关系积理论在特征选择中的应用研究

刘阿力
(商丘师范学院 教育系,河南 商丘 476000)

摘 要:提出了一个适用于海量文本数据集的特征选择方法。该方法利用一个优化的文档频进行特征初选以滤除一些词条来降低特征空间的稀疏性,并利用一个基于关系积理论的属性约简算法消除冗余,从而获得较具代表性的特征子集。实验结果表明,此种特征选择方法效果良好。

关键词: 特征选择;文本分类;文档频;关系积理论;属性约简
中图分类号: TP301 **文献标识码**: A

Study of attribute union theory in feature selection

LIU A Li
(Department of Education, Shangqiu Normal University, Shangqiu 476000, China)

Abstract: It presented a feature selection method which is suitable for massive text data set. The method firstly uses an optimized document frequency to select feature and filtes out some terms to reduce the sparsity of feature spaces, and then employs an attribute reduction algorithm based on attribute union theory to eliminate redundancy, so can acquire the feature subset which are more representative. The experimental results show that the syntaxic method is promising.

Key words: feature selection; text categorization; document frequency; attribute union theory; attribute reduction

在文本分类中,文本通常是以向量形式来表示,其特点是高维性和稀疏性[1]。而在中文文本分类中,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章中的全部词条构成。由于中文的词条总数有二十多万条,这使得其高维性和稀疏性更加明显,这样就大大限制了分类算法的选择空间,降低了分类算法的效率和精度。为此,寻找一种高效的特征选择方法,用以降低特征空间维数、避免维数灾难,提高文本分类的效率和精度,成为文本自动分类中亟待解决的重要问题[2]。

1 常用的文本特征选择方法

常用的文本特征选择方法有 IG、WF、DF 等[3]。这里仅简单介绍以下几种方法,其他请参阅文献[3]。

1.1 词频方法

词频方法 WF(Word Frequency)选择特征时仅考虑特征在文档集中出现的次数。如果某个特征在文本集中出现的次数达到一个事先给定的阈值,则留下该特征,否则删除。该方法的缺点在于仅选择出现频繁的词作为特征,而忽略了出现频率较低的词。

1.2 文档频方法

文档频方法 DF(Document Frequency)选择特征时仅考虑特征所在的文档数。如果某个特征在文本集中存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。该方法的缺点在于它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。这样就产生了一个问题: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,通常情况是文档中仅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。

2 本文方法所用策略

2.1 优化的文档频

通过分析词频方法和文档频方法发现,这两种方法具有互补性。为此提出了一个优化的文档频——基于最小词频的文档频。

定义 1 基于最小词频的文档频:特征 f 的最小词频的文档频是指出现特征 f 次数达到一定数目的文档数,记为 DF_n,其中 n 为特征词在文档中至少出现的次数。

2.2 本文属性约简算法

粗糙集理论是一种研究不精确、不确定性知识的数学工具[4]。属性约简是该理论中一个非常重要的概念,它反映了一个信息系统的本质信息。求解一个信息系统的全部约简或计算出最佳约简都是 NP 难问题[5]。本文基于集合理论,提出了关系积概念和基于关系积的属性约简算法,把信息系统的属性约简过程转化为关系积的运算,减小了对决策表的扫描次数,提高了属性约简的效率。

2.2.1 粗糙集基本知识

定义 2 信息系统[5] 信息系统 S 可以表示为 $S=(U,R,V,f)$,其中 U 为对象集合,$R=C\cup D$ 是属性集合,C 为条件属性集,D 为决策属性集,$V=\underset{r\in R}{U}V_r$ 是属性值的集合,V_r 表示属性 r 的值域。$f:U\times R\to V$ 是一个映射函数,它指定 U 中每一个对象 X 的属性值。信息系统也可用二维表来表示,称之为决策表,其中行代表对象 x_i,列代表属性 r,$r(x_i)$表示第 i 个对象在属性 i 上的取值。

定义 3 信息系统 $S=(U,C\cup D,V,f),P\subseteq C$,则 P 在 U 上的不可分辨关系定义为:

$IND(P)=\{(x,y)|x,yU^2,\forall p\in P,f(x,p)=f(y,p)\}$,$IND(P)$表示把 U 划分成若干个等价类簇,记为 $U/IND(P)$,其中 $U/IND(P)=\{X_1,X_2,\cdots,X_n\}$,$X_i$ 为等价类,$i=1,2,\cdots,n$。

