基于特征辨别能力和分形维数的特征选择方法*

夏晶晶[1],朱颢东[2]
(1.郑州牧业工程高等专科学校 信息工程系,河南 郑州 450011
2.中国科学院成都计算机应用研究所,四川 成都 610041)

摘 要:基于分形维数的属性约简算法与特征辨别能力相结合,提出了一个综合的特征选择方法。该方法利用特征辨别能力进行特征初选,过滤掉一些词条来降低特征空间的稀疏性,以利用所提约简算法消除冗余,从而获得较具代表性的特征子集。实验结果表明,此种特征选择方法效果良好。

关键词:文本分类;特征选择;特征辨别能力;分形维数
中图分类号:TP301 **文献标识码**:A

Feature selection method based on feature distinguishability and fractal dimension

XIA Jing-jing[1], ZHU Hao-dong[2]
(1.Department of Information Engineering, Zhengzhou College of Animal Husbandry Engineering, Zhengzhou 450011, China
2.Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China)

Abstract: The paper combined the reduction algorithm based on fractal dimension with the feature distinguishability and proposed a comprehensive feature selection method. The comprehensive method firstly uses the feature distinguishability to select features and filter out some terms to reduce the sparsity of feature spaces, and then uses the feature reduction algorithm to eliminate redundancy, so can acquire the feature subset which are more representative. The experimental results show that the comprehensive method is promising.

Key words: text categorization; feature selection; feature distinguishability; fractal dimension

文本分类系统通常采用特征集来表示文档,这使得特征向量的维数非常大,有时会达到数十万维。如此高维的特征对于后续的分类过程未必全都重要、有益,而且高维的特征可能会大大增加分类的计算量,使整个处理过程的效率降低,可能产生与小得多的特征子集相似的分类结果[1]。因此,必须对文档的特征向量进一步净化处理,在保持原文含义的基础上,找出既能反映文本内容,又比较简洁的特征向量,特征选择就是为解决上述问题而产生的一个关键选择方法。该方法不仅能够解决上述问题,而且在一定程度上还能消除噪声词语,使文本之间的相似度更加准确,既提高了语义上相关文本之间的相似度,同时降低了语义上不相关的文本之间的相似度[2]。

本文首先分析了几种经典的特征选择方法,提出了特征辨别能力的概念,紧接着把分形维数引入粗糙集并提出了一个基于分形维数的属性约简算法,最后把该约简算法与特征辨别能力结合起来,提出了一个综合的特征选择方法。该方法首先利用特征辨别能力进行特征初选以过滤掉一些词条来降低特征空间的稀疏性,然后利用所提属性约简算法消除冗余,从而使得选择的特征子集具有较低的冗余性、较好的代表性。

1 几种经典特征选择方法

目前常用的文本特征选择方法有 DF、WF、IG、CHI、MI 等[3-9]。

1.1 文档频DF(Document Frequency)

特征的文档频是指在训练语料集中出现该特征的文档数。该方法选择特征时仅考虑特征所在的文档数,如果某个特征在训练语料集中所在的文档数达到一个事先给定的阈值,则留下该特征,否则删除。该方法的缺

*基金项目:四川省科技计划项目(2008GZ0003)

点在于仅考虑特征词在文档中出现与否，忽视了特征在文档中出现的次数。于是产生了一个问题：如果特征词a和b的文档频相同，那么该方法就认为这两个特征词的贡献是相同的，而忽略了它们在文档中出现的次数。但是，通常情况是文档中出现次数较少的词是噪声词，这样就导致该方法所选择的特征不具代表性。

1.2 词频WF(Word Frequency)

特征的词频是指特征在文档中出现的数目。使用该方法选择特征时，特征只有在文档中出现的次数达到一个阈值时，才被保留，否则予以删除。该方法的缺点在于仅选择出现频繁的词作为特征，但是有时候在某个文档中出现频繁的特征对分类贡献并不大。

1.3 信息增益IG(Information Gain)

