面向非平衡文本情感分类的TSF特征选择方法

王 杰[1] 李德玉[1,2] 王素格[1,2]
(山西大学计算机与信息技术学院 太原 030006)[1]
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)[2]

摘 要 非平衡数据中样本数量的不平衡分布往往伴随着特征分布的不平衡,在多数类文本中经常出现的特征,在少数类中却很少出现。针对非平衡数据特征分布的特点,提出了一种新的双边fisher特征选择算法TSF。该方法通过显式地组合正相关和负相关特征,缓解了特征层面的非平衡性,较好地表示了文本的信息。TSF方法在图书评论和COAE2014微博非平衡数据上进行实验,结果验证了该方法是可行的。

关键词 非平衡,文本情感分类,正负相关特征,双边特征选择
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.039

TSF Feature Selection Method for Imbalanced Text Sentiment Classification

WANG Jie[1] LI De-yu[1,2] WANG Su-ge[1,2]
(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)[1]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)[2]

Abstract In the imbalanced datasets, the imbalanced distribution of the samples is often accompanied by the imbalanced distribution of features. The features, which often appear in the majority class, rarely appear in the minority class. According to the characteristics of the imbalanced feature distribution, we proposed a new two-side fisher (TSF) feature selection method. TSF can control combination of positive features and negative features explicitly and tackle the imbalanced problem in the level of feature. Experiments are conducted on the book reviews and COAE2014 imbalanced dataset. Experimental results indicate that TSF is an effective feature selection method for the imbalanced problem.

Keywords Imbalanced, Text sentiment classification, Positive and negative feature, Two-side feature selection

1 引言

随着微博、社交网络、电子商务等互联网应用的迅猛发展,人们越来越习惯于在网络上表达自己的观点和情感,网络上随之出现了大量带有情感信息的文本。这些文本大多以评论、微博的形式存在,通过这些评论文本可以获得用户对产品的情感倾向,如商家可从褒义评论中挖掘用户的关注点和产品的卖点,同时也可以从贬义评论中发现自身的缺点与竞争对手的不足[1]。传统的基于主题的文本分类系统已经无法满足对这些主观文本分析的需求。近年来,相当多的厂家、公司和贸易团体对信息的情感分类有着很强的需求,该领域的研究得到了许多研究者的重视[2]。情感分类已成为一项具有较高实用价值的技术,也是组织和管理数据的有力手段。在此背景下,情感分类作为一种面向主观文本分析的特定任务受到越来越广泛的关注[3],在自然语言处理研究领域也成为一个热点研究方向[3-6]。

根据数据的平衡性可将文本情感分成两类:平衡[4-6]文本数据和非平衡文本数据[7,8]。目前大部分情感分类研究平衡的文本数据,即两类或多类文本数据中各类文本的个数都是相等或近似相等。然而,在实际产品评论中却发现正负类别的样本数目经常相差比较大,即正负类数据的分布不平衡。这种数据的不平衡性将导致传统机器学习分类算法在分类过程中严重偏向多类样本,性能也受到很大损失,分类器小类被大类淹没[9,10]。相对平衡的文本数据处理而言,解决非平衡文本数据的分类问题将更加困难。

一个情感分类任务的成败,首先取决于所利用的特征是否较好地表示了文本的信息。特征选择方法在文本分类研究中占有非常重要的地位[11],如何设计和获取特征是情感分类任务的第一步。传统的情感分类特征选择通常基于平衡的文本,然而对于非平衡情感分类,特征选择方法研究相对较少,如何选择能够更好地刻画少数类文本的特征,是非平衡情感分类任务中亟待解决的问题。

到稿日期:2015-09-22 返修日期:2016-02-10 本文受国家自然科学基金项目(61175067,61272095,61573231,61432011,U1435212),国家"863"高技术研究发展计划基金项目(2015AA015407),山西省回国留学人员科研项目(2013-014),山西省科技基础条件平台计划项目(2015091001-0102)资助。

王 杰(1990-),男,硕士生,主要研究方向为数据挖掘;李德玉(1965-),男,教授,博士生导师,主要研究方向为人工智能等,E-mail:lidy@sxu.edu.cn(通信作者);王素格(1964-),女,教授,博士生导师,主要研究方向为自然语言处理、智能检索等。

