一种改进的高效贝叶斯短信文本分类器

张永军,刘金岭

(淮阴工学院计算机工程学院,江苏 淮安 223003)

[**摘要**] 针对短信分类问题,提出了分类能量空间的概念,将特征词转换为分类能量空间上的一个能量元,以此为基础计算短信的能量特征向量.通过计算短信能量特征向量的领域密度,结合贝叶斯公式输出了短信在不同分类的分类概率.在分类过程中,还对分类概率差别较小的短信采用支持向量机进行了二次分类以提高分类效果.实验结果表明,该分类器模型具有良好的分类效果.

[**关键词**] 短信,文本分类,贝叶斯,支持向量机,分类能量空间

[**中图分类号**]TP181 [**文献标志码**]A [**文章编号**]1672-1292(2014)03-0070-05

An Improved Efficient Bayesian Short Message Text Classifier

Zhang Yongjun, Liu Jinling

(College of Computer Engineering, Huaiyin Institute of Technology, Huai'an 223003, China)

Abstract: A Bayesian classifier model is proposed to classify short message according to its content. The concept of category energy space is introduced and the word feature is converted to an energy unit in category energy space. Then the short message is represented as an energy vector based on its words. To obtain each category's probability, the energy vector density is calculated and brought in Bayesian probability formula. When the category probabilities are not very different, a SVM model is used to reclassify the short message. The experimental results shows that the proposed model is superior to other classification methods in the classification result.

Key words: short message, text classification, Bayesian, SVM, category energy space

随着手机用户的飞速增长,短信已经成为一种非常重要的交流手段.根据工信部的统计,2012 年,全国移动短信发送量达到 8 973.1 亿条[1].但是伴随着短信业务的发展,越来越多的垃圾信息也严重干扰到人们的正常生活.各种诈骗、虚假、色情信息通过短信进行传播,甚至部分不法分子通过短信传播谣言和反动信息,影响了社会稳定和安定团结,与建设和谐社会的总体方针相违背.从技术上建立一个有效的短信分类识别系统,有效地控制垃圾短信的泛滥是很有必要的.

国内外对短信分类的研究主要有 3 类方法[2]:(1)根据发送行为结合黑白名单机制进行判断.例如在一个时间间隔范围中发送短信频次超过某个指定阈值,则认定所发生短信为疑似垃圾短信;(2)基于关键词判断是否为垃圾短信;(3)采用文本分类技术根据短信的内容对短信进行分类.方法(1)、(2)实现简单,但分类效果较差.方法(3)采用文本分类技术实现短信分类,具有良好的效果.国内外学者在采用方法(3)进行短信分类时,主要采用的分类算法有:(1)贝叶斯分类算法[2-14];(2)支持向量机分类算法[5,15-19];(3)决策树分类算法[20].其中,贝叶斯分类算法具有较好的分类效果.

贝叶斯分类算法的 2 个假设前提是:(1)特征词之间是相互独立的;(2)不同的特征词在分类时所起的作用是等同的.有时这 2 个假设并不成立,因此会导致分类误差.该分类器还存在另外一个较为严重的问题,当某个分类中很少出现的特征词出现在待分类短信中时,会带来较大的分类误差.本文提出了一种改进的贝叶斯短信分类算法.算法考虑了特征词对不同分类主题反映能力差别,以分类空间下的领域密度来表达短信在不同分类中的概率分布,并采用支持向量机进行二次判定.

收稿日期:2013-12-11.

基金项目:国家级星火计划项目、农村民生建设信息反馈平台建设项目(2011GA690190).

通讯联系人:张永军,讲师,研究方向:中文信息处理. E-mail:13511543380@139.com

1 短信贝叶斯分类算法

设短信样本空间为 $S=\{<s_1,sc_1>,\cdots,<s_n,sc_n>\}$,其中 s_i 表示短信,$sc_i\in C=\{c_1,c_2,\cdots,c_k\}$,$c_i$ 是短信的分类标签. 定义分类子空间为 $S_i=\{<s_j,sc_j>,sc_j=c_i(1<=i<=k)\}$,用标记 $w_j\in s$ 表示短信 s 中包含有特征词 w_j,$t(w_j,s)$ 表示短信 s 中特征词 w_j 的频度,由于短信内容较短,词重复现象较少,因此可以将 $t(w_j,s)$ 定义为

$$t(w_j,s)=\begin{cases}1, & \text{如果 } w_j\in s\\ 0, & \text{其他}\end{cases}. \tag{1}$$

