

DOI: 10.7652/xjtuxb201810001

超平面距离的非平衡交互文本情感实例迁移方法

田锋^{1,2}, 王媛媛^{1,2}, 吴凡^{1,2}, 郑庆华^{1,2}
(1. 西安交通大学陕西省天地网技术重点实验室, 710049, 西安;
2. 西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 针对非平衡交互文本少数类实例匮乏易导致训练的情感分类模型泛化性能差的问题, 提出基于超平面距离的非平衡交互文本情感实例迁移方法。该方法将在少数类和多数类支持向量之间的源数据集实例作为待迁实例, 并基于目标数据集上的分类超平面构造一个偏移超平面。依据最优信息效用原则基于待迁实例到偏移超平面的距离最短来筛选迁入的实例, 同时通过调节迁入比例控制迁入实例规模生成合成数据集。实验结果表明: 随着迁入实例增多, 合成数据集对原始分布的偏离增大, 所训练的序列最小优化算法(SMO)模型的泛化分类性能呈现先提升后降低的现象, 类似于信息效用的 Wundt 曲线; 与 SMOTE、Subsampling、Oversampling 3 种数据层处理方法相比, 所提方法训练的 SMO、LibSVM、随机森林、代价敏感、CNN 5 个分类模型在少数类识别 F 值上平均获得 11% 的增幅, 且迁入比例最佳范围为 20%~30%, 在有效缓解非平衡特性的同时提高了少数类识别的泛化分类性能。

关键词: 实例迁移; 信息效用; 非平衡分类; 超平面

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 0253-987X(2018)10-0001-07

A Transfer Method of Emotional Instances for Unbalanced Interactive Texts Based on Hyperplane Distance

TIAN Feng^{1,2}, WANG Yuanyuan^{1,2}, WU Fan^{1,2}, ZHENG Qinghua^{1,2}
(1. Shannxi Key Laboratory of Satellite and Terrestrial Network Technology Research and Development, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A transfer method of emotional instances for unbalanced interactive texts is proposed based on hyperplane distance to focus the problem of poor generalization ability of sentiment classification models when they are trained on an unbalanced interactive text dataset that lacks of minority-class instances. The method uses instances of source dataset between support vectors of the minority class and the majority class as the transferrable instances, and constructs an offset hyperplane based on the classification hyperplane on the target dataset. The method uses the principle of optimal information utility to select the transfer instances based on the shortest distance between the instances and the offset hyperplane, and adopts the migration ratio to control the size of the transfer instances and to generate a synthetic dataset. Experiment results show that when transfer instances increase, the deviation of the synthetic dataset from the

original distribution increases, and the generalized classification performance of the trained SMO model rises at the beginning and then decreases after it reaches its maximum, which is similar to the Wundt curve of the information utility. Comparisons with three data layer processing methods (SMOTE, Subsampling and Oversampling) show that five classification models (SMO, LibSVM, random forest, cost sensitive and CNN) trained by the proposed method obtain an average increase of 11% in the F -value of recognizing the minority class, and the optimal range of the migration ratio is [20%, 30%]. It is concluded that the proposed method effectively alleviates the unbalanced characteristics and raises the generalized classification performance of the minority class.

Keywords: instance transfer; information utility; unbalance classification; hyperplane

交互文本是网络用户的重要交流方式之一,如QQ和微博。面向交互文本的情感分析技术在E-learning、舆情分析、人机对话等领域具有重要意义。现实中,非平衡特性在交互文本中普遍存在,即一类情感类别的样本数远远大于其他情感类别的样本数,导致在其上训练的分类模型忽视少数类信息,产生过拟合问题。同时,网络上的交互文本更新速度快,用原始数据集构建的模型对新生数据的分类效果比较差,即形成分类模型泛化性能差的问题。如何提高非平衡情感分类模型的泛化分类性能是当前研究的难点之一。

在处理非平衡问题时,已有的很多方法都是在单一训练数据集上变化,近年来,迁移学习的概念被引入到非平衡分类中^[1-2],此类方法基于目标数据集与源数据集实例间的相似度进行源数据集中迁移实例筛选,进而合成新的数据集并在其上训练出情感分类模型。从源数据集向目标数据集迁移高质量的实例极其重要,但是基于实例迁移的方法在解决非平衡问题时很少考虑泛化问题。

