

基于核 Fisher 判别的中文文本情感分类研究

邢玉娟,李恒杰,曹晓丽,张成文

(甘肃联合大学 电子信息工程学院,甘肃 兰州 730000)

摘要:针对文本情感分类准确率不高的问题,提出基于核 Fisher 判别的文本情感分类方法,判别文本观点是正面还是负面。首先采用向量空间模型对文档进行数据化表示,然后将不同权重计算方法和词性特征选择规则与核 Fisher 判别方法相结合来判别文档的情感观点。实验结果表明:核 Fisher 判别方法在训练的过程中使用了所有的文本特征向量而不是少数几个支持向量,因此比传统支持向量机具有较高的分类准确率,同时不同的权重特征计算方法和词性特征的选取规则对文本情感分类准确率具有较大的影响。

关键词:文本情感分类;核 Fisher 判别;支持向量机;向量空间模型;Fisher 线性判别

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2013)01-0112-03

doi:10.3969/j.issn.1673-629X.2013.01.028

Research of Text Sentiment Classification Based on Kernel Fisher Discriminant

XING Yu-juan, LI Heng-jie, CAO Xiao-li, ZHANG Cheng-wen

(School of Electronics and Information Engineering, Gansu Lianhe University, Lanzhou 730000, China)

Abstract: In view of the problem of low accuracy rate of text sentiment classification, the method of text sentiment classification based on kernel Fisher discriminant was proposed to decide that the text view is positive or negative. VSM was used to digitalize text firstly, and then classification results was achieved based on KFD combined with different weight computing method and different parts of speech feature selection rule. The experiment results showed that KFD was superior to SVM in text sentiment classification problem, and also showed that the method of weight computation and the rule of parts of speech feature selection had big affection on recognition results.

Key words: text sentiment classification; kernel Fisher discriminant; support vector machine; vector space model; Fisher linear discriminant

0 引言

随着网络信息技术的迅速发展,博客、微博、论坛受到广大网民的关注与参与,使得网络在线资源数量不断增多,而这些资源大多都是以文本的方式出现。如何对这些文本进行有效的组织和分类处理,并根据这些文本快速地判断出发表者的情感趋向,成为自然语言处理和人工智能领域的研究热点。

情感分析(Sentiment Analysis),亦称作观点挖掘(Opinion Mining),两者可以互换^[1]。情感倾向性分析就是指通过对评论的文本进行观点挖掘,对给定的文本的相关信息进行搜索,提取关键词,根据关键词采用一定的判决技术判断出文本所表达的观点(肯定、否定)。基于情感的文本分类是情感分析中一个重要

的分支,主要针对文本所表达的情感等主观内容进行分类,判断其是正面还是负面。G Salton 提出向量空间模型(Vector Space Model, VSM)^[2]对文本进行数据化表示,使得各种机器学习算法可以方便地应用于文本情感分类。Bo Pang^[3]最早将机器学习方法应用到文本情感分类中,采用朴素贝叶斯、最大熵和支持向量机(Support Vector Machine, SVM)对电影评论数据进行分析。在文献[4]中,徐军等人采用朴素贝叶斯分类方法和最大熵分类方法进行新闻内容的情感自动分类,同样获得了较好的研究成果。周杰^[5]将 SVM 应用于网络新闻评论情感分析,并将其和 KNN、RBF 网络等方法进行了比较,指出 SVM 的性能远远优于上述两种方法,同时也指出采用机器学习方法得到的分类结果普遍高于人工选择的特征。核 Fisher 判别(Kernel Fisher Discriminant, KFD)^[6]技术是基于 Fisher 线性判别提出的一种非线性分类方法,文献[7]采用 KFD 进行分类,获得了理想的实验结果,指出由于 KFD 在求解中使用了所有的训练样本而不仅仅是一

收稿日期:2012-04-22;修回日期:2012-07-27

基金项目:甘肃省教育基金项目(1113-01);甘肃联合大学科研高水平成果项目(2011GPS-01)

作者简介:邢玉娟(1981-),女,甘肃天水人,硕士生,研究方向为生物特征情感识别。

些特殊样本即支持向量,因此 KFD 的分类性能在某些方面优于 SVM,实验结果验证了这种分类方法的性能优于其他分类技术。

文中主要研究 KFD 与不同权重计算方法和不同词性特征选择方法相结合,来判断文本的情感观点。

1 文本情感向量空间模型

文本的向量空间模型(Vector Space Model, VSM)的核心思想就是将文档映射为高维空间中的一个向量,文件的每一个特征项对应向量的一维,而每一维的权值表示该对应特征项在文本中的重要程度。

