基于RS-SVM的网络商品评论情感分析研究

王 刚 杨善林
(合肥工业大学管理学院 合肥 230009)
(过程优化与智能决策教育部重点实验室 合肥 230009)

摘 要 网络商品评论情感分析对网络购物用户的决策有着重要的帮助,因此,分类准确性的提高一直是网络商品评论情感分析研究关注的重点问题之一。近些年,集成学习理论是提高分类精度的一种有效途径,并已有研究将Bagging、Boosting引入网络商品评论的情感分析领域,但对于Random Subspace集成学习方法关注相对较少。为此,本研究根据网络商品评论情感分析问题的高维度数据特征,提出一个新的网络商品评论情感分析方法RS-SVM。该方法以集成学习中的Random Subspace为基础,选取目前在情感分析领域广泛应用的SVM作为基学习器,通过集成Random Subspace较强的学习能力,进一步提高网络用户评论情感分析的准确程度。最后,在网络商品评论情感分析经典数据库Movie Reviews上进行了实验,结果表明RS-SVM取得了比其它分类器都好的实验结果。

关键词 情感分析,商品评论,集成学习,Random Subspace,SVM

Study of Sentiment Analysis of Product Reviews in Internet Based on RS-SVM

WANG Gang YANG Shan-lin
(School of Management, Heifei University of Technology, Hefei 230009, China)
(The Ministry of Education Key Laboratory of Process Optimization and Intelligent Decision, Hefei 230009, China)

Abstract As product reviews in the internet are helpful for the decision of online shopping, and the classification accuracy of sentiment analysis is one of important problems. Recently, ensemble learning has been proved to be an effective method of enhancing the classification accuracy. Bagging and Boosting have been applied into the sentiment analysis, while Random Subspace is paid less attention to. In this paper, an new method, RS-SVM, was proposed for sentiment analysis based on the characteristic of high dimension of product review's dataset. RS-SVM uses the state-of-the-art SVM as base learner and Random Subspace as ensemble method in order to enhance the accuracy of sentiment analysis. Lastly, experiments based on movie reviews' dataset were conducted to verify the effectiveness of RS-SVM. Experimental results reveal that RS-SVM gets the best classification results compared with other methods.

Keywords Sentiment analysis, Production review, Ensemble learning, Random subspace, SVM

1 引言

2012年全球互联网数据显示,截至到2012年12月,全球网站数量总数达到6.34亿,全球网民的人数达到了24亿。互联网的普及也带动了电子商务这一网络活动的发展,并且随着近些年Web2.0的飞速发展,用户不仅仅可以浏览电子商务网站的内容,也可以方便地将自己对商品的评论发表在互联网上。网络商品评论是指由买家发布的、对产品的性能和使用情况的描述性说明,包含了个人情感和购物体验,具有观点鲜明、表达自由等特点[1]。与线下购物不同,网络购物不能检查商品的质量和品质,因此,对于网络购物而言,网络商品评论就非常重要,这些评论可以帮助顾客进一步了解商品,帮助他们做出正确的决策[1,2]。

当前,随着电子商务的不断普及,网络商品评论的数目也在急剧增长,如何借助技术手段从网络商品评论中获取用户最感兴趣的信息已成为企业界和学术界关注的热点问题,这其中的一个研究方向是对网络商品评论的情感分析研究[1,2]。对网络商品评论的情感分析是指通过分析和挖掘网络用户评论的文本中的立场、观点、情绪等主观信息,对网络用户评论的情感倾向做出判断。目前,对于网络商品评论情感分析的主要方法有基于情感知识的方法和基于数据挖掘的方法[1-3]。其中基于情感知识的方法主要依靠一些已有的情感词典和语言知识,比如SentiWordNet、General Inquire、POS tragger等,来对文本的情感倾向进行分类。这种方法主要以自然语言处理为基础,但由于目前自然语言理解领域还存在一些关键技术需要突破并且基于情感知识的情感分析需要事先构建情感知识库,因此限制了基于情感知识方法的进一步发展。因此,本研究中主要关注基于数据挖掘的情感分析方

