一种基于质心空间的不均衡数据欠采样方法

金 旭[1] 王 磊[1] 孙国梓[1,2,3] 李华康[1,2,3]

(南京邮电大学江苏省大数据安全与智能处理重点实验室 南京 210023)[1]

(江西省经济犯罪侦查与防控技术协同创新中心 南昌 330103)[2]

(数学工程与先进计算国家重点实验室 江苏 无锡 214000)[3]

摘 要 针对目前的分类算法在不均衡数据集上的分类效果不理想的问题,将监督学习和无监督学习相结合,提出了一种基于质心的欠采样——ICIKMDS。在现实应用中,一些数据并不容易获得,或者不同类型的数据本身在数量上就存在着差异性,因此造成了数据集分布的不均,如疾病检测中疾病患者和正常人比例的不均、信用卡欺诈中欺诈用户和正常用户比例的不均等。所提方法很好地解决了数据集不均衡的问题,首先通过求解样本之间的欧氏距离得到初始质心,然后采用k-means算法在大类样本集上进行聚类,使不均衡数据集在分布上更加均衡,有效地改善了分类器的分类效果。所提方法使分类器在测试集小类上的分类准确率远远高于随机欠采样和SMOTE算法,在整个测试集上的准确率几乎与其他算法相同。

关键词 不均衡,欠采样,k-means,SMOTE算法

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.008

Under-sampling Method for Unbalanced Data Based on Centroid Space

JIN Xu[1] WANG Lei[1] SUN Guo-zi[1,2,3] LI Hua-kang[1,2,3]

(Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)[1]

(Collaborative Innovation Center for Economics Crime Investigation and Prevention Technology, Nanchang 330103, China)[2]

(State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi, Jiangsu 214000, China)[3]

Abstract In view of the fact that the classification performance of current classification algorithms is not ideal for the unbalanced dataset, through combining supervised learning and unsupervised learning, this paper proposed a sub-sampling method based on centroid, namely ICIKMDS. In practical applications, some data are not easily to be obtained or different types of data are different in quantity, resulting in uneven distribution of data, such as the disproportion of the sufferer and the normal people in the detection of diseases, the disproportion of the fraud users and normal users in credit card fraud and so on. The new method solves the disproportion problem of dataset well. In this method, the initial centroid is obtained by solving the Euclidean distance between samples, and then the k-means algorithm is used to cluster the large-class sample sets to make the disproportionate dataset more balanced in distribution, effectively improving the effect of classifiers. The proposed method makes the classification accuracy of the classifier much better than that of random under-sampling and SMOTE algorithm on the subclass of test set, and its accuracy on the whole test set has little difference from other algorithms.

Keywords Unbalanced, Under-sampled, k-means, SMOTE algorithm

到稿日期:2017-12-19 返修日期:2018-04-19 本文受国家自然科学基金资助项目(61502247,11501302,61502243),国家博士后科学基金(2016M600434),江苏省自然科学基金(BK20140895,BK20150862),江苏省博士后科研资助计划(1601128B),江西省经济犯罪侦查与防控技术协同创新中心开放基金资助课题(JXJZXTCX-015)资助。

金 旭(1993-),男,硕士生,主要研究方向为自然语言处理;王 磊(1994-),男,硕士生,主要研究方向为自然语言处理和情感分析;孙国梓(1972-),男,博士,教授,CCF会员,主要研究方向为网络空间安全、电子数据取证;李华康(1982-),男,博士,讲师,CCF会员,主要研究方向为智慧城市、大数据应用、互联网安全,E-mail:huakanglee@163.com(通信作者)。

1　引言

在 2000 年的 AAAI 大会和 2003 年的 ICML 会议上，数据集的不均衡问题被作为一个研究热点提出[1-3]。Maloof 等提出大部分的文本分类问题都是数据集不均衡问题。数据集上的各类样本数目存在差异，被称为机器学习上的不均衡问题[4]。一般来说，导致数据集不均衡的原因有两种：1）数据本身就存在类别上的样本数量差异，信用卡欺诈、异常检测、大众中癌症病人的筛查、发现不可靠的电信客户[5]、卫星雷达图像中漏油的检测[5]、学习单词发音[5]等方面的数据不均衡就是这类原因产生的；2）数据的获取很难，例如车牌识别中的汉字。一般地，将含有样例数少的类称为小类，将样例数多的类称为大类[5]。传统的机器学习算法很难解决数据不均衡问题，但该问题又是极其重要的。比如，在银行欺诈检测中，一般而言，大多数的客户都在正常交易，但个别客户的行为可能是欺诈行为，这类行为虽然数目很少，但造成的损失是巨大的[6]。因此，对此类问题进行研究具有重要的价值。