受SVM^[3]方法中超平面构造的思路启发,本文提出一种基于实例同超平面距离的源数据集可迁移实例筛选算法,以解决如何有效地向目标数据集引入新信息的问题。在构造的合成数据集上训练出泛化性能优的分类模型,克服泛化性能弱的难点。

1 相关工作

在现实中非平衡问题广泛存在,目前解决非平衡分类问题主要采用从数据集层面上和算法层面上处理两类方法。数据集层面的方法通过改变训练集的规模与特征分布,降低不平衡度来提高分类性能,典型方法有Oversampling、Subsampling、SMOTE等;算法层面的方法有集成方法、代价敏感学习方法、特征选择方法和单类学习方法等^[4]。然而,上述

方法都是在单一训练数据集上进行构造,信息的更新度不够,在新数据集上分类性能差。

最近,因迁移学习^[5]从一个或多个源数据集中提取知识,给训练数据集带来了新的信息,所以迁移学习应用于文本情感分类逐渐成为研究热点。按照被识别对象的粒度可分为句子级上、段落级上和文档级上的文本情感分类研究。

在句子级上,文本情感分类在社交网络中应用最为广泛^[6]。Yong等提出基于word2vec情感词语义相似性的迁移方式^[7];Wu等提出在迁移之前使用两阶段采样方式,根据共现特征和目标数据集特征筛选合适的数据^[8];田锋等针对多领域类分布不平衡问题提出分领域实例选取与迁移方法,采用目标数据集单一实例最相似 N 个实例无差别迁移的策略^[9]。

在段落级上,文本情感分类在商品、电影评论中应用广泛。Zhang等提出一种新的基于非负矩阵三因子分解的迁移学习,通过链接相似特征簇进行情感分类^[10]。

在文档级上,Li等通过源领域的学习将文档类别信息迁移到词特征上,然后把情感分类信息通过词特征迁移到目标领域的文档^[11];庄福振在此基础上将词特征改进为词特征聚类进行迁移学习^[12]。

虽然上述迁移学习文本情感分类涉及了非平衡分类,改善了目标数据集的特征分布,但是没有考虑交互文本内容与情感随着时间而不断变化、不同时段内数据特征分布差异较大的现象,所训练模型并不能满足对未来数据的分类需求、仍易过拟合,并未缓解泛化问题。

针对此问题,本文拟结合SVM分类超平面特点引入新信息的效用理论来指导筛选高质量迁移实例,提升交互文本情感分类的泛化性能。

2 超平面距离实例迁移方法

2.1 超平面距离实例迁移方法的思想

心理学家威廉·麦克斯·华德特提出的信息效用理论反映了人们对于事物中所含信息的心理反应强度,这种新颖程度(新奇度)只有在一定的范围内才会对人和动物产生激励作用^[13]。这种事物的新奇度与人们心理反应之间的关系可用一条倒 U 型的曲线来表示,即被认为心理学第一定律的 Wundt 曲线,如图 1 所示。本文将信息效用理论引入文本情感分类,利用文本的新奇度与最优信息效用筛选高质量迁移实例。本文假设文本的新奇度和最优信息效用是由源数据集蕴含的知识、目标数据集知识和分类超平面距离 3 个因素决定的。

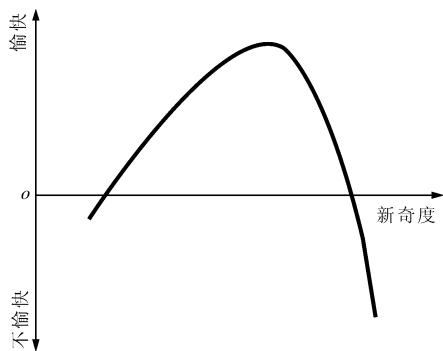


图 1 Wundt 曲线

由于交互文本特征维度高,使用分布之间的差异相加(如 KL 距离)来衡量源数据集中待迁移实例的新奇度不准。本文采用源数据集实例与分类超平面距离的远近作为新奇度的一个维度,即距离分类超平面越远的实例蕴含的旧信息越多,新奇度越低,无益于提升分类性能,向分类超平面附近添加适量的少数类,更有益于提升分类性能。

