基于类别和改进的 CHI 相结合的特征选择方法*

高宝林，周治国†，杨文维，肖泽力
（东北师范大学 信息科学与技术学院，长春 130117）

摘　要：针对传统 CHI 方法的低频词缺陷问题以及传统 CHI 方法是在全局范围内作特征选择，忽略了特征和类别间的相关性问题，提出了改进方法。通过引入类内和类间分布因子，减少了低频词带来的干扰，并且降低了特征词在类间均匀分布时对分类带来的负贡献，同时提出基于类别的特征选择方法。采用随机森林分类算法，将提出的方法应用在微博情感分析领域。实验结果表明，以上方法能够有效地提高微博情感分类的准确率、查全率和 *F* 值。

关键词：卡方检验；特征选择；情感分析；随机森林
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2018)06-1660-03
doi:10.3969/j.issn.1001-3695.2018.06.013

Feature selection method based on combination of category and improved CHI

Gao Baolin, Zhou Zhiguo†, Yang Wenwei, Xiao Zeli
(*College of Information Science & Technology, Northeast Normal University, Changchun* 130117, *China*)

Abstract: On the issues of the low-frequency word defects in traditional CHI method, and traditional CHI method chooses the feature in the global scope, ignoring the correlation between the feature and the category, this paper proposed an improved method. It introduced the intraclass and interclass factors to reduce the interference which was caused by the low-frequency words and the negative contribution of the feature words evenly distributed among classes. At the same time, this paper proposed a feature selection method based on category. Using the random forest classification algorithm, this paper applied the proposed method to Weibo sentiment analysis. The experimental results show that this method can effectively improve the accuracy rate, recall rate and *F*-measure in sentiment classification of Weibo.

Key words: CHI; feature selection; sentiment analysis; random forest

0　引言

微博是一种通过关注机制分享简短实时信息的广播式社交平台，用户可以通过网页、移动终端等客户端，发表最多 140 字的信息并实现与他人共享。微博自问世以来，吸引了大量用户在微博上记录生活、讨论热点话题、表达和分享观点，已成为挖掘人们观点与情感的重要资源[1]，为用户满意度调查、舆情监测、社会学研究等应用提供有效的数据支持。情感分析技术能够自动将文本中表达的情感进行分类。在进行分类前，微博文本需要用特征向量来描述，向量维数过高会提高分类时间代价，降低分类精度，因此特征降维是分类预处理过程中的关键步骤。

在特征选择方面，美国卡内基梅隆大学的 Yang 等人[2]针对文本分类问题，对 IG、DF、MI 和 CHI 等特征选择方法进行了比较，得出 IG 和 CHI 方法分类效果相对较好的结论。熊忠阳等人[3]分析了 CHI 统计方法的不足，将频度、集中度和分散度指标应用到 CHI 统计方法上，对 CHI 统计进行改进；肖婷等人[4]通过引入文档内频度和类内正确度指标对 CHI 统计进行改进；刘海峰等人[5]通过特征项的类内分布、类间分布以及类内不同文本之间分布等角度，对 CHI 模型进行逐步优化；黄源等人[6]通过计算特征词之间的剩余互信息，提出了对卡方检验的选择结果进行优化的方法；肖雪等人[7]通过设置最低词频阈值，去除了部分低频词，减少了低频词带来的干扰；裴英博等人[8]分析了影响传统 CHI 统计方法分类精度的因素，去除了特征项与类别负相关的情况；黄章树等人[9]通过降低负相关低频词在 CHI 特征选择算法中的权重，减小了低频词对模型的影响；张辉宜等人[10]通过计算词概率和文档概率来衡量词文档频繁程度，提出一种基于概率的卡方特征选择方法，在不平衡数据集上具有很好的分类效果；宋阿羚等人[11]根据特征项在文本中的位置信息和词频信息对 CHI 算法作出改进，实验取得了较好的结果；Jin 等人[12]使用样本方差计算词的分布信息，并考虑最大词频信息来改进 CHI 方法，在三个语料库上均取得了较好的结果。CHI 算法是有监督的特征选择算法，它和 IG 特征选择算法在文本分类的性能表现上不相上下，有时候甚至比 IG 特征选择算法更为出色，所以更多地应用于文本分类[13]；但对于多分类问题来说，上面的这些改进措施先计算特征针对每个类别的卡方值，再针对类别求最大值，作为特征相对于整个训练集合的卡方值，这种全局方法忽视了特征和

收稿日期：2017-01-23；**修回日期**：2017-03-09　　**基金项目**：东北师范大学教师教学发展基金项目(15B1XZJ014)

作者简介：高宝林(1989-)，男，吉林通化人，硕士，主要研究方向为情感分析、数据挖掘；周治国(1976-)，男(通信作者)，安徽枞阳人，副教授，博士，主要研究方向为计算机网络安全(zhouzg281@nenu.edu.cn)；杨文维(1994-)，男，江西瑞金人，硕士，主要研究方向为情感分析、数据挖掘；肖泽力(1992-)，男，四川泸州人，硕士，主要研究方向为网络安全.

