DOI：10.7652/xjtuxb201504011

领域实例迁移的交互文本非平衡情感分类方法

田锋[1,2]，兰田[1,2]，CHAO Kuo-Ming[3]，吴凡[1,2]，郑庆华[1,2]，高鹏达[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；3.考文垂大学计算机科学与技术系，CV1 2JH，英国考文垂）

摘要：针对交互文本句子短、成分缺失、多领域下类分布不均衡导致的高维、特征值稀疏、正样本稀少的难点，提出面向目标数据集实例迁移的数据层面采样方法。该方法提出目标数据集和源数据集共性特征的 Top-N 信息增益和值占比函数，选择评价两个数据集实例相似度的特征；提出目标数据集和源数据集特征空间一致性处理方法，克服两者特征空间不一致的问题；提出分领域的实例选取与迁移方法，克服多领域下的类分布不均衡问题。实验结果表明：该方法有效缓解了交互文本的非平衡问题，使支持向量机、随机森林、朴素贝叶斯、随机委员会 4 个经典分类算法的加权平均的接收者运行特征曲线(receiver operating characteristic，ROC)指标提升了 11.3%。

关键词：交互文本；非平衡情感分类；多领域；实例迁移
中图分类号：TP391.1 **文献标志码**：A **文章编号**：0253-987X(2015)04-0067-06

An Unbalanced Emotion Classification Method for Interactive Texts Based on Multiple-Domain Instance Transfer

TIAN Feng[1,2]，LAN Tian[1,2]，CHAO Kuo-Ming[3]，WU Fan[1,2]，
ZHENG Qinghua[1,2]，GAO Pengda[1,2]
（1. Shaanxi 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；3. Department of Computer Science and Technology，Coventry University，Covertry CV1 2JH，UK）

Abstract：A data level sampling method of target dataset-oriented instance transfer is proposed to solve the problem that the characteristics of interactive texts such as short sentences，missing parts of sentences and unbalanced class distribution in multiple-domains result in difficulties of high dimension，sparse eigenvalue in feature space and lack of positive instances. A function is employed to choose features for evaluating the instance similarity between source and target datasets. The function calculates the sum of the information gains of Top-N common features of these two datasets and their proportions in the sum. Moreover，a homogenization processing method is presented for feature spaces of the target dataset and the source dataset to overcome the feature spaces inconsistency between these two datasets. A method for selecting and transferring instances from a domain of source dataset to the corresponding one of target dataset is adopted to solve the problem of unbalanced class distribution in multiple domains. Experimental results show that the proposed method effectively alleviates the unbalanced problem in target dataset.

收稿日期：2014-09-11。 作者简介：田锋(1972—)，男，副教授，博士生导师。 基金项目：国家自然科学基金资助项目(61472315)；国家科技支撑计划资助项目(2013BAK09B01)；教育部"创新团队发展计划"资助项目(IRT13035)；国家留学基金资助项目(20133018)。
网络出版时间：2015-02-10 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20150210.0924.003.html

The proposed method running with four classic classification methods, i. e. support vector machine, random forest, naive Bayes, and random committee, results in an 11.3% improvement in average of weighted receiver operating characteristic curve (ROC).

Keywords: interactive texts; imbalanced sentiment classification; multiple domain; instance transfer

文本交互是社交媒体上的重要交流形式之一，因而交互文本成为情感识别研究的重要载体。现实应用中，文本交互不仅在情感类别分布上不均衡，而且同一数据源的交互文本在不同主题下也存在情感标签分布非均衡的特点，极易导致在其上训练所得的分类模型忽视少数类信息，产生“过拟合”现象，这即是多领域下中文交互文本非平衡多类情感识别问题。解决该问题的难点在于：①需克服交互文本在话轮级别上情感识别的固有难点，例如句子短、成分缺失、非语言符号多等[1]；②需寻找合适的分类模型，目前的方法多是基于对支持向量机(support vector machine, SVM)方法的改进[2]，然而，前期研究[3-4]指出，由于交互文本在话轮级别上语言具有非形式化和特征稀疏等特点，SVM 算法在其上的情感识别性能不佳；③需要在多领域下满足多类情感标签分类的需求，已有的迁移学习方法多数集中在有无情感、正负情感等二分类问题上，针对多情感分类的研究较少。