最优信息效用表示迁入多少源数据集实例时能够达到最佳分类性能。当迁移实例过多,引入过多新奇度低或噪声实例,将使数据集偏离原始分布太远,产生误导作用,使分类器分类性能下降。相反,加入适量新奇度高的实例,能明显提升分类性能。迁移实例比例是决定文本新奇度的另一个维度。

本文受 SVM 分类器超平面构造方法的启发,提出将位于少数类和多数类支持向量之间的源数据集实例作为待迁实例,并基于目标数据集上的分类超平面构造一个偏移超平面。基于待迁实例到偏移超平面的距离最短来筛选迁入的实例,同时采用迁入比例控制迁入实例规模生成合成数据集。该方法旨在通过向非平衡目标数据集加入与偏移超平面一

定距离的源数据集实例,促使超平面向多数类实例方向偏移,减少错分的少数类实例个数。同时,源数据集实例的加入增加了目标数据集少数类信息,可缓解非平衡交互文本特征空间稀疏的问题,改变目标数据集的特征空间分布,提高目标数据集泛化分类性能。图 2 是基于实例与超平面距离的迁移实例筛选示意图。

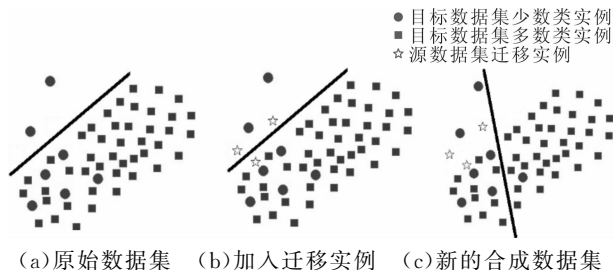


图 2 基于超平面距离的迁移实例筛选示意图

2.2 超平面距离实例迁移方法描述

本文提出的迁移方法首先将目标数据集和源数据集实例合并提取特征,包括语法特征、统计特征、uni-gram 和 bi-gram 词频特征(基于中文情感词汇本体构建词典),使源数据集实例可以直接迁移到目标数据集中,完成预处理;然后,使用基于超平面距离的迁移实例筛选算法在源数据集上筛选满足条件的实例加入目标数据集构成合成数据集。

基于实例与超平面距离的迁移实例筛选方法的核心是如何利用新奇度来迁移实例。本文中新奇度考虑实例与超平面的距离和迁入实例比例两个变量。由于加入位于超平面附近的实例对超平面的影响比较明显,故本文设置位于支持向量之间的实例为待迁实例。首先,将由目标数据训练得到的线性核 SVM 分类器命名为 C,通过间隔最大化学习得到的分类超平面为

$$w^* x + b^* = 0 \quad (1)$$

相应的分类决策函数为

$$f(x) = \text{sign}(w^* x + b^*) \quad (2)$$

假设将支持向量与 SVM 分类超平面的距离视为 1,多数类支持向量所在的超平面距离分类超平面的距离表示为 1,少数类表示为 -1。使用训练好的分类器 C 对源数据集中的实例进行类别预测,计算源数据集中每个实例 (x_i, y_i) 与从分类器 C 构造的偏移超平面的距离公式如下

$$d_i = w^* x_i + b^* - \beta \quad (3)$$

式中: β 是控制偏移超平面位置的距离阈值,取值集合为 $\{-1, -0.5, 0, 0.5, 1\}$ 。阈值取 -1 时选取的偏移超平面是少数类支持向量所在的超平面,阈值

取1时选取的偏移超平面是多数类支持向量所在的超平面。

由于迁移实例个数在不同规模数据集上不同,本文以迁移实例占目标数据集少数类的比例为度量,记为迁入比例 γ ,根据迁入比例计算迁移实例数 M 。将计算出的源数据集中每个实例与偏移超平面的距离进行从小到大的排序,选择前 M 个实例加入目标数据集构成合成数据集。

3 基于超平面距离的实例迁移实验

3.1 实验数据集

本实验以二分类问题(正负面情感)为研究对象,采用的数据集为微博数据集、BBS论坛数据和QQ数据集。微博数据集来自新浪微博网友互动,BBS论坛数据来自某校内论坛用户互动,均是涵盖多种类型话题的交互短文本。QQ数据集来自腾讯QQ上某学习小组的交互短文本。

