

**2017级本科生科研实践论文**

**中医药数据特征选择分析平台**

学 号： 201701014011

姓 名： 赖智豪

学 院： 计算机学院

指导老师： 杜建强、李郅琴

专 业： 计算机科学与技术

完成时间： 2020年1月7号

**科研实践论文独创性声明**

本人郑重申明：

1．所呈交的科研实践论文，是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得江西中医药大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了致谢。  
  
  
  
2．江西中医药大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交科研实践论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权江西中医药大学办理。  
  
  
  
3.本人完全了解江西中医药大学有关保留、使用科研实践论文的规定，即:学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

科研实践论文作者（签名）：赖智豪

2020年 1 月7 日

**摘 要**

近年来，随着科学技术的发展，我国中医药事业的现代化正在不断推进。火热的数据挖掘等技术也正应用于我国中医药领域。中医药领域的数据大多呈现高维度小样本的特征，高维度小样本数据直接用于分析研究，不论是精度还是效率都不太理想。原因是高维度的数据计算复杂度太大，而且往往存在较多冗余和无关特征数据。高维度数据的数据挖掘是一项极具挑战性的任务，容易引发“维数灾难”与过拟合[1]。针对这个问题，特征选择为数据挖掘更加有效的应用于高维度数据提供了强有力的保障。

以推动中医药事业的现代化，提高中医药领域工作者的工作效率为出发点，本次科研实践，经过需求分析，相关文献的整理等过程，实现了一个中医药数据特征选择分析平台。本平台特征选择采用了两个特征选择算法：一是基于文献[2]加以修改而实现的四阶段特征选择算法(four-Staged Feature Selection),以下简称FSFS算法；二是Lasso算法。FSFS算法是一种新颖的混合式特征选择方案，Lasso算法主要是用于和FSFS算法作对比，结果表明FSFS算法具有一定的价值。

**关键词：**特征选择；FSFS算法；降维；中医药数据

**ABSTRACT**

In recent years, with the development of science and technology, the modernization of Chinese traditional medicine is advancing. Hot data mining and other technologies are also being applied in the field of traditional Chinese medicine in China. Most of the data in the field of traditional Chinese medicine show the characteristics of high-dimensional and small samples. The high-dimensional and small sample data are directly used for analysis and research, and neither the accuracy nor the efficiency is ideal. The reason is that the computational complexity of high-dimensional data is too large, and there are often more redundant and irrelevant feature data. Data mining of high dimensional data is a challenging task, which may lead to "dimension disaster" and overfitting [1]. In view of this problem, feature selection provides a strong guarantee for the more effective application of data mining to high dimensional data.

With the purpose of promoting the modernization of traditional Chinese medicine and improving the work efficiency of workers in the field of traditional Chinese medicine as the starting point, this scientific research practice has realized a platform for selecting and analyzing the characteristics of traditional Chinese medicine data through the process of demand analysis and relevant literature sorting. This platform Feature Selection adopted two Feature Selection algorithms: one is based on the literature [2] the four phases of modified and implemented Feature Selection algorithm (four - Staged Feature Selection), hereinafter referred to as the FSFS algorithm; The second is Lasso algorithm. FSFS algorithm is a novel hybrid feature selection scheme. Lasso algorithm is mainly used to compare with FSFS algorithm. The results show that FSFS algorithm has certain value.

**Key Words:** Feature selection; FSFS algorithm; Dimension reduction; TCM data

# 引言

## 研究背景和意义

中医药是我国医学科学的瑰宝，数千年来为中华民族的繁荣昌盛作出了重要贡献。经年来，国家对中医药事业发展的重视力度不断加强。党的十八大以来，习近平总书记吧中医药事业发展放在全面深化改革、进一步扩大对外开放的战略高度，融入实现“两个一百年”奋斗目标、实现中华民族伟大复兴中国梦的伟大实践，多次作出重要指示。习近平总书记指出：“中医药学是我国各族人民在长期生产生活和同疾病做斗争中逐步形成并不断丰富发展的医学科学，是我国具有独特理论和技术方法的体系”，“凝聚着深邃的哲学智慧和中华民族几千年的健康养生理念及其实践经验”，强调“中医药学是中国古代科学的瑰宝，也是打开中华文明宝库的钥匙”。

我国中医药政策体系和战略逐步完善，中医药在国家医药顶层设计中的地位越发高企和吃重，我国推动中医药走向现代化和国际化的决心不容置疑。推动我国中医药走向现代化和国际化需要充分利用当今的科学技术。目前比较火热的技术有数据挖掘，机器学习，深度学习等，将中医药与其他领域交叉融合，有利于促进两个领域的发展。其中，数据挖掘技术已经相对成熟，它已经广泛应用于各个领域。我国也正将数据挖掘应用于中医药领域。将数据挖掘技术应用于中医药数据分析有利于提升效率，更加充分的发掘出数据的价值。在这个大数据时代，中医药数据更加趋向于高维度小样本的特征。 但是高维度数据的数据挖掘是数据挖掘所面临的一个极具挑战的任务，容易引发“维数灾难”和过拟合。而特征选择正是针对这类问题所进行的。对数据特征选择就是一个数据降维的过程。特征选择能够去除冗余，无关以及噪声特征，能够减少训练或者运行时间、提高模型精确度。对中医药数据特征选择算法进行研究，能够推动数据挖掘在中医药领域的发展。同时，基于特征选择算法，实现中医药数据特征选择分析平台，将有利于提高中医药科研工作者的工作效率。

