基于相关性和冗余度的联合特征选择方法

周 城 葛 斌 唐九阳 肖卫东
(国防科技大学信息系统工程重点实验室 长沙 410073)

摘 要 比较研究了与类别信息无关的文档频率和与类别信息有关的信息增益、互信息和 χ^2 统计特征选择方法,在此基础上分析了以往直接组合这两类特征选择方法的弊端,并提出基于相关性和冗余度的联合特征选择算法。该算法将文档频率方法分别与信息增益、互信息和 χ^2 统计方法联合进行特征选择,旨在删除冗余特征,并保留有利于分类的特征,从而提高文本情感分类效果。实验结果表明,该联合特征选择方法具有较好的性能,并且能够有效降低特征维数。

关键词 文本情感分类,联合特征选择,相关性,冗余特征
中图法分类号 TP391 **文献标识码** A

Joint Feature Selection Method Based on Relevance and Redundancy

ZHOU Cheng GE Bin TANG Jiu-yang XIAO Wei-dong
(Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China)

Abstract Based on a comparative study of four feature selection methods, including document frequency(DF) unrelated to class information, and information gain(IG), mutual information(MI) and chi-square statistic(CHI), which are related to class information, we analyzed the disadvantages of combining these two kinds of methods directly and proposed a joint feature selection method based on relevance and redundancy to joint DF and one of IG, MI and CHI. This approach aims to eliminate redundant features, find useful features for classification and consequently improve the accuracy of text sentiment classification. The results of the experiment show that the proposed method can not only improve the performance but also reduce the feature dimension.

Keywords Text sentiment classification, Joint feature selection, Relevance, Redundant feature

1 引言

随着因特网在全球范围内的飞速发展,越来越多的网民通过这种渠道来表达观点、传播思想,但仅靠人工的方法难以应对网上海量信息的收集和处理。由于情感分类可以在一定程度上解决网上各种评论信息杂乱的现象,方便用户准确地定位所需信息,因此文本情感分类已成为一项具有较大实用价值的关键技术,是组织和管理数据的有力手段。然而,文本数据的高维性给分类任务带来了巨大的挑战,降维已成为必须解决的关键技术问题。目前,特征选择已成为降维的一个重要手段。特征选择通过删除无关、冗余特征,提高了分类任务的效率,改善了预测精确度等性能指标,同时增强了分类结果的可解释性。

常用的特征选择方式有文档频率(Document Frequency, DF)、信息增益(Information Gain, IG)、互信息(Mutual Information, MI)和 χ^2 统计(Chi-square Statistic, CHI),其中信息增益和卡方统计的效果较好[1]。以上各种特征选择方法都具有自身的优、缺点[2]。代六玲等人[3]考察了上述4种特征选择方法,分析了单独使用它们时分类效果不好的原因在于,利用类别信息的特征选择方法对低频词存在不同程度的倚重,同时通过实验证明组合的特征抽取方法不但明显提高了分类的精度,而且显著缩短了分类器的训练时间。李玉鑑等[4]将DF与CHI相结合,不仅保留了CHI统计方法能够考虑特征词项与类别相关性的优点,而且可利用文档频率DF值来过滤低频词,降低了CHI对低频词的倚重,从而有效地选取识别能力强的词汇。但是他们都是简单地将DF和CHI等特征选择方法的选取结果做交集而得到最终特征,一方面不能过滤冗余特征,另一方面过滤了低频词中的一些重要特征。

本文考虑设计一种基于相关性和冗余度的联合特征选择算法 RRJFS(Relevance and Redundancy based Joint Feature Selection algorithm),将DF方法分别与IG、MI和CHI结合进行特征选择,目的是消除冗余特征并保留低频词中的重要特征。实验的结果证明本文设计的方法可行。

2 特征选择方法

在文本的情感分类中,特征选择有利于去除噪音以及冗

到稿日期:2011-05-06 返修日期:2011-08-18 本文受国家自然科学基金(60903225),国防科技大学优秀研究生创新基金(S100502)资助。
周 城(1987—),男,硕士生,主要研究方向为Web挖掘和信息可视化,E-mail:zhoucheng301@163.com;葛 斌(1979—),男,硕士,讲师,主要研究方向为语义分析;唐九阳(1978—),男,副教授,硕士生导师,主要研究方向为P2P计算、无线传感器网络;肖卫东(1968—),男,教授,博士生导师,主要研究方向为信息管理、信息可视化、智能决策技术。