实验共两组数据集,第一组的目标数据集来自微博短文本,命名为weibo,共5 065条,非平衡比(少数类实例个数:多数类实例个数)为1:4,稀疏度约为 1.6974×10^{-4} ,记为 D_{weibo} 。源数据集来自BBS论坛数据中的3 137条负面情感实例,命名为BBS。第二组的目标数据集来自QQ数据集,命名为QQ,共4 778条,非平衡比为1:4,稀疏度约为 9.4258×10^{-5} ,记为 D_{QQ} 。源数据集来自微博数据集中学习主题的4 268条负面情感实例,命名为w_study。两组的测试数据集分别为和两个目标数据集不同时段的数据集,各为1 045条和959条实例。Syn_W数据集表示第一组实验中的weibo数据集和BBS数据集构成的合成数据集,非平衡比为1:3.3,记为 $D_{\text{syn_w}}$ 。Syn_Q数据集表示QQ数据集和w_study数据集构成的合成数据集,非平衡比为1:3.3,记为 $D_{\text{syn_q}}$ 。

3.2 实验步骤

合成数据集生成时,在迁移实例中为了确定最优信息效用,设置不同的参数执行实验实现对目标数据集引入新信息的比例控制策略。该策略首先采用目标数据集训练分类器,每次筛选一定比例的实例加入目标数据集形成合成数据集。在每个合成数据集上训练分类模型,比较不同迁入比例下的泛化分类性能,确定最优信息效用。设置方法中的一些参数为:迁入比例参数以10%为步长,从10%递增至300%,使数据集达到平衡。实例与超平面的距离 d 分别取-1.0,-0.5,0.0,0.5,1.0。

对比数据集及实验方法:使用SMOTE^[14]、Subsampling^[4]、Oversampling^[4]方法分别生成的3类数据集,其非平衡比例均为1:1,分别记为 D_{SMOTE} 、 D_{Sub} 、 D_{Over} 。在目标数据集以及各个合成数据集上使用SMO、LibSVM、RandomForest(RF)、代价敏感的SMO(Cost+SMO)、代价敏感的LibSVM(Cost+LibSVM)和代价敏感的RandomForest(Cost+RF)、CNN^[15]共7种分类方法进行泛化性能测试实验。

3.3 性能评价指标与实验结果

本文选择加权受试者工作曲线下的面积(加权ROC)、少数类(负面情感类) F 值和多数类(正面情感类) F 值为非平衡交互文本情感分类的评价指标。加权ROC准确反映了真正率和假正率之间的关系,设为 R ,是非平衡问题分类准确性的综合代表。 F 值是情感分类准确率和召回率的综合评价指标。

按如上步骤的实验结果如图3~4、表1~6所示。

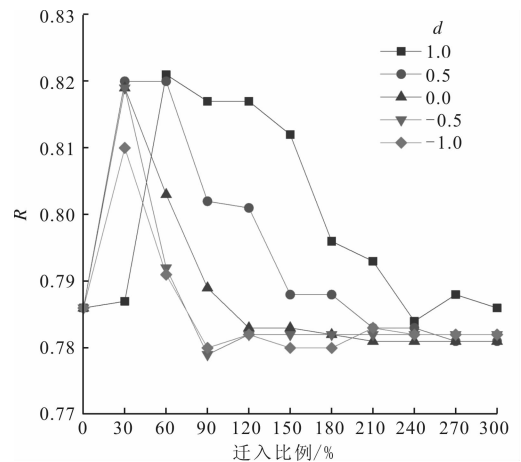


图3 不同距离参数下泛化分类性能 R 随迁入比例的变化

图3是不同距离参数下随着迁入比例增加,Syn_W数据集在SMO分类方法下泛化分类性能加权ROC值的变化图。随着迁入比例越来越大,所训练分类模型的 R 呈现出先快速增长、后缓慢下降的现象。图3证实迁移文本带来的新信息在一定程度上才能对非平衡情感分类产生正面作用,比较符合信息效用理论。

图4是在不同距离参数下随着迁入比例变化时的分类超平面图。采用余弦计算超平面法向量 $\mathbf{w}$ 与初始超平面法向量 $\mathbf{w}_0$ 的夹角 θ ,用夹角的变化衡量超平面变化。由图4可以看出,当迁入比例达到一定值时,超平面均不再改变,形成稳定状态。该现