本文主要在特征层面上，通过显式地调整正负相关特征的比例，提出了一种新的特征选择方法——TSF，从而缓解了特征的非平衡性，较好地表示了文本的信息。

2 相关工作

近年来，针对非平衡数据集的文本情感分类研究主要集中于数据层的重采样算法和特征选择方法的优化算法。

数据层重采样算法的主要思想是减少类别间样本数量的非均衡程度，包括下采样算法和上采样算法两类。下采样方法主要有简单的随机下采样[10]和启发式下采样，后者包括Kuba提出的One-sided selection算法[12]、Wang提出的BRC(Boundary Region Cutting)算法[8]等。上采样算法包括随机复制少数类样本和人工合成少数类样本的过抽样技术SMOTE(Synthetic Minority Over-sampling Technique)[13]等。也有学者将二者结合以达到类别间样本数量平衡的目的[14]。

传统的特征选择通常用于平衡的文本分类问题，Yan等[15]提出了OCFS(Optimal Orthogonal Centroid Feature Selection)特征选择方法，且在平衡语料上的效果优于信息增益(IG)和卡方统计CHI；Wang等[16]提出的fisher特征选择在两个平衡数据上证明了其有效性；代六玲等[17]考察了文档频率(DF)、IG、互信息(MI)、CHI 4种不同的特征选取方法在6类平衡数据上的表现。

在非平衡领域特征选择算法的优化方面，主要是改进传统特征选择方法或构造新的特征选择方法使其在特征选择阶段对小类特征具有公平性。早期学者Mladenic等[18]比较了数据非均衡下的各种特征选择算法，重点分析并比较了IG、文本证据权(WET)及优势率(OR)等方法，验证了在传统文本分类中表现良好的IG特征选择算法在非均衡数据集下并未表现出较好的分类结果。Chen[19]提出了FAST(Feature Assessment Sliding Thresholds)特征选择算法，在小样本分类中表现出较好的分类效果，但其需要对每个特征训练一个分类器，增加了算法的复杂度。Yin[20]提出了一种基于类别分解的特征选择方法，同时提出一种基于Hellinger距离的特征选择方法，在非平衡数据集上取得了不错的分类效果。任永功[21]通过分别计算每类特征的信息增益值，弥补了数据集不平衡时信息增益选取特征覆盖不全的缺点，提高了分类的效果。Hiroshi Ogura等[22]比较了3类不同的特征选择算法，并实验验证了第一类特征选择算法与第三类具有相同的性能。Zheng等[23]指出相关系数CC (Correlation coefficient)和OR算法为单边准则特征选择方法，而IG和CHI特征选择则为同时考虑了特征的类别正负特性的双边准则特征选择方法，前者忽略了具有一定价值的负相关特征，而后者并不能保证特征的正负相关性的最优结合，进而提出显式组合特征正负相关特性的一种特征选择算法iIG和iCHI，但仅对几种特征选择方法做了讨论，且所用的语料不是情感分类语料。本文针对非平衡文本情感分类，参考Zheng等[23]提出的组合特征算法，在fisher特征选择的基础上，提出了新的特征选择选择方法TSF。

3 TSF特征选择方法

3.1 正相关和负相关的定义

定义1(正相关) 出现在目标类文本而不出现在其他类文本的特征，称该特征与目标类正相关。

定义2(负相关) 不出现在目标类文本而出现在其他类文本的特征，称该特征与目标类负相关。

特征与类的关系如表1所列。

表1 特征与类的关系

	C_j	$\sim C_j$
t_k	A	B
$\sim t_k$	C	D

其中，A表示属于C_j类且包含t_k的文档的频数；B表示不属于C_j类，但包含t_k的文档的频数；C表示属于C_j类，但不包含t_k的文档的频数；D表示不属于C_j类且不包含t_k的文档的频数。