目前，这个方面的研究面临着各种各样的难题。1）不均衡问题上的方法的评判准则：一般地，在均衡数据上都是将分类精度作为准则，而将其作为不均衡问题上的评判标准则是不适合的，比如在二分类问题中，分类器会将预测出的待测样本作为大类，由于大类中的样本量占据了很大的分量，因此得到的分类精度也相应很高。2）样本的缺失：由于小类所含信息有限，因此难以确定小类数据的分布，从而造成识别小类样本较为困难[6]。

针对数据集不均衡问题，相关学者进行了大量的研究。有些学者利用数据分布来解决这个问题，即我们提到的采样技术，通过过采样和欠采样使数据分布达到均衡，但是过采样容易产生过拟合问题，而欠采样容易产生数据信息的丢失问题。一些学者在现有的分类算法上进行改进，使实际中的分类面向着理想中的超平面靠近，最后达到基本重合。Núñez 等人对 SVM 算法进行改进，提出了一种低成本的后处理策略来改善 SVM 在小类上的分类性能[7]。Boosting 算法也是通过改进分类器而产生的一种算法，其中最具代表性的是 AdaBoost 算法，它对错分样本增加权重而对正确分类样本减少相同比例的权重，但分类效果不佳。学者 BaiPeng 也是基于 SVM 算法进行改进，提出了一种 C-SVM 模型，用来解决数据集不均衡问题。它增大了对错分少类样本的惩罚参数，同时减小了对错分多类样本的惩罚参数，达到了不错的效果。还有一些学者通过改进特征选择算法来提升分类器在不均衡数据集上的分类性能。在不均衡的数据集上用传统的特征选择算法进行处理时，比较容易得到大类样本集中的特征，而小类样本集中的特征往往很少被选择，由于某些特征选择算法一般会倾向高频词。这样得到的特征集就不具代表性，比较倾向于大类，而对于小类样本的分类效果就会表现得不佳。在特征选择算法优化方面一般有以下 3 种方法：特征选择算法的改进、构造新的特征选择算法和特征选择结合分类器的方法。对于特征选择算法的改进，Abdellatif 等人提出的 ARCID 算法主要强调从不均衡的数据集中进行信息抽取，而小类中的抽取信息不会严重影响分类器的预测准确率，从而提高了整体的分类性能[8]。You 通过平衡因子对 IG 特征选择算法进行了改进，降低小类负相关的特征所带来的干扰，有效地提高了分类效果，其通过加入类别倾向性因子并借鉴 TF-IDF 中的 IDF 定义，提出了 ICF（逆类别频）的概念对 CHI 算法进行了改进，有效地提高了分类效果。对于构造新的特征选择算法，Liu 构造了一种基于 Hellinger 距离的新特征选择算法，其距离越大，特征代表性越好，由于 Hellinger 距离对类别的倾斜不敏感，因此得到了较好的分类效果。

本文提出的方法着眼于采样的技术层面，对大类中的样本进行欠采样。首先找到大类样本的初始质心，然后使用 k-means 算法进行聚类，最终得到 k 个簇。选择每个簇中与其质心相似度最大的那个样本组成大类最终的样本集合。该方法在遵循大类样本数量减少的前提下，最大限度地保证了大类样本中的信息含量[9]。得到初始质心的主要过程如下：首先分别计算样本集中的每个样本与其他样本的欧氏距离之和，再将距离之和最小的样本作为第一个初始聚类质心，然后将与第一个初始聚类质心距离最大的样本作为第二个初始质心，最后将与第一个和第二个质心距离之和最大的样本作为第三个初始质心；依次类推，将会得到 k-means 算法的初始质心集合。第一个初始质心是通过求解得到的，而不是随机选择的，因为随机选择的样本可能是一个边缘样本，边缘样本会影响最终的聚类效果。该算法的另一个重要步骤是不把最终质心作为最终大类样本集，而是选择经过 k-means 算法得到的 k 个簇中与其质心相似度最大的那个样本组成最终大类样本集，这样能提高分类器的分类效果。实验结果表明了本文提出的方法在小类的分类准确率上优于其他方法的准确率。