针对以上问题与难点，本文提出了针对中文交互文本非平衡情感分类的多领域实例迁移方法，其特点是：①提出迁移实例的特征筛选策略，即依据 Top-N 信息增益和值占比函数值的贪婪算法，选取共有特征集中对非平衡数据分类贡献权重大的共有特征，作为目标数据集 T 和源数据集 S 中用于实例相似性评价的特征集；②提出特征空间一致性处理方法，即目标数据集数据特征和源数据集被迁移数据特征之间的特征空间一致性处理方法，克服 T 和 S 数据集的特征不一致问题；③提出分领域实例迁移的新训练集生成方法；④所提方法适用于多类标签的情感分类问题。

1 相关工作

非平衡数据分类是机器学习领域中的一个富有挑战性的问题。类标记样本数据分布(简称类分布)的不平衡使得分类器在训练过程中严重偏向多数类，从而导致其少数类的识别性能急剧下降。目前处理这一问题主要包含以下几类方法：数据层面的采样、代价敏感学习、特征选择、特征权值调整和单分类学习[1]。

数据层面的采样主要包括欠采样和过采样两类基础方法。欠采样通过从多数类中采样得到一部分数据，从而使类分布达到平衡。过采样通过将少数类样本多次采样或者采用直接复制增加的方式使类分布达到平衡。文献[5-6]讨论了这两种方法在处理非平衡问题上的劣势：欠采样会导致数据的丢失；过采样会增加模型训练时间，同时也会产生过拟合现象。

代价敏感学习的主要思路是赋给误判少数类和误判多数类不同的代价权值，使分类器更看重少数类[7-8]。

特征选择的思路是挑选更偏向少数类的特征来更好地学习少数类知识[1]。Wang 等将基于边界区域的切割算法应用于两类情感分类问题[9]。

特征权值调整是通过赋予对少数类更重要的特征以较大的权值来纠正分类器对多数类的偏移，从而解决非平衡问题[10]。

单分类学习主要应用于数据严重非平衡的情况，例如信息过滤和欺诈检测，是一种忽略其他类别，只用某一类的数据来训练模型的处理方式[1]。

上面提到的解决非平衡问题的方法都是针对单一数据集。受近年来利用辅助数据思想的启发，本文拟借鉴迁移学习的思想，展开中文交互文本中非平衡情感识别的研究。

2 多领域实例迁移问题及所提算法

2.1 多领域实例迁移问题表述及解决思路

多领域实例迁移所面临的问题是：多领域实例迁移的目标数据集 T 中，各领域的样本数据规模大小不同，且总体上在 3 个待学习情感类标签(正面、负面、平静)上的数据规模差异较大(少数类与多数类的比值一般小于 1∶3到 1∶10)，如图 1 所示。

多领域实例迁移的目标是：从源数据集 S 各个领域中筛选一定数目的实例迁移到目标数据集对应领域中，形成新的目标数据集 D，使得新数据集的各类别在数据规模上基本接近，从而提高目标数据集在各个领域上的情感分类精度。此过程如图 1 所示，其中

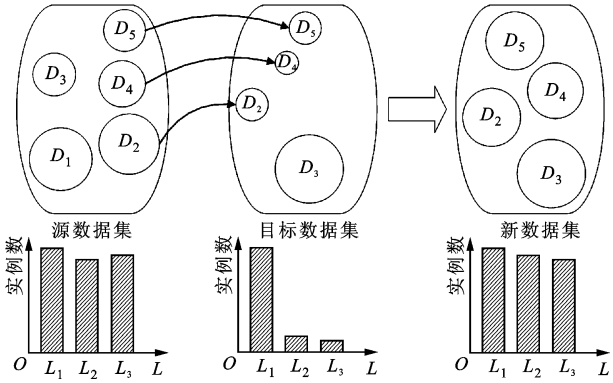

图 1　多领域实例迁移实例图

源数据集有 5 个领域 D_1、D_2、D_3、D_4 和 D_5，目标数据集有 4 个领域 D_1、D_2、D_3 和 D_4。源数据集和目标数据集有相同的待识别类别 L_1、L_2 和 L_3。通过迁移使得新数据集 D 的各类别数据规模近似。