信息增益是指特征在某类文档中出现前后的信息熵之差，该差用平均信息量表示。信息增益的缺点在于不但考虑了特征出现的情况，而且还考虑了特征未出现的种情况。即使某个特征不在文本中出现也可能对判断文本类别有所贡献，但实验证明，这种贡献十分微小，尤其是在样本分布和特征分布失衡的情况下，某些类别中出现的特征词在全部特征词的比例很小，较大比例的特征词在这些类别中是不存在的，也就是此时的信息增益中特征不出现的部分占绝大优势，这将导致信息增益的效果大大降低。

2 特征辨别能力

如果一个特征对某个类的贡献较大，那么该特征对这个类的辨别能力应该较强。为此，本文定义了特征对类别的辨别能力，简称特征辨别能力。

定义1 特征辨别能力 表示特征 f_i 对类别 c_i 的辨别能力，用 $Feature-Distinguishability(f_i,c_i)$ 表示。由于一个类别的特征词有多个，因此可用以下公式来表示特征辨别能力：

$$Feature-Distinguishability(f_i,c_i)=\sum_{k=1且k\neq j}^{m}\left(\frac{DF_n(f_i,c_j)-DF_n(f_i,c_k)}{\sum_{j}DF_n(f_i,c_j)}\right)^2 \quad (1)$$

其中 m 为属于类别 c_j 的特征个数，$DF_n(f_i,c_j)$ 为在类别 c_j 的文本训练集中出现特征 f_i 的次数不小于 n 的文本数。经分析可知，$Feature-Distinguishability(f_i,c_j)$ 不但考虑了特征出现的文档数，而且还考虑了特征在文档中出现的次数，把文档频和词频进行了有机的结合。$Feature-Distinguishability(f_i,c_j)$ 越大则表明特征 f_i 对类别 c_j 的辨别能力也就越大，那么该特征的分类能力也就越强，即该特征也就越重要。

3 本文属性约简算法

粗糙集RS(Rough Sets)理论是由PAWLAK Z在20世纪80年代初提出来的一种新的处理不精确、不相容、不完全和不确定知识的软计算工具。其本质就是在保持分类能力不变的前提下，通过知识约简导出问题的分类规则[10]。目前RS已被广泛应用于机器学习、决策分析、数据挖掘、过程控制、智能信息处理等领域[11]。

属性约简是粗糙集的核心内容之一，现已出现了大量的属性约简算法，例如以属性重要度为基础的属性约简算法[9]、以信息论为基础的属性约简算法[12]。但是这些约简算法执行效率低且不一定能够得到最小约简。基于分形维数的属性约简算法可以有效地改变这一状况[13-14]。本文利用数据集的分形维数进行属性约简。

3.1 相关基本知识[15]

如果一个数据集在所有的观察尺度下都具有自相似性，即一个数据集的部分分布有着与整体分布相似的结构或特征，则称该数据集是分形的。

定义2 嵌入维 数据集中的数据点所在欧式空间的维数称为数据集的嵌入维，即一个数据集中属性的个数。

定义3 固有维 一个数据集的固有维是指一个数据集所表示的空间对象的实际维数。

一般地说，空间对象的维数(固有维)不会超过所在欧式空间的维数(嵌入维)。例如，所有欧氏空间的直线不论嵌入维是二维或是三维，其固有维都是一维的。

定义4 分形维 嵌入维数等于 n 的数据集可视为 n 维空间中的点。用边长为 $r(r\in(r_1,r_2))$ 的 n 维立方体分割数据集，记落入第 i 个立方体中的数据点的数目为 C_i。则分形维 D_q 计算如下：

$$D_q=\frac{1}{q-1}\times\frac{\partial\log\sum_i D_i^q}{\partial\log r},r\in(r_1,r_2) \quad (2)$$

其中 D_0 称为分形维；当 q 趋近于1时，D_1 称为信息分形维；当 $q=2$ 时，D_2 称为相关分形维。D_2 描述了随机选择的两个数据点的距离落在某一范围内的概率，因此相关分形维 D_2 的改变意味着数据集中数据点分布的变化。在实际应用中，计算数据集的相关分形维几乎是不可能的，计盒维数常被用来近似估计相关分形维数。