本文受国家自然科学基金(71101042),高等学校博士学科点专项科研基金(20110111120014),中国博士后科学基金(2011M501041,2013T60611),国家重点基础研究发展计划(973计划)(2013CB329603),合肥工业大学政治理论研究中心课题(2012HGXJ0392)资助。

王 刚(1980-),男,副研究员,主要研究方向为商务智能、信息管理与信息系统等,E-mail:wgedison@hfut.edu.cn;杨善林(1948-),男,教授,主要研究方向为信息系统、决策分析。

法。基于数据挖掘的网络商品评论情感分析方法不需要事先构建情感知识库，而主要依靠数据挖掘的分类方法来对文本中的情感进行分析，主要使用的分类方法有朴素贝叶斯(Naive Bayes，NB)、支持向量机(Support Vector Machine，SVM)等[1,2]。

对于网络商品评论情感分析，已有研究表明在众多数据挖掘分类方法中，SVM取得了目前较好的分析结果。但对于网络商品评论的情感分析问题来讲，分类准确性的提高是该类研究关注的重点问题之一。近些年，在数据挖掘领域，集成学习理论是提高分类精度的一种有效途径，已在许多领域显示出其优于单个分类器的良好性能，已成为近年来数据挖掘领域一个重要的研究方向[4,5]。并已有研究将Bagging、Boosting引入网络商品评论的情感分析领域，但对于Random Subspace集成学习方法关注相对较少。为此，本研究根据网络商品评论情感分析问题的高维度数据特征，提出一个新的网络商品评论情感分析方法RS-SVM。该方法以集成学习中的Random Subspace为基础，选取目前在情感分析领域广泛应用的SVM作为基学习器，通过集成Random Subspace较强的学习能力，进一步提高网络用户评论情感分析的准确程度。最后，通过在情感分析领域的经典数据集Movie Reviews上的实验表明，RS-SVM取得了比其它分类器都好的实验结果，并且采用RBF核函数的RS-SVM取得了最好的实验结果。

2 基于RS-SVM的网络商品评论情感分析模型

2.1 SVM

SVM是20世纪90年代Vapnik等根据统计学习理论中结构风险最小化提出的一种机器学习方法，自提出就得到了广泛的重视与发展[6]。SVM算法的基本思想是：通过训练给定的训练集得到一个超平面，该超平面使得不同类别的样本之间与最大的分类相间隔。其中线性可分的支持向量原理如图1所示。

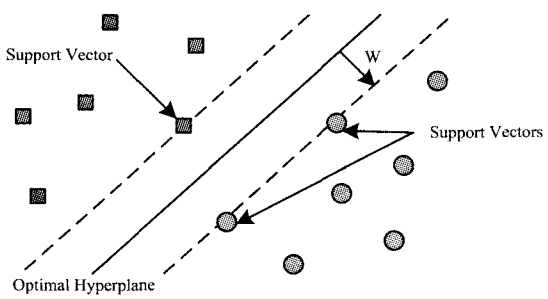

图1 线性可分的支持向量机

对于一个给定学习问题的训练集：(x_i,y_i)，$i=1,2,\cdots,n$，$x\in R^d$，$y\in\{+1,-1\}$，求解最优分类平面问题可以转换为一个有约束的二次线性规划问题：

$$\min \frac{1}{2}\|w\|^2 \text{ S.T. } y_i(w\cdot x_i+b)\geqslant 1,i=1,2,\cdots,n \quad (1)$$