余特征。常用的特征选择方法有以下几种。

2.1 文档频率

文档频率指词条 t 在训练语料中出现该词条的文档数。文档频率通过设置阈值去掉了低频词，当低频词为噪音时，可提高分类效果。但低频词也可能带有很大的信息量，这时直接去掉低频词会损失一部分特征，而影响分类效果。但是文档频率方法具有实现简单、算法复杂度低等优点，能够胜任大规模的分类任务[1]。我们将在中文情感分类的环境中重新检验 DF 的有效性。

2.2 信息增益

信息增益通常指该特征在文本中出现前、后的信息熵之差，用来衡量特征中包含的类别信息。对于词条 t 和文档类别 c，IG 考察 c 中出现和不出现 t 的文档频数，来衡量 t 对于 c 的信息增益。采用如下的定义：

$$IG(t)=-\sum_{i=1}^{m}P(c_i)\log P(c_i)+P(t)\sum_{i=1}^{m}P(c_i|t)\log P(c_i|t)+P(\bar{t})\sum_{i=1}^{m}P(c_i|\bar{t})\log P(c_i|\bar{t}) \quad (1)$$

式中，$P(c_i)$ 表示 c_i 类文档在语料中出现的概率，$P(t)$ 表示语料中包含词条 t 的文档的概率，$P(c_i|t)$ 表示文档包含词条 t 时属于 c_i 类的条件概率，$P(\bar{t})$ 表示语料中不包含词条 t 的文档的概率，$P(c_i|\bar{t})$ 表示文档不包含词条 t 时属于 c_i 的条件概率，m 表示类别数。

对在语料中出现的每个词条计算其信息增益值时，从原始特征空间中移除低于特定阈值的词条，而保留高于阈值的词条作为表示文档的特征。IG 不但考虑了类别信息，而且考虑了低频词对分类结果的影响，因此一般情况下分类效果较好，虽然其统计花费较大。但是当类分布和特征分布不平衡情况严重的时候，该特征选择方法对分类效果并不好。

2.3 互信息

互信息是信息理论中一种衡量两个变量间相互关系的方法。$P(c_i)$ 可看作属于类 c_i 的文本占整个文本集的比例。当词条 t 依赖于类别 c_i 时，互信息较大；当词条 t 与类别 c_i 相互独立时，互信息等于 0；互信息还可能是负值，表示两者是负相关的。对于所有类别，词条 t 的平均互信息[5]可通过式(2)计算：

$$MI(t)=\sum_{i=1}^{m}P(c_i)\log\frac{P(t|c_i)}{P(t)} \quad (2)$$

式中，m 为类别数。MI 方法将低于特定阈值的词条从原始特征空间中移除，降低特征空间的维数，保留高于阈值的词条。MI 的优点是考虑了低频词带有信息量的情况，缺点是低频词的互信息比常用词的互信息高，过于倾向低频词，分类效果不好。

2.4 CHI 统计

CHI 统计方法度量词条 t 和文档类别 c 之间的相关程度，并假设 t 和 c 之间符合具有一阶自由度的 χ^2 分布[6]。词条对于某类的 χ^2 统计值越高，它与该类之间的相关性就越大，其携带的类别信息也越多。令 N 表示训练语料中的文档总数，c 为某一特定类别，t 表示特定的词条，A 表示属于 c 类且包含 t 的文档频数，B 表示不属于 c 类但是包含 t 的文档频数，C 表示属于 c 类但是不包含 t 的文档频数，D 是既不属于 c 也不包含 t 的文档频数，则 t 对于 c 的 CHI 值由式(3)计算：

