基于改进的 CHI 统计方法在文本分类中的应用①

黄章树, 叶志龙

(福州大学 经济与管理学院, 福州 350108)

摘 要: 随着文本分类技术的发展与成熟, 越来越多的企业将其应用到客户投诉分类系统中, 并获得了一定的效果. 针对传统卡方统计方法偏向于选择出负相关低频噪音词, 将改进的 CHI 统计方法运用到文本特征选择, 通过降低负相关低频词在特征选择算法中的权重, 减小其对模型的影响. 最后, 对某省通信公司的业务投诉文本进行实验, 结果表明该模型和方法是有效的, 能更准确地对业务投诉工单进行分类, 从而为后续问题的分析提供数据支持.

关键词: 业务投诉; 文本分类; 特征选择; 卡方统计方法

Application of Text Categorization Based on Improved CHI-Square Statistic Method

HUANG Zhang-Shu, YE Zhi-Long

(Department of Economics and Management, Fuzhou University, Fuzhou 350108, China)

Abstract: With the development and maturity of text classification technology, more and more enterprises have applied it to the customer complaint classification system, and obtained the certain effect. Given that the CHI-square Statistic methods tend to choose negative words, so an improved CHI statistical method is applied to the text feature selection, which means reducing the weight of negative words in the feature selection algorithm and minimizing the impact on the model. Finally, an experiment is performed on the complaint text of a communications company business. The result shows that the model and method are effective, and can be more accurate for the classification of business complaints, so as to provide data support for the follow-up problem analysis.

Key words: business complaints; text classification; feature selection; CHI-square statistic method

现如今各大企业为了更好地适应市场竞争, 更全面地满足客户需求, 纷纷将企业的重心转向客户服务, 并通过不断提升自身服务来提高客户满意度. 而客户抱怨与客户投诉是企业发现自身不足与问题的一个重要途径, 也是企业提升自身服务与客户满意度的一个重要方向. 但是目前大部分企业在处理客户投诉问题时, 主要还是以人工鉴别、分类的方式为主. 这种方式效率低、成本高、结果一致性不高. 如果不能够及时、准确的处理客户投诉问题, 可能会给相关部门的后续处理与决策带来影响, 甚至给企业造成不可估量的损失. 因此, 如何及时、准确地对客户投诉问题进行分类, 已经成为一个亟需解决的问题.

目前国内外一些学者对维度约减的特征选择算法进行了一系列的研究. 常见的有卡方特征选择(CHI-square, CHI)、互信息(Mutual Information, MI)、信息增益(Information Gain, IG)和基尼指数(Gini index)等. 朱颢东等采用文档词频对特征集进行初选, 之后根据 K-均值的聚类结果对特征集进一步选择[1]. 孟佳娜等提出一种基于贡献度 FCD(feature contribution degree)的特征选择方法. 特征在某个类别的 FCD 是特征出现在该类别的文本数和出现在所有类别的文本数的比值[2]. Foithong 根据粗糙集和互信息理论, 提出了一种融合了两种理论的特征选择算法[3]. 刘赫等根据最大边界相关理论, 将 CHI 统计量和信息新颖度的思想相结合并用于特征的选择[4]. Qinbao Song 利用图论的聚类方法将特征分为集群, 接着从每个类中选取最

① 收稿时间:2016-02-18;收到修改稿时间:2016-03-22 [doi:10.15888/j.cnki.csa.005393]