该研究的核心思想是：鉴于多领域实例迁移源数据集 S 和目标数据集 T 识别任务相同，故采用两个数据集在共同分类任务上的特征空间 $\Omega(F|T)$ 和 $\Omega(F|S)$ 的相似性来评价迁移实例。当 $\Omega(F|T)=\Omega(F|S)$ 时，此类问题相对容易解决。但是，通常源数据集和目标数据集存在共有特征和独有特征[3]，特征空间不等。若依据已有研究[4]，中文交互文本数据的共有特征为交互文本的句法结构特征、交互特征和频次特征；独有特征是 Ngram 特征。Ngram 特征是指数据集中的词语以及词语组合，具有很强的数据依赖性。因而，实现该思路的难点是如何评价两个数据集在共同分类任务上特征空间 $\Omega(F|T)$ 和 $\Omega(F|S)$ 中特征的相似性及其分类性能，以及如何克服两个数据集各自独有特征引起的差异，因此需要解决 3 个问题：①源数据集与目标数据集的共性发现与选择问题；②源数据集的实例向目标数据集的可迁移性问题；③被迁移实例特征与目标数据集实例特征的同质化问题。

针对以上问题，本文提出了一种基于贪婪策略的多领域实例迁移方法，该方法分 3 个步骤完成。

步骤 1　从源数据集对应领域中筛选迁移的实例。该阶段首先解决共有特征的发现与选择问题，其次解决迁移实例的选择问题，它包含两个子问题：①需要迁移实例的个数；②源数据集的实例向目标数据集的可迁移性问题。对于问题①，其目标是使各个类别上的实例数量基本一致，克服其非平衡特性。对于问题②，要考虑从源数据集中迁移哪些实例到目标数据集对应领域中，才能尽可能减少对目标数据集的特征分布的改变，同时提高其识别精度。

步骤 2　源数据迁移实例与目标数据实例特征空间一致性处理，该步骤是用于解决特征的同质化问题。

步骤 3　把迁移实例按领域与情感标签分别归入目标数据集，在迁移后所得的新目标数据集 D 上训练分类器。

2.2　源数据集和目标数据集共有特征选择

在中文交互文本的情感识别问题中，文献[4]通过多组对比实验得出，基于决策树的分类算法在解决这个问题上表现很好，并且通过计算得出共有特征信息增益在两个数据集中均占有较大比重。本文通过共有特征信息增益在数据集中所占比重的大小来挑选代表实例的特征，其流程如下：①分别计算数据集中各个特征的信息增益，并将特征按信息增益值降序排列；②在源数据集特征序列中标记共有特征出现的位置；③对源数据集每个共有特征标记位置，计算该标记位置之前所有共有特征的信息增益和值与该标记位置之前的所有特征的信息增益和值的比值（共有特征信息增益比），挑选比值最大者得到代表实例的特征集合。

2.3　利用余弦相似度计算规则从源数据集中筛选被迁移实例

余弦相似度是自然语言处理中常用的相似度计算方式，它通过将文件表示成向量形式来计算实例的相似性。本文采用基于共有特征的余弦相似度得分来衡量源数据集中的实例与目标数据集对应领域实例的相似度，从而评价该实例的可迁移性。该算法可以分为下面 3 个步骤。

步骤 1　将每一个实例表示成所选共有特征值的向量形式，并对共有特征进行归一化处理。特征归一化流程包括两个步骤：①分类变量属性/特征数值化处理，其中针对分类变量属性/特征，本文采用直接数值替换的方式进行数值化处理，数值取值从 0 开始，依次递增，例如特征 conjunction 中共有 8 个值，分别为{none, adversative, casual, hypothetical, coordinate, comparison, undertake, alternative}，本文将其分别对应为{0,1,2,3,4,5,6,7}；②数值特征归一化处理，本文采用常用的最大最小值归一化处理[11]。

步骤 2　计算源数据集某领域中某一情感类别的实例在目标数据集对应领域中对应情感类别的实例中的总体余弦相似度得分。一般认为，两个实例越相似，则它们的总体余弦相似度越高。设 $L=\{l_1,l_2,\cdots,l_N\}=\{l_p|p=1,2,\cdots,N\}$ 表示数据集分类类别标签，N 表示分类任务的类别数；$D=\{d_1,d_2,\cdots,d_M\}=\{d_k|k=1,2,\cdots,M\}$ 表示数据集中的领域（主题）集合，M 表示其领域（主题）个数，具体公式为

$$s(S_{d_k}^{l_p}(i)) = \sum_{i=1}^{m} \cos(S_{d_k}^{l_p}(i), T_{d_k}^{l_p}(j))/m \quad (1)$$