类别间的相关性[3~12]。

针对以上问题,本文提出基于类别和改进的 CHI 相结合的特征选择方法。

1　CHI 特征选择方法

CHI 统计方法假设特征项 t 与类别 c 之间的非独立关系类似于具有一维自由度的 χ^2 分布,t 对于 c 的 CHI 统计量可计算为

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

其中:A 表示包含特征项 t 且属于类别 c 的文档频数;B 表示包含特征项 t 但不属于类别 c 的文档频数;C 表示属于类别 c 但不包含特征项 t 的文档频数;D 表示既不属于类别 c 也不包含特征项 t 的文档频数;N 表示语料中的文档总数。

CHI 统计方法用来度量特征项 t 和类别 c 之间的相关程度。特征项 t 与类别 c 相互独立时,$\chi^2(t,c)=0$。特征项 t 与类别 c 的相关性越强,$\chi^2(t,c)$ 的值就越大,此时特征项 t 所包含的与类别 c 相关的鉴别信息就越多。

2　基于类别和改进的 CHI 相结合的特征选择方法

2.1　类内分布因子

引入类内分布因子是基于这样的考虑:对于两个特征 t_1 和 t_2,在类别 c 中出现的词频数量都是 100,但是 t_1 均匀出现在 100 条微博文本中,而 t_2 集中出现在 20 条微博文本中。这样,t_1 在类别 c 中的分布更为广泛,能很好地代表类别 c,而 t_2 集中出现在少数微博文本中,不能很好地代表类别 c。所以,在特征选择时应更偏向于 t_1。同时,类内分布因子还能很好地解决微博文本分类中的低频词问题。由于微博文本最多只有 140 个文字,通过统计分析,低频词绝大多数是那些只出现在一条或几条微博文本中的词。因此,类内分布因子的计算方法为

$$\alpha=\frac{N_t}{N} \tag{2}$$

其中:α 表示类内分布因子;N_t 表示类别 c 中包含特征 t 的微博文本数量;N 表示类别 c 中全部微博文本数量。在计算时,α 越大,表示特征 t 的表征效果越好,分类能力也就越高。

2.2　类间分布因子

引入类间分布因子是基于这样的考虑:一个好的特征应该集中出现在某一个类别中,而不是均匀地分布在所有类别中。因此,类间分布因子的计算方法为

$$\beta=\frac{C_t}{C+1} \tag{3}$$

其中:β 表示类间分布因子;C_t 表示类别 c 中包含特征 t 的微博文本数量;C 表示包含特征 t,但不属于类别 c 的微博文本数量,加 1 是为了防止 $C=0$。

由以上定义可知,对于某一特征 t,其 α 和 β 值越大,则对微博文本的分类越有用,区分能力越强。所以,本文在 CHI 统计方法的基础上乘上类内和类间分布因子,对 CHI 统计方法作出改进,得到改进公式如下:

$$\chi^2(t,c)=\frac{N(AD-BC)^2\times\alpha\times\beta}{(A+C)(B+D)(A+B)(C+D)} \tag{4}$$

2.3　基于类别的特征选择

传统的 CHI 方法是把训练集合的文本数据进行分词后,形成候选特征集合,然后根据式(1)计算候选特征词相对于各个类别的卡方值,再把得到的卡方值求概率和或是最大值,作为该特征词相对于整个训练集合的最后卡方值。然后根据最后的卡方值进行降序排列,选取前 N 个特征词作为最后的特征集合。

本文在进行特征选择时,将训练集合的微博文本数据按类别进行分词,形成每个类别的候选特征集合,使用式(4)分别计算每个类别的候选特征词的卡方值,按降序排列,最后选取每个类别的前 N 个特征词作为最后的特征集合。如果选取的各个类别间的特征词有重复的,则只取一次。

具体算法流程如下:

输入:类别集合 C,特征选择维度 N,训练集文本 D。

输出:特征词集合 K。

a)将训练集文本 D 进行分词;

b)按照标注为类别 c_i 生成原始特征词集合 K_i;

c)针对类别集合 C 中的每个类别 c_i($c_i\in C$),使用式(4)计算其原始特征词集合 K_i 中每个特征词的卡方值;

d)将原始特征词集合 K_i 中的特征词按照卡方值降序排列;

e)分别从各个类别的原始特征词集合 K_i 中,选取排名前 N 的特征词放入集合 K 中;

f)将集合 K 中重复的特征词保留一个,其余删除;

g)输出特征词集合 K,算法结束。

3　实验和结果分析

3.1　数据准备

本文的实验数据是通过新浪微博提供的官方接口下载到的微博文本数据,通过人工判断将数据分为四个类别,即 happy、angry、sad 和 disgust[14]。每个类别各有6 250 条数据,共 25 000条。

在预处理阶段,分词使用的是中国科学院的汉语分词系统 NLPIR[15]。分词后,去除停用词和链接等无用信息。

在特征选择阶段,分别使用本文提出的方法(C-ICHI)、传统的 CHI 方法(CHI)、文献[8]提出的式(2)方法(CHI-IMP)和式(6)方法(CHI-FCDI),并使用 TF-IDF 算法为特征进行加权计算。

在将微博文本抽象成以特征词的权重为分量的特征向量时,由于微博文本是短文本,最多只有 140 字,会造成特征向量中存在大量的 0 值,给分类结果造成影响。所以在抽象前,先将特征词按照词性进行划分,分为名词、动词、形容词和其他词四种,其他词中还包括对分类有很大帮助的表情符号。这样一条微博文本会表示为具有四列数据的特征向量,其中每列数据会包含多个特征词,求出这些特征词的平均权重,作为该列的最终值。

3.2　评价标准

本文中使用的评价标准是文本分类中普遍使用的性能评价指标:准确率、查全率、F 值。

$$准确率 = \frac{正确分类的文本数}{被分类器识别为该类的文本数}$$

$$查全率 = \frac{正确分类的文本数}{测试集中该类文本总数}$$

$$F\text{ 值} = \frac{准确率 \times 查全率 \times 2}{准确率 + 查全率}$$

本文将微博文本分为四个类别,在计算以上三个性能指标时,先分别计算每个类别的,然后求出四个类别的准确率、查全率和 F 值的平均值,作为最后的实验结果。

3.3 实验结果分析

本文实验是在 WEKA[16] 平台上实现的,采用随机森林分类算法,参数为默认参数,测试集从实验数据集中随机抽取 20% 和 40%,特征维度为 1 000、1 500、2 000。实验结果如图 1 ~6 所示。

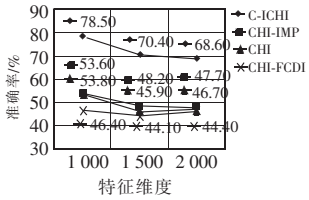

图 1 测试集占 20%的准确率

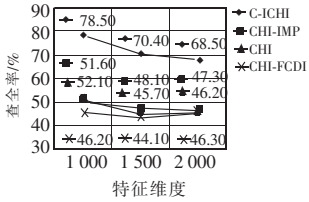

图 2 测试集占 20%的查全率

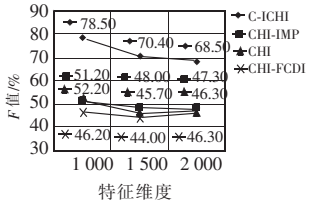

图 3 测试集占 20%的 F 值

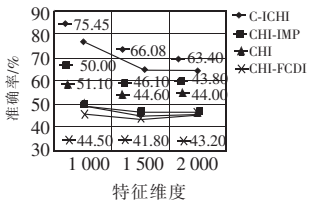

图 4 测试集占 40%的准确率

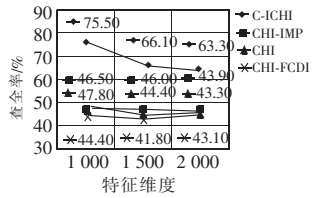

图 5 测试集占 40%的查全率

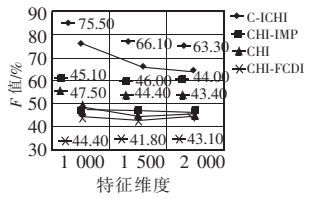

图 6 测试集占 40%的 F 值

从图 1 ~6 的实验结果可以看出:

a) C-ICHI 方法的准确率、查全率和 F 值这三项指标都要优于其他三种方法,这主要是因为其他三种方法是基于全局的特征选择和加权。不同于文献[8]实验时所使用的包含不同领域数据的语料集,微博文本的各个类别间并不存在每个类别独有的、对分类有很大贡献的特征词,几乎所有的有效特征词都会在各个类别中出现,所以造成了全局的特征选择和加权方法使得这些特征词的类别区分度下降,导致分类准确率等指标不高。而 C-ICHI 方法能够根据每个类别本身的特点,选择更能代表本类别的特征。即使是同时出现在四个类别中的特征词,也会得到不同的权值,这样就提高了分类的准确性等指标。

b) 随着特征维度的上升,C-ICHI 方法的准确率等各项指标在下降。通过分析四个类别的候选特征集合,在特征维度为 1 000 维时,这 1 000 个特征基本上是候选特征集合中最为有效的,它们的类内分布相对广泛,类间分布相对不均衡。当特征维度上升时,后续加入的特征词既有高频但类间分布均衡的词,也有低频词,这样的特征词会降低分类的准确性。

从上面的实验可以看出,本文提出的方法不仅能够降低特征的维度,而且还能提高微博文本情感分类的总体性能。与文献[10]的方法相比较,在测试集占 20% 时,准确率提高了 14.2%,查全率提高了 25.2%,F 值提高了 20.2%。

4 结束语

传统的 CHI 特征选择方法忽视了特征和类别间的相关性,同时 CHI 特征选择方法还存在低频词缺陷。针对以上问题,本文提出基于类别和改进的 CHI 相结合的特征选择方法。将该方法应用在微博文本情感分析领域,与传统的 CHI 方法和其他改进的 CHI 方法比较。实验结果表明,该方法能够有效地提高微博情感分类的准确性等指标,要优于传统的 CHI 方法和其他改进的 CHI 方法。

参考文献:

[1] Alexander P, Patrick P. Twitter as a corpus for sentiment analysis and opinion mining[C]//Proc of the 7th International Conference on Language Resources and Evaluation. Valletta: ELRA Press, 2010: 1320-1326.

[2] Yang Yiming, Pedersen J. A comparative study on feature selection in text categorization[C]//Proc of the 4th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997: 412-420.

[3] 熊忠阳,张鹏招,张玉芳. 基于 χ^2 统计的文本分类特征选择方法的研究[J]. 计算机应用,2008,28(2):513-514,518.

[4] 肖婷,唐雁. 改进的 χ^2 统计文本特征选择方法[J]. 计算机工程与应用,2009,45(14):136-137,140.

[5] 刘海峰,苏展,刘守生. 一种基于词频信息的改进 CHI 文本特征选择[J]. 计算机工程与应用,2013,49(22):110-114.

[6] 黄源,李茂,吕建成. 一种基于开方检验的特征选择方法[J]. 计算机科学,2015,42(5):54-56,77.

[7] 肖雪,卢建云,余磊,等. 基于最低词频 CHI 的特征选择算法研究[J]. 西南大学学报:自然科学版,2015,37(6):137-142.

[8] 裴英博,刘晓霞. 文本分类中改进型 CHI 特征选择方法的研究[J]. 计算机工程与应用,2011,47(4):128-130,194.

[9] 黄章树,叶志龙. 基于改进的 CHI 统计方法在文本分类中的应用[J]. 计算机系统应用,2016,25(11):136-140.

[10] 张辉宜,谢业名,袁志祥,等. 一种基于概率的卡方特征选择方法[J]. 计算机工程,2016,42(8):194-198,205.

[11] 宋阿羚,刘海峰,刘守生. 基于位置及词频信息的优化 CHI 文本特征选择方法[J]. 计算机科学与应用,2015,5(9):322-330.

[12] Jin Chuanxin, Ma Tinghuai, Hou Rongtao, *et al*. Chi-square statistics feature selection based on term frequency and distribution for text categorization[J]. IETE Journal of Research, 2015, 61(4): 1-12.

[13] 尚文倩. 文本分类及其相关技术研究[D]. 北京:北京交通大学,2007.

[14] Zhao Jichang, Dong Li, Wu Junjie, *et al*. MoodLens: an emotion-based sentiment analysis system for Chinese tweets[C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 1528-1531.

[15] 中国科学院计算技术研究所. NLPIR[EB/OL]. (2016-11-15)[2017-01-01]. http://ictclas.nlpir.org/.

[16] The University of Waikato. WEKA[EB/OL]. (2016-12-01)[2017-01-01]. http://www.cs.waikato.ac.nz/ml/weka/.