具代表性的特征作为特征子集[5]. 冀俊忠等给出了一种类别加权策略以强化小类别的特征, 然后设计了类别方差统计策略来凸显含有丰富类别信息的特征, 最后将两种策略相融合实现联合特征选择[6]. Fan Min 提出一种基于测试成本约束的特征选择方法, 该方法能够为分类选取饱含信息又低成本的特征子集[7]. 李湘东等结合分布斜率训练集自身的特点, 提出基于改进的 LDA 模型的特征选择方法[8]. 段洁等重新定义了邻域粗糙集的下近似和依赖度计算方法, 构造了基于邻域粗糙集的多标记分类任务的特征选择算法[9]. 吴树芳等针对话题特征提取方法 ITF-IDF 没有考虑类别信息的缺点, 提出改进的互信息计算方法 CMI 和 DCMI[10]. 黄贤英等根据词项的 TF-IDF、词性与词长因子构造综合评估函数, 对微博短文本进行特征词选择[11]. 唐立力采用 K-means 聚类对前三层逐层实现特征词提取, 最后使用 Aprori 算法找出第四层的最大频繁项集作为第四层的特征词集合[12]. 樊小超等加入词频信息、文档频率信息以及类别相关度因子, 提出一种基于改进的互信息特征加权方法[13]. 刘帅等针对面向聚类的特征选择算法效率和效果无法兼顾, 提出了一种基于邻域分析的加权特征选择算法 ENFSA[14]. 王连喜等定义了一种特征平均相关度的度量方法, 提出了基于特征聚类的特征选择方法 FSFC[15].

广泛应用的文本表示方法存在两个特点, 一个是特征维度高, 另一个是稀疏度高. 这两个特点不但会降低分类效率、加大计算成本, 还可能引起过度拟合现象. CHI 统计方法是文本分类中常用的特征选择方法, 但是 CHI 统计方法由于考虑了特征项与类别的负相关, 在实际应用中, 偏向于选择出负相关的特征词. 而结果表明, 很多负相关词都是低频噪音词. 基于此, 本文提出一种改进的 CHI 统计特征选择方法, 旨在通过降低负相关低频词在特征选择算法中的权重, 减小其对模型的影响.

1 业务投诉文本分类模型构建

1.1 文本分词

分词之前需要将业务词库中的词添加到分词词库, 这样才能保证分词结果的正确性. 接着根据停用词表进行停用词确定, 再将文本中出现的停用词去除.

同义词合并, 将文本中具有相同意思的不同词语用固定统一的词语替换. 同义词合并和去停用词一样, 需要根据同义词表进行同义词合并, 将文本中出现的同义词替换为统一的词语.

1.2 特征选择

通过对 CHI 特征选择算法的分类结果分析, 发现影响其分类效果的主要是因为它考虑了特征项和类别的负相关关系, 并且实验发现 CHI 特征选择算法对低频词的倚重较大. 在实际的文本集中, 存在一定数量的低频词, 而仅有少数的低频词和类别存在比较强的相关性. 大部分低频词都是噪声词, 不应该被选取到特征集. 对于只有在几个少数的类别中出现而与大部分类别都为负相关关系的低频噪声词, CHI 特征选择算法通常会给予比较高的评价, 从而影响到模型的分类效果. 尤其是在类别分布不平衡的时候, 其影响十分明显.

基于上述存在的问题, 本文对 CHI 特征选择算法进行了改进, 旨在去除或减小负相关特征项的影响, 改进后的方法见式(1):

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

CHI 特征选择算法的改进是在原 CHI 特征选择算法的基础上再乘以因子 $A/(A+C)$, 其作用是, 特征项在计算与各类别的 $\chi^2(w,c)$ 时, 只考虑本类别出现过的特征词, 对于不在本类别文本中出现的特征词都不考虑, 即 $A=0$, 代入式(1)可得 $\chi^2(w,c)=0$. 且对于本类别文本中出现频率较小的词, 即 $A/(A+C)$ 较小, 其 $\chi^2(w,c)$ 也会比较小, 对于出现频率较大的词, 即 $A/(A+C)$ 较大, 其 $\chi^2(w,c)$ 也会比较大.

一般特征项的 CHI 值是其对所有类别的 CHI 平均值或最大值. 在改进的 CHI 特征选择算法上, 本文规定特征项的 CHI 值为其对所有类别的 CHI 最大值.

