结合类别相关性和辨识集的特征选择方法

王加龙[1],朱颢东[2]

(1.商丘师范学院 实验设备管理中心,河南 商丘 476000;
2.中国科学院成都计算机应用研究所,四川 成都 610041)

摘 要:介绍了基于辨识集的属性约简算法,把该属性约简算法同类别相关性结合起来,提出了一个综合的特征选择方法。该综合方法使用类别相关性进行特征初选,并用所提属性约简算法消除冗余。实验结果表明此种特征选择方法能够获得较具代表性的特征子集。

关键词:特征选择;类别相关性;粗糙集;辨识集;属性约简

中图分类号:TP301 **文献标识码**:A

Feature selection method combined category correlation with discernible sets

WANG Jia Long[1], ZHU Hao Dong[2]

(1.Experimental Equipment Management Center, Shangqiu Normal University, Shangqiu 476000, China;
2.Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China)

Abstract: This paper presented a new attributes reduction algorithm based on discernible sets. It combined the attribute reduction algorithm with the category correlation and proposed a comprehensive feature selection method. The comprehensive method uses the category correlation to select primary features and employs the proposed attribute reduction algorithm to eliminate redundancy. The experimental results show that the comprehensive method can acquire the feature subsets which are more representative.

Key words: feature selection; category correlation; rough set; discernible set; attribute reduction

文本分类是文本挖掘中一个重要的研究内容,其根据某种方法自动为未知类别的文档分配类别[1]。在中文文本分类中,通常采用词条作为最小的独立语义载体,原始的特征空间可能由出现在文档中的全部词条构成。而中文的词条总数有20多万条,这就使得文本特征空间具有高维特性和稀疏性,这些特性大大限制了分类算法的选择、降低了分类算法的性能。因此,寻求一种有效的特征选择方法,以降低特征空间维数、提高分类的效率和精度,成为中文文本分类中需要首先解决的核心问题[2]。为此,本文提出了一个综合性特征选择方法。该方法首先利用类别相关性进行特征选择,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 常用的文本特征选择方法

常用的文本特征选择方法有MI、WF、DF等[3]。本文仅介绍WF和DF,其他请参阅参考文献[3]。

1.1 词频WF(Word Frequency)

某个特征的词频是指该特征在一篇文档出现的次数。基于词频的方法往往选取在某类别中比其他类别更频繁出现的词作为特征词,而忽视了词在不同文档中的出现情况。

1.2 文档频DF(Document Frequency)

某个特征的文档频是指出现该特征的文档次数。文档频方法仅考虑特征词在文档中出现与否,忽视了在文档中出现的次数。由此带来的问题是:如果2个特征词的文档频相同,那么其在文档中出现多次和仅出现1次的相关度相同。而文档中仅出现1次的词经常是噪声

词。文档频评估函数的理论假设是稀有单词不含有用信息、太少而不足以对分类产生影响、或者是噪音，所以可以删去。显然它在计算量上比其他评估函数小得多，但在实际运用中效果却很好。在实际运用中一般并不直接使用文档频，而常作为评判其他评估函数的标准。

2 本文方法思想

上述提到的两种特征词选择方法存在共同的缺点，那就是它们在选择特征时仅依靠权重作为选择的标准，而没考虑特征词之间的潜在关系，从而使得选出的特征子集存在冗余，不具备较好的代表性[4]。为了解决这个问题，本文首先把文档频和词频结合起来提出了类别相关性方法，紧接着把粗糙集引入本文并提出了基于辨识集的属性约简算法，最后把该属性约简算法同类别相关性结合起来提出了一个综合性特征选择方法。

3 本文方法所用策略

3.1 类别相关性

如果一个特征对某个类别贡献较大，那么该特征应该在该类中集中出现，而不是分散地出现在各类文档中。为此，本文定义了特征与类别的相关性，用于表现特征对类别的贡献度。

定义 1 类别相关性 表示特征 f_i 与类别 c_j 的相关程度，用 $\rho(f_i,c_j)$ 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类别相关性：

$$\rho(f_i,c_j)=\sum_{k=l\ 且\ k\neq j}^{m}\left(\frac{MinWF_n(f_i,c_j)-MinWF_n(f_i,c_k)}{\sum_i MinWF_n(f_i,c_j)}\right)^2 \quad (1)$$