假设文档 D_i , 采用 VSM 可以将其表示为: $\mathbf{d}_i = (w_{1i}, w_{2i}, \dots, w_{ni})$ 。其中 w_{ji} 表示文档 D_i 中出现词 w_j 的权重。

权重的计算^[8]主要有以下四种方法:

(1) 二值(Binary)法:如果文档中出现词 w_j 其权重为 1, 否则为 0。

$$w_{ji} = \begin{cases} 0 & tf_{ji} > 0 \\ 1 & \text{其它} \end{cases}$$

(2) 绝对词频(TF)法: $tf_{ji} = \frac{\text{freq}_{ji}}{|\mathbf{d}_i|}$, 表示 w_j 在文档 D_i 中出现的次数。

(3) 文档频率(IDF)法: $w_{ji} = \log \frac{N}{n_i}$, 其中 N 表示文档集的数目, n_i 表示文档集中词 w_j 出现的文档数。

(4) TF-IDF 法:

$$w_{ji} = tf_{ji} \times idf_j = \frac{\text{freq}_{ji}}{|\mathbf{d}_i|} \times \log \frac{N}{n_i}$$

由此可见, VSM 将文档集合映射到了高维空间的向量, 每一个向量就表示一篇文档, 向量的维数表示文档的词条数, 而点的各个维的坐标值就是词的权重 w_{ji} 。VSM 将文本数字化为向量形式, 便于各种机器学习方法进行处理, 如今引起越来越多的关注。

2 基于核 Fisher 判别的文本情感分类

基于 KFD 的文本情感分类的过程是: 首先对语料进行预处理, 包括语料的分词、词性标注, 然后根据词性选择相应的特征, 紧接着在特征集上建立向量空间模型, 生成数字化的文本特征向量, 最后由 KFD 进行二元判断, 得出正面或负面的结果。其系统框图如图 1 所示。

KFD 是 Fisher 线性判别(Fisher Linear Discriminant, FLD)的非线性扩展, 它的核心思想是通过一个

非线性映射 Φ 将原始特征空间映射到一个新的特征空间 H , 在新的特征空间 H 中使用 FLD 进行分类^[9]。

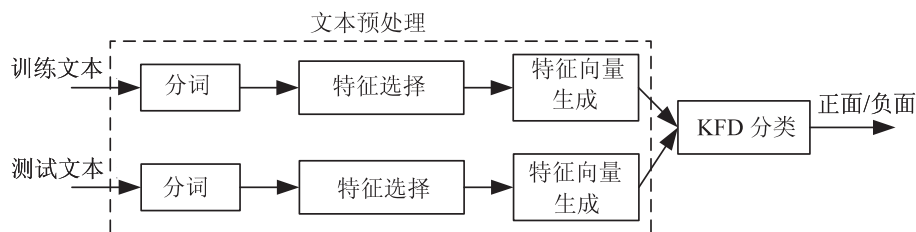
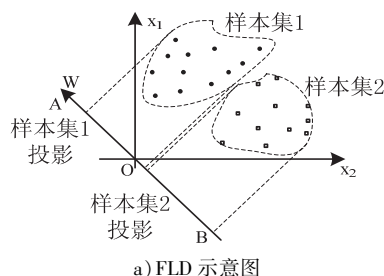
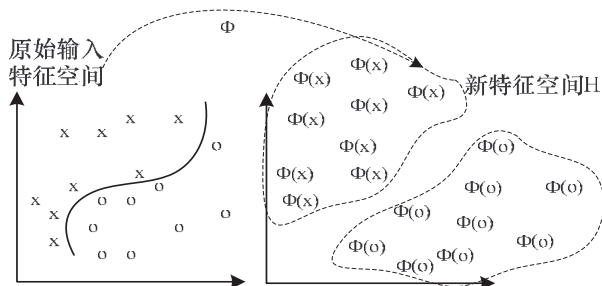