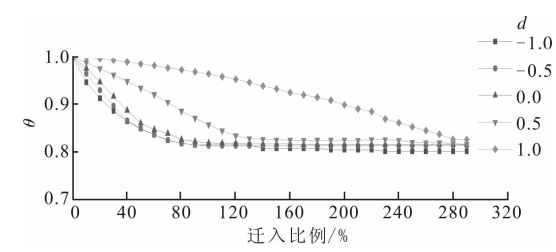


图 4 不同距离参数下的分类超平面图

象是由于源数据集中满足距离约束的迁移实例个数有限,即新奇度高的实例有限。迁入比例较小时,其恰好能全部迁移进入目标数据集,迁入比例过大时,新奇度低的实例对超平面的影响越来越小。

在 weibo 和 QQ 数据集上分别使用本文方法获得最佳合成数据集与使用 SMOTE、Subsampling、Oversampling 方法处理获得平衡数据集,5 种分类模型进行泛化分类性能测试的 R 、少数类 F 值和多数类 F 值结果见表 1~表 6。由表 1 和表 4 的黑体数值可以看出,本文方法在代价敏感的 RandomForest 上的分类结果与对比方法中最好的 SMOTE

+RandomForest 方法相比, R 有 2% 的微弱下降,但是由表 2 和表 5 的黑体数值可以看出,本文方法在少数类 F 值上均有显著上升,平均增幅为 11%。在 QQ 数据集上,本文方法在代价敏感的 RandomForest 上与 SMOTE 方法对比,将少数类 F 值提升 21%。由表 3 和表 6 的黑体数值可以看出,与目标数据集相比,在所有对照方法上多数类 F 值均有明显下降,但是本文方法中迁移实例使多数类 F 值略微下降,很少会出现微弱上升的现象,可能与新信息的引入使分离间隔更准确有关。由于训练数据集的规模不够大,同时存在非平衡问题,使用深度学习 CNN 方法进行分类并未表现出很好的泛化分类性能。从整体上来看,由于两个目标数据集特征稀疏度不同,在 weibo 数据集上使用 SMOTE、Subsampling 和 Oversampling 方法使数据集达到平衡后少数类 F 值出现下降的现象,这是将更多多数类错分为少数类导致的过拟合问题,而本文方法并未出现过拟合,分类结果较好。

表 1 weibo 数据集上不同方法 R 的比较

数据集	R						
	SMO 方法	LibSVM 方法	RF 方法	Cost+SMO 方法	Cost+LibSVM 方法	Cost+RF 方法	CNN 方法
D_{weibo}	0.783	0.823	0.884	0.786	0.812	0.883	0.765
D_{SMOTE}	0.791	0.794	0.890				
D_{Sub}	0.757	0.771	0.868				0.768
D_{Over}	0.758	0.780	0.867				0.762
$D_{\text{syn_w}}$				0.809	0.816	0.875	0.780

表 2 weibo 数据集上不同方法的少数类 F 值的比较

数据集	少数类 F 值					
	SMO 方法	LibSVM 方法	RF 方法	Cost+SMO 方法	Cost+LibSVM 方法	Cost+RF 方法
D_{weibo}	0.670	0.669	0.568	0.558	0.605	0.621
D_{SMOTE}	0.570	0.570	0.608			
D_{Sub}	0.513	0.531	0.554			
D_{Over}	0.521	0.549	0.608			
$D_{\text{syn_w}}$				0.676	0.687	0.614

表 3 weibo 数据集上不同方法的多数类 F 值的比较

数据集	多数类 F 值					
	SMO 方法	LibSVM 方法	RF 方法	Cost+SMO 方法	Cost+LibSVM 方法	Cost+RF 方法
D_{weibo}	0.925	0.896	0.903	0.780	0.832	0.868
D_{SMOTE}	0.784	0.792	0.900			
D_{Sub}	0.694	0.727	0.779			
D_{Over}	0.731	0.766	0.849			
$D_{\text{syn_w}}$				0.913	0.914	0.853

