学校编码：10384　　分类号_____密级____

学号：X2005223035　　UDC___

厦门大学

硕　士　学　位　论　文

基于广义小样本的数据挖掘方法研究

Research on the Data Mining Technique Based on Generalized Small-Samples

游文杰

指导教师姓名：吉国力　教授

王永忠　高工

专　业　名　称：控　制　工　程

论文提交日期：2009 年 5 月

论文答辩时间：2009 年 5 月

学位授予日期：2009 年　月

答辩委员会主席：_____

评　　阅　　人：_____

2009 年 5 月

厦门大学学位论文原创性声明

本人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其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

200　　年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

200　年　　月　　日

摘　要

数据挖掘涉及人工智能、模式识别、机器学习、统计学等领域，不同领域利用各自不同的技术和方法对数据挖掘进行研究，将不同领域的理论、技术进行融合是数据挖掘研究一个方法。小样本数据挖掘则是数据挖掘的一个研究方向，针对高维小样本数据如何构造挖掘算法及其效率是小样本数据挖掘的核心问题。

文章在数据挖掘前人研究基础上，提出广义小样本概念，对广义小样本数据挖掘的理论进行研究，从统计方法与统计学习角度出发，研究了小样本数据挖掘中的偏最小二乘回归、支持向量分类与回归机的相关理论。文章先对经典统计方法作了论述，指出小样本情形下算法的不足，提出了基于小样本的统计推断及其优化。进而阐述了偏最小二乘回归方法的数学原理及其小样本挖掘中的具体步骤，并定义了三个指标：自变量解释增益(*IEG*)、因变量解释增益(*DEG*)与变量投影重要性(*VIP*)，以实现基于 *PLS* 的特征选择；最后引入统计学习原理，介绍了支持向量机若干算法及其广义小样本的挖掘。

文章最后尝试*PLS*和*SVM*对广义小样本数据的信息融合。针对肿瘤亚型微阵列数据分类问题，实现基于*PLS*与*SVM*的二类分类与多类分类问题的应用；及其基于*PLS*与*SVM*的回归问题的仿真。从而实现对广义小样本数据挖掘。

第一章介绍数据挖掘与统计方法研究关系，提出广义小样本概念及小样本时的数据挖掘，并介绍本文主要工作。

第二章对数据挖掘中的经典统计方法及相关理论作了综述，指出小样本情形下算法的不足，并对小样本的统计推断做了优化。

第三章研究了小样本数据挖掘中的偏最小二乘回归模型、算法，并其相应实现步骤，提出了变量解释增益的概念。

第四章引入统计学习原理，介绍了支持向量机若干算法及其实现，实现广义小样本数据挖掘。

第五章尝试 *PLS* 和 *SVM* 对广义小样本数据的信息融合。针对肿瘤亚型微阵列数据分类问题，实现基于 *PLS* 与 *SVM* 的二分类与多分类问题的应用，所得数量少且识别强的特异基因子集；以及基于 *PLS* 与 *SVM* 的回归问题的数值仿真。

关键词：广义小样本，偏最小二乘，支持向量机，肿瘤亚型微阵列

Abstract

Data Mining included Artificial Intelligence, Pattern Recognition, Machine Learning, Statistics and other fields, different realms researched it using different techniques and methods. Data integration between theories and technologies in different areas was a researching method of data mining. New studying direction in data mining was Small Samples mining, which aimed at high-dimensional and small-sample data. How structured the mining algorithms and how enhanced the efficiency of algorithms were the core problems.

The article studies on the theory of data mining the basis of previous result, defined Generalized Small-Sample, from the viewpoint of statistical methods and statistical learning, it researched the theories on *PLS* algorithm *SVC* and *SVR* in Generalized Small-Sample. First, the article discussed methods in the classical statistics, it pointed out the fault of algorithm on small-samples, and statistical inference and optimization in small-samples was proposed. Furthermore, the article expounded mathematical principle of *PLS* regression and its mining approach in small-samples, and three indicators of *IEG*, *DEG* and *VIP* were defined in order to constructed a novel feature selection algorithm based on *PLS*; finally the statistical learning theory was introduced, some *SVM* algorithms were come true and achieved thorough mining in Generalized Small-Sample.

At the end of article, it attempted the information fusion of *PLS* and *SVM* on small-samples. Using the instance of cancer subtype microarray data classification, it achieved the two-category and multi-classification application on *PLS* and *SVM*, and simulated the regression problem based on *PLS* and *SVM*. And it realized the mining in small-sample.