其中 m 为类别的个数，$MinWF_n(f_i,c_j)$ 是指在类别 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。$\rho(f_i,c_j)$ 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。$\rho(f_i,c_j)$ 越大则表明特征 f_i 与类别 c_j 的相关联程度越大，那么该特征的分类能力也就越强，即该特征也就越重要。

3.2 基于辨识集的属性约简

粗糙集理论是一种研究不精确、不确定性知识的数学工具，属性约简是该理论中一个非常重要的概念，反映了决策表的本质信息。差别矩阵的属性约简算法[5-8]是粗糙集理论中一类经典的属性约简算法。然而，经分析发现，这类算法随着问题规模的增大，存放差别矩阵的空间和算法执行时间的代价都很大，并不适用于海量文本特征的约简。参考文献[5-8]中给出了相关改进的算法，但仍要存放差别矩阵，参考文献[7-8]中虽将差别矩阵转换成特征矩阵，但特征矩阵的存放和计算与差别矩阵并无两样。为解决上述问题，本文提出了辨识集的定义，进而给出了基于辨识集的属性约简的定义。同时证明了该定义与基于差别矩阵的属性约简定义是等价的。在此基础上，设计了一个新的属性约简算法，由于这一算法在求属性约简的过程中不用生成差别矩阵和大量的无用元素，因而大大减少了存储量和计算量，从而提高了算法的效率。

3.2.1 问题的提出

定义 2 决策表 $S=(U,C,D,V,f,d)$，U 的差别矩阵是一个 $n\times n$ 的对称矩阵，$\boldsymbol{M}_{n\times n}=(m_{ij})$，其元素定义为

$$m_{ij}=\{a|a\in C,f(x_i,a)\neq f(x_j,a)\cap(\exists s\in D,f(x_i,s)\neq f(x_j,s))\} \quad (2)$$

其中 $i,j=1,2,\cdots n$

在基于差别矩阵的属性约简算法中，常常是先求出差别矩阵，然后再根据某一启发信息选取一个属性放入属性约简中，再在差别矩阵的元素中删除所有包含该属性 a 的元素，直至差别矩阵为空[5-8]。这个过程存在如下缺点：

(1)存放差别矩阵的空间可能很大，例如，当对象个数为 1 000 000 单元，条件属性的个数为 100 单元时，则存放差别矩阵的最大空间为 $100\times1\ 000\ 000\times(1\ 000\ 000-1)/2=5\times1\ 013$ 单元，极大地影响了算法的效率；

(2)在所存放的元素中有很多是重复的，造成了存储空间的极大浪费。因为在属性约简算法中，显然要删除包含某一元素的所有元素，由于这些元素的存放占用了大量的空间，删除时就要花费大量的比较时间，显然这对算法的运行也是很不利的。

3.2.2 改进的属性约简算法原理

定义 3 决策表 $S=<U,C,D,V,f>$ 中 $\forall x_i\in U$，记

$$D(U,U)=m|m=\{p\in C:f(x_i,p)\neq f(x_j,p)\ 且\ \exists d\in D,f(x_i,d)\neq f(x_j,d),i<j=1,2\cdots n\} \quad (3)$$

称 $D(U,U)$ 为属性集 C 的辨识集。记

$$D(U,c)=m|c\in m,m=\{p\in C:f(x_i,p)\neq f(x_j,p)\ 且\ \exists d\in D,f(x_i,d)\neq f(x_j,d),i,j=1,2\cdots n\} \quad (4)$$

称 $D(U,c)$ 为属性 $c\in C$ 的辨识集。

定理 在决策表 $S=<U,C,D,V,f>$ 中，$\boldsymbol{M}_{n\times n}=(m_{ij})$ 为决策表的辨识矩阵，$D(U,C)$ 为决策表的辨识集，则有：$\forall m_{ij}\in\boldsymbol{M}_{n\times n}$ 且 $m_{ij}\neq\phi\Leftrightarrow m_{ij}\in D(U,C)$。