2.2.2 关系积理论相关基本知识

定义 4 信息系统 $S=(U,C\cup D,V,f)$,设 $P_1\subseteq C,P_2\subseteq C$,设 $R_UNION(P_1)$和 $R_UNION(P_2)$分别为它们对 U 所导出的等价类,则对于属性集 $P_3=P_1\cup P_2$ 所对应的等价类 $R_UNION(P_1\cup P_2)$称为 $R_UNION(P_1)$和 $R_UNION(P_2)$的关系积[6],记为 $R\_UNION(P_1,P_2)$。$R_UNION(P_1)$称为一元关系积,$R_UNION(P_1,P_2)$称为二元关系积,$R_UNION(P_1,P_2,P_3)$为三元关系积,$R_UNION(P_1,P_2,\cdots P_n)$为 m 元关系积。

关系积概念的实质是两种划分对集合的联合划分,与集合的交集不同,集合的交集取的是两个集合的公共部分,而关系积只是对集合的一次重新划分[6]。

定理 1 关系积运算满足交换率、结合率、分配率等集合运算。

证明根据集合的有关性质,定理显然成立。

根据定理 1 可以降低属性组合的计算数量和利用次级关系积生成新关系积,也就是 $R_UNION(P_1,P_2,P_3)$可以通过 $R_UNION(P_1,P_2)$和 $R_UNION(P_2,P_3)$的关系积运算生成。

定理 2 如果某元关系积的任一元素是决策属性集的子集,则可把该元素删除。

证明:$\forall i$,设 $R(i,j)=R_UNION(P_1,P_2,\cdots P_i)(j)$为这个 i 元关系积的第 j 个元素(其实是个集合)。如果 $S(i,j)\cap R\_UNION(D)=S(i,j)$,那么该元素的任何子集仍然属于同一决策属性子集,对该元素(子集)的进一步划分已没有必要,可以不考虑该子集。所以可以把该元素从 i 元关系积中删除。

推论 1 由于单元素子集一定是决策属性集的子集,因此某元关系积的单元素子集可以直接删除。

定理 3 如果某元关系积通过定理 2 运算后为空集,则该元关系积就是一个属性约简。

证明 根据定理 2 可知,删除的子集是决策属性的子集,如关系积为空,则可知该元关系积的子集一定全部包含在决策属性集中,即 $POS_c(D)=U$,则该元关系积就是一个属性约简。

2.2.3 基于关系积理论的属性约简算法

经典的 Pawlak 约简算法是一种自顶向下的属性约简算法。这类算法需要反复地对决策表进行重组,生成新的决策表,然后根据新的决策表判断是否获得了最小约简,这需要占用大量的计算资源,也降低了算法的效率。引入关系积理论后完全可以根据决策表提供的各属性等价关系及其它们的关系积对属性进行约简,从而获得最小约简属性。由于高元关系积可由次元关系积和一元关系积的集合运算生成(定理 1),不需要对决策表重新进行组织和扫描,把对表扫描转化为集合运算,节省了大量 I/O 开销,极大地提高了算法的效率。

基于关系积理论的属性约简算法简单描述如下:

Input:$S=<U,C\cup D,V,f>$

Output:C 的一个约简 Red,初始 Red=C

(1)$\forall c\in C$,计算其一元关系积 $R_UNION(c)$并计算决策属性集 D 的一元关系积 $R\_UNION(D)$;

(2)根据定理 2 和推论 1,检查一元关系积的子集是否包含在决策属性集中,如成立,则可在一元关系积中删除该子集;

(3)根据定理 3,检验是否获得属性的最小约简,如果获得最小约简,则 goto(6);

(4)$k=2$;

(5)While($k\neq|C|$)

{ For $i=1$ to $|R\_UNION(C_{k-1})|$

For $j=i$ to $|R\_UNION(C_1)|$

$R\_UNION(C_k)=R\_UNION(C_{k-1}(i))\cap R\_UNION(C_1(j))$;根据定理 1,求 k 元关系积。其中 C_{k-1} 表示有 $k-1$ 个条件属性组成的集合,$R\_UNION\ (C_{k-1})$ 表示 $k-1$ 元关系积的集合,$R_UNION(C_{k-1}(i))$表示 $R_UNION(C_{k-1})$的第 i 个 $k-1$ 元关系积,$R_UNION(C_1(j))$表示为第 j 个一元关系积