记特征词集合 $W=\{w_1,w_2,\cdots,w_m\}$,根据朴素贝叶斯分类器,短信 s 的分类概率可以表示为:

$$p(c_j|s)=\frac{p(c_j)p(s|c_j)}{\sum_{j=1}^{k}p(c_j)p(s|c_j)}=\frac{p(c_j)\prod_{i=1}^{m}p(w_i|c_j)^{t(w_i,s)}}{\sum_{j=1}^{k}p(c_j)\prod_{i=1}^{m}p(w_i|c_j)^{t(w_i,s)}}=\frac{p(c_j)\prod_{w_i\in s}p(w_i|c_j)}{\sum_{j=1}^{k}p(c_j)\prod_{w_i\in s}p(w_i|c_j)}. \tag{2}$$

从式(2)可以看出,若某个特征词 w_i 在某个分类 c_j 中出现频度较低,即 $p(w_i|c_j)$ 较小,将会导致 $p(c_j|s)$ 被显著降低,分类器的输出倾向于 $\bar{c}_j$.

2 基于能量强度的贝叶斯分类器

直观上,不同的特征词在短信分类过程中所起的作用是有区别的. 例如词“好礼”相比于词“现在”具有更高的分类价值. 另一方面,同一个词对不同分类的主题体现也不同.

定义 1(分类能量空间) 对于一个 k 分类问题,分类能量空间是一个 k 维度空间,每个子分类对应一个维度.

定义 2(分类能量元) 每个特征词 w 可以转换为分类能量空间中的一个特征能量元,转换后的分类能量元是分类能量空间中的一个向量,该向量的第 i 个分量体现了特征词 w 对分类 c_i 的主题体现能力. 用标记 $\boldsymbol{wen}(w)=(\boldsymbol{wen}_i(w))^{\mathrm{T}}$ 表示特征词 w 所对应的能量元.

定义 3(短信的分类能量特征向量) 短信 s 的分类能量特征向量是分类能量空间中的一个向量,该向量表示为

$$\boldsymbol{sen}(s)=\sum_{w_j\in s}\boldsymbol{wen}(w_j). \tag{3}$$

在能量空间中,短信 s 周围 c_i 分类中的样本数可以用以 s 对应的能量特征向量为中心,ε 球半径范围内所包含的 c_i 分类中样本数短信 s 表示:

$$neighbor_i(s)=\sum_{s_j\in S_i}I(dis(\boldsymbol{sen}(s_j),\boldsymbol{sen}(s))\leqslant\varepsilon), \tag{4}$$

式中 $I(x)$ 是示性函数,当 x 为真时取 1,否则取 0. ε 为距离阈值,$dis(\boldsymbol{sen}(s_j),\boldsymbol{sen}(s))$ 表示向量 $\boldsymbol{sen}(s_j)$ 与向量 $\boldsymbol{sen}(s)$ 的距离,定义为:

$$dis(\boldsymbol{sen}(s_j),\boldsymbol{sen}(s))=1-\frac{\boldsymbol{sen}(s_j)\cdot\boldsymbol{sen}(s)}{|\boldsymbol{sen}(s_j)||\boldsymbol{sen}(s)|}. \tag{5}$$

短信 s 在能量空间中的能量密度用向量 $\boldsymbol{density}(S)$ 标记,其中第 i 分量为

$$density_i(s)=\frac{neighbor_i(s)}{|S_i|}. \tag{6}$$

将式(2)中的 $p(s|c_j)$ 用式(6)代替,可得:

$$p(c_i|s)=\frac{p(c_i)density_i(s)}{\sum_i p(c_i)density_i(s)}. \tag{7}$$

3 特征词的能量元表示

特征词在分类 c_i 中的能量表示了特征词对分类主题的体现能力,常用的局部特征提取算法如信息增

益,CHI 等方法所计算得到的特征度量反映了特征词在分类中的相对重要程度,研究表明信息增益和 CHI 具有较好的效果[21-23],信息增益方法反映了特征词的全局重要程度,而 CHI 方法则可以反映在某个分类中特征词的重要程度,即局部重要程度. 因此两者的组合相比于单个方法的应用更能反应特征词的分类能量分量. 因此在本研究中将能量元能量分量定义为:

$$wen_i(w)=\mathrm{CHI}(w,c_i)*\mathrm{IG}(w), \tag{8}$$

其中

$$\mathrm{CHI}(w,c_i)=\frac{N\times(AD-CB)^2}{(A+B)\times(B+D)\times(A+B)\times(C+D)}. \tag{9}$$

$$\mathrm{IG}(w)=-\sum_i P(c_i)\log_2 P(c_i)+\alpha P(w)\sum_i P(c_i|w)\log_2 P(c_i|w)+(1-\alpha)P(\bar{w})\sum_i P(c_i|\bar{t})\log_2 P(c_i|\bar{t}), \tag{10}$$

式(9)中 N 表示样本中短信的总数,A 表示分类 c_i 中含有特征词 w 的短信数,B 表示分类 c_i 中不含有 w 的短信数,C 表示不属于分类 c_i 且含有特征词 w 的短信数,D 表示不属于分类 c 且不包含特征词 w 的短信数. 式(10)中 α 为调节参数,反映了特征词 w 出现和不出现的重要程度,实验表明,该参数取 0.2 较为合适.

