 *c*-均值算法是目前理论基础较完善、应用较广泛和研究较充分的一种模糊聚类算法，但是它仍存在一些薄弱环节．特别是基于 Mahalanobis 距离的 FCM 算法，收敛速度较慢．本文提出一种改进的算法，通过对算法迭代中聚类中心的移动轨迹图的分析和使用 iris 数据集进行的 200 次实验证明，改进的算法可有效地加快算法的收敛速度，并提高识别率．

由于超盒在规则表示上的方便性，近年来，不少学者开始研究基于超盒表示的分类方法．本文结合模糊聚类提出一种基于模糊聚类的快速分类器算法，构造方法简单、易于理解；针对使用递增学习方法进行训练数据的超盒分类器的共同缺点——分类结果受训练样本排列次序影响，本文提出另一种结合 FCM 的二阶段分类器构建方法，通过与 FMMC 算法的实验对比证明该算法确实有效．将上述两种算法应用于模式识别中取得了一定的成效．

关键词：数据挖掘；模糊 *c*-均值；超盒分类器．

ABSTRACT

Data mining is an information extraction activity whose goal is to discover hidden and useful facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.

As a new subject, data mining has drawn the attention of many researchers. Clustering and classification, two of the most important branches of data mining, have been used in many fields such as pattern recognition, classification and recognition of images, etc…

In this paper, the author summarized the theory, methods and techniques used in data mining, as well as current status of research in this field around the world.

The main focus of the paper was on Fuzzy c-means (FCM) algorithm and hyperbox-based classifier methods.

While theoretically well developed and widely used, there are still some drawbacks in FCM, for example, the huge execution time for Mahalanobis-distance-based FCM. In this paper, the author proposed a modified algorithm(MFCM) to improve the convergent time of FCM-based hyperelliptical clusters and tested it on Fisher's famous iris data set. The outcomes showed that MFCM can decrease the iterative times and quicken the convergent speed effectively.

In recent years, more and more people focus their research on hyperbox-based classifiers due to the easy extraction of IF-THEN rules from hyperbox. The author also proposed two algorithms to generate hyperbox-based classifiers and tested them on a heart disease data set. The outcomes proved that the algorithms were effective and efficient.

Keywords: Data mining; Fuzzy *c*-means; Hyperbox-based classifier;

目　录

Contents

第一章　绪论

1.1　数据挖掘概述

1.1.1 引言

在当今的信息社会时代中，信息已经成为社会的三大重要资源之一，如何有效地管理和利用这些信息资源成为人们关注的焦点与技术发展的趋势．大量信息在给人们带来方便的同时，也带来了一系列问题：如何从大量信息中提取和选择所需数据、信息的真伪辨析、信息组织形式的一致化等等．计算机的快速发展给信息处理带来了巨大的方便，但随着人们对信息的需求和要求的不断增长，对数据处理技术的要求也不断的提高，数据挖掘（Data Mining, DM）即是由此而发展起来的一项新技术．

数据挖掘出现于 20 世纪 80 年代末，最早是以知识发现——从数据库中发现知识（Knowledge Discovery in Database, KDD）的研究起步的．数据挖掘是 KDD 中的一个重要步骤，但在通常的应用中，并不区分二者的概念．KDD 一词首先出现在 1989 年的人工智能国际会议上，之后 KDD 的研究得到迅速发展，自 1995 年起，知识发现与数据挖掘国际学术会议每年均召开一届，此外还有相关的地区性国际大会定期或不定期召开．目前，数据库中的知识发现和数据挖掘技术已成为研究热点和焦点，在国外已达到一定的水平并投入应用领域中，商品化的 KDD 软件工具已投放到市场，如 IBM 公司的 IBM DB2 Intelligent Miner、美国 Business Objects 公司的 Business Miner 等. 国内从事数据挖掘的研究起步较晚，但近年来许多高校、科研院所在这一领域的研究也取得一定的成绩．

数据挖掘的应用领域非常广泛，如生物医学、零售、金融、气象、电

子商务等等.

1.1.2 数据挖掘的定义

对数据挖掘的定义，不同的学者有不同的表述形式，归结起来，数据挖掘是一个处理过程，是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的，人们事先不知道的，但又是潜在的有用的并且最终可理解的信息和知识的过程．[1]

1.1.3 数据挖掘的功能

数据挖掘的任务是从数据集中发现模式，模式有许多种，按功能可分为两大类：预测型（Predictive）和描述型（Descriptive），而在实际应用中，往往根据模式的实际作用进行细分，如分类、聚类、回归等．数据挖掘所处理的数据类型也很丰富，包括数值数据、文本数据、关系数据库、Web网页等等．一般来说，数据挖掘的功能与被挖掘的数据类型有关，某些功能只能应用于特定数据类型中，而某些功能则可以应用于多种数据类型中．数据挖掘的功能具体包括以下内容[2]．

- 概念描述

概念描述（Concept Description，或称为类描述 Class Description）是通过对与某类对象关联数据的汇总、分析和比较，对此类对象的内涵进行描述，并概括这类对象的有关特征．这种描述是汇总的、简洁的和精确的，并且能够对数据分析起到重要作用．概念描述可以通过数据特征化（data characterization）和数据区分（data discrimination）两种方法获得．前者是对目标类数据的一般特征或特性的汇总，生成的描述一般是该类中所有对象的共性；后者是将目标类数据的一般特性与一个或多个对比类数据的一般特性进行比较，生成的描述一般是目标类和对比类中对象的共性．