学校编码：10384　　分类号_____密级_____
学　　号：23220111153241　　UDC _____

厦门大学

硕　士　学　位　论　文

高维小样本数据的特征选择研究及其稳定性分析

Research on Feature Selection and Stability Analysis for High Dimensionality Small Sample Size Data

宁永鹏

指导教师姓名：周绮凤　副教授

专　业　名　称：模式识别与智能系统

论文提交日期：2014 年 4 月

论文答辩时间：2014 年 5 月

学位授予日期：2014 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2014 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）：

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）：

年　　月　　日

摘 要

随着生物信息学、图像处理、文本挖掘等大规模数据挖掘问题的不断涌现，数据挖掘的研究对象越来越复杂，对象的特征维数也越来越高。在现实生活及科学研究中产生了大量的高维小样本数据，如果直接利用这些高维小样本数据进行数据挖掘，容易出现维数灾难问题。通过特征选择，可以删除高维小样本数据中的冗余特征和噪声特征，从而降低学习算法的时间和空间复杂度，避免维数灾难。

已有的特征选择方法主要侧重于特征选择结果的高分类性能或者聚类性能，而忽略了特征选择结果的稳定性。特征选择的稳定性问题对于高维小样本数据的数据挖掘与机器学习过程是非常重要的，不稳定的特征选择结果将带来很多歧义，难以获取可以理解的真实特征。本文以高维小样本数据的特征选择及其稳定性为研究对象，做了如下主要工作：

1. 通过大量地阅读特征选择及其稳定性的相关文献，系统地介绍了特征选择稳定性的概念、意义，详细地整理了已有的稳定性度量方法，对现有的稳定性特征选择方法进行整理研究，为后续的研究打下基础。

2. 提出了一种高维小样本数据的特征选择方法——基于随机森林的递归聚类消除特征选择方法RF-RCE. RF-RCE是在SVM-RCE以及ISVM-RCE的基础上提出的。RF-RCE在ISVM-RCE的框架上使用随机森林的特征重要性给特征评分，由于随机森林在处理高维小样本数据集上的优越性，使得RF-RCE在保持分类准确率和特征选择的稳定性的基础上，极大地提高了特征选择的时间效率，并且能够解决ISVM-RCE不能解决的超高维数据集。

3. 为了提高特征选择的稳定性，本文系统地整理并分析了特征选择不稳定的原因，并进行了大量的实验验证，此外本文引入了一种新的稳定性度量方法，该度量方法同时考虑了基于特征子集和特征排序的稳定性度量方法。在已有的稳定特征选择方法的研究基础上，本文提出了一种基于随机森林思想的稳定特征选择方法——随机集成特征选择方法REFS，通过在多个高维小样本数据集上进行实验，验证了所提方法的有效性。

关键词：高维小样本；特征选择；稳定性；随机森林

Abstract

With the rapidly development of bioinformatics, image processing, text mining and other large-scale data mining problems, the study of data mining is more complex. In real life and scientific research, a lot of high dimensionality small sample size data were generated, if we use these high dimensionality small sample size data for data mining directly, it will prone to the curse of dimensionality. Feature selection can reduce the dimensionality of high dimensionality small sample size data by remove redundancy features and noise characteristics, improve the classification accuracy, reduce the algorithm complexity, and avoid the curse of dimensionality.

Existing feature selection methods ignore the stability of feature selection, while feature selection primarily focuses on the classification performance and clustering performance. Stability of feature selection is the insensitivity of the result of a feature selection algorithm to variations to the training set. Stability of feature selection is very important for data mining and machine learning process of high dimensionality small sample size data, unstable feature selection results will bring a lot of ambiguity, and it is difficult to get the understandable feature subset. This paper researches on the feature selection and its stability for high dimensionality small sample size data. The main contributions are summarized as follow:

1. In this paper, we review feature selection models, and review some proposed methods and approaches that aim to stabilize feature selection results. We also review the approaches to evaluate stability of feature selection method. In addition, the stability measurements are systematically reviewed in this paper.

2. This paper proposes a feature selection method, RF-RCE (Random Forests Recursive Cluster Elimination) feature selection, for high dimensionality small sample size data. RF-RCE is proposed based on SVM-RCE and ISVM-RCE, and RF-RCE use the Random Forest variable importance to score feature. Because of the superiority of the random forest to deal with high dimensionality small sample size, RF-RCE greatly improve the computational efficiency of ISVM-RCE, meanwhile it

achieves the close classification accuracy and stability. Also RF-RCE can solve the ultrahigh-dimensional data which ISVM-RCE cannot be resolved.