Chapter I it introduced statistical methods and data mining and its relation, The concept of Generalized Small-Sample was proposed and the data mining of it was studied, at the same time it presented chiefly work on this paper.

Chapter II it reviewed the methods of classical statistics and related theories of the data mining, and it pointed out the fault of algorithms on Small-Samples, and the statistical inference in Small-Samples was optimized.

Chapter III it studied the mathematic model and algorithms of *PLS* regression based on data mining in Small-Sample and the corresponding implement steps.

Chapter IV it introduced statistical learning theory, and some *SVM* algorithms and implementation were discussed. The data mining in Small-Samples was realized.

Chapter V it attempted the information fusion of Small-Samples based on *PLS* and *SVM*. Using the instance of cancer subtype microarray data classification, it achieved the two-category and multi-classification application on *PLS* and *SVM*, and simulated the regression problem based on *PLS* and *SVM*.

Key words: Generalized Small-Sample; Partial Least-Squares (*PLS*); Support Vector Machines (*SVM*); Cancer Subtype Microarray Data

目　录

Contents

第一章　前 言

数据挖掘核心模块技术包括数理统计、人工智能、机器学习，经历几十年发展，近来开始把数据挖掘中的一些研究工作由统计方法来完成，并认为最好的策略是将统计方法与数据挖掘有机的结合起来。

统计学(*statistics* 习惯称数理统计)研究有效地收集、分析和解释数据，以提取信息、建立模型，并进行推断、预测和决策的方法和理论。人类在生产、社会和科学活动中通过实验、观测和调查获得数据(各种资料)，再从数据中获得知识。科技、经济和工农业生产的发展是其源泉与动力，也是其目的和归宿。统计学的本质、特征决定其具有广泛应用性和极强交叉性。目前使用的一些经典数据挖掘技术(如 *CART* 和 *CHAID* 等)都来自统计技术。在数据挖掘中的概率、独立性、偶然性和过适应性等概念也都来源于统计技术。统计类数据挖掘技术是数据挖掘技术中较为成熟的一种，包括数据的聚集与度量技术、回归技术、聚类挖掘技术和最近邻域挖掘技术等。

1.1　小样本数据挖掘研究意义

1.1.1　广义小样本

统计方法的精确、易理解的优点已被广泛关注。统计分析工具是的一种处于知识发现工具和信息处理工具之间的数据挖掘工具。经典的统计(估计)方法中，一般数学模型结构为已知，训练样本用来估计参数。这种方法很大局限性是它需要已知样本的分布形式，也就是需要大的样本容量，而实际问题中这需要花费很大的代价或者是根本就不可行；同时，传统统计学是研究样本容量趋于无穷大时的渐近理论，作为一门学科其研究基础是建立在大样本特性之上，现有学习方法也是基于此假设，但实际应用中，训练样本是有限的，甚至为小样本情形。

所谓广义小样本，是指样本容量 n 小于其变量个数 p，表现为高维数据小样本情形。广义小样本是一相对概念，其实质是信息冗余与高噪声，其建模方法的有效性体现在小样本数据潜在信息的充分挖掘，在最大化数据有用信息量的情况下去除冗余与噪声。许多数据挖掘算法在广义小样本时效率下降甚至失效。近年来，

为解决这类高维小样本分类问题而提出一些新方法，其中通过特征选择与特征提取可以将其转换为低维数据。因此，构造有效的特征选择方法是广义小样本的一个研究方向。在分类算法中，有效的特征选择与特征提取可以减少冗余信息与噪声对分类的影响。

1.1.2 小样本研究意义

在高新技术和国防科技中，由于产品价格和试验费用十分昂贵，在各种环境条件下所获得的试验数据中，能作为来自同总体样本的样本量相当小，而且大部分是不完全数据。在许多复杂问题中，样本量的绝对数并不算小，但其相对于数据的维数或参数个数而言，样本量就相当小。对这一类问题，经典的精确统计方法并不适用，而大样本理论的统计推断又精度较差。

许多高维数据相对其维数而言，样本量过小。如 20 世纪 90 年代 *DNA* 微阵列基因芯片,该技术使得研究人员可以同时测定成千上万个基因的表达水平,得到大量微阵列数据。而该数据的特点是样本容量较小(一般是十几或几十个) ,而变量数(基因)非常多(一般是几百、几千甚至是几万个)。微阵列数据一方面提供了极其丰富、详细的信息,但是在另一方面,这种高维小样本数据对随后的统计分析工作带来了前所未有的困难。再如，互联网的快速发展，网上出现大量文档数据，自动文本分类也成为处理海量数据的不可或缺关键技术，其中对使用向量空间模型的分类器的最主要困难是高维的特征空间，因此采用高效的文本特征选择是首要的任务。对于这一类的问题，统计工作者和实际领域中的数据分析工作者都进行了许多研究，有些方面已经形成了一些比较有效的方法，但很多方面还处在探索阶段，缺乏系统有效的方法，更缺乏完整的系统理论。