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)+(C+D)} \quad (3)$$

对于多类问题，分别计算词条 t 对于每个类别的 CHI 值，取它们的平均值，可用式(4)计算词条 t 对于整个语料的 CHI 值：

$$\chi^2_{avg}(t)=\sum_{i}^{m}P(c_i)\cdot\chi^2(t,c_i) \quad (4)$$

式中，m 为类别数，$P(c_i)$ 为训练集中出现 c_i 类网页文档的概率。从原始特征空间中移除低于特定阈值的词条，保留高于该阈值的词条作为文档表示的特征。统计方法相对复杂，其优点是考虑了特征词项与类别之间的相关性，缺点是对低频词的倚重大。

3 联合特征选择算法 RRJFS

前述的特征选择方法都存在或多或少的问题，单纯地使用要么不能很好地过滤低频词中的噪音词，要么不能有效地选取较强类别信息的词项。文献[3]先使用 DF 移除低于一定阈值的低频词，消除 IG、CHI 或 MI 对低频词的倚重，再使用 IG、CHI 或 MI 从剩余词条中移除类别信息较低的噪音词。但是，简单地将 DF 和 CHI 等特征选择方法的选取结果做交集而得到最终特征，一方面不能过滤冗余特征，另一方面过滤了低频词中的一些有利于分类的特征。因此，本文考虑设计一种通用的联合特征选择方法，以实现性能互补，并过滤冗余特征。

3.1 相关概念定义

为了形式化描述联合特征选择算法，引入以下几个定义。Yu 等人基于 Markov Blanket 的概念[7]，给出了冗余特征、近似 Markov Blanket 和支配特征的定义[8]。

定义 1(Markov Blanket) 给定特征 f_i 属于特征集 S，假设 $M_i\subset S(f_i\notin M_i)$，$M_i$ 是 f_i 的 Markov Blanket，当且仅当 $P(S-M_i-\{f_i\},C|f_i,M_i)=P(S-M_i-\{f_i\},C|M_i)$。

Markov Blanket 过滤方法指的是：假设 G 是当前所选的特征子集(初始条件下 $G=S$)，如果 G 中存在 f_i 的一个 Markov Blanket，那么就从 G 中删除 f_i，由此引出冗余特征的定义。

定义 2(冗余特征) 假设 G 是当前所选择的特征子集，一个特征是冗余的并且应该从 G 中删除，当且仅当该特征是弱相关的，并且在 G 中有一个 Markov Blanket M_i。

如果 M_i 是特征 f_i 的 Markov Blanket，说明了 M_i 不仅要包含 f_i 关于类别 C 的信息，还要包含 f_i 关于其它所有特征的信息。由定义 2 可知，如果当前特征集中存在一个子集 M_i 是 f_i 的 Markov Blanket，那么若后续每步继续根据定义删除特征，仍然可以在后续的特征集中找到 f_i 的一个 Markov Blanket。该性质保证了特征在整个过滤过程中都是冗余的。当某个冗余特征被删除后，即使继续删除一些其它特征，该特征仍然是冗余的[7]。

定义 3(近似 Markov Blanket) 如果特征 f_i 与类别 C 之间的相关性用 $R_{i,c}$ 表示，两个特征 f_i 和 f_j 之间的相关性用 $R_{i,j}$ 表示，那么称 f_j 是 f_i 的一个近似 Markov Blanket(approximate Markov blanket)，当且仅当 $R_{j,c}\geqslant R_{i,c}$ 和 $R_{i,j}\geqslant R_{i,c}$ 同时成立。

定义 4(支配特征) 如果相关特征 f_j 在当前特征集中

找不到任何一个近似 Markov Blanket，则称特征 f_j 是支配(predominant)特征。

参照文献[8]中选用的相关性度量方法，本文亦使用 Symmetrical Uncertainty[9]作为相关性度量方法，计算定义 3 中的 $R_{i,c}$ 和 $R_{i,j}$。