式中：$T_{d_k}^{l_p}(j)$表示目标数据集 l_p 类别、d_k 领域的一个实例，$j=1,2,\cdots,n$，表示目标数据集该领域中该类别实例有 n 个；$S_{d_k}^{l_p}(i)$表示源数据集 l_p 类别、d_k 领域的一个实例，$i=1,2,\cdots,m$，表示源数据集该领域中该类别实例有 m 个；$s(S_{d_k}^{l_p}(i))$为针对实例 $S_{d_k}^{l_p}(i)$的基于共性特征的余弦相似度得分，其中余弦函数是计算两个实例在共性特征上归一化后的余弦相似度。

步骤 3 对源数据集中各领域上各情感类别的实例，分别依据基于共性特征的余弦相似度得分按降序排列，优先迁移排在前面的实例。

2.4 特征空间一致性处理

此阶段解决迁移而来的实例与目标数据集中实例的同质化问题。源数据集和目标数据集之间既存在共有特征，又存在独有特征，这导致从源数据集中迁移而来的实例无法直接和目标数据集中的实例进行训练，因此需要将迁移的实例和目标数据集实例进行一致性处理，使两者的特征空间一致。本文中的源数据集和目标数据集特征空间的差异体现在 Ngram 特征上，此类独有特征的类型均为数值型，故利用其在对应领域中取得的基于共性特征的余弦相似度得分 s 和目标数据集实例来统一迁移实例特征空间，即源数据集中迁移实例的 Ngram 特征和目标数据集实例的 Ngram 特征维度相同，其中，共有特征可以直接用于新实例中。具体步骤如下。

步骤 1 分别计算目标数据集中各个领域各个类别中独有 Ngram 特征的平均值

$$\overline{N_T^{l_p d_k}} = \sum_{j=1}^{n} N_T^{l_p d_k}(j)/n \quad (2)$$

式中：$N_T^{l_p d_k}(j)$为目标数据集 d_k 领域、l_p 类别上的第 j 个实例在 Ngram 特征上的值。

步骤 2 利用迁移实例自身的 Ngram 值和其对应的目标数据集中平均 Ngram 特征值，并结合其总体余弦相似度，构造出新的迁移实例的 Ngram 特征，其特征空间与目标实例一致

$$N_n(S_{d_k}^{l_p}(i)) = \overline{N_T^{l_p d_k}} s(N_S^{l_p d_k}) + N'^{l_p d_k}_S(i) \quad (3)$$

式中：$N_S^{l_p d_k}(i)$为源数据集中 d_k 领域、l_p 类别上第 i 个实例在 Ngram 特征上的取值；$N'^{l_p d_k}_S(i)$为用零在后面填补到与目标数据集中实例的 Ngram 特征维数相同的特征取值。

2.5 实例合并与模型训练

经过上述两步的处理，将源数据集实例迁移到目标数据集，克服了目标数据集的非平衡。实例合并遵照下面两个原则：源数据集中某领域的某类别数据只能迁移一次，因为共有特征上余弦相似度较大的实例多次迁移，易造成过拟合问题；使目标数据集的各个领域内不同情感类别实例的个数趋近相同，即尽可能克服目标数据集中各个领域上的非平衡性。

3 实 验

3.1 实验步骤

步骤 1 收集实验语料。实验所用数据集 Linux_QQ 为腾讯 QQ 群上某学习小组的 5 123 条中文交互文本数据，数据集 Xjtu_BBS 为在西安交通大学兵马俑 BBS 上收集的 9 957 条中文交互文本数据。对数据进行情感和领域标注来得到统计信息，结果表明：Linux_QQ 为非平衡数据，且非平衡问题在各个领域（主题）内同时存在；Xjtu_BBS 为平衡数据，存在大量 Linux_QQ 中的少数类实例，所以将 Linux_QQ 作为目标数据集 T，以 Xjtu_BBS 为源数据集 S。

步骤 2 按 2.2 节中的步骤挑选代表实例的共有特征，并按照 2.3 节中的步骤在各个领域分别计算其总体相似度，确定各个领域的迁移实例。

步骤 3 按 2.4 节中的步骤对迁移实例进行特征空间一致性处理。

步骤 4 按 2.5 节中的步骤分领域分别合并实例，形成新的训练数据集，记为 Immigration。

步骤 5 在数据集 Immigration、Linux_QQ、经过子采样（Subsampling）方法处理过的数据集 Subsampling 1800 以及经过合成少数类过采样技术（synthetic minority over-sampling technique，SMOTE）处理过的数据集 SMOTE 上，分别采用随机森林（random forest，RF）、支持向量机（support vector machine，SVM）、随机委员会（random committee，RC）和朴素贝叶斯（naïve Bayes，NB）4 个分类算法进行实验，并比较实验结果。实验结果评价指标采用接收者运行特征曲线（receiver operating characteristic，ROC）[11]，设为 R_{OC}。该指标被认为是评价非平衡处理效果的最佳指标。