根据优化理论可得最终的决策函数为：

$$f(x)=\text{sgn}(\sum_{i=1}^{n}a_i^* y_i x_i\cdot x+b^*) \quad (2)$$

SVM真正有价值的应用是在解决非线性可分的问题上，Vapnik等人将核空间的理论成功引入到非线性可分问题上，将低维输入空间的数据通过非线性映射函数ϕ映射到高位空间，从而把分类问题转化为线性可分问题。由于寻优函数和分类函数都只是涉及训练样本之间的内积运算，因此高维特征空间只需要进行内积的运算，而这种内积运算又可以通过原输入空间中对核函数$K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$的计算实现。最终得到的决策函数为：

$$f(x)=\text{sgn}(\sum_{i=1}^{n}a_i^* y_i K(x_i,x_j)+b^*) \quad (3)$$

要构造出一个具有良好性能的SVM，核函数的选择很关键，关于核函数的选择包括两部分：一是核函数类型的选择；二是确定核函数类型后相关参数的选择。满足Mercer条件的对称函数都可以作为核函数，目前，常用的核函数有以下几种[6]：

(1)线性核函数(Liner)：$K(x_i,x_j)=x_i\cdot x_j$；

(2)多项式核函数(Polynomial)：$K(x_i,x_j)=[(x_i\cdot x_j)+1]^d$；

(3)径向基核函数(RBF)：$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$。

2.2 基于RS-SVM的情感分析模型

集成学习是近年来机器学习领域的研究热点之一，它针对同一问题使用多个学习器进行学习，并使用某种规则把各个学习结果进行整合，从而获得比单个学习器更好的学习效果的一种机器学习方法，其中集成学习中的每一学习器也被称为基学习器。较早开展集成学习研究的是Dasarathy和Sheela[7]，之后Hansen和Salamon通过研究发现，通过训练多个神经网络并将其结果按照一定的规则进行组合，就能显著提高整个学习系统的泛化能力[8]。与此同时，Schapire通过构造性方法证明了可以将弱学习算法提升成为强学习算法[9]，而这个过程就是集成学习算法Boosting的雏形，并且正是由于这个定理，以及在以上早期研究的带动下，关于集成学习的研究迅速开展起来，理论和应用成果不断涌现，使其成为近十年来机器学习领域最主要的研究方向之一[4,5]。

相对于单个学习器，集成学习方法具有更强的泛化能力，可以更好地解决非均衡数据分类问题。构造集成学习的方法有很多，主要分为基于数据划分的方法(Data Partitioning Methods)和基于特征划分的方法(Attribute Partitioning Methods)。其中基于数据划分的方法通过处理训练样本产生多个样本集，分类器运行多次，每次使用一个样本集。基于数据划分的方法主要有Bagging和Boosting等[9-11]。基于特征划分的方法把输入特征划分成子集，用作不同分类器的输入向量，每次使用一个特征子集。基于特征划分的方法主要有Random Subspace等[12]，相对于基于数据划分方法，由于Random Subspace主要通过随机抽取特征子集来构建基学习器，因此其也更适合于高维度问题。

目前已有大量研究将Bagging和Boosting应用到情感分析领域，并取得比较好的结果。但对于基于特征划分的方法Random Subspace关注得比较少。对于网络商品评论情感分析问题而言，Random Subspace方法相对于Bagging和Boosting更适合。主要原因在于网络商品评论情感分析问题的一个重要特征：和文本分类问题一样，网络商品评论情感分析问题的分类特征往往成千上万，并且存在一定的噪声。高维度和噪声问题一直是困扰分类器的两个问题，相对于基于数据划分的方法，Random Subspace更能通过将分类特征划分为不同子集的方式来减轻高纬度和噪声问题。

Random Subspace是在高维度下，提高单个学习器效能及稳定性的重要方法，借由组合多个不同特征的单个学习器来增加分类信息的互补性，再透过组合的策略，能有效提高分

类精度。Random Subspace 在人脸识别、图像检索等高纬度问题上已经有了成功的应用，但目前在情感分析领域受关注的还比较少，为此，本研究提出基于 RS-SVM 的网络商品评论情感分析，即通过集成 SVM 和 Random Subspace 的优势对网络商品评论进行情感分析。首先，对网络商品评论经过分词、去停用词、词干提取等操作后得到原始语料集：$(x_1, y_1), (x_2, y_2), \cdots, (x_k, y_k)$，其中 x_i 为分类特征，y_i 为分类标签；然后，将原始语料集分为训练集 $D_{Train} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$ 和测试集 $D_{Test} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，让 RS-SVM 在训练集上学习得到 $H(x)$，然后再在测试集对得到的 $H(x)$ 进行测试。其中 RS-SVM 算法的学习过程如图 2 所示。