3.2 算法 RRJFS 描述

算法 FCBF[8] 基于定义 2 中对冗余特征性质的陈述，由支配特征构成近似 Markov Blanket，显式地去除冗余特征。本文通过改进 FCBF 算法，提出了能够删除冗余特征的联合特征选择算法 RRJFS。本算法基于相关性和冗余度，将文档频率与信息增益、互信息以及 CHI 统计 3 种方法中的一种进行联合，文中分别称它们为 RR_DF-IG、RR_DF-MI 和 RR_DF-CHI。表 1 以 RR_DF-IG 为例对算法进行说明，其中候选特征集 F 基于 4 种词性的词(名词、动词、形容词、副词)来表示特征[10]，包括正面(positive)文档集中的特征集 $S_p(F_1,F_2,\cdots,F_M,C_p)$ 和负面(negative)文档集中的特征集 $S_n(f_1,f_2,\cdots,f_N,C_n)$。

表 1 算法 RRJFS_DF-IG

输入：候选特征集 $S=S_p(F_1,F_2,\cdots,F_M,C_p)\cup S_n(f_1,f_2,\cdots,f_N,C_n)$
输出：最优特征子集 S_{best}

```
1 begin
2 计算 S 的 DF 值，选取值较大的词项形成特征集 SDF；计算 S 的 IG 值，选取值较大的词项形成特征集 SIG
3 令 SDF−p=SDF∩Sp(F1,F2,…,FM,Cp)，SIG−p=SIG∩Sp(F1,F2,…,FM,Cp)
4 令 SDF−n=SDF∩Sn(f1,f2,…,fN,Cn)，SIG−n=SIG∩Sn(f1,f2,…,fN,Cn)
5 在 SDF−p 和 SIG−p 中计算 Fi 和 Cp 的相关性 Ri,p，对 SDF−p 和 SIG−p 进行降序排列得到两个序列，分别记为 S1 和 S2
6 在 SDF−n 和 SIG−n 中计算 fi 和 Cn 的相关性 Ri,n，对 SDF−n 和 SIG−n 进行降序排列得到两个序列，分别记为 S3 和 S4
7  for z=1 to 4 do begin // z 为第 3,4 步中 S1、S2、S3 和 S4 的下标
8  Fj= getFirstElement(Sz) // 选取 Sz 中的第一个元素作为 Fj
9    do begin
10     Fi= getNextElement(Sz,Fj) // 选取 Sz 中 Fj 的下一个元素作为 Fi
11     if(Fi≠NULL)
12       do begin
13         if (Ri,j≥Ri,c) // 当 Sz 为 S1 和 S2 时，Ri,c=Ri,p(第 2 步)，当 Sz 为 S3 和 S4 时，Ri,c=Ri,n(第 3 步)
14         delete Fi from Sz// 由于是降序排列，这时已满足 Rj,c≥Ri,c，所以 Fi 是 Fi 的近似 Markov Blanket
15         Fi= getNextElement(Sz,Fi)
16       end until(Fi==NULL)
17     Fj= getNextElement(Sz,Fj)
18   end until(Fj==NULL)
19 end
20 Sbest=S1∪S2∪S3∪S4
21 end
```

为了更进一步说明算法是如何执行的，图 1 表述了上述算法第 7 步到第 19 步的执行过程。根据第 7 步，初始化 $z=1$，即选择 S_1。第 8 步时，将 F_1 作为支配特征，依据近似 Markov Blanket。从第 9 步到第 16 步，删除了 F_2 和 F_3。根据第 17 步，将特征 F_4 作为新的支配特征，跳回第 10 步。由于在第 11 步判断 F_i 为空，跳到第 17 步，这时 F_j 亦为空，故该循环结束。跳到第 7 步，接着选择 S_2，执行步骤 8 到 18，直至 for 循环结束。最后到第 20 步，得到最优特征子集 $S_{best}=\{F_1,F_4,F_5,F_6,f_1,f_2,f_3,f_5\}$，算法结束。需要说明的是，由于已经根据相关性进行了降序排列，因此 S_z 中第一个特征一定是支配特征。