4　分类过程

给定待分类短信 s,通过式(7)计算得到的分类概率表示为向量 $\boldsymbol{p}(s)=(p(c_i|s))^{\mathrm{T}}$,一种简单的分类方法是取最大 $p(c_i|s)$ 值所对应的 c_i 作为 s 的分类标签. 考虑一个 3 分类问题,若 $p(c_1|s)=80\%$,另 2 个分类概率各为 10%,显然判定 s 的分类为 c_1 较为合理. 但是,当 $p(c_1|s)=34\%$,另外 2 个分类概率各为 33%,分类器的输出如果简单地判定为 c_1 可能会导致较大误差.

本文所研究的分类器的分类过程分为 2 个阶段:(1)初步判定阶段;(2)二次判定阶段. 在初步判定阶段,将 $p(c_i|s)$ 进行从大到小排序,然后用 $gap=\max(p(c_i|s))-second\max(p(c_i|s))$ 求得最大 $p(c_i|s)$ 与次大 $p(c_i|s)$ 的差值. 如果 $gap>\theta$,则判断短信 s 属于 $\underset{c_i}{\mathrm{argmax}}p(c_i|s)$;否则进入二次判定阶段. 在二次判定阶段,考虑训练集中的样本$<s_i,sc_i>$,采用式(7)计算 s_i 的分类概率向量为 $\boldsymbol{p}(s_i)=(p(c_i|s_i))^{\mathrm{T}}$,将训练集可以转换为矩阵:

$$\boldsymbol{P}=\begin{bmatrix} p(c_1|s_1) & \cdots & p(c_k|s_1) & sc_1 \\ \vdots & \cdots & \vdots & \vdots \\ p(c_1|s_n) & \cdots & p(c_k|s_n) & sc_n \end{bmatrix}.$$

将矩阵 $\boldsymbol{P}$ 的每一行视为一个样本,sc_i 是样本的对应输出,用样本的分类标签表示. 采用支持向量机模型(线性核函数,默认参数)对矩阵 $\boldsymbol{P}$ 训练得到判别函数 $f(\boldsymbol{p}(s)):R^k->\{c_1,c_2,\cdots,c_k\}$. 给定待分类短信 s,根据式(7)计算得到分类概率向量 $\boldsymbol{p}(s)$,然后应用判别函数 $f(\boldsymbol{p}(s))$ 输出短信的分类标签.

5　实验

本文的实验环境为:神舟精盾 K790S 笔记本(CPU:2.4 GHz,内存:8 G);软件环境为:MyEclipse 2013+Weka 3.7.9 开发包. 在未做特别说明情况下,分类参数均采用 Weka 的默认参数.

5.1　实验数据及评价指标

本文采用江苏某通信公司所提供的 2 套数据进行人工标注,形成了训练集和测试集,训练集包含样本短信 20 822 条,测试集包含短信 6 642 条,分布情况如表 1 和表 2 所示.

采用全局查准率(p),查全率(r)和 F_1 作为实验的评价指标.

5.2　实验过程

实验中采用了 5 个分类学习机,将贝叶斯分类器标记为 NB,径向基网路学习机记为 RBF,决策树分类器记为 J48,支持向量机(采用线性核函数)记为 SVM,本文所研究的分类器记为 ENB. 采用 CHI 特征提取算法对特征词进行过滤,特征提取后保留 3 000 个特征词. 然后再将短信表示为 0 ~ 1 向量模型,保存为 Weka 所能识别的 arff 格式文件,再通过 Weka 平台进行实验.

在实验过程中,首先确定式(4)中参数 ε 和分类阈值 θ 的值,表3列出了对 ε 和 θ 的不同组合所产生的实验结果.