由表1可以得出：A和D表示t_k与C_j正相关，B和C表示t_k与C_j负相关。

3.2 fisher特征选择

fisher特征选择[16]的主要思想是选择那些使得文本类间的离散度达到最大，同时文本类内的离散度达到最小的特征。特征t_k的fisher值如式(1)所示：

$$F(t_k)=\frac{(E(t_k|P)-E(t_k|N))^2}{D(t_k|P)+D(t_k|N)} \tag{1}$$

其中，$E(t_k|P)$和$E(t_k|N)$为t_k在正负两类的均值，$D(t_k|P)$和$D(t_k|N)$为t_k在正负两类的方差。

可以根据表1将式(1)整理为式(2)。

$$F(t_k)=\frac{(A*D-B*C)^2}{(A+C)*(B+D)*(D(t_k|P)+D(t_k|N))} \tag{2}$$

3.3 TSF特征选择方法

通常情况下，使用特征函数$\mathrm{sgn}(t_k)=\mathrm{sgn}(A*D-B*C)$的值表示特征$t_k$关于$C_j$的正负相关性，$\mathrm{sgn}(t_k)$值为正表示正相关，其值为负表示负相关。然而在非平衡问题中，多数类文本的数量远多于少数类文本的数量，负相关特征很难达到和正相关一样的最大值，即$B*C$的值与$A*D$的值往往是不能比拟的，这将导致正负相关特征的不平衡性。

单边特征选择是将正相关特征和负相关特征分别计算，一般单边特征选择只选择正相关特征，但负相关特征在分类问题中可以帮助分类器排除那些无关的文本。双边特征选择是隐式选择的正相关和负相关的特征。如果是平衡数据集，双边特征选择将选择数量大致均衡的正负相关特征，但在非平衡数据中，正负相关特征的不平衡性将导致双边特征选择的选择正相关特征偏多，这对于分类将是不利的。

fisher特征选择本质上也是双边特征选择，即隐式地考虑了正相关和负相关特征。在非平衡数据中式(2)夸大了t_k的正相关性，将过多地选择一部分不能表示少数类文本的特征，而舍弃一部分能够表示多数类文本的特征，从而导致分类器性能下降。

TSF方法在式(2)的基础上添加特征函数$\mathrm{sgn}(t_k)=\mathrm{sgn}(A*D-B*C)$，得式(3)：

$$F(t_k)=\frac{\mathrm{sgn}(A*D-B*C)*(A*D-B*C)^2}{(A+C)*(B+D)*(D(t_k|P)+D(t_k|N))} \tag{3}$$

可以看出式(3)中$F(t_k)$的取值较式(2)多了负值，即变形后的式(3)为单边特征选择，正相关和负相关特征可分别计算，进而可以显式地组合正负相关特征。

3.4 基于TSF方法的特征选择过程

假设从m个候选特征中选取n个特征，将式(3)中的特

征值按降序排列，然后采用如下策略：

(1)正序选择最大的 r 个正相关特征；

(2)逆序选择最小的 $n-r$ 个负相关特征；

(3)将(1)和(2)得到的特征进行组合，其中正相关特征的比例为 $ratio=r/n$；

(4)最后在训练集上调整 $ratio$，得到最佳的 n 维特征。

4 实验设计

4.1 实验数据

本文使用了 3 个非平衡数据集：COAE 2014 评测中翡翠和保险两个领域的微博数据、当当网上的图书数据。数据的详细信息如表 2 所列。

表 2 3 个数据集的信息

数据	多数类		少数类		倾斜度	
	样本	特征	样本	特征	样本	特征
翡翠	1961	6675	283	1992	6.9	3.4
保险	1725	5614	451	3078	3.8	1.8
图书	1486	2652	424	1670	3.5	1.6

表 2 依次列出了 3 个数据中多数类和少数类的样本数和特征数，并且分别计算了二者的倾斜度。从表中可以得出 3 组数据不仅存在样本数量的不平衡，也存在特征的不平衡。

4.2 分类器和评价指标

支持向量机(SVM)[24]是建立在小样本统计学习理论的 VC 维理论和结构风险最小化原理基础上的一种新型机器学习方法。由于其优越的泛化能力，在机器学习领域得到了广泛的应用，因此本文使用 SVM 作为分类器。