Input: Data set $D_{Train} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$;

Base learning algorithm SVM;

Number of random subspace rate k;

Number of learning rounds T

Process:

For t=1,2,…,T:

$D_{Train}^{t} = RS(D_{Train}, k)$; % Random generate a subspace sample from D_{Train}

$h_t = SVM(D_{Train}^{t})$; % Train a base learner h_t from the subspace sample

end

Output: $H(x) = \text{argmax}_{y \in Y} \sum_{t=1}^{T} 1(y = h_t(x))$; % the value of 1(α) is 1 if α is true and 0 otherwise

图 2　RS-SVM 算法

3　实验设计

为了验证 RS-SVM 在网络商品评论情感分析中的有效性，本文选取经典的语料库 MovieReviews 作为实验原始语料。该语料库包括 1000 个正面评论和 1000 个负面评论，分别存储在 POS 和 NEG 两个文件夹下。

实验的评价指标采用目前文本情感分类领域常用的评价指标：分类精度(Classification Accuracy)，定义如下：

$$ClassificationAccuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

本研究采用的实验环境——计算机 CPU：Intel Core 2 Duo，内存 2GB，操作系统 Microsoft Windows XP，软件 WEKA3.7.0。首先，采用向量空间表示模型表示商品评论，使用 WEKA 自带的 StringToWordVector 函数，剔除停用词，并将原始的文本语料转化为 WEKA 所识别的 ARFF 文件格式，最终获得 1165 个分类特征。然后在实验中选取了目前在情感分析领域常用的分类器 ME、DT、NB、SVM 作为基础分类器，分别选取 WEKA 下的 Logistic 模块(WEKA 下的 multinomial logistic regression 实现)、J48 模块(WEKA 下的 C4.5 实现)、NavieBayes 模块和 SMO 模块来具体实现 ME、DT、NB、SVM 算法，选取 Bagging 模块、ADBoostM1 模块和 RandomSubSpace 模块来具体实现 Bagging、Boosting 和 Random Subspace 算法，模型中各算法参数无特殊说明，均取默认值。

为了提高实验结果的可信性和有效性，实验过程使用 10 倍交叉验证法，即将初始样本集划分为 10 个近似相等的数据集，每个数据集中属于各分类的样本所占的比例与初始样本集中的比例相同，在每次实验中用其中 9 个数据集组成训练集，用剩下的 1 个数据集作为测试集，轮转一遍进行 10 次试验，因此下文的实验结果均为 10 倍交叉验证的平均值。整体实验流程如图 3 所示。