## 国内外研究现状

特征选择在国内外都有广泛的研究。基于评价准则分为过滤式和封装式。

对于过滤式特征选择(Filter)，最为常见的有卡方检验(CHI-square )、信息增益( Information Gain )、互信息(Mutual Information)、几率比(Odds Ratio) 、NGL系数(NGL Coefficient) 、GSS系数( GSS Coefficient ) [3]、基于模糊度量的特征选择(Ambiguity Measure，AM )[4]、改进的基尼指数( Improved Gini index )[4]、最佳项( Best Term )[5]、基于泊松分布的度量(Measure using Poisson distribution)[6]、基于二项假设检验的特征选择（Bi-Test）[7]、基于综合度量的特征选择（CMFS）[8]等。张平[9]基于相关性，提出加权相关性WR，同时提出了一种新颖的条件相关特征选择算法CRFS(Conditional Relevance Feature Selection),并在理论上证明了它的优越性，除此之外，从不同角度，在实验上取得了较好的结果。张尧[10]基于互信息提出了一种基于最小条件和最小条件冗余的特征选择算法(MCRMCR,Minimum Conditional Relevancy and Minimum Conditional Redundancy)和WCRFR(Weight Composition of Feature Relevancy)算法，并在理论与实际中证明这两个算法可以提高分类器的分类准确率。

封装式特征选择算法(Wrapper)依赖具体的分类器算法,虽然该方法的计算复杂度较高，但所选的特征子集的规模相对较小，分类性能较好。现在有很多分类算法被应用到封装式特征选择算法中，并取得了较高的分类效果，例如决策树[11]、Fisher 判别分析结合遗传算法[12]、遗传算法结合神经网络[13]、支持向量机(SVM)[14]等。近年来，有很多学者提出了过滤式特征选择方法和封装式特征选择方法相结合的特征选择算法。Huang[15]分别利用多种过滤式特征选择算法选择特征子集，然后利用基于多个分类算法的投票策略来预测股票走势。Foithong[16]提出了基于互信息和粗糙集的封装式特征选择算法。

针对高维小样本数据的特征选择，国内外对其也有所研究。张靖[[17]等针对高维小样本数据会导致“维数灾难”的问题，提出一种嵌入特征选择方法K-split Lasso来降维，提高分类模型的精度，解决计算复杂度高的问题。针对高维小样本数据会导致“过拟合”问题，结合过滤方法和嵌入方法的优点，并在此基础上提出一种新的混合特征选择方法GSIL，目的是从高维数据中选出具有强类别区分能力的特征子集，解决“过拟合”问题。张凯[18] 提出一种基于分组的互信息过滤型特征选择算法。通过考虑单个特征与类别的相关性,同时在以往典型特征选择算法的基础上同时加入了对特征冗余局部性的考量,并提出一种基于分组的互信息过滤型特征选择算法(Group based Filtering Feature Selection Algorithm,GBFS)。在多个数据集上的实验表明,两种方法在选出较小子集的同时可以有效的提高分类精度。

# 相关技术与算法介绍

## 特征选择简介

## 特征选择相关算法

### Losso

### FSFS

# 系统设计

## 需求分析

## 产品设计

# 系统设计规范

## 程序设计规范

### 运用范围

本系统采用Python语言在PyCharm平台进行开发，适用于安装了Anaconda2-4.2.0和Python2.7、PyCharm-2017.1.3以上版本并配置了相应的运行环境的电脑上运行。可用于高维小样本实验数据（维数较高的），且均具有多重线性相关和无关的还有冗余的数据，而观测数据样本量又较少时，及一切想运用融入CI\_AMB进行去无关和冗余的多元线性回归建模的实际问题中。

### 数据格式

程序的入口一般是先打开系统，然后导入用户数据，用户的数据格式要求为文本文件(\*.txt;\*.xlsx)。导入的实验数据先是放在表格中。

### 数据操作

程序导入数据后，导入的实验数据先是放在表格中，要求用户选择数据，用户选择数据对应的变量名进行导入和导出，可以选择一部分变量也可以选择全部的变量，实现变量的筛选。同时用户也可以对导入的实验数据进行复制、粘贴、删除、剪切等操作。界面输出的数据（如最大信息系数、均方根误差等）均可以选中这些数据进行复制。