表1 训练集样本分布

Table 1 The distribution of train examples

		样本数(条)
分类标签	促销广告	4 937
	代开税票	2 061
	色情	476
	诈骗	1 490
	招工	176
	理财	368
	反动	266
	祝福问候	428
	其他	10 620
合计		20 822

表2 测试集短信分布

Table 2 The distribution of test examples

		样本数(条)
分类标签	促销广告	1 687
	代开税票	225
	色情	123
	诈骗	672
	招工	28
	理财	46
	反动	109
	祝福问候	128
	其他	3 624
合计		6 642

观察表3可知,参数 ε 对分类效果的影响最大,这是由于该参数表明了具有相似分布的短信余弦距离阈值. 当参数 ε 取0.015,分类阈值 θ 取0.3时,分类效果最佳,实验中采用该参数组合进行下一步实验.

表3 ε 和 θ 的不同组合的实验结果

Table 3 The experimental results of ε and θ's different combination

ε	θ	实验结果 F_1	ε	θ	实验结果 F_1
0.005	0.2	0.492	0.02	0.2	0.774
	0.3	0.501		0.3	0.783
	0.4	0.504		0.4	0.785
	0.5	0.482		0.5	0.779
0.01	0.2	0.701	0.025	0.2	0.541
	0.3	0.712		0.3	0.552
	0.4	0.705		0.4	0.548
	0.5	0.698		0.5	0.534
0.015*	0.2	0.863			
	0.3*	0.878			
	0.4	0.872			
	0.5	0.866			

注:* 表示分类效果最优的参数组合

5.3 实验结果与分析

表4列出了5种分类学习机的实验结果.

可以看出,本文所提出的分类学习机在 p,r 和 F_1 指标上均优于其他4个分类学习机算法,主要的原因在于,(1):式(7)弱化了贝叶斯分类的2个分类假设,并通过领域密度的方式计算短信在分类中的概率,降低了单个特征词对最终计算结果的影响;(2):分类过程中,对于通过式(7)计算得到的各分类概率差别不明显时,通过支持向量机算法进行二次判定,进一步提高了分类效果.

表4 5种分类学习机实验结果

Table 4 The experimental results of 5 classifers

分类算法	实验结果		
	p	r	F_1
NB	0.841	0.865	0.853
RBF	0.846	0.823	0.834
J48	0.839	0.857	0.848
SVM	0.853	0.861	0.857
ENB	0.874	0.882	0.878

6 结束语

本文将特征词定义为分类能量空间上的能量元,考虑了短信在分类能量空间上的能量密度分布情况,以卡方和信息增益为基础计算特征词能量元. 在能量元的基础上,考量短信在不同分类空间中的能量分布,提出了一种改进的贝叶斯短信文本分类器,实验结果表明了该分类器分类效果优于其他常用分类器. 但本文提出的算法需要计算能量密度,因此在分类速度上有所降低. 今后需要对如何提高分类速度开展更加深入的研究工作.

[参考文献](References)

[1] 新浪科技. 2012年我国短信量同比增2%人均发送量下滑[R/OL]. [2013-1-28]. http://tech.sina.com.cn/t/2013-01-28/00538020096.shtml.
Sina Tech. SMS quantity increased is 2% and per capita volume has declined in China in 2012[R/OL]. [2013-1-28]. http://tech.sina.com.cn/t/2013-01-28/00538020096.shtml. (in Chinese)

[2] 陈功平,沈明玉,王红,等. 基于内容的短信分类技术[J]. 华东理工大学学报:自然科学版,2011,37(6):770-774.
Chen Gongping, Shen Mingyu, Wang Hong. SMS classification technology based on content[J]. Journal of East China University of Science and Technology: Natural Science Edition, 2011, 37(6): 770-774. (in Chinese)

[3] 李继刚. 短信自动分类技术研究与应用[D]. 上海:东华大学计算机科学学院,2011.
Li Jigang. Study and application of SMS automatic classification[D]. Shanghai: Computer Science & Technology College,