证明：对比辨识矩阵的定义 2 和辨识集的定义 3，很明显定理成立。

根据定理，本文基于辨识集的属性约简算法为：

(1)初始约简属性集 $B=$NULL；

(2)求得 $D(U,C)$；

(3) $\forall m\in D(U,C)$ 则 $|m|=1$，把 m 加入到 B，即 $B=\{m\}+B$，$D(U,C)=D(U,C)-\{d|m\in d,d\in D(U,C)\}$；

//* 相当于求核

(4) $\forall b\in C-B$，选择 $\mathrm{MAX}\{|\{d|b\in d,d\in D(U,C)\}|\}$(若不止一个，可根据具体情况选择其一)，$D(U,C)=D(U,C)-\{b|m\in d,d\in D(U,C)\}$，$B=\{b\}+B$；

(5)若 $D(U,C)$ 为 null，输出 B，算法结束；否则转向(4)。

3.2.3 改进的属性约简算法效率分析

新算法的最大优点是没有生成无用的元素，因为在

(2)中获得了核属性,这样做可使得(3)开始就生成较小的搜索空间,显然这可以提高算法的效率,在(3)中生成 $D(U,C)-\{b\}$,其意义是凡是能由属性 b 区分的元素就不用生成,这一步则相当于在基于差别矩阵的属性约简算法中删除差别矩阵中所有包含属性 b 的元素,由于在新算法中没有生成这样的元素,当然也就用不着删除,这就大大地压缩了占用的存储空间,算法的效率得到极大地提高。

3.2.4 改进的属性约简算法例证

使用参考文献[5]中的决策表,采用本文算法与基于差别矩阵[5-8]的属性约简算法(称旧算法)进行比较,表 1 为其对应的差别矩阵,表 2 为本文算法所对应的辨识集。

表 1 旧算法对应的差别矩阵

O	OT	OTH	$OTHW$	TH	OTH	THW	OTW	OH
OW	OTW	$OTHW$	OTW	THW	$OTHW$	TH	OT	OHW
$OTHW$	THW	W	O	OW	TW	OT	OTH	OTW
OT	OT	THW	$OTHW$	TH	OH	HW	OW	OTH
OTW	W	THW	OTH	$OTHW$	HW	TH	OW	$OTHW$

表 2 本文算法对应的辨识集

$O,OT,OTH,OTHW,TH,THW,OTW,OH,OW,OHW,W,TH,HW$

采用本文算法,第 1 步生成 $D(U,C)=\{O,OT,OTH,OTHW,TH,THW,OTW,OH,OW,OHW,W,TH,HW\}$,需要比较的次数为:$(9+9+3\times3+6+2+5+5)\times5=45\times5=225$。选择核时 $D(U,C)$ 内部比较的次数为 13 次,各个核生成 $\{d|m\in d,d\in D(U,C)\}$ 时需比较 $13\times2=26$ 次(有 2 个核),核选择后,$B=\{W,O\}D(U,C)=\{TH\}$;第 2 步选择 T 或 H 只需要比较 1 次就行了,$B=\{T,W,O\}$ 或 $B=\{H,W,O\}$,$D(U,C)=\{\}$,算法结束。因此本文算法总的比较次数为:225+13+26+1=265 次。每个属性元素存储时占 1 个存储单元,则本文算法只需要 30 个存储单元。而在参考文献[7]中分析的旧算法的比较次数为 447 次,存储单元需 116 个。由此可见本文改进的算法大大提高了算法的效率。

4 本文方法描述

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 以及最小文档数阈值 ω;

(1)计算 f_i 的 $\rho(f_i,c_j)$;

(2)若 $\rho(f_i,c_j)<\omega$ 则把 f_i 从 T 中删除,否则 f_i 保留;

(3)若 T 中还存在没考察的元素则转到(1);

(4)若 C 中还存在没考察的类别则转到(1);

(5)将上述各类别所选的特征合并为 1 个特征集;

(6)将(5)得到的特征集以及标有类的训练集组织成为一个决策表:$S=<U,R=C\cup D,V,f>$,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 本文特征选择方法实验验证