图1 基于 KFD 的文本情感分类系统框图

FLD 根据最大化类间离散度最小化类内离散度的准则, 确定原始向量的投影方向, 使各类之间最大程度的分离, 从而达到正确的分类。图 2 为 FLD 和 KFD 的原理示意图。



a) FLD 示意图



b) KFD 示意图

图2 FLD 和 KFD 示意图

假设 $X_1 = \{\text{正面样本}\}$ 样本数为 C_1 , $X_2 = \{\text{负面样本}\}$ 样本数为 C_2 , $C_1 + C_2 = R$ 。

在新特征空间 H 中目标函数为 $J(w) = \frac{\mathbf{w}^T \mathbf{S}_b^\Phi \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^\Phi \mathbf{w}}$,

其中 $\mathbf{S}_b^\Phi = (\mathbf{m}_1^\Phi - \mathbf{m}_2^\Phi)(\mathbf{m}_1^\Phi - \mathbf{m}_2^\Phi)^T$ 和 $\mathbf{S}_w^\Phi = \sum_{i=1}^2 \sum_{x \in X_i} (\Phi(x) - \mathbf{m}_i^\Phi)(\Phi(x) - \mathbf{m}_i^\Phi)^T$ 分别是样本在 H 中的类间离散度矩阵和类内离散度矩阵, $\mathbf{m}_i^\Phi = \frac{1}{C_i} \sum_{j=1}^{C_i} \Phi(x_j^i)$, $i=1, 2$ 是样本的均值向量, w 是投影方向。

根据核函数理论以及广义 Rayleigh 熵并忽略比例因子, 最大化目标函数可得 H 中任一向量 $\Phi(x)$ 在 Fisher 判定最优方向上的投影为 $\mathbf{w}^T \Phi(x) = \sum_{i=1}^N a_i K(x_i, x)$ 。通过训练一线性 SVM^[10] 得到合适的阈值 b , 从而在空间 H 中的最终分类判别函数为:

$$f(x) = \text{sgn}[\mathbf{w}^T \Phi(x) + b] = \text{sgn}[\sum_{i=1}^N a_i K(x_i, x) + b]$$

根据此判别函数进行二元判断,得出文本观点分类结果(正面/负面),文中采用径向基核函数^[11]。

3 实验结果与分析

3.1 实验数据语料预处理

实验采用中国科学院计算技术研究所谭松波博士提供的中文文本情感分析语料库。该语料库包含三种语料集:

- (1)酒店评论数据正负类各 2000 篇;
- (2)笔记本电脑评论数据正负类各 2000 篇;
- (3)书籍评论数据正负类各 2000 篇。

采用 ICTCLAS 汉语分析系统对所有的语料文本进行分词,标记词性。常见的语义倾向词的表示有:名词(N)、形容词(A)、副词(D)、动词(V)、成语(I)和习惯语(L)。如语料“外观,键盘,性能和电池都挺不错,但就是进不了系统,郁闷啊。硬盘嘎吱响,怀疑有问题。虽然楼上有人讲 F9 导入系统可以解决问题,但验货的时候可不知道,京东的售后也没提这个方法,直接换货了,还是换个新的保险”经过 ICTCLAS 分词系统处理后结果如下:

“外观/n ,/w 键盘/n ,/w 性能/n 和/c 电池/n 都/d 挺/d 不错/a ,/w 但/c 就/d 是/v 进/v 不/d 了/u 系统/n ,/w 郁闷/a 啊/y 。/w 硬盘/n 嘎/o 吱/o 响/v ,/w 怀疑/v 有/v 问题/n 。/w 虽然/c 楼上/s 有人/r 讲/v F9/x 导入/v 系统/n 可以/v 解决问题/n ,/w 但/c 验货/v 的/u 时候/n 可/c 不/d 知道/v ,/w 京东/n 的/u 售/v 后/f 也/d 没/d 提/v 这个/r 方法/n ,/w 直接/a 换/v 货/n 了/y ,/w 还/d 是/v 换/v 个/q 新/a 的/u 保险/n”

文中不考虑标点符号和助词如“的”对语料情感的影响,将其去除。在每个语料集中随机选择一半作为训练语料,另一半作为测试语料。

3.2 实验结果及分析

实验 1:不同权重计算方法分析比较。

使用 Binary、TF、IDF 和 TF-IDF 四种方法分别计算特征权重,实验结果如表 1 所示。

表 1 不同权重计算方法的分类准确率

权重计算方法	分类方法					
	KFD			SVM		
	酒店评论	笔记本电脑评论	书籍评论	酒店评论	笔记本电脑评论	书籍评论
Binary	83.34%	83.37%	81.85%	81.91%	81.64%	79.49%
TF	80.16%	81.92%	78.49%	79.43%	80.33%	78.69%
IDF	80.54%	82.13%	80.77%	80.12%	81.29%	78.81%
TF-IDF	83.09%	84.31%	82.52%	81.36%	82.57%	80.39%