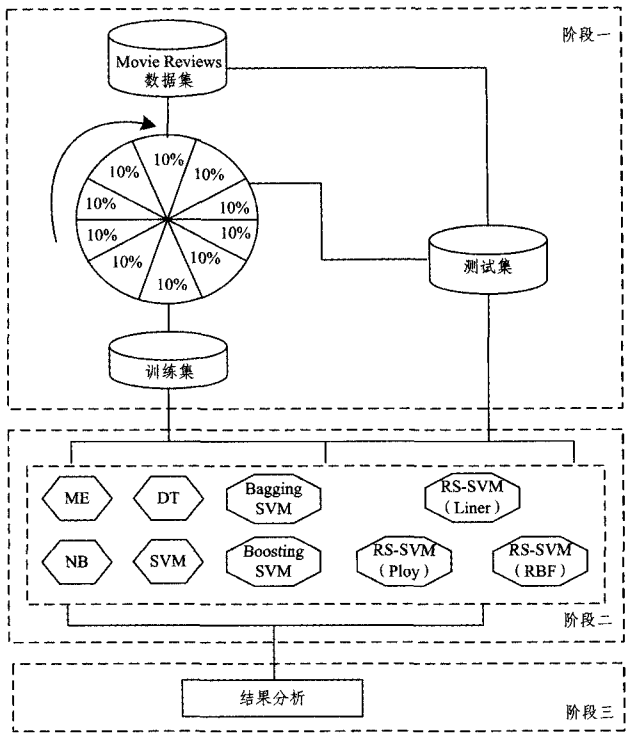

图 3　整体实验流程

4　结果分析

根据上节实验设计，得到表 1 所列的实验结果，Max，SD，Mean 分别表示实验结果的最大值、方差和均值。

表 1　不同方法在 Movie Reviews 数据集上的分类精度

分类方法	Max	SD	Mean
ME	64.00%	3.25%	55.91%
DT	73.50%	3.53%	66.37%
NB	89.00%	2.92%	80.32%
SVM	88.50%	2.80%	79.86%
Bagging SVM	88.50%	2.33%	81.53%
Boosting SVM	86.00%	2.30%	80.05%
RS-SVM (Liner Kernel)	89.50%	2.64%	82.54%
RS-SVM (Poly Kernel)	90.50%	2.86%	83.38%
RS-SVM (RBF Kernel)	89.00%	2.46%	83.85%

4.1　实验结果整体分析

如表 1 所列，比较的 9 种方法中：(1)RS-SVM 取得了较好的分类结果，其中采用 RBF 作为核函数的 RS-SVM 取得了最好的分类结果，分类精度达到 83.85%，其余两个 RS-SVM 的分类精度也分别达到 82.54%(Liner Kernel)和 83.38%(Ploy Kernel)。(2)Bagging SVM 和 Boosting SVM 也都取得了比 SVM 要好的分类精度，这个结果也说明了集成学习在情感分析中的有效性。进一步，Boosting SVM 的结果相对较差，并且通过后面的显著性统计检验(表 2)可以看出，Boosting SVM 和 SVM 的结果在统计上不存在差异，主要原因在于相对 Bagging 和 Random Subspace，Boosting 更容易受到噪声的干扰，并且情感分析的数据还存在高维度特性。(3)对于 4 个单个学习器，NB 取得最好的分类精度 80.32%，这

也和已有研究结论相一致，主要原因在于 NB 在文本分类中具有较好的分类效果，情感分析从本质上说也属于文本分类，因此 NB 取得了比较好的分类结果。

4.2 显著性检验

通过上述分析我们看到 RS-SVM 在网络商品评论情感分析中应用的有效性。为了确保以上分析不是偶然得到的，我们使用配对 t 检验对上述结果进行统计检验，表 2 为统计检验结果，其中 RS-SVM(L)、RS-SVM(P)、RS-SVM(R)分别表示采用线性核函数、多项式核函数和径向基核函数的 RS-SVM。

表 2 显著性统计检验结果

	DT	NB	SVM	Bagging SVM	Boosting SVM	RS-SVM (L)	RS-SVM (P)	RS-SVM (R)
ME	20.704**	59.758**	55.376**	67.781**	64.008**	59.639**	58.043**	69.700**
DT	—	41.648**	33.863**	41.235**	38.567**	40.821**	43.913**	41.463**
NB		—	−3.916**	0.326	−2.856**	3.624**	5.896**	6.808**
SVM			—	4.605**	1.774	10.901**	9.657**	11.916**
Bagging SVM				—	−5.647**	3.894**	5.969**	7.531**
Boosting SVM					—	6.792**	8.891**	10.488**
RS-SVM (L)						—	2.198*	3.743**
RS-SVM (P)							—	1.119

Notes: * P-values significant at alpha=0.05; ** P-values significant at alpha=0.01.