本文采用少数类的召回率(RP)、总体准确率(Acc)、宏平均 F 值(F_macro)和 ROC 曲线下面积(AUC)4 种评价指标。混淆矩阵[25]如表 3 所列。我们进行了 5 折交叉验证实验来验证方法的有效性。

表 3 混淆矩阵

	被分为少数类	被分为多数类
实际为少数类	TP	FN
实际为多数类	FP	TN

$$RP(Recall)=\frac{TP}{TP+FN} \tag{4}$$

$$Acc(Accuracy)=\frac{TP+TN}{TP+FN+FP+TN} \tag{5}$$

$$F(F\_macro)=\frac{2PR}{P+R} \tag{6}$$

式(6)中 P 为宏平均准确率，R 为宏平均召回率，P 和 R 的计算公式如式(7)、式(8)所示。

$$P=\frac{(\frac{TP}{TP+FP}+\frac{TN}{TN+FN})}{2} \tag{7}$$

$$R=\frac{(\frac{TP}{TP+FN}+\frac{TN}{FP+TN})}{2} \tag{8}$$

ROC 曲线(Receiver Operating Characteristic Curve)是从医疗分析领域引入的一种新的分类模型性能评判方法，其横轴为分类器的假阳率，纵轴为真阳率。通常采用 ROC 曲线下面积 AUC(Area Under the ROC Curve)来度量分类模型的好坏，面积越大，分类器的性能越好。本文采用 scikit-learn 工具包计算 AUC 值。

4.3 实验方案

本文设计了 3 个实验来验证 TSF 特征选择方法的有效性。

实验 1 TSF 方法中 $ratio$ 对实验结果的影响，并选择各个维度下最优的 $ratio$。

实验 2 将本文提出的 TSF 与 fisher、iIG、iCHI[23]和基于 Hellinger 距离[20]的特征选择方法在各维度进行比较实验。

实验 3 使用统计检验方法，估计本文提出的方法 TSF 对于上述方法性能提升的显著性程度。

5 实验结果与分析

实验 1 TSF 在图书数据各维度下各个指标的实验结果如表 4－表 7 所列(各个特征维度的最优值通过加粗表示)。

由表 4－表 7 可以得出以下几点：

(1)只使用正相关特征($raito=100\%$)和只使用负相关特征($ratio=0\%$)，在各个维度下 4 种指标值均不能取得最优，而 TSF 方法通过调整 $ratio$ 可以弥补这两种单边特征选择的缺点。

(2)随着维度增加，各个指标最好值时的 $ratio$ 呈递减趋势，最终趋于一个比较稳定的值。

保险和翡翠的实验结果同图书类似，由于篇幅限制，保险和翡翠数据只列出 TSF 在各个维度下取得最优时的 $ratio$，如表 8 所列。

表 4 TSF 在图书数据各维度下 RP 的实验结果

特征维度	ratio										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
100	0.000	0.588	0.660	0.652	0.667	**0.683**	0.640	0.643	0.633	0.581	0.443
500	0.407	0.726	**0.738**	0.731	0.698	0.705	0.690	0.683	0.669	0.633	0.417
1000	0.448	**0.752**	0.731	0.726	0.710	0.714	0.707	0.683	0.667	0.643	0.383
1500	0.467	**0.752**	0.740	0.729	0.724	0.710	0.702	0.698	0.669	0.633	0.410
2000	0.460	**0.762**	0.743	0.719	0.729	0.705	0.700	0.688	0.657	0.638	0.450

表 5 TSF 在图书数据各维度下 Acc 的实验结果

特征维度	ratio										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
100	0.500	0.764	0.800	0.804	0.807	**0.813**	0.796	0.800	0.796	0.767	0.700
500	0.658	0.830	**0.840**	0.837	0.825	0.826	0.821	0.818	0.811	0.793	0.688
1000	0.667	**0.843**	0.838	0.837	0.829	0.835	0.830	0.819	0.804	0.793	0.663
1500	0.674	0.843	**0.850**	0.836	0.836	0.829	0.827	0.824	0.810	0.789	0.674
2000	0.664	**0.856**	0.842	0.833	0.840	0.820	0.824	0.817	0.799	0.785	0.695