表 4 QQ 数据集不同方法 R 的比较

数据集	R						
	SMO 方法	LibSVM 方法	RF 方法	Cost+SMO 方法	Cost+LibSVM 方法	Cost+RF 方法	CNN 方法
D_{QQ}	0.550	0.571	0.890	0.724	0.722	0.865	0.702
D_{SMOTE}	0.702	0.714	0.892				
D_{Sub}	0.726	0.742	0.781				0.735
D_{Over}	0.717	0.710	0.841				0.721
D_{syn_q}				0.735	0.723	0.872	0.725

表 5 QQ 数据集不同方法的少数类 F 值的比较

数据集	少数类 F 值					
	SMO 方法	LibSVM 方法	RF 方法	Cost+SMO 方法	Cost+LibSVM 方法	Cost+RF 方法
D_{QQ}	0.194	0.263	0.175	0.503	0.523	0.520
D_{SMOTE}	0.491	0.511	0.332			
D_{Sub}	0.500	0.532	0.504			
D_{Over}	0.526	0.516	0.415			
D_{syn_q}				0.529	0.511	0.549

表 6 QQ 数据集不同方法的多数类 F 值的比较

数据集	多数类 F 值					
	SMO 方法	LibSVM 方法	RF 方法	Cost+SMO 方法	Cost+LibSVM 方法	Cost+RF 方法
D_{QQ}	0.892	0.893	0.875	0.769	0.753	0.894
D_{SMOTE}	0.801	0.825	0.884			
D_{Sub}	0.747	0.807	0.794			
D_{Over}	0.850	0.847	0.875			
D_{syn_q}				0.815	0.796	0.900

根据上文信息效用理论中的新奇度和最优信息效用进行迁移,在非平衡比较小的数据集上,迁移位于非平衡超平面与少数类支持向量所在平面之间的实例,当迁入比例为 20%~30%左右时,可获得泛化分类性能优的模型。

4 结 论

针对交互文本存在非平衡、特征维度高、特征值稀疏问题,导致高性能分类模型难构建,本文提出基于超平面距离的非平衡交互文本情感实例迁移方法。在信息的信息效用理论启发下,迁移富含新信息的少数类实例缓解非平衡特性,利用迁入比例控制策略改变特征分布、缓解特征值稀疏问题,获得了泛化分类性能优的模型。与 Subsampling、Oversampling 等基于目标数据集的采样方法不同,本文方法在一定程度上缓解了这些方法中存在的训练模型过拟合问题。通过两组大量的对比实验证明:与

SMOTE、Subsampling、Oversampling 3 种数据层处理方法和 SMO、LibSVM、随机森林、代价敏感、CNN 等分类方法相比,本文方法所训练的分类模型在加权 ROC 指标微弱降低的代价下,可获得少数类识别 F 值上的较大增幅。

本文的研究思路不仅适用于 SVM 分类,同样可用于其他分类方法,具有可拓展性。但是,本文仅用实验结果验证了理论方法的有效性和可信性,后续将结合随机矩阵理论,进一步探索该方法的机理,以及最优合成数据集的参数估计方法,并推广应用。

参考文献:

[1] TIAN Feng, WU Fan, CHAO Kuo-Ming, et al. A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews [J]. Electronic Commerce Research and Applications, 2016, 16(3): 66-76.