在表 1 中,TF-IDF 特征权重是四种特征权重中识别准确率最高的,但是这四种特征权重的差距并不是很大,这主要是因为具有正面或负面的语义倾向的词

语只要在一句话或文章中出现,就决定了这句话或文章的语义倾向,与该词的出现次数无关,并且语料库中的语句较短,且具有明显的语义倾向,正面和负面的词语很少重复出现。在后续的实验中采用 TF-IDF 特征权重计算方法。

实验 2:不同词性特征选择规则分析比较。

形容词、名词、副词的语义倾向在很大程度上决定了文档的情感分类^[12],因此将它们的不同组合作为特征词的选择规则,实验结果如表 2 所示。

表 2 不同词性选择特征规则的分类准确率

词性选择规则	分类方法					
	KFD			SVM		
	酒店评论	笔记本电脑评论	书籍评论	酒店评论	笔记本电脑评论	书籍评论
N+D+A	82.52%	81.33%	83.44%	79.09%	80.11%	80.41%
D+A	82.27%	82.29%	83.98%	80.29%	79.54%	79.91%
A	80.01%	79.42%	78.21%	78.41%	76.19%	77.02%

从表 2 中可以看出,副词和形容词作为特征词的性能是三种特征词选择方法中最好的。“D+A”特征的维数比“N+D+A”特征的维数要小,其分类性能优于“N+D+A”特征,在书籍评论语料中采用 KFD 的分类准确率达到 83.98%。采用“D+A”特征,降低了 VSM 的维度,进而在后续的计算中可以大大地降低计算复杂度,节省时间。而“A”特征的性能最差,这主要是因为只有形容词作为特征时,特征数量太少,并且很多的形容词是名词共同出现时,才具有语义倾向,或者与不同的名词共同出现具有不同的语义倾向,这样导致分类阶段的误差。

表 1 和表 2 表明不管是不同的权重,还是不同的词性选择,KFD 的分类准确率都高于 SVM,在 TF-IDF 权重中,KFD 对笔记本电脑评论语料的分类准确率达到 84.31%,在选取“D+A”词性特征时,KFD 对书籍评论语料的分类准确率达到 83.98%。KFD 在训练的时候使用所有的特征向量,而 SVM 在训练的时候仅使用少数的支持向量,这样势必使得 KFD 的识别准确率要高于 SVM。

4 结束语

通过对文本建立向量空间模型,应用核 Fisher 判别方法进行中文文本情感分类,取得了较为理想的实验结果。对四种不同特征权重计算方法在 KFD 和 SVM 下的分类结果进行了分析比较,TF-IDF 方法获得了较优的准确率;在不同词性特征选取规则下,“D+A”特征不仅获得了较高的准确率,并且拥有较低的维度,降低了后续阶段的计算复杂度。同时,采用向量空间模型表示的文本,全部数据以矩阵形式表示,因此在

(下转第 118 页)

择关联规则是一种行之有效的办法,这种选择方式更合理、更容易。

外部环境的变化会导致保留权重和价值区分度发生相应的变化,用户可以及时提供给数据挖掘相关人员,及时的调整参数,重新生成与外部环境相适应的关联规则,用经济价值度进行排序,方便用户从经济价值的角度进行取舍。

4 结束语

基于外部环境的关联规则挖掘重新定义信任度,实现保留具有潜在价值的关联规则的目的,并且根据经济价值度对关联规则进行排序,为用户取舍关联规则提供依据,降低因环境剧烈变动导致损失的可能性。

这种挖掘方式考虑了外部环境的变化和用户的经济效益,有很大的现实意义,但也有不足之处,例如保留权重是人为设定的,容易产生误差,进而影响结果,所以对于如何更客观的设定权重,得到有价值的关联规则,还需要进一步的深入研究。

参考文献:

- [1] 谭 华,谢 赤,储慧斌. 基于模糊关联规则的股票市场交易规则抽取[J]. 系统工程,2007,25(4):92-97.
- [2] 井福荣,谢辅爱. 关联规则在网站结构优化中的改进算法[J]. 计算机系统应用,2007(1):43-46.
- [3] 李雄飞,李 军. 数据挖掘与知识发现[M]. 北京:高等教育出版社,2003.
- [4] Li G, Hamilton H J. Basic association rules[C]//Proceed-

(上接第 114 页)

向量空间模型中包含非常多的 0 值,即该向量空间模型是一种稀疏矩阵,因此在后续的研究工作中会重点关注特征类别区分能力度量以及特征数据稀疏处理的问题。

参考文献:

- [1] 赵妍妍,秦 兵,刘 挺. 文本情感分析[J]. 软件学报,2010,21(8):1834-1848.
- [2] Salton G, Wang A, Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(11):613-620.
- [3] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques [C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. [s. l.]:[s. n.], 2002:79-86.
- [4] 徐 军,丁宇新,王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报,2007,21(6):95-100.
- [5] 周 杰. 基于机器学习的网络新闻评论情感分类研究[J]. 计算机应用,2010,30(4):1011-1014.

ings of the 4th SIAM International Conference on Data Mining. Orlando, USA:[s. n.], 2004:166-177.

- [5] Bastide Y, Pasquier N, Taouil R, et al. Mining minimal non-redundant association rules using frequent closed itemsets [C]//Proceedings of the 1st International Conference on Computational Logic. Berlin, German: Springer, 2000:972-986.
- [6] Xu Yue, Li Yuefeng. Mining non-redundant association rules based on concise bases[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2007, 21(4):659-675.
- [7] Loglisci C, Malerba D. Mining multiple level non-redundant association rules through two-fold pruning of redundancies [C]//Proceedings of MLDM. [s. l.]:[s. n.], 2009:251-265.
- [8] Cheng J, Ke Y P, Ng W. Effective elimination of redundant association rules[J]. Data mining and knowledge discovery, 2008, 16(2):221-249.
- [9] 陈 茵,闪四清,刘 鲁,等. 最小冗余的无损关联规则集表述[J]. 自动化学报,2008,34(12):1-7.
- [10] 李 帆,夏士雄,张 磊,等. 基于语义划分的多层次关联规则冗余处理方法[J]. 微电子学与计算机,2010,27(8):252-260.
- [11] 韦素云,吉根林,曲维光. 关联规则的冗余删除与聚类[J]. 小型微电子计算机系统,2006,27(1):110-113.
- [12] 文 拯,梁建武,陈 英. 关联规则算法研究[J]. 计算机技术与发展,2009,19(5):56-58.
- [13] 张广路,雷景生,吴兴惠. 一种改进的 Apriori 关联规则挖掘算法[J]. 计算机技术与发展,2010,20(6):84-88.
- [6] Cevikalp H, Neamtu M, Wilkes M. Discriminative Common Vector Method with Kernels[J]. IEEE Transactions on Neural Networks, 2006, 17(6):1550-1565.
- [7] Mikat S, Fitscht G, Weston J. Fisher discriminant analysis with kernels [C]//Neural Networks for Signal Processing IX. New York:IEEE, 1999:41-48.
- [8] 咎红英,郭 明,柴玉梅,等. 新闻报道文本的情感倾向性研究[J]. 计算机工程,2010,36(15):20-22.
- [9] Deng Weihong, Hu Jiani, Guo Jun. Robust Fisher Linear Discriminant Model for Dimensionality Reduction[C]//International Conference on Pattern Recognition (ICPR 2006). [s. l.]:[s. n.], 2006:699-702.
- [10] 徐庆伶,汪西莉. 一种基于支持向量机的半监督分类方法[J]. 计算机技术与发展,2011,21(10):115-117.
- [11] Yang Shu, Yan Shuicheng, Zhang Chao. Bilinear Analysis for Kernel Selection and Nonlinear Feature Extraction[J]. IEEE Transactions on Neural Networks, 2007, 18(5):1442-1452.
- [12] 徐淑坦. 基于改进 RBF 神经网络的文本情感分类研究[D]. 长春:吉林大学,2011.

基于核 Fisher 判别的中文文本情感分类研究

作者: [邢玉娟](#), [李恒杰](#), [曹晓丽](#), [张成文](#)
作者单位: [甘肃联合大学 电子信息工程学院, 甘肃 兰州 730000](#)
刊名: [计算机技术与发展](#)
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2013(1)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjz201301030.aspx