表 6　TSF 在图书数据各维度下 F 值的实验结果

特征维度	ratio										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
100	0.000	0.783	0.814	0.820	0.821	**0.825**	0.813	0.817	0.814	0.788	0.733
500	0.684	0.838	**0.849**	0.846	0.837	0.837	0.833	0.830	0.825	0.810	0.725
1000	0.686	**0.849**	0.847	0.846	0.839	0.845	0.840	0.833	0.817	0.808	0.698
1500	0.692	0.850	**0.860**	0.844	0.845	0.839	0.839	0.836	0.824	0.805	0.706
2000	0.681	**0.863**	0.849	0.843	0.850	0.830	0.835	0.828	0.812	0.798	0.725

表 7　TSF 在图书数据各维度下 AUC 的实验结果

特征维度	ratio										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
100	0.777	0.899	0.915	0.924	**0.927**	0.925	0.922	0.927	0.924	0.906	0.854
500	0.759	0.938	0.937	**0.942**	0.937	0.935	0.935	0.940	0.927	0.914	0.830
1000	0.748	**0.937**	0.936	0.936	0.931	0.930	0.931	0.928	0.916	0.912	0.795
1500	0.753	**0.941**	0.937	0.930	0.932	0.927	0.924	0.917	0.911	0.900	0.767
2000	0.737	**0.942**	0.935	0.931	0.930	0.921	0.915	0.908	0.901	0.892	0.779

表 8　TSF 在保险和翡翠数据各个维度下取值最优时的 *ratio*

特征维度	保险				翡翠			
	RP(%)	Acc(%)	F(%)	AUC(%)	RP(%)	Acc(%)	F(%)	AUC(%)
100	50	50	50	50	60	60	50	40
500	20	20	50	20	20	20	20	20
1000	30	30	30	30	10	10	10	10
1500	20	20	20	20	10	10	10	10
2000	10	30	30	30	10	10	10	10

实验 2　根据 4.3 节中实验 2 的设置，各方法在各个指标下的结果如图 1－图 12 所示。

图 1－图 12 中的 TSF 在各个维度上取实验 1 得出的最优 *ratio* 时的结果，iIG 和 iCHI 各个维度的最优值选取与 TSF 方法相同。

从图 1－图 12 可以得出：

(1)TSF 在非平衡数据集上明显高于 fisher 以及文献[27]提出的基于 Hellinger 距离的特征选择方法。以图书数据特征维度 1500 为例，TSF 在 *RP* 指标下比 fisher 和 Hel 均提高约 8%，*Acc* 指标下提高 4%和 5%，F 值指标下提高 3%和 5%，*AUC* 指标下提高 2%和 3%。在保险和翡翠数据集上，TSF 和这两种方法的对比也有类似的结果。说明在非平衡数据中，显示的调整正负相关特征的 TSF 方法优于隐式的选择特征的 fisher 和 H 距离的方法。

