基于二次熵的互信息特征选取方法的研究*)

刘丽珍[1,2] 宋瀚涛[2] 陆玉昌[3]

(首都师范大学 北京100037)[1] (北京理工大学 北京100081)[2] (清华大学 北京100084)[3]

摘 要 随着全球网络的普及应用,大量没有统一结构和管理的在线资源急需进行处理,高效的网页自动分类方法是从网上海量信息中提取所需信息的关键技术,特征选取又是文本分类挖掘的重要基础,本文以广义信息论为理论基础,提出了基于二次熵的互信息特征选取方法,独立评估特征集中的每个特征,分析特征和类别的关系,从高维的特征空间中选取出对文本分类有效的特征,降低了文本特征空间的维数,提高了文本分类的性能。

关键词 Web文本分类,特征选取,互信息

The Research of Mutual Information Feature Selection Method Based on Quadratic Entroy

LIU Li-Zhen[1,2] SONG Han-Tao[2] LU Yu-Chang[3]

(Information Engineering College, Capital Normal University, Beijing 100037)[1]

(Department of Computer, Beijing Institute of Technology, Beijing 100081)[2] (Department of Computer, Tsinghua University, Beijing 100084)[3]

Abstract With the global prevalence of the network application, there are so many resources on line that have no uniform structures and managements. They need to be processed as quickly as possible. The method of network pages automatically classification with high efficiency is the key technology, which can abstract needed information from vast network information. For feature selection is the important foundation the field of text classification mining. We use generalized information theory as the theory base to present the method of regarding quadratic entropy mutual information (QEMI) as the feature selection. The method can value every feature that has concentrate feature, and analyze the relationships between features and classes to get good features, which help to effectively classify texts, from high dimensionality feature space, and can also decrease the dimensionality of text feature space. So it can improve the performance of text classification.

Keywords Web text classification, Feature selection, Mutual information

1 引言

Web是一个巨大的、分布广泛的全球信息中心,但Web页面十分复杂[1],为了更好地管理Web站点的数据,使其高效地为网上用户提供有效的信息服务[2],进行Web文本自动分类技术的研究是十分必要的。文本分类是将每个文本自动分配到预先定义好的类中,其中一个主要的问题就是高维的特征空间,这些高维的特征集对分类学习未必全是重要的和有益的,而且这种情况对于许多的学习算法来说是难以处理的。

高维的特征集不仅影响分类学习精度,而且还会加剧机器学习的负担,占用较多的时间但却只能得到较少的相关特征,因而在不影响特征分类准确度的情况下,通过特征选取(feature selection)减少文本描述空间的高维特征数量是很有必要的[3]。

特征选取早在KDD出现以前就在机器学习和模式识别领域中得到了广泛的关注,这个过程是对原始数据进行变换,得到最能反映分类本质的特征。在有的文献中将特征选取作为特征的形成过程,有的指从形成、选择或变换得出有效特征的全过程。本文中的特征选取是指从高维的特征空间中选取出对于分类最有效的特征,以达到降低特征空间维数,提高分类效率的目的。

特征选取借助统计学、信息论等方法分析特征和类别的关系,选择出对分类最有意义的特征,建立特征子集[4],以便更有效地设计分类器。针对Web上的大量文本形成的高维特征集,许多特征子集的选取方法都很难达到理想的效果,目前我们常采用的方法有评估函数法、主成份分析法和模拟退火算法等。

向量降维以及对文本向量权值的调整通常是通过在训练数据集上的统计来计算每一维的某种特征值,根据指标值的高低决定是否保留相应的字或词,或者对对应维的权值进行加权,从而实现特征选择。在实际中存在多种筛选特征项的算法[5],如根据词和类别的互信息量判断、根据词熵判断、根据距离判断等不纯度度量,作为一种好的特征选取方法在一定程度上能提高分类的精确度[6]。

2 基于二次熵的互信息(Quadratic Entropy Mutual Information, QEMI)特征选取方法

在文本分类的特征选取中,互信息(Mutual Information, MI)衡量的是某个特征词和类别之间的统计独立关系,从互信息函数的分析中可以得出,由于互信息函数没有假设变量之间存在任何关系,因而它不仅能反映变量之间的线性相关

*)基金项目:973国家重点基础研究项目(G1998030414)。刘丽珍 副教授,博士,主要研究领域为数据仓储及知识发现;宋瀚涛 教授,博士生导师,主要从事多媒体与信息管理技术、网络通信技术的研究;陆玉昌 教授,从事数据集成和知识发现等的研究。

性，而且能反映变量之间的非线性相关性[7,8]。另外文本向量空间中的高维特征集事实上是很稀疏的，所以用互信息通过变量的分布函数及其积分函数，找到一个比较合适的目标函数来选取特征，要比直接使用互信息精度高[9~11]。