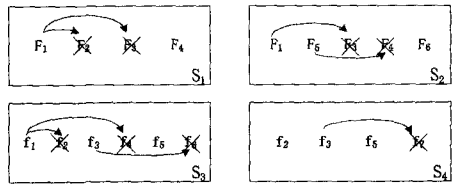

图 1 冗余特征的删除

4 实验结果及分析

4.1 实验设置

实验中，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，并去掉其中的停用词。本文选取名词、动词、形容词、副词作为候选特征，因为情感倾向分类主要取决于情感词，而情感词一般属于以上几种词性。文档表示采用向量空间模型(Vector Space Model，VSM)[11]，特征权重用 TF-IDF(Term Frequency-Inverse Document Frequency)[12]来计算。常用的 4 种特征选择方法使用中科院文本分类演示系统 TxtCat 提供的方法，联合特征选择方法由本课题组编码实现，代码用 Java，在 JDK 1.6 的环境下实现。实验使用支持向量机(Support Vector Machine，SVM)作为文本分类器，在实现 SVM 分类模型时使用了台湾大学的 LibSVM 软件包。实验的数据集采用中国科学院计算技术研究所谭松波博士提供的中文情感挖掘语料 ChnSentiCorp-Htl-unba-10000[13]，内容为酒店评论。原数据集为非平衡语料，共 10000 篇评论，其中 7000 篇为正面评论，3000 篇为负面评论。经过处理后，留下正、负评论各 3000 篇。

4.2 实验结果

4.2.1 性能测试

为了证明算法 RRJFS 的有效性，在实验中将 RR_DF-IG、RR_DF-MI、RR_DF-CHI 与文献[3]中提到的简单组合特征选择方法 DF+IG、DF+MI 和 DF+CHI 以及 DF、IG、MI 和 CHI 方法的性能进行比较(采用最通用的性能评价方法：召回率 R(Recall)、准确率 P(Precision)和 F_1 评价，其中 $F_1=\frac{2RP}{R+P}$)。经过预处理后，分别在正、负评论训练集中抽取 2/3 的数据作为训练集，余下的作为测试集。基于支持向量机的方法得到最终的分类结果，如表 2 所列。实验选取文档频率、信息增益、互信息和 CHI 统计 4 种方法按值降序排列后排在前面 40%的特征项。

表 2 特征选择方法对比实验结果

特征选择方法 / 宏平均值	DF	IG	MI	CHI	DF+IG	DF+MI	DF+CHI	RR_DF-IG	RR_DF-MI	RR_DF-CHI
召回率	0.8055	0.8045	0.7185	0.8020	0.8305	0.8030	0.8250	0.8505	0.8345	0.8530
准确率	0.8058	0.8055	0.7196	0.8030	0.8328	0.8054	0.8270	0.8523	0.8357	0.8543
F_1 值	0.8057	0.8050	0.7191	0.8025	0.8316	0.8042	0.8260	0.8514	0.8351	0.8537

在单独使用特征选择方法的比较中，可以看出 DF 和 IG 方法效果最好，MI 效果最差。主要原因在于中文的特征空间维数较高，导致许多低频词的出现，其中有些与情感类别相关，有些则是噪声词。使用类别信息的特征选取方法在不同程度上倚重低频词，最终使分类效果变差。结果证实，文献[3]所提到的简单组合特征选择方法 DF+IG、DF+MI 和DF+CHI确实能够提高分类效果，这是因为其降低了单独使用 IG、MI 和 CHI 时对低频词的倚重。其中 MI 和 DF 组合后提高最为显著，宏平均 F_1 值($Macro\text{-}F_1$)从原来的 0.7191 提高到 0.8042，提高了近 9 个百分点。原因在于 MI 过于倚重低频词(将词频作为分母)，从而使用 MI 导致许多噪声词未被过滤掉，极大地影响了实验结果。通过与 DF 联合后，降低了低频词的影响，效果得到显著提高。IG 和 CHI 通过与 DF 组合后，提高了近 3%，这主要是 IG 和 CHI 本身效果较好，同时对低频词的倚重程度不如 MI 显著。

通过联合特征选择方法后，分类性能在组合的基础上提高了近 3 个百分点，这是因为 DF+IG、DF+MI 和 DF+CHI 只能简单地通过取交集将 DF 与 IG、MI 或 CHI 的选取特征集合进行组合，不能过滤一些冗余特征，而且由于 DF 的作用过滤了低频词中的一些重要类别特征。而 RR_DF-IG、RR_DF-MI 和 RR_DF-CHI 在考虑相关性和冗余度的情况下过滤了冗余特征并保留低频词中的重要特征，从而能够得到更好的分类效果。