人工标记的两个数据集 Xjtu_BBS 和 Linux_QQ 中不同主题（生活、学习、友情、爱情、家庭）上的情感标签分布情况为：Xjtu_BBS 中生活主题占所有话轮 75%的比重，而 Linux_QQ 中学习主题占所有话轮超过 75%的比重。在主题分布上，Xjtu_BBS 中友情、爱情和家庭的话轮数比 Linux_QQ 中丰富很多，尤其是友情主题，它在 Xjtu_BBS 中所占比例近似于学习主题。

3.2 实验结果

3.2.1 步骤 2 实验结果　源数据集为 Xjtu_BBS，目标数据集为 Linux_QQ，其中共有特征有 40 维，利用上述算法挑选出 18 维代表实例的共有特征分别为{length，oneWord，emotionGraph，negWord，punNum，mimeticExist，adjBelongAdver，posWord，emotionVer，adjBelongAtt，advBelongAdver，nxExist，conjunction，interjectionExist，maxEverySeten，FreAdv，FreVerb adjBelongcomplement}。

后文中用 $F=\{f_1,\cdots,f_{18}\}=\{f_c\mid c=1,\cdots,18\}$ 代表 T 和 S 集合实例的共有特征。表 1 和表 2 统计了按照实验步骤 2 从源数据集各个领域分别迁移的实例个数，以及迁移后所得的新数据集(Immigration)的各个领域以及类别上的实例个数。

表 1　不同主题下各标签从源数据集迁移的实例个数

主题	实例个数		
	负面	正面	平静
生活	485	432	0
学习	820	1 285	0
爱情	19	16	0
友情	17	11	0

表 2　迁移后数据集实例在领域和类别上的分布

主题	实例个数		
	负面	正面	平静
生活	545	545	545
学习	1 155	1 832	3 466
爱情	22	22	22
友情	18	18	18

3.2.2 实验结果与分析　表 3～表 6 为不同非平衡数据处理方法在 4 种分类算法上的实验结果对比。

经过实例迁移后得到的新数据集(Immigration)，其实例在各领域和类别上的分布如表 2 所示。由表可见，目标数据集中有 3 个领域实例数目相同，学习领域受限于源数据集的实例分布，其实例分布没有达到完全平衡，但总体比值由 1∶10 提升到了 1∶3以上，极大地缓解了该领域的非平衡性。从整体上看，所提方法对正面和负面两类情感的识别性能均有提升，这是以损失一定的平静情感类别的识别准确率和召回率为代价的。大多数情况下 ROC 曲线值都有相应地提升，这也说明这种代价是值得的。

从情感分类加权平均指标上看，见表 3，多领域实例迁移方法在 Random Forest 分类算法上取得了最好的实验结果。由表可见，由于非平衡处理策略主要考虑的是照顾少数类分类结果，因此 RF、SVM、NB 等算法在 Subsampling 数据集上的识别指标略差于在 Linux_QQ 上的表现，而这些方法在 SMOTE 和 Immigration 数据集上的性能均优于在 Linux_QQ 上的实验结果。

表 3　非平衡数据处理方法实验结果(加权平均)

数据集	加权平均的 R_{OC}			
	RF	RC	SVM	NB
Linux_QQ	0.791	0.791	0.801	0.801
Subsampling1800	0.670	0.670	0.673	0.673
SMOTE	0.796	0.796	0.702	0.702
Immigration	0.925	0.916	0.851	0.851

从少数类情感的分类指标上看，见表 4、表 5，针对少数类别，SMOTE 和 Immigration 数据集表现稳定，4 个分类算法均在其上提高了情感分类的性能。虽然 Subsampling 数据集在 RF 算法上有明显的性能提升，但在其他算法上却表现一般或者变差。同时，在对少数类分类的实验中，多领域实例迁移算法生成的 Immigration 数据集表现最佳。

表 4　非平衡数据处理方法实验结果对比(负面)