For $i=1$ to $|R\_UNION(C_k)|$

{ If$|R_UNION(C_k(i))|=1$ or

$R\_UNION(C_k(i))\cap R\_UNION(D)=R\_UNION(C_k(i))$

Then $R\_UNION(C_k)=R\_UNION(C_k)-\{R\_UNION(C_k(i))\}$;

根据定理 2 和推论 1,删除没有必要进一步划分的元素;$R_UNION(C_k(j))$表示为第 j 个 k 元关系积

If $|R\_UNION(C_k(i))|=0$ then

Red=$C_k(i)$goto (6);//根据定理 3 可知获得一个最小约简

}

$k=k+1$;

}

(6)输出 Red,算法结束。

本算法从关系积的概念出发，把对决策表的扫描转化为集合的运算,只对决策表扫描一次,就可生成一个最小约简属性集,避免了对决策表的频繁操作。并且在生成各阶关系积时,为了避免属性组合的爆炸问题,本算法根据关系积的特点提出了一些启发式,从而降低关系积运算的次数。

2.2.4 简单实例

已知决策表 S,由于篇幅有限具体就不表示了。其中对象集 U={1,2,3,4,5,6,7,8,9,10,11,12,13,14};D={d};

C={c_1,c_2,c_3,c_4};$R\_UNION(C)$={{1},{2},{3},{4},{5},{6},{7},{8},{9},{19},{11},{12},{13},{14}};

$R\_UNION(D)$={{1,2,6,8,14},{3,4,5,7,9,10,11,12,13}};

$R\_UNION(c_1)$={{1,2,8,9,11},{3,7,12,13},{4,5,6,10,14};

$R\_UNION(c_2)${{1,2,3,13},{4,8,10,11,12,14},{5,6,7,9}};

$R\_UNION(c_3)${{1,2,3,4,8,12,14},{5,6,7,9,10,11,13}};

$R\_UNION(c_4)${{1,3,4,5,8,9,10,13},{2,6,7,11,12,14}}.

按照本文算法计算如下：

(1) 因为各属性关系积元素不为空，故一阶关系积不能构成最小约简;

(2)下面计算二阶关系积：

$R\_UNION\ (c_1,c_2)$={{1,2},{8,11},{9},{3,13},{12},{7},{4,10,14},{5,6}};

$R\_UNION(c_1,c_3)$={{1,2,8},{9,11},{3,12},{7,13},{4,14},{5,6,10}};

$R\_UNION(c_1,c_4)$={{1,8,9},{2,11},{3,13},{7,12},{4,5,10},{6,14}};

$R\_UNION(c_2,c_3)$={{1,2,3},{13},{4,8,12,14},{10,11},{5,6,7,9}};

$R\_UNION(c_2,c_4)$={{1,3,13},{2},{4,8,10},{11,12,14},{5,9},{6,7}};

$R\_UNION(c_3,c_4)$={{1,3,4,8},{2,12,14},{5,9,10,13},{6,7,11}}.

(3)根据定理 2 和推论 1 及定理 3,分别对各关系积进行化简,化简后的二元关系积为：

$R\_UNION(c_1,c_2)$={{8,11},{4,10,14},{5,6}};

$R\_UNION(c_1,c_3)$={{4,14},{5,6,10}};

$R\_UNION(c_1,c_4)$={{1, 8, 9}, {2, 11}};

$R\_UNION(c_2,c_3)$={{1,2,3},{4,8,12,14},{5,6,7,9}};

$R\_UNION(c_2,c_4)$={{1,3,13},{4,8,10},{11,12,14},{6,7}};

$R\_UNION(c_3,c_4)$={{1,3,4,8},{2,12,14},{6,7,11}}.

(4)由于所有二阶关系积元素不为空,也不构成最小约简。同理计算三元关系积及其化简,化简后的三元关系积如下：

$R\_UNION(c_1,c_2,c_3)$={{4,14},{5,6}};

$R\_UNION(c_1,c_2,c_4)$={},此时算法结束,条件属性集的一个最小约简子集为{c_1,c_2,c_4}。

其实本文算法稍加改进，就可以求出所有的最小约简集以及核属性。

本文中 $R_UNION(c_1,c_2,c_4)$也为空,条件属性集的另一个最小约简子集为{c_1,c_3,c_4}。核属性={c_1,c_4}。

3 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 $f_j \in T_j$,给定最小词频数 n 以及最小文档数阈值 ω):

(1) 计算 f_j 的 $DF_n(f_j,c_j)$;