分类器是由后验概率来确定的，因此我们通过特征的后验概率分布衡量它对分类的有效性，即：$p(c_i|x)=1/n$；假设先验概率未知，那么我们无法确定样本的类别，因而错误概率为：$P_e=(n-1)/n$；如果使得特征组有：$p(c_i|x)=1$并且$\forall\ j\neq i$ $p(c_i|x)=0$则 $p_e=0$。所以当后验概率分布越集中时，错误概率越小；后验概率越接近均匀分布时，错误概率越大。为了给出定量衡量后验概率分布集中程度的指标，我们借助了广义信息论中的熵函数的概念[12]，将二次熵函数运用于互信息方法中进行特征选取。

由广义信息论知广义熵定义如下：

$$J_n^a[p(c_1|x),p(c_2|x),\cdots,p(c_n|x)]$$

$$=(2^{1-a}-1)^{-1}[\sum_{i=1}^{n}p^a(c_i|x)-1] \quad (1)$$

其中 a 是一个正参数，$a\neq 1$。不同的 a 值可以得到不同的熵分离度量，当 $a=2$，就得到了二次熵定义[12]。

广义熵具有对称的特点，$\hbar_n(p_1,p_2,\cdots,p_n)=\hbar_n(p_2,p_1,\cdots,p_n)=\cdots=\hbar_n(p_n,\cdots,p_1)\geqslant 0$，如果 $p_k=1$而且 $p_i=0,(1\leqslant i\leqslant n,i\neq k)$则 $\hbar_n(p_1,p_2,\cdots,p_n)=0$，

$\hbar_{n+1}(p_1,p_2,\cdots,p_n,0)=\hbar_{n1}(p_1,p_2,\cdots,p_n)$；

对于任意的概率分布：

$$p_i\geqslant 0,(i=1,2,\cdots,n),\sum_{i=1}^{n}p_i=1,$$

有 $\hbar_n(p_1,p_2,\cdots,p_n)\leqslant\hbar_n(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n})$

对于所有的事件，熵函数是连续函数。为了评估所选取的特征，向量空间每一点的熵函数都要进行计算，在熵函数值较大的空间部分，不同类的样本必然在较大的程度上互相重叠，所以熵函数的期望值可以表征类别的分离程度，作为选取特征分类性能的评价指标。

综合以上理论，类别和特征的二次熵函数如下：

$$H(C)=-\log\sum_{i=1}^{n}p(c_i)^2 \quad (2)$$

$$H(X)=-\log\int_x p(x_j)^2dx \quad (3)$$

在互信息中，使用二次熵函数度量互信息量还需要求出度量变量的概率密度函数的分布，将每个特征 x_i 都作为中心点，σ 是 x_i 的方差，求其概率密度分布为：

$$p(x_i)=\sum_{i=1}^{n}G(x_i-x_j,\sigma I) \quad (4)$$

结合熵函数的期望值等理论，得出基于二次熵的互信息在两个变量 c_i 和 x_j 之间的度量为：

$$MI(C,X)=\log\frac{(\sum_{i=1}^{n}\int_x P(c_i,x_j)^2dx_j)(\sum_{i=1}^{n}\int_c p(c_i)^2p(x_j)^2dx_j)}{(\sum_{i=1}^{n}\int_x p(c_i,x_j)p(c_i)p(x_j)dx_j)^2} \quad (5)$$

用基于二次熵的特征变量集 X 和类别变量集 C 之间的互信息特征选取方法，对每个特征 x_j，求其二次熵互信息做评估，去除阈值之下的词条，达到高维特征空间的降维目的。

3 实验与分析

实验数据来源于从 Internet 下载的4800篇关于计算机(800篇)、教育(800篇)、法律(800篇)、艺术(800篇)、体育(800篇)和军事(800篇)的六个类别的中文网页，对数据进行处理后形成目标数据集，在此基础上进行基于 Web 文本分类的特征选取。我们采用经典的 KNN 分类方法和 Naïve Bayes 分类方法对主成份分析法(PCA)、互信息(MI)、二次熵互信息(QEMI)和 χ^2统计(CHI)四种特征选取的方法进行实验比较，结果如图1、图2所示。