本文第 2 节呈现了相关研究，比较详细地介绍了采样技术以及采样技术中比较典型的两种方法：SMOTE 算法[10]、随机欠采样；第 3 节详细介绍了本文提出的方法；第 4 节介绍对比实验和所采用的数据集，以及几种度量实验性能的方法；最后总结全文。

2　相关工作

目前用于解决数据集不均衡的方案一般包含 3 种[11]。这 3 种方案分别从采样、算法和特征选择[4,12-13]3 个方面有效地提高了不均衡数据集上的分类效率。已有算法一般是以牺牲大类分类的准确率来提高对小类样本分类的准确率，比如从算法的层次上来说，通过改变分类面的方向来对支持向量机算法进行改进，使它向着大类样本的方向倾斜[14]。本文主要介绍采样技术，然后介绍采样技术中的两种典型方法：SMOTE 算法和随机欠采样。采样技术作为数据集预处理的一种方法，用来改变数据集的分布。采样技术分为过采样和欠采样两种类型。过采样是在小类样本中加入新的样本，使其在分布上与大类样本达到平衡；欠采样技术在原有大类样本中剔除一些样本，使之减少到小类的样本数量。但是，过采样一般存在着过拟合的问题，欠采样容易删除掉一些对分类

贡献较大的样本。下面介绍过采样中的一种典型方法——SMOTE 算法和欠采样中最具代表性的一种方法——随机欠采样。

2.1 SMOTE 算法

在合成小类的过采样技术中，SMOTE 算法是最具代表性的算法[10,15-16]。该算法是 Chawla 等在过采样的基础上提出的，是基于随机过采样算法的一种改进方案。随机过采样采取简单复制样本的策略来增加小类样本，这样容易产生模型过拟合的问题，即模型学习到的信息过于特别而不够泛化；而 SMOTE 算法采用的是人工合成数据的策略，这在一定程度上解决了模型过拟合问题。但该算法容易产生两个方面的问题：1)k 值的选择具有一定的盲目性；2)使少数样本集产生分布边缘化问题。文献[5]在 SMOTE 算法的基础上作出了改进，提出了一种叫作 Borderline-SMOTE 的模型。该模型通过找到小类样本集边界上的样本集，再基于 SMOTE 算法人工合成新的样本，有效地提高了小类样本分类的准确率。

2.2 随机欠采样(RDS)

随机欠采样是欠采样中比较典型的一种方法，其主要思想是从大类样本集中随机删一些样本，直到和小类样本在数量上达到平衡。由于是从大类样本集中随机剔除一些样本，这样容易造成重要信息的流失，可能会删除掉一些对分类具有较大正面影响的样本，而那些噪声样本或者对分类没有作用的样本依然留在大类样本集中，从而影响分类的准确率。

3 基于质心的欠采样

本文方法的主要思想是在大类样本集上使用 k-means 算法进行聚类，使大类样本集中的样本数量和小类样本集中的样本数量大致相等，即在大类样本集上使用欠采样，这样就基本解决了数据分布不均的问题，最后在处理过的数据集上使用机器学习算法进行模型的学习，从而完成分类。k-means 算法一般对初始质心特别敏感，随机选取初始质心的机制，导致该算法的聚类结果不稳定，容易陷入局部最优而非全局最优，因为 k-means 算法一般采用误差平方和准则函数(见式(1))作为目标函数，而此算法的目的就是通过极小化目标函数来得到聚类结果。由于误差平方和准则函数在空间状态上是一个非凸函数，而非凸函数一般存在着多个极小值，因此随机选取初始质心进行聚类一般会得到次优解。

$$E=\sum_{j=1}^{k}\sum_{x_i\in c_j}|x_i-c_j|^2 \tag{1}$$

其中，k 为簇的个数，x_i 为簇质心 c_j 的数据对象。

本文提出的方法 ICIKMDS 最关键的部分是在大类样本集上使用了基于 k-means 算法的欠采样方法并使用 k-means 算法对初始质心进行求解，即本文不采用在样本集中随机选取样本作为初始质心的机制，而是通过求解来得到初始质心。假设数据集为 $D=\{(x_1,y_1),\cdots,(x_N,y_N)\}$，其中 $x_i=(a_{i1},\cdots,a_{ik})$，$N$ 为数据对象的个数，k 为聚类的个数，通过 FindInitialCenter(D, k)找到 k 个初始聚类质心集合 $C=\{c_1,c_2,\cdots,c_k\}$。求解过程如下：