3.2 基于分形维数的属性约简算法

决策表 $S=<U,A=C\cup D,V,f>$ 其中 U 可以看作 $E=|C\cup D|$ 维空间中的点集，可用于区分数据对象的属性数目不超过 E。由此，可以通过计算 E 维空间中的点集的分形维来估计 S 应包含的最少属性数。可通过两个计算步骤来约简 S 的属性集：(1)计算包括所有 E 个属性的分形维，称为 WFD(全分形维)；(2)剔除其中的一个属性，再计算剩余的 $E-1$ 个属性的分形维，称为 PFD(部分分形维)，共计算出 E 个部分分形维 $PFD_i,i=1,2,\cdots,E$，从这些部分分形维中选择最接近 WFD 的 PFD_i 将其对应的 a_j 删除，同时令 $WFD=PFD_i$ 在 $A=A-\{a_j\}$ 中继续上述的步骤，直到剩余属性数目与 S 的分形维数相同。具体算法描述如下：

Input：$S=<U,A=C\cup D,V,f>$

Output：C 的最小约简

(1)计算出决策表 S 的计盒维数 WFD;令 $WFD_0=WFD$。

(2)$\forall a_i \in A$,计算 U 在 $A-\{a_i\}$ 上的 PFD_i。

(3)从 $PFD_1, \cdots, PFD_{|A|}$ 中选择最接近 WFD 的一个 PFD_j,令 $WFD=PFD_j$,$A=A-\{a_j\}$。

(4)如果 $|A|>WFD_0+1$,则转(2)。

(5)输出 A,算法结束。

在该算法中,计算有 N 个元组的决策表的分形维的时间复杂度为 $O(N\times N)$;从 $|A|$ 个属性中选择一个属性并剔除需要扫描 $|A|$ 次对象集 U,如果剔除 $K(K<|A|)$ 个属性,则需要扫描 $(K\times(2|A|-K+1))/2$ 次,所以整个算法的时间复杂度为 $O(|A|\times K\times N\times N)$。

4 本文特征选择方法基本步骤

设 T 为原始特征集,C 为类别集,对于 $\forall c_j \in C$,设 c_j 的训练文档集为 DS_j,其原始特征集 $T_j=T$,c_j 的特征词选择算法如下:

对于每个 $f_i \in T_j$,给定最小词频数 n 以及特征辨别能力阈值 ω。

(1)计算 f_i 的 $Feature-Distinguishability(f_i, c_j)$。

(2)若 $Feature-Distinguishability(f_i, c_j)<\omega$ 则把 f_i 从 T_j 中删除,否则 f_i 保留。

(3)若 T_j 中还存在没考察的元素则转到步骤(1)。

(4)若 C 中还存在没考察的类别则转到步骤(1)。

(5)将上述各类别所选的特征合并为 1 个特征集。

(6)将步骤(5)得到的特征集以及标有类的训练集组成一个决策表:$S=<U, R=C\cup D, V, f>$,使用本文提出的属性约简算法进行属性约简。

(7)对得到的特征子集进行微调,以突出那些对分类贡献比较大的特征词,然后输出特征集。

5 实验例证

5.1 实验语料库

进行文本分类方面的实验,语料库的选择是非常重要的,选择的原则是国内外使用广泛、权威标准和规范。这样使得实验和国内外同行的试验结果具有可比性,同时也便于分析实验数据、算法的优劣。

在中文文本分类方面,经过分析、比较,本文选用的分类语料库是复旦大学中文文本分类语料库。语料文档全部采自互联网,可以从网上免费下载,网址为:http://www.nlp.org.cn/categories/default.php?cat_id=16。复旦大学中文文本分类语料库中包含 20 个类别,分为训练文档集和测试文档集两个部分。每个部分都包括 20 个子目录,相同类别的文档存放在一个对应的子目录下;每个存储文件只包含 1 篇文档,所有文档均以文件名作为唯一编号。共有 19 637 篇文档,其中训练文档 9 804 篇,测试文档 9 833 篇;训练文档和测试文档基本按照 1:1 的比例来划分。去除部分重复文档和损坏文档后,共保留有文档 14 378 篇,其中训练文档 8 214 篇,测试文档 6 164 篇,跨类别的重复文档没有考虑,即一篇文档只属于一个类别。该语料库中的文档的类别分布不均匀。其中,训练文档最多的类 Economy 有 1 369 篇训练文档,而训练文档最少的类 Communication 有 25 篇训练文档;同时,训练文档数少于 100 篇的稀有类别共有 11 个。训练文档集和测试文档集之间互不重叠。本文只取前 10 个类的部分文档,其类别文档分布如表 1 所示。