1.3 文本向量化

本文的文本表示方法选择 VSM. 将上面通过改进后的 CHI 特征选择算法筛选得到的特征项作为 VSM 的列矩阵变量. 特征向量的权重将根据 TF-IDF 算法计算. TF-IDF 的计算公式如下所示.

$$w_{ik}=tf_{ik}\times idf_{ik} \quad (2)$$

式(2)中, tf_{ik} 为给定特征项 t_k 在文本 d_i 中出现的频率, idf_{ik} 为特征项 t_k 的逆向文档频率, 其认为特征项 t_k 在文本集中出现的范围越广, 该特征项就越不重要, 计算如下所示:

$$idf_{ik}=\log\left(\frac{N}{n_k}+\alpha\right) \tag{3}$$

式(3)中，N 为文本集中的总文本数；n_k 为包含特征项 t_k 的文本总数，α 为一个可调节的参数，通常取 0.01. 因为考虑到文本长度对特征项的权重值的影响，一般会用式(4)对权重值进行归一化处理.

$$w_{ik}=tf_{ik}\times idf_{ik}=\frac{tf_{ik}\times\log\left(N/n_k+0.01\right)}{\sqrt{\sum_{j=1}^{N}\left(tf_{jk}^2\times\log\left(N/n_k+0.01\right)^2\right)}} \tag{4}$$

1.4 分类器训练

本文采用支持向量机作为模型的分类算法. 为了把支持向量机的分类功能应用在本文的多类分类问题上，需对其使用方法进行改进. 如果直接从理论上将支持向量机扩展为支持多类分类的支持向量机，该优化问题含有的变量个数为 $(M-1)\times l$ 个,当样本量 l 比较大时,这样一个大规模的规划问题不便于直接求解，模型训练时间较长，且分类精度没有优势，所以该方法应用较少. 比较简单的方法是将多类分类问题转化为多个二类分类问题，从而使用基本的支持向量机进行分类.

鉴于业务投诉文本的类别数量较大，如果采用“一对一分类”方法，模型代价高. 因此，本文采用“一类对其余类”(One-Versus-Rest)方法来解决多类分类问题. “一类对其余类”算法构造出 $M-1$ 个二值 SVM 子分类器. 其中第 i 个子分类器将 M 类中的第 i 类样本标为正类，除第 i 类以外的其它 $M-1$ 类样本标为负类. 这种方法的优势是，它只需要训练 $M-1$ 个二值子分类器,测试时对于每个待分类样本，也只需计算 $M-1$ 个决策函数值即可得到分类结果,相对于普通方法的 $(M-1)\times l$ 个变量，“一类对其余类”方法更加高效，降低了运算复杂度. 因此，其所需分类时间相对较少. 第 i 个支持向量机需要解决下面的最优超平面:

$$\begin{aligned}&\min_{w,h}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i\\&\begin{cases}y_i\left[(\omega\cdot\Phi(x_i))+b\right]\geq 1-\xi_i, i=1,2,...,l\\ \xi_i\geq 0, i=1,2,...,l\end{cases}\end{aligned} \tag{5}$$

最后，得到模型的判别函数:

$$f(x)=\operatorname{sgn}\left(\max_{k=1}^{m}\left(g^k(x)\right)\right) \tag{6}$$

对于预测样本 x 判定其属于第 k 类，其中 k 为 $g^1(x),...,g^m(x)$ 中最大者的上标.

2 模型评估

对文本分类的结果进行评估是一个复杂的问题，因为作为一个文本分类系统，影响其性能的因素有很多，如所使用的语料库，文本的表示方法等. 先规定以下几个符号的定义，如表 1 所示.

表 1 符号定义表

	实际属于该类别	实际不属于该类别
判定属于该类别	a	b
判定不属于该类别	c	d

1) 类召回率. 类别 i 的召回率:

$$m_i=\frac{a}{a+c} \tag{7}$$

2) 类准确率. 类别 i 的准确率:

$$q_i=\frac{a}{a+b} \tag{8}$$

3) 宏平均召回率，即所有类别的平均召回率. 宏平均召回率:

$$m=\frac{\sum_{i=0}^{k}m_i}{k} \tag{9}$$

k 代表类别的数目.