本实验使用的数据集由人民网(http://www.people.com.cn/)下载的一些新闻材料组成,这些新闻材料发表日期范围为 2007~2009 年。共下载 10 类新闻组,其文档分布情况如表 3 所示。文本表征词典根据训练文档的正文(忽略所有的报头)生成。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,原始特征维数高达 21 092。本实验选用线性支持向量机作为基准分类器。

表 3 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	575	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

实验使用的软件工具 Weka 是纽西兰的 Waikato 大学开发的数据挖掘相关的一系列机器学习算法。其网址为:http://www.cs.waikato.ac.nz/ml/weka/。实验使用的计算工具为 MATLAB 7.0。本文算法中各参数需要反复试验才能得到,经试验算法中各参数最后设置如下:$n=3$,$\omega=0.09$。为便于比较,在实验中测试了 4 种特征选择方法:使用本文的方法、互信息(MI)、x^2 统计量(CHI)、信息增益(IG)。为评价实验效果,实验中选择分类正确率和召回率作为评价标准。

图 1 所示为 4 种方法在准确率方面的仿真对比结果,图 2 所示为 4 种方法在召回率方面的仿真对比结果。

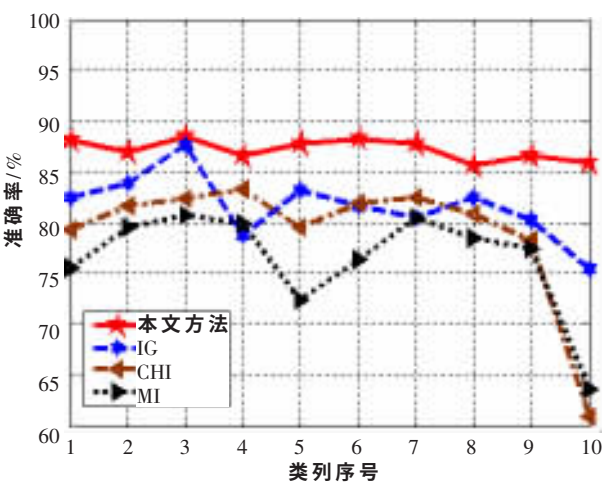

图 1 准确率对比结果

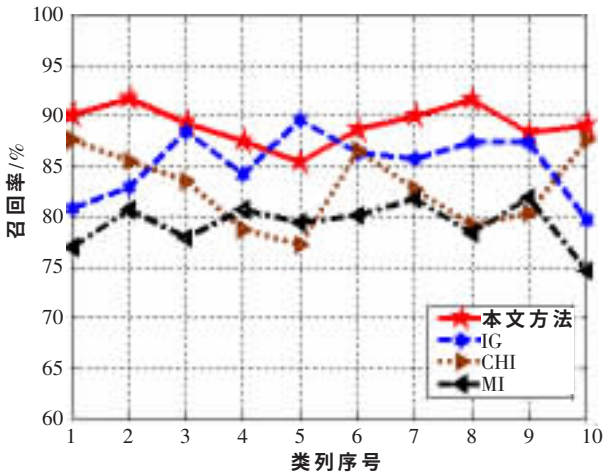

图 2 召回率对比结果

图 1、图 2 为 4 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。由于本文方法首先利用类别相关性方法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，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。由于 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。因此，本文所提的方法是有效的，在文本挖掘中有一定的实用价值。

实验证明，本文特征选择方法与 3 种经典特征选择方法(IG、CHI、MI)相比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，从而使得本文方法在文本分类中有一定的使用价值，同时也为中文文本特征选择提供一种思路。

参考文献

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

[2] 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.

[3] 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.

[4] 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.

[5] 徐章艳，杨炳儒，宋威等.一种快速计算 HU 差别矩阵的属性约简算法[J].小型微型计算机系统，2008，29(10)：1820-1827.

[6] 周创德，田卫东.基于约束函数的差别矩阵及其求核算法[J].计算机工程，2008，34(15)：60-62，66.

[7] 林晓淑，叶东毅.一种基于扩展差别矩阵的规则获取方法[J].计算机科学，2008，35(3)：231-233.

[8] 赵卫东，戴伟辉.基于特征矩阵的决策表约简研究[J].系统工程理论与实践，2003，23(3)：65-69.

(收稿日期：2009-07-07)