(2) 若 $DF_n(f_j,c_j)<\omega$ 则把 f_j 从 T 中删除,否则 f_j 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一个特征集;

(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。

4 本文特征选择方法实验验证

4.1 实验语料库

本文选用的分类语料库是复旦大学中文文本分类语料库。该语料库由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组构建,语料文档全部采自互联网,它可以从网上免费下载。

该库中包含 20 个类别,分为训练文档集和测试文档集两个部分。本文只取前 10 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

4.2 实验环境及参数设置

实验设备是一台普通计算机。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。实现语言是 Java，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,它是开源项目,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为 n=2,ω=5。

4.3 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)等 2 种特征选择方法对后续文本分类精度的影响,因此本实验在各种特征选择方法后采用相同的分类器对文本进行分类。本实验中使用 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。

4.4 实验结果

表 2 总结了 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优化的文档频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,从而获得较具代表性的特征子集,因此效果最佳;由于 IG 方法受样本分布影响,在样本分布不均匀的情况下,它的效果就会大大降低,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,因此总体效果次之;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,因此 CHI 方法比 MI 方法效果要好。总的来说,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。

表 2 实验结果

	本文方法		IG		CHI		MI	
类别	准确率/%	召回率/%	准确率/%	召回率/%	准确率/%	召回率/%	准确率/%	召回率/%
经济	89.23	90.34	82.52	80.83	79.31	87.67	75.63	76.99
体育	87.12	91.28	83.88	82.93	81.71	85.60	79.54	80.78
计算机	89.47	89.76	87.64	88.43	82.41	83.51	80.71	77.91
政治	87.09	87.89	78.78	84.29	83.29	78.80	79.99	80.72
农业	87.45	89.64	83.27	89.67	79.56	77.23	72.48	79.45
环境	88.67	88.93	81.67	86.42	81.93	86.56	76.42	80.13
艺术	87.48	89.39	80.55	85.81	82.51	82.78	80.51	81.81
太空	86.31	91.42	82.46	87.47	80.84	79.23	78.57	78.47
历史	88.63	88.61	80.33	87.39	78.34	80.42	77.45	81.92
军事	86.71	89.41	75.53	79.73	60.94	87.67	63.67	74.71
平均率	87.82	89.67	81.66	85.30	79.08	82.95	76.50	79.29

本文首先讨论了几种经典特征词选择方法,总结了它们不足,然后把文档频和词频结合起来提出了一个优化的文档频,紧接着把粗糙集引入进来并提出了一个新的基于关系积理论的属性约简算法,最后把该属性约简算法同优化的文档频结合起来,提出了一个综合性特征选择方法。由于该方法首先利用优化的文档频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,本文特征选择方法同 3 种经典特征选择方法“互信息”和“x^2 统计量”以及信息增益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,从而使其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, SANCHEZ D, VILA M A.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] PAWLAK Z. Rough sets [J]. International Journal of Information and Computer Sciences,1982, 11(5):341-383.

[5] 曾黄麟.智能计算[M].重庆:重庆大学出版社,2004.

[6] 焦吉成,高学东 ,王元璞,等. 关系积理论及属性约简算法[J].山东大学学报(工学版),2008,38(2):112-117.

(收稿日期:2009-04-28)

NEC 电子推出支持 RF 遥控标准“ZigBee RF4CE”的 16 位微控制器产品
~实现全球最高水平低功耗的 RF 无线通信的全闪存微控制器~

近日,NEC 电子推出三款满足家电 RF 遥控国际标准“ZigBee RF4CE”,且在 RF 发送及接收时的低功耗达到全球最高水平的 16 位微控制器产品。新产品于即日起开始提供样品。

新产品在 56 pin、8 mm 的 QFN 封装中,集成了低功耗的 16 位微控制器 78K0R-L 及低功耗 RF 发送接收电路,根据闪存容量大小不同分为 64 KB“μPD78F8056”、96 KB“μPD78F8057”及 128 KB“μPD78F8058”三款产品。

新产品样品,以集成 128 KB 闪存、8 KB RAM 的“μPD78F8058”为例,价格为 500 日元/个。预计 2010 年上半期开始量产,2011 年该三款产品的产量预计将达 120 万个/月。

此外,NEC 电子及合作伙伴即日起开始提供 RF4CE 软件开发套件、RF4CE 库、微控制器软件开发环境、评价板。

(NEC 电子供稿)