4) 宏平均准确率，即所有类别的评价准确率. 宏平均准确率

$$q=\frac{\sum_{i=0}^{k}q_i}{k} \tag{10}$$

召回率反映了分类模型的完备性；准确率则反映了分类模型的准确性.

3 实验

3.1 数据来源

本文实验所使用的数据来源于某省通信公司 2013 年的客户业务投诉文本，包括公司在营业厅、网厅、客户服务热线等各渠道的投诉数据. 该公司的业务投诉文本有 172 个细分小类，共匹配到 3165 条样本，平均每个类别匹配到 18 条样本. 为了测试分类模型的准确性，将各类别的样本分别随机选取 60%，共 1903 条，作为训练样本；剩下的 40%，共 1262 条，作为测试样本的一部分.

3.2 数据预处理

在数据预处理过程中，发现数据中存在以下几种

异常情况. 首先是工单记录不规范和错误. 其次是记录模板未删除. 客户服务人员在记录客户投诉内容后, 没有将记录模板删除, 导致记录模板保留在客户投诉内容里面. 而记录模板中包含有大量的业务词和特征词等, 如果不将其从客户的投诉内容中删除, 必然影响模型分类的准确性. 因此, 需要对样本进行异常值纠正、删除其中的异常内容.

在数据梳理过程中, 发现数据中存在一定数量的重复数据, 也存在一个客户在不同时间上的重复投诉. 重复数据会破坏样本的均衡性, 对特征选择, 权重计算都会产生影响, 并最终影响模型的准确性. 因此, 需要对训练样本进行去重处理, 去除训练样本中重复的数据.

3.3 模型构建

本文采用改进的 CHI 特征选择算法, 对分词结果进行特征选择. 根据式(1)计算各个特征的 CHI 值, 然后按 CHI 值由大到小对特征进行排序, 并选取 CHI 值排名靠前的 30%的特征作为最终的模型特征. 在剩下的 70%中, 如果存在业务词库中的词, 也将其添加到模型特征空间中. 最后得到 2406 个特征词, 部分结果如表 1 所示.

表 1 文本特征词

特征词	特征编码	CHI 值	类别编码
费用	1	6.124468364	504020107
多收	2	5.613387958	904021200
未生效	3	5.437153536	701020300
……	……	……	……

确定特征后, 根据向量空间模型的思想将训练样本向量化, 并根据式(4)计算特征权重, 得到的向量空间矩阵是高维稀疏矩阵. 为了节省存储空间, 以 key-value 格式存储训练样本向量. 确定模型训练的输入数据之后, 根据式(5)分别对样本各类别进行最优超平面训练, 并得到其对应的判别函数. 通过 5-flods 交叉验证方法对一定范围内的 C 和 g 进行 grid 搜索, 从而得到全局最优的参数.

4 实验结果分析及模型评估

为了评估某省通信公司业务投诉文本分类模型以及验证改进的 CHI 特征选择算法的有效性, 本文通过应用不同的特征选择算法进行对照实验. 实验组使用改进的 CHI 特征选择算法进行特征选择, 对照组则分别使用 CHI 统计和信息增益特征选择算法进行特征选择. 测试数据为 3.1 节准备的 1262 条测试样本. 该公司的业务投诉文本有 172 个细分小类, 其中 10 个类别的实验结果如图 1 所示.

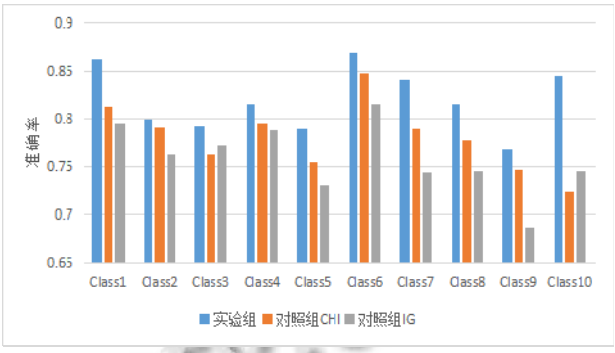

图 1 部分类别实验结果

如图 1 所示, 实验组在这 10 个类别上的准确率相比另外两个对照组均有所提升. 模型整体测试结果如表 2 所示.