表 1 文档分布

类别	训练文档数目	测试文档数目
经济	480	419
体育	584	489
计算机	628	591
政治	573	482
农业	547	435
环境	405	371
艺术	510	286
太空	506	248
历史	466	468
军事	74	75

5.2 实验环境及参数设置

实验设备是一台普通计算机:操作系统为 Microsoft Windows XP Professional (SP2),CPU 规格为 Intel (R) Celeron(R) CPU 2.40 GHz,内存 512 MB,硬盘 80 GB。

进行中文分词处理时,采用的是中科院计算所开源项目"汉语词法分析系统 ICTCLAS"系统。

实验使用的软件工具是 Weka,这是新西兰的 Waikato 大学开发的数据挖掘相关的一系列机器学习算法。实现语言是 Java 软件可以直接调用,也可以在代码中调用。Weka 包括数据预处理、分类、回归分析、聚类、关联规则、可视化等工具,对机器学习和数据挖掘的研究工作很有帮助。

本算法中各参数需要反复试验各参数最后设置如下:$n=3$,$\omega=0.09$。

5.3 实验所用分类器及其评价标准

本实验旨在比较本文方法与信息增益(IG)、X^2 统计量(CHI)、互信息(MI)三种特征选择方法对后续文本分类精度的影响,因此实验采用相同的分类器对文本进行分类。实验中使用 KNN 分类器来比较这几种特征选择方法(K 设置为 10)。

为了评价实验效果,实验中选择分类正确率和召回率作为评价标准:准确率=$a/(a+b)$,作为所判断的文本与人工分类文本吻合的文本所占的比率;召回率=$a/(a+c)$,作为人工分类结果应有的文本与分类系统吻合的文本所占的比率。在实际应用中,查准率比查全率重要。其中 a、b、c 代表相应的文档数,其含义如表 2 所示。

5.4 实验结果

图 1、图 2 表明了四种方法在所选数据集上的分类准确率和召回率,从总体上看,本文方法>IG>CHI>MI。由于本方法首先利用特征辨别能力进行特征初选以过

表 2 二值联表

	真正属于此类	真正不属于此类
判断属于此类	a	b
判断不属于此类	c	d

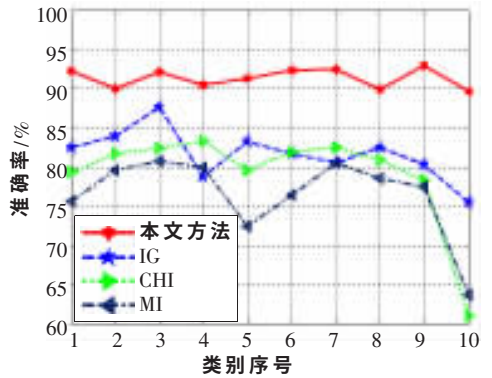

图 1 准确率对比结果

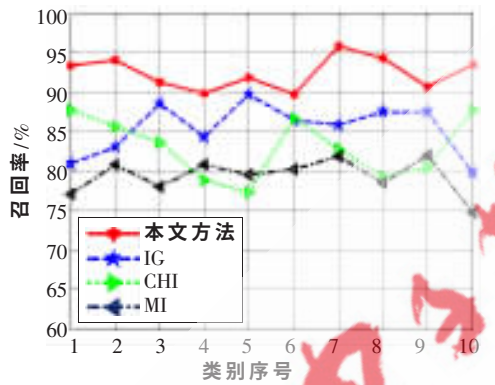

图 2 召回率对比结果

滤掉一些词条来降低特征空间的稀疏性，然后利用所提属性约简算法消除冗余，从而获得较具代表性的特征子集，所以效果最佳；由于 IG 方法受样本分布影响，在样本分布不均匀的情况下，其效果就会大大降低，但从整体上看本文所选样本分布相对均匀，只有极个别相差较大，所以总体效果次之；MI 方法仅考虑了特征发生的概率；而 CHI 方法同时考虑了特征存在与不存在时的情况，所以 CHI 方法比 MI 方法效果要好。总体来说，本文所提的方法是有效的，在文本分类中有一定的实用价值。