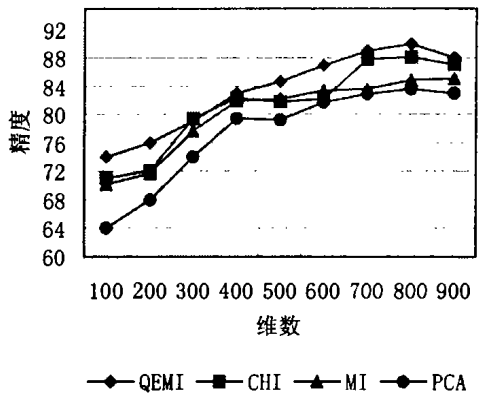

图1 用 KNN 分类方法测试四种特征选取方法

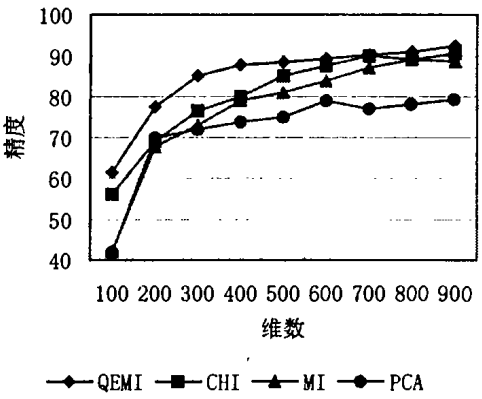

图2 用 Naïve Bayes 分类方法测试四种特征选取方法

3.1 基于实验数据的比较与分析

在实验中，我们针对特征集选用 QEMI、χ^2统计、PCA 和 MI 四种特征选取方法，选出对分类有区分价值的特征词条，在不同的阈值下得到不同维数的特征子集，并进行分类精确度的比较。实验结果表明，QEMI 特征选取方法在精度上高于其它几种同类方法，分类精度随着特征子集的维数增大而呈现良好趋势，但到了一定的维数高度，精度提高的幅度越来越小，逐渐趋向平缓，然后渐渐开始出现下降的趋势。这点很符合我们的理论研究，因为特征选取就是为了解决特征集维数太高而影响了文本分类性能的问题，所以当选取的特征维数大到一定程度，精度自然就会逐渐降低，但特征子集的维数也不能太少，如果少到不能反映文本内容特点的程度，也同样影响分类的精度。

另外，从以上实验中我们还进一步得出，在相同的特征选取情况下，Naïve Bayes 分类算法比 KNN 具有更好的分类效果。

3.2 基于实验方法的比较与分析

主成份分析法 PCA(Principal Component Analysis)是特征选取常用的方法之一，它的矩阵分解方法能够揭示更多有关变量主要方向的信息，而在数据方法中，由于数据被直接操

(下转第168页)

Show 中的系统设备枚举器(System Device Enumerator)枚举出所有插在 USBHub 上 USB 摄像头，得到各摄像头别名(Moniker)；b)进行静态图像捕捉，采用 DirectShow 中 ICaptureGraphBuilder2接口进行具体的图像捕捉控制，顺序选取步骤 a 所得到的摄像头设备别名，并创建其捕捉滤波器(Capture Filter)，从而获取各焊缝图像。

结论 USB 接口具有可扩展的、热插拔和即插即用等优良特性，已经成为计算机和外设接口的工业标准之一。基于该接口的焊缝图像采集系统，因此具有扩展性强，系统简洁可靠，编程容易等特点，并且已经成功应用于轿车副车架焊缝质量检测系统中。

参考文献

1 Universal serial bus specification [EB]. Compaq, Intel, Microsoft, NEC, Revision 1.1, 1998
2 武安河，周利莉. Windows 设备驱动程序开发务实[M]. 北京：电子工业出版社，2001
3 Cant C，孙义. Windows WDM 设备驱动程序开发指南[M]. 机械工业出版社，2000
4 乔林. Visual C++6.0高级编程技术 DirextX 篇[M]. 中国铁道出版社，2000
5 Kathywp. 使用 Sample Grabber 过滤器捕获图像. CSDN 文档中心，2001
6 刘涛. 基于 DirectShow 的流媒体解码和回放. Yesky 文档资源，2002

(上接第136页)

纵，所以可以自适应地实现数据的调整，不必重用所有数据，尤其是用在实时应用程序和计算开销极为宝贵的超高维问题上更显其优越性。但它的问题在于矩阵方法中要使用奇异值分解对角化矩阵求解方差-协方差，这种方法的时间复杂度是 $O(mn^2+n^3)$。如果求解特征方程使用 Hotelling 的幂方法，时间复杂度为 $O(mn^2+dn^3)$，因此矩阵方法的复杂度至少是按照数据的二次方增长。当 n 很大而时间很宝贵时，矩阵方法就有些不切实际了，这种情况下人们往往采用 PCA 的神经网络的方法，这是个数据方法，使用 Hebbian 学习规则，用 Hebbian 项 $y_i x_k$ 驱动 a_i^k 逼近特征向量 a_i，反 Hebbian 项 $y_i y_i a_i^k$ 负责保持权重界限[12]。Hebbian 规则确定 $m\times n$ 维数据，d 个主要成份的复杂度为 $O(dmn)$，但其复杂度的常数依赖于收敛所需的时间，也就是在很大程度上依赖于学习率 β_k，如果这个学习参数调整不适度，则会极大地影响特征的选取工作。