表 2 模型测试结果

	宏平均准确率(%)	宏平均召回率(%)
改进 CHI 特征选择算法	83.63	78.34
CHI 特征选择算法	81.52	74.58
信息增益特征选择算法	78.12	76.36

从表 2 可以看出, 改进的 CHI 特征选择算法, 在宏平均准确率和宏平均召回率上均比未改进的 CHI 和信息增益特征选择算法好一些. 再从隶属度方面来看, 改进的 CHI 特征选择算法整体上都比未改进的 CHI 和信息增益高. 因此, 与 CHI 和信息增益特征选择算法相比, 改进的 CHI 特征选择算法选择出来的特征更具有代表性. 部分结果如表 3 所示.

表 3 隶属度样本表

工单编码	改进的 CHI	CHI	信息增益
d131590000130	0.994295721	0.920094634	0.725812577
d131590000106	0.827923572	0.761130071	0.800917099
d131590000020	1.13460531	0.725812577	0.855136045
d131590000181	0.779802956	0.705022278	0.725812577
d131590000029	1.113932235	0.689437844	0.827631678
……	……	……	……

5 结语

及时、准确地对客户投诉问题进行分类, 减少客户投诉带来的损失, 对企业具有重要意义. 本文针对 CHI 特征选择算法对低频词的倚重较大, 通过降低负相关低频词在特征选择算法中的权重, 对 CHI 特征选

择算法进行改进, 减小算法对低频词的倚重. 最后将改进后的算法用于某省通信公司业务投诉文本分类模型的特征选择, 结果表明本文构建的业务投诉文本分类模型能够取得较满意的分类效果, 提高了模型的准确度.

参考文献

1 朱颢东,钟勇.基于并行二进制免疫量子粒子群优化的特征选择方法.控制与决策,2010,25(1):53–58.

2 孟佳娜,林鸿飞,李彦鹏.基于特征贡献度的特征选择方法在文本分类中应用.大连理工大学学报, 2011,51(4):611–615.

3 Foithong S, Pinngern O, Attachoo B. Feature subset selection wrapper based on mutual information and rough sets. Expert Systems with Applications, 2012, 39(1): 574–584.

4 刘赫,张相洪,刘大有,李燕军,尹立军.一种基于最大边缘相关的特征选择方法.计算机研究与发展,2012,49(2): 354–360.

5 Song Q, Ni J, Wang G. A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans. on Knowledge and Data Engineering, 2013, 25(1): 1–14.

6 冀俊忠,吴金源,吴晨生,杜芳华.基于类别加权和方差统计的特征选择方法.北京工业大学学报,2014,40(10): 1593–1602.

7 Min F, Hu Q, Zhu W. Feature selection with test cost constraint. International Journal of Approximate Reasoning, 2014, 55(1): 167–179.

8 李湘东,曹环,黄莉.基于分布偏斜训练集的特征选择方法研究.情报理论与实践,2015,38(4):139–144.

9 段洁,胡清华,张灵均,钱宇华,李德玉.基于邻域粗糙集的多标记分类特征选择算法.计算机研究与发展,2015,52(1): 56–65.

10 吴树芳,徐建民,朱杰.基于互信息的话题特征选择方法研究.情报杂志,2015,34(4):160–164.

11 黄贤英,陈红阳,刘英涛,熊李媛.一种新的微博短文本特征词选择算法.计算机工程与科学,2015,37(9):1761–1767.

12 唐立力.基于信息熵与动态聚类的文本特征选择方法.计算机工程与应用,2015,51(19):152–157.

13 樊小超,张重阳,邓雄伟.基于互信息的文本特征加权方法.计算机工程与应用,2015,51(13):145–148.

14 刘帅,杨英杰,刘武越.一种面向聚类的加权特征选择算法.计算机应用研究,2015,32(12):3596–3599.

15 王连喜,蒋盛益.一种基于特征聚类的特征选择方法.计算机应用研究,2015,32(5):1305–1308.