Donghua University,2011. (in Chinese)
[4] 綦科,谢冬青. 基于内容的短信分类系统的设计与实现[J]. 广州大学学报:自然科学版,2011,10(5):43-47.
Qi Ke,Xie Dongqing. Implement of classification system of short message based on text content[J]. Journal of Guangzhou University:Natural Science Edition,2011,10(5):43-47. (in Chinese)
[5] 张兢,候旭东,吕和胜. 基于朴素贝叶斯和支持向量机的短信智能分析系统设计[J]. 重庆理工大学学报:自然科学版,2010,24(1):77-81.
Zhang Jing,Hou Xudong,Lv Heshen. Journal of chongqing university of technology[J]. Journal of Chongqing University of Technology:Natural Science Edition,2010,24(1):77-81. (in Chinese)
[6] Ganiz M C. Higher order Naïve Bayes:a novel non-IID approach to text classification[J]. IEEE Transactions on Knowledge and Data Engineering,2011,23(7):1 022-1 034.
[7] Zhang Haijun. Textual and visual content-based anti-phishing:a Bayesian approach[J]. IEEE Transactions on Neural Networks,2011,22(10):1 532-1 546.
[8] Tak-Lam Wong,Wai Lam. Learning to adapt web information extraction knowledge and discovering new attributes via a Bayesian approach[J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(4):523-536.
[9] Belem D. Content filtering for SMS systems based on Bayesian classifier and word grouping[C]//Network Operations and Management Symposium(LANOMS),Quito:IEEE Press,2011:1-7.
[10] Uysal,Alper Kursat. Detection of SMS spam messages on mobile phones[C]//Signal Processing and Communications Applications Conference(SIU),Mugla:IEEE Press,2012:1-4.
[11] Vahora S,Hasan M,Lakhani R. Novel approach:Naïve Bayes with vector space model for spam classification[C]//2011 Nirma University International Conference,Ahmedabad Gujarat:Nirma University Press,2011:1-5.
[12] Gunal S,Ergin S,Gunal E S. Detection of SMS spam messages on mobile phones[C]//2012 20th Signal Processing and Communications Applications Conference(SIU),Mugla:IEEE Press,2012:1-4.
[13] Han Kyoungsoo,Rrim Haechang,Sung Hyon Myaeng. Some effective techniques for Naive Bayes text classification[J]. IEEE Transactions on Knowledge and Data Engineering,2006,18(11):1 457-1 466.
[14] Khemapatapan C. Thai-English spam SMS filtering[C]//Communications(APCC),Auckland:IEEE Press,2010:226-230.
[15] 宋艳艳. 基于内容分类的垃圾短信拦截系统的研究[D]. 哈尔滨:哈尔滨理工大学测控技术与通信工程学院,2012.
Song Yanyan. Research on spam message interception system based on content classification[D]. Harbin:Measurement and Control Technology & Communication engineering College,Harbin University of Science and Technology,2012. (in Chinese)
[16] 李慧,叶鸿,潘学瑞,等. 基于 SVM 的垃圾短信过滤系统[J]. 计算机安全,2012,13(6):34-38.
Li Hui,Ye Hong,Pan Xuerui. Spam messages filtering system based on SVM[J]. Computer Security,2012,13(6):34-38. (in Chinese)
[17] 冯鸥鹏. 垃圾短信过滤中字特征与词特征对过滤效果的比较研究[D]. 北京:北京邮电大学计算机学院,2011.
Feng Oupeng. A comparative study of chinese character feature and word feature in SMS spam filtering[D]. Beijing:School of Computing,Beijing University of Posts and Telecommunications,2011. (in Chinese)
[18] 徐易. 基于短文本的分类算法研究[D]. 上海:上海交通大学电子信息与电气工程学院,2010.
Xu Yi. Research of text classification algorithm based on short text[D] Shanghai:Electronic Information and Electrical Engineering College,Shanghai Jiao Tong University,2010. (in Chinese)
[19] 龚垒. 基于支持向量机的垃圾短信过滤方法研究[D]. 焦作:河南理工大学计算机科学与技术学院,2011.
Gong Lei. The research of filtering methods of spam messages based on SVM[D]. Jiaozuo:Computer Science & Technology College,Henan Polytechnic University,,2011. (in Chinese)
[20] 刘庆瑜. 基于决策树分类的手机垃圾短信过滤器的设计与实现[D]. 杭州:浙江工业大学计算机科学与技术学院,2011.
Liu Qingyu. Design and implementation of mobilephone garbage SMS filters based on sorting algorithm of decision tree[D]. Hangzhou:Computer Science & Technology College,Zhejiang University of Technology,2011. (in Chinese)
[21] 熊忠阳,蒋健,张玉芳. 新的 CDF 文本分类特征提取方法[J]. 计算机应用,2009,29(7):1 755-1 757.
Xiong Zhongyang,Jiang Jian,Zhang Yufang. New feature selection approach(CDF) for text categorization[J]. Journal of Computer Applications,2009,29(7):1 755-1 757. (in Chinese)
[22] Yang Y,Pederson J O. A comparative study on feature selection in text categorization[C]//Proceedings of the 14th International Conference on Machine Learning. San Francisco:Morgan Kaufmann,1997:412-420.
[23] Forman G. An Extensive empirical study of feature selection metrics for text classification[J]. Special Issue on Variable and Feature Selection,2003,8:1 289-1 305.

[责任编辑:顾晓天]