1)计算数据集 D 中的每两个样本之间的欧氏距离：

$$Dist(x_i,x_j)=\sqrt{(a_{i1}-a_{j1})^2+(a_{i2}-a_{j2})^2+\cdots+(a_{ik}-a_{jk})^2} \tag{2}$$

2)得到数据集 D 中每两个样本之间的欧氏距离之和：

$$sumDist(x_i)=\sum_{j\in(D-x_i)}Dist(x_i,x_j) \tag{3}$$

3)根据 $x_{\text{First}}=\underset{x_i\in D}{\operatorname{argmin}}(sumDist(x_i))$ 得到 x_{First}，并将其作为第一个初始质心 c_1。由于这个样本与其他样本的欧氏距离之和最小，因此它不会是噪声样本，而且必属于负类样本集中某个簇中的一个样本，可以把它作为一个初始质心。

4)$k-1$ 个质心的求解过程为：根据 $x_{\text{Second}}=\underset{x_j\in(D-x_{\text{First}})}{\operatorname{argmax}}(Dist(x_j,x_{\text{First}}))$ 得到 x_{Second}，并将其作为第二个质心 c_2，然后根据 $x_{\text{third}}=\underset{x_k\in(D-(x_{\text{First}}\cup x_{\text{Second}}))}{\operatorname{argmax}}(Dist(x_k,x_{\text{First}})+Dist(x_k,x_{\text{Second}}))$ 得到 x_{third}，将其作为第三个质心 c_3，依次类推，得到 k 个初始质心$\{c_1,c_2,\cdots,c_k\}$。

下面具体介绍本文提出的方法——基于质心的欠采样。以二分类为例，假设正类(小类)样本集为 $X^+=\{(x_1{}^+,y_1{}^+),\cdots,(x_m^+,y_m^+)\}$，负类(大类)样本集为 $X^-=\{(x_1{}^-,y_1{}^-),\cdots,(x_n^-,y_n^-)\}$，且 $n>>m$，首先将簇的个数设为 $k(k=m)$，在负类样本集上进行聚类，得到 m 个样本为负类的新样本集 $X^-=\{(x_1{}^-,y_1{}^-),\cdots,(x_m^-,y_m^-)\}$。

首先介绍 k-means 算法中的样本和质心的相似度计算公式和质心更新公式，假设每个样本表示为 $x_i=(a_{i1},\cdots,a_{ik})$，每个质心表示为 $c_j=(b_{j1},b_{j2},\cdots,b_{jk})$，那么样本 x_i 与每个质心 c_j 的相似度同样采用欧氏距离度量，距离越小，相似度就越大。计算公式如下：

$$sim(x_i,c_j)=\frac{x_i\cdot c_j}{\|x_i\|\cdot\|c_j\|}=\frac{\sum_{h=1}^{k}a_{ih}\cdot b_{jh}}{\sqrt{\sum_{h=1}^{k}a_{ih}^2}\cdot\sqrt{\sum_{h=1}^{k}c_{jh}^2}} \tag{4}$$

每个簇的质心 c_j 的更新公式如下：

$$c_j{}'=\frac{\sum_i x_i}{\text{属于当前质心 } c_j \text{ 的样本数}} \tag{5}$$

其中，x_i 为属于当前质心 c_j 的样本。

5)选取每个簇中与其相似度最大的那个样本，得到 m 个样本的样本集，即负类的新样本集 $X^-=\{(x_1{}^-,y_1{}^-),\cdots,(x_m^-,y_m{}^-)\}$。将得到的负类数据集 X^- 和 X^+ 组成数据集，用支持向量机进行训练得到模型，然后进行测试集的验证。

本文提出的 ICIKMDS 算法的具体描述如算法 1 所示。

算法 1 ICIKMDS 算法

输入：数据集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_{n+m},y_{n+m})\}=X^+\cup X^-$，其中小类样本集 $X^+=\{(x_1^+,y_1^+),\cdots,(x_m^+,y_m^+)\}$，大类样本集 $X^-=\{(x_1^-,y_1^-),\cdots,(x_n^-,y_n^-)\}$，且 $n\gg m$