数据集	负面情感的 R_{OC}			
	RF	RC	SVM	NB
Linux_QQ	0.753	0.711	0.060	0.060
Subsampling1800	0.760	0.722	0.198	0.198
SMOTE	0.942	0.919	0.222	0.222
Immigration	0.940	0.928	0.738	0.738

表 5　非平衡数据处理方法实验结果对比(正面)

数据集	正面情感的 R_{OC}			
	RF	RC	SVM	NB
Linux_QQ	0.703	0.680	0.230	0.230
Subsampling1800	0.701	0.679	0.264	0.264
SMOTE	0.865	0.839	0.273	0.273
Immigration	0.913	0.908	0.755	0.755

从平静类情感的分类结果来看，见表 6，各个分类器在没有经过处理的非平衡数据 Linux_QQ 上大多表现最好。这验证了其非平衡特性对多数类的倾向和过拟合问题。在 Subsampling 方法中，4 个

分类器的表现均最差,分析其原因为欠采样造成了大量信息丢失。

综合上述实验结果可以得出,多领域实例迁移的策略在中文交互文本非平衡情感分类问题上表现良好,该方法有效缓解了交互文本的非平衡问题,使加权平均的接收者运行特征曲线指标提升了11.3%,其与 Random Forest 分类方法的组合取得了最好的表现。

表 6 非平衡数据处理方法实验结果对比(平静情感)

数据集	平静情感的 R_{OC}			
	RF	RC	SVM	NB
Linux_QQ	0.725	0.701	0.968	0.968
Subsampling1800	0.728	0.707	0.696	0.932
SMOTE	0.883	0.861	0.723	0.939
Immigration	0.926	0.916	0.864	0.851

4 总 结

针对交互文本中的非平衡问题,本文提出多领域实例迁移方法,分步实现了迁移实例的特征筛选、特征空间的一致性处理以及分领域实例迁移的新训练集生成方法,适用于多个情感标签分类问题。经对比实验验证,该方法在解决中文交互文本中的非平衡问题时效果较佳,有效地缓解了交互文本的非平衡性,提高了识别精度,为多领域且目标数据集和源数据集特征空间非同质情况下的实例迁移提供了一种解决思路。

本文所提出的多领域实例迁移方法在进行特征筛选时,主要对目标数据集和源数据集的共性特征通过信息增益的加和运算进行选择。下一步工作可以考虑在特征选择的过程中,充分发现目标数据集和源数据集中的分类知识相似度,提出基于知识的迁移学习算法,努力在识别性能上得到突破。

参考文献:

[1] 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.

[2] DAI Wenyuan, YANG Qiang, XUE Guirong, et al. Boosting for transfer learning [C]// Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007: 193-200.

[3] TIAN Feng, GAO Pengda, LI Longzhuang, et al. Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems [J]. Knowledge-Based Systems, 2014, 55: 148-164.

[4] TIAN Feng, LIANG Huijun, LI Longzhuang, et al. Sentiment classification in turn-level interactive Chinese texts of e-learning applications [C]// Proceedings of the 2012 IEEE 12th International Conference on Advanced Learning Technologies. Piscataway, NJ, USA: IEEE, 2012: 480-484.

[5] BARANDELA R, VALDOVINOS R M, SÁNCHEZ J S, et al. The imbalanced training sample problem: under or over sampling? [C]// Proceedings of the IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition. Berlin, Germany: Springer, 2004: 806-814.

[6] CHAWLA N V. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure [C]// Proceedings of the International Conference on Machine Learning from Imbalanced Datasets: Ⅱ. New York, USA: ACM, 2003: 1-8.

[7] SUN Yanmin, KAMEL M S, WONG A K C, et al. Cost-sensitive boosting for classification of imbalanced data [J]. Pattern Recognition, 2007, 40(12): 3358-3378.

[8] ZHOU Zhihua, LIU Xuying. Training cost-sensitive neural networks with methods addressing the class imbalance problem [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63-77.

[9] WANG Suge, LI Deyu, ZHAO Lidong, et al. Sample cutting method for imbalanced text sentiment classification based on BRC [J]. Knowledge-Based Systems, 2013, 37: 451-461.

[10] LIU Ying, LOH H T, SUN Aixin. Imbalanced text classification: a term weighting approach [J]. Expert Systems with Applications, 2009, 36(1): 690-701.

[11] HAN Jiawei, KAMBER M. 数据挖掘概念与技术 [M]. 2 版. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2006: 216-217.

(编辑 武红江)