## 程序流程规范

当用户选择数据后，需要选择导入算法中的自变量和因变量，这里用户必须得选择自变量和因变量且都不能为空，否则将不能继续进行下去。选择完自变量和因变量后，单击模型参数设置进入CI\_AMB窗口可以直接进行开始建模。如果没有点开始建模就直接点击生成图形，将会弹出请先点击开始建模按钮对话框。这样设置是为了要符合逻辑。建模完成后在文本框上出现在不同K下的RMSE以及K=5时的处理数据的过程，然后点击生成图形可以在图上非常清晰的看出不同K下的RMSE并得出K=5时RMSE有最小值。用户还可以查看过滤无关特征后的候选特征子集和最优特征子集的特征。然后点击功能面板，用户可以点击过滤无关特征比较图和剔除冗余特征比较图模型图，然后看到原始特征子集分别与候选特征子集和最优特征子集的RMSE的比较。

## 命名规范

良好的命名规范是一种习惯，有利于团队的合作开发。为了更加清楚的展现程序及团队的便捷合作，本系统所有的命名采用了一系列有规范的命名方式并与其程序本身的功能相适应，关系清晰，普通人都非常容易看懂它的含义。

### 文件命名

系统中Python文件命名，都是由单词组成，且这些实现一个界面的文件一般都会放在一个文件夹中。.py文件一般由几个单词组成，例如第一个单词为CI\_AMB，第二个单词为Main或Model1或Model2，表示基于CI\_AMB的算法。同时文件名必须保证唯一性，可以根据不同的算法和内容给予不同的命名。

### 类命名

采用大小写混合的方式，名字的首字母使用大写，即Python类的命名规则；类名是个名词，大写的字母用来间隔，其它的字母为小写，即驼峰格式形式。

例如主窗口的类名：MainFrame。第一单词为Main，第二单词为Frame，首字母大写，表示主窗口的类。还有其他的类的命名：PageOne、PageTwo、PageThree。表示第一页选项卡、第二页选项卡和第三页选项卡。

### 控件命名

在此系统中，每个控件都是由其代表的意义，用大写字母开头，其他字母为小写，大写字母分隔单词的方式命名。

如控件按钮名：Button3、Button6，或者其它控件名：text\_top01、textContents001、OpenButton、CheckBox等。

### 方法命名

方法命名的规则：在导入数据的界面中根据方法的作用命名方法。在CI\_AMB的界面中第一个单词是以On开头；之后的单词为该方法对应的控件名，则以大写字母作为开头。

如一些控件按钮和方法所对应的方法名：On\_Button3、judge、readtxt等。

# 平台的实现

## 开发环境简介

## 系统界面展示

# 实验结果与分析

## 结果

## 分析

# 总结与展望

## 本文总结

## 展望

参考文献

* 1. 王 翔 ，胡学钢 .高维小样本分类问题中特征选择研究综述，2017．09
  2. Ayça Çakmak Pehlivanlı, A novel feature selection scheme for high-dimensional data sets: four-Staged Feature Selectionl, 2016，Vol. 43, No. 6, 1140–1154.
  3. SEBASTIANI F. Machine Learning in Automated Text Categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.
  4. MENGLE S S R, GOHARIAN N. Ambiguity Measure Feature-Selection Algorithm [J]. Journal of the American Society for Information Science and Technology, 2009, 60(5): 1037-1050.
  5. FRAGOUDIS D, MERETAKIS D, LIKOTHANASSIS S. Best terms: an efficient feature-selection algorithm for text categorization [J]. Knowledge and Information Systems, 2005, 8(1): 16-33.
  6. OGURA H, AMANO H, KONDO M. Feature selection with a measure of deviations from Poisson in text categorization [J]. Expert Systems with Applications, 2009, 36(3, Part 2): 6826-6832.
  7. YANG J, LIU Y, LIU Z, et al. A new feature selection algorithm based on binomial hypothesis testing for spam filtering [J]. Knowledge-Based Systems, 2011, 24(6): 904-914.
  8. YANG J, LIU Y, ZHU X, et al. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization [J]. Information Processing and Management, 2012, 48(4): 741-754.
  9. 张平.基于过滤法的信息论特征选择算法研究,2018.
  10. 张尧.基于互信息的特征选择方法研究,2019.
  11. HSU W H. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning [J]. Information Sciences, 2004, 163(1-3): 103-122.
  12. SIEDLECKI W, SKLANSKY J. A note on genetic algorithms for large-scale feature selection [J]. Pattern Recognition Letters, 1989, 10(5): 335-347.
  13. LI L, WEINBERG C R, DARDEN T A, et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method [J]. Bioinformatics, 2001, 17(12): 1131-1142.
  14. SHIMA K, TODORIKI M, SUZUKI A. SVM-basedfeature selection of latentsemanticfeatures [J]. Pattern Recognition Letters, 2004, 25(9): 1051-1057.
  15. HUANG C J, YANG D X, CHUANG Y T. Application of wrapperapproach and compositeclassifier to the stocktrendprediction [J]. Expert Systems with Applications, 2008, 34(4): 2870-2878.
  16. FOITHONG S, PINNGERN O, ATTACHOO B. Feature subset selection wrapper based on mutual information and rough sets [J]. Expert Systems with Applications, 2012, 39(1): 574-584.
  17. 张靖.面向高维小样本数据的分类特征选择算法研究,2014.
  18. 张凯.高维小样本数据的互信息特征选择方法研究