1.2 小样本数据挖掘研究进展

小样本数据处理方法之一：降维技术。从变量变换方法分为:线性降维(主成分分析,偏最小二乘法,切片逆回归)和非线性降维(等容特征映射,多维尺度)；从是否利用类别信息分为:有监督降维(偏最小二乘法,切片逆回归)和无监督降维(主成分分析,等容特征映射,多维尺度)等。近几十年来出现了自助法、随机逼近、鞍点逼近及其它高阶渐近逼近等方法。对于较复杂的数据和模型，基于样本容量较

小的数据，设法给出尽可能精确的统计推断，是一个重要的研究方向。另一方面，在实际问题中除来自所研究的总体的直接数据之外，还可能有其他一些与之有关的数据，也包含一些有关所研究总体的信息。如何充分挖掘这些数据，对于小样本问题是十分重要的。贝叶斯(*Bayesian*)分析是融合不同来源信息的较好方法，困难在于如何客观地确定先验分布，避免先验分布中的主观成分。这就是近些年讨论较多的客观贝叶斯方法。经验非线性方法，如人工神经网络(*ANN*)等，这等方法的基本思想是利用训练样本建立非线性模型，克服了传统参数估计方法的困难。但是这种方法缺乏统一的数学理论。统计学习理论(*SLT*)是一种研究小样本下机器学习规律的新理论，不仅考虑了对渐近性能的要求，而且追求在现有有限信息条件下得到最优结果。

1.3 本文内容提要

文章在数据挖掘前人研究基础上，对小样本数据挖掘的理论进行研究，从统计方法与统计学习角度出发，研究了小样本数据挖掘中的偏最小二乘回归、支持向量分类与回归机的相关理论。文章先对经典统计方法作了论述，指出小样本情形下算法的不足，提出了基于小样本的统计推断及其优化。进而阐述了偏最小二乘回归方法的数学原理及其小样本挖掘中的具体步骤；最后引入统计学习原理，介绍了支持向量机若干算法及其实现广义小样本数据挖掘。文章最后尝试 *PLS* 和 *SVM* 对广义小样本数据的信息融合。针对肿瘤亚型微阵列数据分类问题，实现基于 *PLS* 与 *SVM* 的二分类与多类分类问题的应用；及其基于 *PLS* 与 *SVM* 的回归问题的仿真。从而实现对广义小样本数据挖掘。

第二章　数据挖掘中(传统)统计方法

从传统意义上讲，统计分析方法不是数据挖掘。统计技术是由数据驱动的，并用来发现模式和建立预测模型。从用户的角度，在解决数据挖掘问题时，有时会先用统计方法去试着解决问题，在数据预处理时也非常有用。统计分析是利用统计学原理，对总体中的样本数据进行分析得出描述和推断该总体信息和知识的方法，这些信息和知识揭示总体中的内部规律。数据挖掘统计分析方法是统计分析方法学的延伸和扩展，就数据挖掘算法本身，很大一部分可以从数理统计中获得理论解释。因统计方法有成熟的数学基础，可以很好的对数据进行解释，在数据挖掘中有大量的运用，在数据挖掘中传统的统计方法有参数估计、假设检验、回归分析、主成分分析、判别分析、聚类分析等。

2.1　数据挖掘常用统计方法及数学基础[6,8,11]

2.1.1　线性模型参数估计

2.1.1.1　参数估计概念

统计抽样研究的目的是用样本的信息推断总体特征，这叫统计推断。统计推断包括：参数估计和假设检验。参数估计是用样本指标（称为统计量）估计总体指标（称为参数）。

用样本均数估计总体均数以及用样本率估计总体率。

点估计：直接用样本的统计量估计总体参数的估计值的方法称为点估计。

点估计的方法简单，缺点是没有考虑抽样误差，而抽样误差在抽样研究中是不可忽视的。

区间估计：按一定的概率估计总体参数所在的范围的方法。

置信区间：总体参数的所在范围，该区间以一定的概率（如 95%或 99%）包含总体参数。

2.1.1.2　计算置信区间

总体均数 m 的置信区间计算，根据总体标准差 s 是否已知及样本容量 n 的大