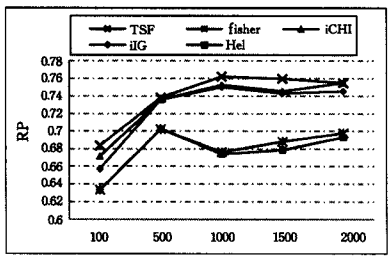

图 1　5 种特征选择方法在图书数据上 *RP* 的对比结果

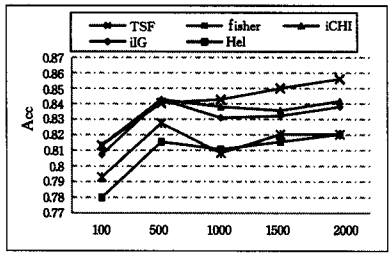

图 2　5 种特征选择方法在图书数据上 *Acc* 的对比结果

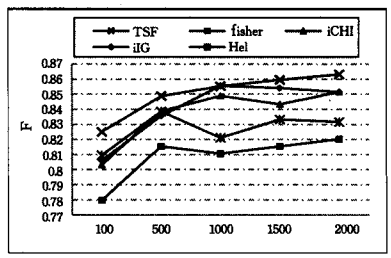

图 3　5 种特征选择方法在图书数据上 F 的对比结果

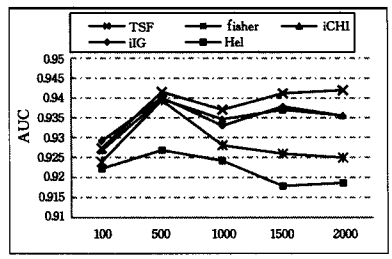

图 4　5 种特征选择方法在图书数据上 *AUC* 的对比结果

(2)TSF 方法比 iIG 和 iCHI 略有提高。在图书数据中，TSF 与 iIG 和 iCHI 方法在 *RP*、*Acc* 和 *AUC* 3 个指标下当特征维度低于 500 维时基本持平，在 500－2000 维时，TSF 性能要优于 iIG 和 iCHI；在 F 值指标下，TSF 在除 1000 维外的各个维度下均优于其他两种方法。在保险数据集中，TSF 与 iIG 和 iCHI 的分类效果在 *RP*、*Acc* 和 F 值指标下几乎相同，但在 *AUC* 指标下要略优于这两种方法。在翡翠数据中，TSF 在 *RP* 和 *Acc* 指标下，当特征维度在 1000 维以下时均优于 iIG 和 iCHI，当高于 1000 维度时，与这两种方法基本持平；在 F 值和 *AUC* 指标下 TSF 在各个维度均优于 iIG 和 iCHI。

