甄志龙 ◆ 著

文本分类中的特征选择方法研究

Research on Feature Selection Methods in Text Categorization

吉林大学出版社

长师范学院学术著作出版基金资助出版

文本分类中的特征选择方法研究

甄志龙/著

吉林大学出版社

图书在版编目(CIP)数据

文本分类中的特征选择方法研究 / 甄志龙著. — 长春：吉林大学出版社, 2016.11
ISBN 978-7-5677-8218-1

Ⅰ. ①文… Ⅱ. ①甄… Ⅲ. ①数据采集-特征抽取-选择方法-研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 284608 号

文本分类中的特征选择方法研究

甄志龙 著

责任编辑：张树臣 责任校对：张树臣 封面设计：张沭沉
吉林大学出版社出版、发行 长春科普快速印刷有限公司 印刷
开本：787×1092 1/16 2016 年 11 月 第 1 版
印张：9.25 字数：160 千字 2016 年 11 月 第 1 次印刷
ISBN 978-7-5677-8218-1 定价：46.00 元

社址：长春市明德路 501 号 邮编：130021
发行部电话：0431-89580028/29
网址：http://www.jlup.com.cn
E-mail:jlup@mail.jlu.edu.cn

前 言

现代社会步入了一个信息时代，各种信息资源无处不在，作为主要信息载体之一的文本信息以超乎人们想象的速度不断地涌现。大量的文本信息超过了人们可以直接处理的范围。为了有效地管理这些信息，自动文本分类技术自然地受到人们的普遍关注，它是解决海量文本信息的有效手段。自动文本分类技术可以帮助我们将大量的文本自动地按照类别或主题的方式进行存储、组织和管理，便于文本的检索、阅读和处理，已经成为信息检索、文本挖掘和自然语言处理等领域中的主要研究方向之一。

向量空间模型是文本分类中标准的文本表示模型，它将每篇文本表示成以抽取的单词或词组的权重为分量的向量形式。问题是文本向量通常都是成千上万维，如果直接处理这些高维数据必然会带来维数灾难、过拟合、降低文本分类精度等诸多问题，所以特征降维是自动文本分类过程中关键的一步。特征选择是实现特征降维的有效途径之一，它不仅有助于降低文本分析的工作量，还可以提高文本分类的效率和效果。因此，研究文本分类中特征选择方法是现实的需要，具有学术研究意义和实际使用价值。

文本频率、信息增益和卡方统计量是最有代表性的文本特征选择方法。在这三种方法思想的启发下，本课题从概率与数理统计、信息论和模糊集理论三方面出发，提出了几种新的文本分类中的特征选择方法。

（1）利用类标记信息提出了类文本频率的概念，类文本频率能够体现文本反映特征在各个类别中的差异情况。定义类文本频率的极差和变异系数量化特征在类上差异性的大小，进而判断特征的重要性。

（2）考虑特征选择中的三种关键因素定义了一个新的关于特征出现时

条件概率分布的泛函作为评价特征的准则。

（3）在信息论中，相对熵是度量距离的有效方法之一。利用相对熵的含义定义了特征的相对熵且构造了一个新的卡方统计量检验特征与类别之间的相关性，提出新的特征评价函数。

（4）为了更好地利用权重信息，本文从特征项与类之间关系的角度，提出了一种基于模糊关系的特征选择方法。

（5）将神经网络敏感性的思想应用于特征选择领域中，以径向基函数网络为模型，通过对每一维输入特征的变化而引起的输出变化的敏感性衡量特征的贡献。

内容提要

文本分类面临的一个主要问题是高维的特征空间，特征空间的维数可达数万维，甚至几十万维，这必定会给文本分类带来不利的影响，特征选择是实现特征降维的有效途径。文本频率、信息增益和卡方统计量是常用的、有效的文本分类特征选择方法。本文以文本分类为研究背景，特征选择为主要研究内容。分析了上述三种文本特征选择方法存在的一些问题，从不同角度提出了一些文本分类中的特征选择方法。另外，本文利用神经网络的敏感性解决特征选择的问题。本文的工作和主要贡献包括以下几方面：

1. 由于文本频率的统计粒度略粗，可能会误删重要的特征及不宜进行大幅度降维。因此，本文提出了基于类文本频率的特征选择方法。它是一种有监督的特征选择方法，首先定义特征的类文本频率用于反映特征在类上的差异，其次给出了类文本频率的极差和变异系数来量化特征在各类上差异性的大小，进而评价该特征对分类的重要性。实验表明，该方法提高了文本分类的性能，取得了很好的效果。

2. 采用信息增益进行特征选择时，特征不出现对于文本分类的贡献远比它所带来的干扰小，从而降低了信息增益特征选择方法的性能。针对这一问题，本文仅考虑特征出现的情况，分析特征出现的概率、特征出现时条件概率分布和特征出现的类别数三种因素对特征选择的作用，从概率统计角度提出了一种面向特征概率分布的文本特征选择方法。与信息增益不同，这种新方法不是基于信息熵的度量而是采用特征出现的条件概率分布的离散度概念来度量特征的重要性。通过在标准的文本数据集上的实验表明该方法具有较好的稳定性和很高的文本分类性能优势。

3. 利用卡方统计量进行特征选择要从局部打分到全局打分，因此既需要局部打分准则又需要全局打分准则。本文分别从这两步入手，提出了基于相对熵的卡方统计量特征选择方法。一方面，提出特征的相对熵定义用来区分特征与每个类的相关性，并利用特征的相对熵作为特征的全局打分准则。实验表明特征相对熵全局评价准则取得了更好的分类效果。另一方面，特征和类别被看作是一对随机变量，利用它们在独立和非独立情况下联合熵的差值研究特征与类别之间的相关性。经过理论推导这个差值的计算表达式就是相对熵，从而给出了相对熵卡方统计量，实验结果也显示了该方法的有效性。

4. 对已有特征选择方法分析后发现，这些方法只利用了文本数的信息来选择特征，没有考虑特征项的权重。本文把特征项权重、文本数和类别信息结合在一起，通过计算词频权重和对数熵权重的基础上，从特征项与类的模糊关系角度出发，给出了一种基于模糊关系的特征选择方法。在 Reuter－21578 数据集上进行了实验测试，并与信息增益和卡方统计量作了比较。实验结果表明该方法能够挑选出有用的特征，并且提高分类的精度。

5. 神经网络的敏感性主要研究当网络中参数发生变化时，网络的输出会产生怎样的变化。本文将神经网络敏感性应用到特征选择领域，以径向基函数网络为模型，提出了一种基于 RBF 网络敏感性的特征选择方法。该方法通过每一维输入特征的微小扰动而引起输出变化的敏感性来衡量特征的贡献。我们把 RBF 网络的敏感性定义为均方偏差，并采用数值积分的方法推导出敏感性的近似计算表达式。在人工数据集、UCI 数据集上大量的实验都验证了将敏感性用于特征选择问题是可行的、有效的。

关键词： 文本分类，特征选择，类文本频率，特征分布，离散度，相对熵，模糊关系，敏感性

目 录

第一章

第二章

第三章

第四章

第五章

第六章

第七章

第八章

图清单

表清单

第一章

DIYIZHANG