本文首先简单分析了几种经典的特征选择方法，总结了它们的不足，然后提出了特征辨别能力的概念，紧接着把分形维数引入粗糙集，并提出了一个基于分形维数的属性约简算法，最后把该属性约简算法与特征辨别能力结合起来，提出了一个综合的特征选择方法。由于该方法首先利用特征辨别能力进行特征初选以过滤掉一些词条来降低特征空间的稀疏性，然后利用所提属性约简算法消除冗余，从而获得较具代表性的特征子集。实验证明，本文特征选择方法与三种经典特征选择方法“互信息”、“x^2 统计量”以及信息增益相比，具有较高的准确率和召回率，为后续的知识发现算法减少了时间与空间复杂性，从而使得本方法在文本分类中有一定的使用价值。

参考文献

[1] DELGADO M, MARTIN M J, SANCHEZ D, et al. Mining text data: special features and patterns[A]. In proceedings of ESF exploratory workshop[C]. London: U.K, Sept, 2002, 32-38.

[2] 申红，吕宝粮，内山将夫，等.文本分类的特征提取方法比较与改进[J].计算机仿真，2006，23(3):221-224.

[3] 朱颢东，钟勇.一种新的基于多启发式的特征选择算法[J].计算机应用，2009，29(3):849-851.

[4] YANG Y, PEDERSEN J O. A comparative study on feature selection in text categorization [A]. In: Fisher DH, ed. Proc. of the 14th Int'l Conf. on Machine Learning(ICML'97) [C]. Nashville: Morgan Kaufmann Publishers, 1997:412-420.

[5] 张海龙，王莲芝.自动文本分类特征选择方法研究[J].计算机工程与设计，2006，27(20)：3838-3841.

[6] 宋枫溪，高秀梅，刘树海.统计模式识别中的维数削减与低损降维[J].计算机学报，2005，28(10):1915-1922.

[7] 周茜，赵明生，扈曼，等.中文文本分类中的特征选择研究[J].中文信息学报，2004，18(3)：17-23.

[8] 胡佳妮，徐蔚然，郭军，等.中文文本分类中的特征选择算法研究[J].光通讯研究，2005(3)：44-46.

[9] 寇苏玲，蔡庆生.中文文本分类中的特征选择研究[J].计算机仿真，2007，24(3):289-291.

[10] 胡寿松，何亚群.粗糙决策理论与应用[M].北京：北京航空航天大学出版社，2006.

[11] LIANG Ji Ye, CHIN K S, CHUANGYIN D, et al. A new method for measuring uncertainty and fuzziness in rough set theory [J]. International Journal of General Systems, 2002,31(4):331-342.

[12] 柴慧芳. 粗糙集下基于信息熵的知识约简算法研究[D].昆明：昆明理工大学，2007.

[13] TRAINA C, TRAINA A, WU L, et al. Fast feature selection using fractal Dimension[A]. In: C. Faloutsos, ed. Proc.of XV brazilian symposium on Databases, Paraila, Brazil: Springer, 2000:78-90.

[14] YAN Guang Hui, LI Zhan Huai, YUAN Liu.The practical method of fractal dimensionality reduction based on Z-ordering technique [C]//LI X, ZAIANE O R, LI Z.Proceedings of the Second International Conference on Advanced Data Mining and Applications. Berlin Heidelberg: Springer-Verlag, 2006:542-549.

[15] 杨光俊.分形的数学[M].昆明：云南大学出版社，2002.

(收稿日期：2009-06-03)

作者简介：

夏晶晶，女，1982 年生，助教，主要研究方向：智能信息处理、网络设计。

朱颢东，男，1981 年生，博士，主要研究方向：软件过程技术与方法、文本挖掘。