实验 3　本文在 5 倍交叉验证的基础上采用统计检验方法，评估本文提出的方法 TSF 和其他几种方法在各个指标下是否有显著的差异。

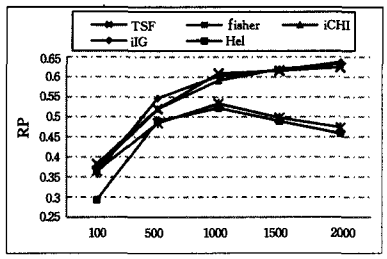

图 5　5 种特征选择方法在保险数据上 *RP* 的对比结果

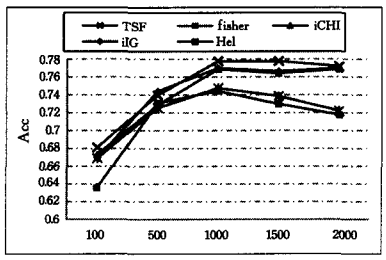

图 6　5 种特征选择方法在保险数据上 Acc 的对比结果

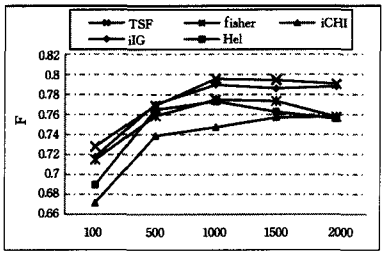

图 7　5 种特征选择方法在保险数据上 *F* 值的对比结果

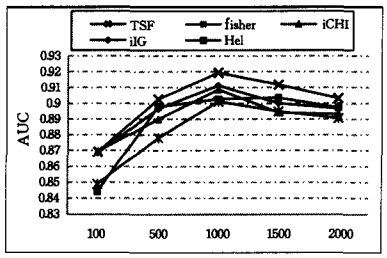

图 8　5 种特征选择方法在保险数据上 *AUC* 的对比结果

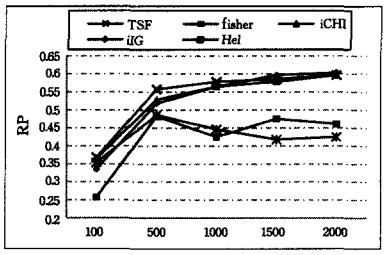

图 9　5 种特征选择方法在翡翠数据上 *RP* 的对比结果

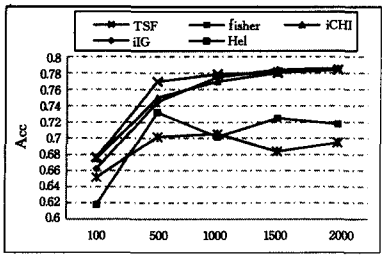

图 10　5 种特征选择方法在翡翠数据上 *Acc* 的对比结果

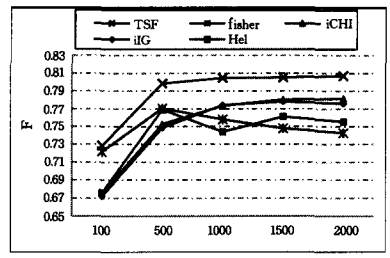

图 11　5 种特征选择方法在翡翠数据上 *F* 的对比结果

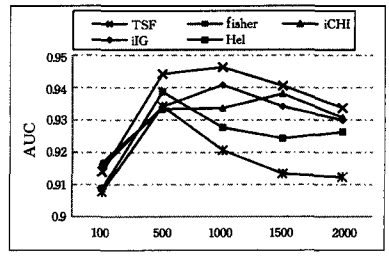

图 12　5 种特征选择方法在翡翠数据上 *AUC* 的对比结果

使用配对的 t 检验来评估方法差异的显著性。以特征维度 1000 维为例，3 种数据集下的 t 检验所得 *P* 值的平均结果如表 9 所列。

表 9　3 种数据集下 t 检验所得 *P* 值的平均结果

指标	iCHI	iIG	fisher	Hel
RP	0.1727	0.0742	**0.0087**	**0.0088**
Acc	0.1316	**0.0429**	**0.0211**	**0.0154**
F	0.0724	0.0609	**0.0393**	**0.0159**
AUC	0.1264	0.1134	0.1246	**0.0150**

从表 9 中可以看出，在 $\alpha=0.05$ 时，TSF 方法较 Hel 的性能在各个指标下均有显著提高。TSF 在除了 *AUC* 的其他指标下，较 fisher 的性能也有显著的提高。对于 iCHI 和 iIG，TSF 性能提升的显著性不明显，但其 *P* 值均在 0.10 左右。

结束语　面对非平衡情感分类任务，本文提出了一种新的特征选择方法 TSF，该方法能显式地控制正负相关特征的组合比例，从而有效地缓解了特征不平衡性。在 3 组非平衡数据集上 TSF 明显优于 fisher 以及基于 Hellinger 距离的特征选择方法，较 iIG 和 iCHI 略有优势，证明了 TSF 方法的有效性。

虽然将 TSF 用于非平衡特征选择取得了一定的效果，但有些问题仍然需要思考，例如：结合数据分布更细比例的选择组合阈值、解决 TSF 特征选择的冗余性以及将 TSF 特征选择用于解决多类非平衡问题等。

参 考 文 献

[1] Lv Yun-yun, Li Yang, Wang Su-ge. A method for chinese opinion sentence identification based on the ensemble classifier with bootstrapping[J]. Journal of Chinese Information Processing, 2013, 27(5): 84-92(in Chinese)
吕云云，李旸，王素格. 基于 BootStrapping 的集成分类器的中文观点句识别方法[J]. 中文信息学报，2013，27(5)：84-92

[2] Tang Hui-feng, Tan Song-bo, Cheng Xue-qi. Research on sentiment classification of chinese review based on supervised machine learning techniques[J]. Journal of Chinese Information Processing, 2007, 21(6): 88-94(in Chinese)

（下转第 224 页）

ness Economy,2014(3):101-102(in Chinese)
王佳乐.搜索引擎的文本聚类研究[J].商业经济,2014(3):101-102
[5] Ulrike L. A tutorial on spectral clustering [J]. Statistics and Computing,2007,17(4):395-416
[6] Jia H,Ding S,Xu X,et al. The latest research progress on spectral clustering[J]. Neural Computing and Applications,2014,24(7/8):1477-1486
[7] Ding C,He X,Zha H,et al. Spectral min-max cut for graph partitioning and data clustering[C]// Proc. of the IEEE Intl. Conf. on Data. Mining. 2001:107-114
[8] David G,Averbuch A. Spectral CAT:categorical spectral clustering of numerical and nominal data[J]. Pattern Recognition,2012:416-433
[9] Calinski T,Harabasz J. A dendrite method for cluster analysis [J]. Communications in Statistics,1974,3:1-27
[10] Lin C R,Chen M S. A robust and efficient clustering algorithm based on cohesion self-merging[C]// Proc of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2002:582-587