3. In order to improve the stability of feature selection, this paper systematically collates and analyzes the causes of instability of feature selection. This paper also introduces a new stability metrics, which taking into account the feature subset and feature ranking. Moreover, this paper proposes a stable feature selection method based on random forests-REFS (Random Ensemble Feature Selection). By conduct experiments on a lot of high dimensionality small sample size data verify the effectiveness of the proposed method.

Key words: high dimensionality small sample size; Feature selection; Stability; Random Forests

目 录

Contents

第一章 绪论

1.1 研究背景及意义

数据挖掘是目前人工智能和数据库领域研究的热点问题，所谓数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知的并存在潜在价值的信息的非平凡过程[1]。但是随着生物信息学、图像处理、文本挖掘等大规模数据挖掘问题的不断涌现，数据挖掘的研究对象越来越复杂，对象的特征维数也越来越高，在现实生活及科学研究中产生了大量的高维小样本数据。大多数高维小样本数据的特征空间中存在许多冗余特征和噪声特征，这些特征一方面可能降低分类或聚类的精度，另一方面会大大增加学习的时间及空间复杂度。为了从大量的高维小样本数据中挖掘出有用的知识，特征选择已成为高维小样本数据分类、聚类或者回归中的关键问题。

特征选择是模式识别、机器学习和数据挖掘等领域的一个热门研究课题，受到广泛的重视。特征选择是从原始特征集中挑选出一些最有效的特征以降低特征空间维数的过程[2]。特征选择不改变原始特征空间的性质，只是从原始特征空间中选择一部分重要的特征，组成一个新的低维空间，其本质就是一个寻优的过程[3]。对于机器学习问题，一个好的学习样本是训练分类器的关键，样本中是否含有不相关或冗余信息直接影响着分类器的性能，因此研究有效的特征选择方法至关重要。

特征选择通常被视为研究数据挖掘问题的第一个步骤，相当于数据预处理过程，特别是对于高维小样本数据，通过特征选择，研究者可以去除大量冗余和不相关的特征，这样可以有效地降低特征空间的维数，从而提高对目标函数的预测性能，降低数据挖掘问题的分析成本。此外，经过特征选择得到的特征更容易被人理解，有利于挖掘底层数据中蕴藏的有用信息。

特征选择方法从研究之初到现在，已经有了很多成熟的方法，使用特征选择方法可以找到具有较好可分性的特征子集，可以有效地降低数据挖掘的时间和空间复杂度。但是，现有的特征选择方法主要侧重于特征选择结果的高分类性能或者聚类性能，而忽略了特征选择结果的稳定性。特征选择的稳定性是指特征选择

结果对训练集变化的不敏感性[4]，该问题对于高维小样本数据的数据挖掘与机器学习过程更为重要，例如在基因表达数列研究中，当收集的样本发生变化时，特征选择方法得到的基因子集或者基因重要性排序结果差别较大，那么专家就会对基因选择的结果产生疑虑，因此特征选择的稳定性也是至关重要的。近年来，特征选择的稳定性受到越来越多的重视，成为了目前的一个研究热点。由于特征选择的稳定性度量独立于分类模型，因此稳定的特征选择方法不一定可以取得良好的分类准确度。在实际应用中，进行特征选择通常是为了后续分类算法的进行，所以我们要将特征选择的稳定性度量和分类的准确率综合起来考虑，以得到一个稳定性强分类准确率高的配置。因此，寻求能保持分类性能且稳定的特征选择算法是我们的主要研究任务。

1.2 国内外研究现状

1.2.1 特征选择的研究现状

特征选择是模式识别、机器学习和数据挖掘等领域的一个热门研究课问题，特征选择的主要作用体现为：减轻维数灾难、提高泛化能力、加快学习过程、决定相关特征和特征空间的维数约简。

自从上个世纪 60 年代起就有研究者对特征选择问题进行研究，主要研究的是训练样本类别已知的问题，即有监督的特征选择研究。实际研究中，主要有三种特征选择问题[5]：(1) 从原始的特征集中选择给定个数的特征，最小化分类器的错误率，这属于无约束组合优化问题；(2) 对于人为设定的允许分类错误率，寻找维数最小的特征子集，这属于有约束组合优化问题；(3) 在特征子集的维数和分类错误率之间进行折中。这三种特征选择都属于NP-hard问题，除了穷举搜索方法之外，其他方法都不能保证得到最优解。在原始特征空间维数较低的情况下，尚可用穷举法解决，但对于高维问题，穷举法基本上是不可行的。

近年来，特征选择的研究呈现出多样化和综合性的趋势，根据特征选择方法是否独立于后续的分类算法，目前已有的特征选择方法可分为过滤式(Filter)、封装式(Wrapper)和嵌入式(Embedded)三种[6]。Filter方法与后续分类算法无关，直接利用训练数据的统计性能评估特征。Wrapper方法利用后续分类算法的性能作为