目前使用评估函数进行特征选取越来越普遍，特征选取算法通过构造一个评估函数的方法，选取预定数目的最佳特征作为特征子集的结果。在各种评估方法中，每一种方法都有一个选词标准，遵从这个标准，从文本集的所有词汇中选取出有某个限定范围的特征词集。因为评估函数的构造不是特别复杂，适用范围又很广泛，所以越来越多的人们喜欢使用构造评估函数来进行特征的选取，其中 χ^2 统计相似于互信息和信息增益，分类精度比其它的评估函数相对较高，它的计算有二次复杂度，而且是规格化评价，其不足是对于低频词来说是不可靠的。

互信息 MI 是信息论的一个基本概念，是两个随机变量统计相关性的测度，所以我们通常用 MI 作为特征词和类别之间的测度，如果特征词属于该类的话，它们的互信息量最大。

MI 的线性特征选取不同于统计技术中的主成份分析法 PCA，区别在于 MI 考虑特征和类之间的依存，而 PCA 考虑的是数据的协方差，所以 PCA 对于分类中要求的具有最优判别能力的特征选取发挥不了最佳的作用，但它在减少数据噪声方面却有着别的方法取代不了的优势。MI 作为特征选取的标准被用来指导设计线性特征选取器，而且在相同计算复杂度的情况下有比传统的 PCA 好的执行效果，在许多实验中 MI 被证明是一个比较好的特征选取标准，所以非常适合于文本分类中特征和类别的配准工作。

但 MI 的不足之处在于得分非常受词条的边缘概率的影响，从 $MI(w,c)=\log p(w|c)-\log p(w)$ 中可以看出，MI 评估对于有相等条件概率 $p(w|c)$ 的词，有时稀有词比常用词的得分还要高。基于二次熵的互信息 $QEMI(C,X)$ 克服了互信息 $MI(c_i,x_j)$ 的随机性，是一个确定的量，因此可以作为信息的整体测度，另外根据费诺不等式 $P\geqslant\frac{H(C)-MI(C,X)-1}{\log N}$ 得出，QEMI 比 MI 最大化的计算复杂度要小，所以可以高效地用在基于分类的特征选取上。

结束语 Web 文本挖掘有着十分广阔的应用前景，由于在 Web 文本中可利用的信息十分有限，传统的信息处理机制无法处理海量的电子文档，仅仅通过分词和词频统计还不能解决大量潜在的有价值信息的选取问题，所以通过有效的特征选取方法，评估确定具有代表性的文本特征项是 Web 文本分类挖掘的重要基础，特征选取结果的优劣直接影响着文本分类模型的性能。

本文提出的基于二次熵的互信息方法针对文本分类挖掘能够找到具有区分价值的特征词条，如果结合上 Web 文档的格式化标记的结构特点，还可以进一步提高特征选取算法的效率和分类的精度，同时，合理的阈值选取也是提高算法精度和改善分类性能的不可缺少的环节。

参考文献

1 Linoff G S, Berry M J a. Mining the web. America, 2001, 348
2 Mena J 著. Data Mining your website. America, 2000, 368
3 Wang Shi, Gao Wen. Web data mining. 计算机科学，2000，27(4)：237～240
4 Jing Li-ping, Huang Hou-kuan. Web text mining and feature selection. Computer and Information Technique, 2002, 10(1): 1～5
5 Han Ja-wei, Meng Xiao-feng. The research of web mining. Journal of Computer Research and Development, 2001, 38, (4): 405～414
6 Tang Qing, Yang Bing-ru. The research of implement of text mining system on web. 计算机科学，2003，30(1)：60～62
7 Chen Yun, Zhou Liang. Information Theory and Coding. Electron Industry Press, 2002
8 Torkkola K, Campbell W. Mutual information in Learning feature transformations. In: Proc. of Intl. Conf. on Machine Learning, Stanford, CA, USA, June 2000
9 Torkkola K. Onlinear Feature Transforms Using Maximum Mutual Information. In: Proc. of the IJCNN, 2001
10 Torkkola K, Campbell W M. Mutual Information in Learning Feature Transformations. In: Proc. 17th Intl. Conf. on Machine Learning, 2000
11 Torkkola K. On Feature Extraction By Mutual Information Maximization. http://citeseer.nj.nec.com, 2002
12 Bian Zhao-qi, Zhang Xue-gong. Pattern Identify. Tsinghua University Press, 2000
13 Zhang Min, Ma Shao-pin. The research of information distributing characteristic and searches strategy based on web text searching. Beijing University, 2003. 137～144