通过表 2 的显著性统计检验结果可以看出，除了相对于 RS-SVM(P)以外，RS-SVM(R)在置信度 99%以上得到了较好的分类结果，并且，RS-SVM(P)和 RS-SVM(L)相对于其它方法，也都得到统计意义上较好的分类结果。

4.3 参数分析

进一步，RS-SVM 的一个重要参数就是 Random Subspace Rate，该参数表示在构造子学习器时，随机挑选分类特征的比率，其取值直接对 RS-SVM 的分类精度产生影响，下面我们就对该参数的取值进行分析，分别取 0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9，得到图 4 所示的结果。

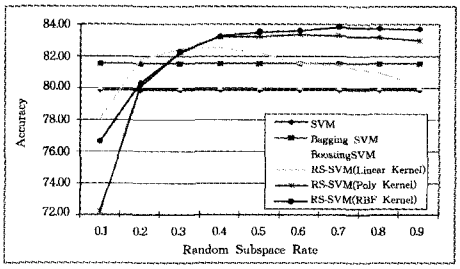

图 4 分类精度提高结果分析

如图 4 所示，RS-SVM(L)、RS-SVM(P)、RS-SVM(R)分别在 0.5，0.6，0.7 处取得了最好的分类精度 82.23%，83.38%和 83.85%，并且在 0.1 处，都取得了比 SVM 还要差的分类结果，当 Random Subspace Rate 大于等于 0.2 后，都取得了比 SVM 要好的分类结果，此结果说明 RS-SVM 的 Random Subspace Rate 参数取值范围比较广。具体而言，针对不同的核函数，RS-SVM 的分类行为不太相同，相对于采用线性核函数的 RS-SVM，采用多项式核函数和径向基核函数的 RS-SVM 结果稳定性较好，因此在实际应用中我们推荐采用多项式核函数和径向基核函数的 RS-SVM。

结束语 随着 Web2.0 的不断普及，人们可以更加容易地在互联网上发表自己的意见，这也促成了社交媒体的出现，给相对较难把握商品质量的网络购物带来了便利，用户可以通过浏览其他用户对此商品的评价，来辅助自己进行决策。

但当前对于商品的评论信息也迅速增长，如何有效地获取并使用这些评论正成为当今电子商务企业以及终端客户日益关注的问题。为此，网络商品评论情感分析技术应运而生，人工智能掘领域大量学者提出了大量模型和方法来解决以上问题。本研究根据网络商品评论情感分析问题的高维度数据特征，提出一个新的网络商品评论情感分析方法 RS-SVM，实验表明 RS-SVM 取得了比其它分类器都好的实验结果。在进一步的研究中，一方面，需要在更广泛的数据集上验证本研究的结论，另一方面，也需要对 RS-SVM 做更深入的研究。由于目前 Random Subspace 仅仅采用了随机抽取分类特征的方法，如何构造更为有效的分类特征抽取方法是未来的一个重要研究方向。

参 考 文 献

[1] 张紫琼，叶强，李一军. 互联网商品评论情感分析研究综述[J]. 管理科学学报，2010，13(006)：84-96

[2] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and trends in information retrieval, 2008, 2(1/2): 1-135

[3] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848

[4] Polikar R. Ensemble based systems in decision making [J]. Circuits and Systems Magazine, IEEE, 2006, 6(3): 21-45

[5] Dietterich T. Ensemble methods in machine learning [J]. Multiple classifier systems, 2000: 1-15

[6] Vapnik V N. The nature of statistical learning theory [M]. Springer Verlag, 2000

[7] Dasarathy B V, Sheela B V. A composite classifier system design: concepts and methodology [J]. Proceedings of the IEEE, 1979, 67(5): 708-713

[8] Hansen L K, Salamon P. Neural network ensembles [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001

[9] Schapire R E, Freund Y, Bartlett P, et al.. Boosting the margin: A new explanation for the effectiveness of voting methods [J]. The annals of statistics, 1998, 26(5): 1651-1686

[10] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140

[11] Freund Y, Schapire R E. Experiments with a new boosting algorithm [M]. Morgan Kaufmann Publishers, Inc, 1996

[12] Ho T K. The random subspace method for constructing decision forests [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844