(上接第 210 页)

唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究[J].中文信息学报,2007,21(6):88-94
[3] Pang B,Lee L,Vaithyanathan S. Thumbs up?:sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics,2002:79-86
[4] Liu S M,Chen Jun-huan. A multi-label classification based approach for sentiment classification[J]. Expert Systems with Applications,2015,42(3):1083-1093
[5] Li Dong,Wei Fu-ru,Liu Shu-jie,et al. A statistical parsing framework for sentiment classification[J]. Computational Linguistics,2015,14(2):293-336
[6] Zhang Dong-wen,Xu Hua,Su Zeng-cai,et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications,2015,42(4):1857-1863
[7] Chawla N V,Japkowicz N,Kotcz A. Editorial:Special issue on learning from imbalanced data sets[J]. SIGKDD Explorations Newsletters,2004,6(1):1-6
[8] Wang Su-ge,Li De-yu,Zhao Li-dong,et al. Sample cutting method for imbalanced text sentiment classification based on BRC [J]. Knowledge-Based Systems,2013,37:451-461
[9] Su Jin-shu,Zhang Bo-feng,Xu Xin. Advances in machine learning based text categorization[J]. Journal of Software,2006,17(9):1848-1859(in Chinese)
苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848-1859
[10] Japkowicz N,Stephen S. The Class Imbalance Problem:A Systematic Study[J]. Intelligent Data Analysis,2002,6(5):429-449
[11] Chandrashekar G, Sahin F. A survey on feature selection methods [J]. Computers & Electrical Engineering,2014,40(1):16-28
[12] Kubat M, Matwin S. Addressing the curse of imbalanced training sets:one-sided selection[C]// Proceedings of the 14th International Conference on Machine Learning. 1997:179-186
[13] Wang B X,Japkowicz N. Imbalanced data set learning with synthetic samples[C]// Proc. IRIS Machine Learning Workshop. 2004:19
[14] Zhu Ming,Tao Xin-min. The SVM classifier for unbalanced data based on combination of RU-Undersample and SMOTE [J]. Information Technology,2012,1:39-43
[15] Yan Jun, Liu Ning, Zhang Ben-yun, et al. OCFS: optimal orthogonal centroid feature selection for text categorization[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2005:122-129
[16] Wang Su-ge,Li De-yu,Song Xiao-lei,et al. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification[J]. Expert Systems with Applications,2011,38(7):8696-8702
[17] Dai Liu-ling,Huang He-yan,Chen Zhao-xiong. A comparative study on feature selection in Chinese text categorization [J]. Journal of Chinese Information Processing,2004,18(1):26-32 (in Chinese)
代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004,18(1):26-32
[18] Mladenic D,Grobelnik M. Feature selection for unbalanced class distribution and naive bayes[C]// ICML. 1999:258-267
[19] Wasikowski M,Chen Xue-wen. Combating the small sample class imbalance problem using feature selection[J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(10):1388-1400
[20] Yin Liu-zhi,Ge Yong,Xiao Ke-li,et al. Feature selection for high-dimensional imbalanced data[J]. Neurocomputing,2013,105:3-11
[21] Ren Yong-gong,Yang Rong-jie,Yin Ming-fei,et al. Information-gain-based text feature selection method[J]. Computer Science,2012,39(11):127-130(in Chinese)
任永功,杨荣杰,尹明飞,等.基于信息增益的文本特征选择方法[J].计算机科学,2012,39(11):127-130
[22] Ogura H,Amano H,Kondo M. Comparison of metrics for feature selection in imbalanced text classification[J]. Expert Systems with Applications,2011,38(5):4978-4989
[23] Zheng Zhao-hui,Wu Xiao-yun,Srihari R. Feature selection for text categorization on imbalanced data[J]. ACM SIGKDD Explorations Newsletter,2004,6(1):80-89
[24] Fan R E,Chen P H,Lin C J. Working set selection using second order information for training support vector machines[J]. The Journal of Machine Learning Research,2005,6:1889-1918
[25] He Hai-bo,Garcia E. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Engineering,2009,21(9):1263-1284