- [2] AL-STOUHI S, REDDY C K. Transfer learning for class imbalance problems with inadequate data [J]. Knowledge & Information Systems, 2016, 48 (1): 201-228.
- [3] CORTES C, VAPNIK V. Support-vector network [J]. Machine Learning, 1995, 20(3): 273-297.
- [4] HE H, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.
- [5] PAN S J, YANG Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(10): 1345-1359.
- [6] BAGHERI H, ISLAM M J. Sentiment analysis of Twitter data [EB/OL]. (2017-12-16) [2018-01-05]. <https://cn.arxiv.org/ftp/arxiv/papers/1711/1711.10377.pdf>.
- [7] YONG R, WANG C, HE X. A transfer learning based boosting model for emotion analysis [C]//Proceedings of the IEEE International Conference on Big Knowledge. Piscataway, NJ, USA: IEEE, 2017: 264-269.
- [8] WU H, JIN Q. Improving emotion classification on Chinese microblog texts with auxiliary cross-domain data [C]//Proceedings of the International Conference on Affective Computing and Intelligent Interaction. Piscataway, NJ, USA: IEEE, 2015: 821-826.
- [9] 田锋, 兰田, CHAO Kuo-Ming, 等. 领域实例迁移的交互文本非平衡情感分类方法 [J]. 西安交通大学学报, 2015, 49(4): 67-72.
TIAN Feng, LAN Tian, CHAO Kuo-Ming, et al. A unbalanced motion classification method for interactive texts based on multiple-domain instance transfer [J]. Journal of Xi'an Jiaotong University, 2015, 49(4): 67-72.
- [10] ZHANG W, ZHANG H, WANG D, et al. Transfer learning by linking similar feature clusters for sentiment classification [C]//Proceedings of the IEEE International Conference on Tools with Artificial Intelligence. Piscataway, NJ, USA: IEEE, 2017: 1019-1026.
- [11] LI Tao, SINDHWANI V, DING C, et al. Knowledge transformation for cross-domain sentiment classification [C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2009: 716-717.
- [12] 庄福振. 迁移学习中文本分类算法研究 [D]. 北京: 中国科学院大学, 2011: 31-46.
- [13] 戴昌钧. 信息效用函数与 Wundt 曲线 [J]. 高校应用数学学报, 1991(2): 241-252.
- [14] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.
- [15] KIM Y. Convolutional neural networks for sentence classification [EB/OL]. (2014-09-03) [2018-01-05]. <https://arxiv.org/abs/1408.5882>.

[本刊相关文献链接]

- 张国和, 黄凯, 张斌, 等. 最大稳定极值区域与笔画宽度变换的自然场景文本提取方法. 2017, 51(1): 135-140. [doi: 10.7652/xjtuxb201701021]
- 李扬, 潘泉, 杨涛. 基于短文本情感分析的敏感信息识别. 2016, 50(9): 80-84. [doi: 10.7652/xjtuxb201609013]
- 王菲菲, 杨扬, 蒋飞, 等. 面向用户话题相似性特征的链路预测方法. 2016, 50(8): 103-109. [doi: 10.7652/xjtuxb201608017]
- 范正光, 屈丹, 闫红刚, 等. 借助音频数据的发音字典新词学习方法. 2016, 50(6): 75-82. [doi: 10.7652/xjtuxb201606012]
- 孙立远, 周亚东, 管晓宏, 等. 利用信息传播特性的中文网络新词发现方法. 2015, 49(12): 59-64. [doi: 10.7652/xjtuxb201512010]
- 刘凯, 张立民, 孙永威, 等. 利用深度玻尔兹曼机与典型相关分析的自动图像标注算法. 2015, 49(6): 33-38. [doi: 10.7652/xjtuxb201506006]
- 赵煜, 邵必林, 边根庆. 一种融合词序信息的多粒度文本话题情感联合模型. 2014, 48(11): 103-108. [doi: 10.7652/xjtuxb201411018]
- 杜友田, 辛刚, 郑庆华. 融合异构信息的网络视频在线半监督分类方法. 2013, 47(7): 96-101. [doi: 10.7652/xjtuxb201307018]
- 孙艳, 周学广, 付伟. 无监督的主题情感混合模型研究. 2013, 47(4): 120-125. [doi: 10.7652/xjtuxb201301023]
- 田丰, 桂小林, 杨攀, 等. 采用类别相似度聚合的关联文本分类方法. 2012, 46(12): 6-11. [doi: 10.7652/xjtuxb201212002]
- 杨攀, 桂小林, 田丰, 等. 一种高效的用于话题检测的关键词元聚类方法. 2012, 46(10): 24-28. [doi: 10.7652/xjtuxb201210005]
- 霍战鹏, 魏正英, 张梦, 等. 手机短信远程控制灌溉系统. 2012, 46(10): 36-41. [doi: 10.7652/xjtuxb201210007]
- 豆增发, 高琳. 利用膜粒子群优化和信息熵的医学文本特征选择. 2012, 46(4): 45-51. [doi: 10.7652/xjtuxb201204008]
- 薛峰, 周亚东, 高峰, 等. 一种突发性热点话题在线发现与跟踪方法. 2011, 45(12): 64-69. [doi: 10.7652/xjtuxb201112012]

(编辑 武红江)