4.2.2 扩展性测试

为了进一步证明算法 RRJFS 的有效性，通过改变 DF、IG、MI 和 CHI 4 种方法的选取范围来测试算法的可扩展性。图 2 给出了 RR_DF-IG、RR_DF-MI 和 RR_DF-CHI 在不同特征数目下的 $Macro\text{-}F_1$。其中横轴表示把所有特征项按权值降序排列后按比例从前面选取特征项的百分比；纵轴表示分别计算选取前 $k\%$($k=5,10,15,\cdots,40$)个特征项时对应的 $Macro\text{-}F_1$。

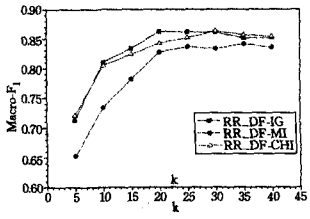

图 2 宏平均 F_1 值的比较

由图 2 可得，算法 RRJFS 的效果较好，随着特征的增加分类性能逐渐得到提高。当保留的特征项的比例达到某个值(20%左右)后，上升的趋势明显减缓，达到稳定状态。比例增加到 30%后，分类性能略有下降。另外，RR_DF-IG 和 RR_DF-CHI 的效果相当，并明显优于 RR_DF-MI。RR_DF-IG 可以在选取较少特征项(20%)的情况下取得最好的效果，$Macro\text{-}F_1$ 约为 87%。

结束语 本文考察了自动分类中常用的 4 种特征选择方法：文档频率、信息增益、互信息和 χ^2 统计，分析了它们的优、缺点，提出了一种基于相关性和冗余度的联合特征选择算法 RRJFS。文中通过将 SVM 方法应用于中文文本情感性分类任务，针对这 4 种特征选择方法、3 个组合特征选择方法[3]和本文提出的基于算法 RRJFS 的联合特征选择方法进行了对比实验。结果表明，采用本文提出的联合特征选择方法能取得较好的分类效果。

但是，我们也应看到，本文实验所使用的语料态度比较明确，用词较为规范，语言朴实简单，要将情感分类做到实用，其实还面临许多困难。其中在中文文本情感分类时考虑句类(汉语语法传统上按照语气标准把句子分为陈述句、疑问句、祈使句和感叹句，称为句类)和否定词的影响以及如何处理文本中存在正反态度并存现象的情感分类，将是下一步重点解决的问题。

参 考 文 献

[1] Yang Y. A Comparative Study on Feature Selection in Text Categorization[C] // Proceedings of the 14th International Conference on Machine Learning(ICML-97). Nashville: Morgan Kaufmann, 1997: 412-420

[2] 孙国菊，张杰. 中文文本分类的特征选取评价[J]. 哈尔滨理工大学学报，2005，10(1)：76-78

[3] 代六玲，黄河燕，陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报，2004，18(01)：26-32

[4] 李玉鑑，周玉珍，操卫平. 基于 DF 和 CHI 的联合特征提取方法及其应用[J]. 北京工业大学学报，2008，34(9)：995-1000

[5] Lewis D D. Feature Selection and Feature Extraction for Text Categorization[C] // Proceeding of Speech and Natural Language Workshop. San Francisco, USA: Morgan Kaufmann, 1992: 212-217

[6] Dunning T E. Accurate methods or the statistics of surprise and coincidence [C] // Proceedings of Computational Linguistics. 1993: 61-74

[7] Koller D, Sahami M. Toward Optimal Feature Selection[C] // Proceedings of the 13th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1996: 284-292

[8] Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research, 2004, 5: 1205-1224

[9] Press W H, Teukolsky S A, Vetterling W T, et al. Numerical Recipes in C[M]. Cambridge: Cambridge University Press, 1988

[10] 唐慧丰，谭松波，程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报，2007，21(6)：96-100

[11] Salton G. A Vector Space Model for Automatic Indexing[J]. Communication, 1975, 18(11): 613-620

[12] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer, 1995

[13] 谭松波. 中文情感挖掘语料-ChnSentiCorp[EB/OL]. http://www.searchforum.org.cn/tansongbo/corpus-senti.htm