输出：均衡的数据集 S^*

1. 利用上文中的 FindInitialCenter(X^-, m)，获取 k(k=m)个初始聚类中心$\{c_1,c_2,\cdots,c_m\}$；

[1] http://www.datafountain.cn/#/competions/268/data-intro

2. for each example $x_i \in X^-$，i=1，2，…，n

3.　根据式(4)选取$\{c_1, c_2, \cdots, c_m\}$中与 x_i 相似度最大的 c_j，并将其隶属于这个簇；

4. end for

5. 根据式(5)更新质心，得到新的质心集合$\{c_1', c_2', \cdots, c_m'\}$；

6. 重复步骤 2－步骤 5，直到质心不发生变化或者达到收敛要求时，得到最终的质心集合$\{c_{1\text{-final}}, c_{2\text{-final}}, \cdots, c_{m\text{-final}}\}$；

7. 选取各个簇中与其质心相似度最大的样本，即根据式(6)得到当前簇中的一个代表样本，因此根据步骤 6 中的 m 个最终质心得到 m 个代表样本组成最终的大类样本集 $X^-{}'$和 X^+，从而构成最终的样本集 S^*。

$$x^{-\prime} = \underset{x_i, x_i \in c_{i-final}}{\arg\max} sim(x_i, c_{i-final}) \tag{6}$$

ICIKMDS 算法中，步骤 1)得到 k-means 聚类算法的初始质心集合；步骤 2)－步骤 6)使用 k-means 算法在大类样本集不断地进行迭代直至收敛，从而得到大类样本集最终的聚类质心集合；步骤 7)得到一个均衡的数据集，使用支持向量机等各类分类算法进行训练得到模型，然后利用此模型在测试集中进行验证。

4　实验评估与结果分析

4.1　数据集

本文采用 3 个中文情感语料数据集，分别是京东的 Jingdong_NB_4000，当当的 Dangdan-g_Book_4000 和 ChnSentiCorp_htl_ba_4000，其中分别含有正、负类语篇各 2000 篇。京东的语料内容是客户对自己在京东上所买的电脑进行的评论，当当的语料内容是书友对其在当当上购买的书进行的评论，最后一个语料所涉及的领域是酒店、电脑与书籍，3 个数据集[1)]的信息描述如表 1 所列。

表 1　3 种数据集的属性描述

Table 1　Attribute description of three kinds of datasets

ID	数据集	正例数	负例数
1	Jingdong_NB_4000	2000	2000
1	Dangdang_Book_4000	2000	2000
1	ChnSentiCorp_htl_ba_4000	2000	2000

本文将语料中的负类作为大类，正类作为小类，先在正类和负类中的 2000 条评论中随机地各取 200 条作为测试集，然后将负类中剩下的 1800 条评论作为训练集的负类(大类)样本集，且设置正、负类的不平衡比例为 k，然后在正类中随机地选取 $1800/k$ 条评论作为训练集中的正类(小类)样本集，接着进行分类器模型的训练，最后在测试集上进行测试。每次进行 5 次重复的实验，然后取它们的平均值作为最终的实验结果，k 分别取 10，8，6，4。

4.2　实验性能评估

一般使用分类的正确率来对均衡数据集上分类的性能进行评估，然而这对于不均衡数据集上的分类并不公平。由于不均衡数据集上的大类样本数占了很大的比例，而且分类器基本上偏向于大类，因此在大类样本上的分类基本都是正确的，这样就会造成分类器的正确率依然很高。本文用两个度量指标来进行算法性能的评估，分别是少数类的准确率 pr 和整体性能评估 $g\text{-}mean$[15]。首先介绍混淆矩阵，如表 2 所列。表 2 中 TP 表示正类样本被分为正类的数量，FP 表示正类样本被分为负类的数量，TN 表示负类样本被分为负类的数量，FN 表示负类样本被分为正类的数量，文中正类表示小类，负类表示大类。由于现实中的实例一般都比较侧重于追求小类分类的准确率，但同时整体的分类准确率也不能太低，因此采用式(7)和式(8)[15,17]进行性能的度量。

$$pr = \frac{TP}{TP+FP} \tag{7}$$

$$g\text{-}mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{8}$$

表 2　混淆矩阵

Table 2　Confusion matrix

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

4.3　实验结果分析

实验用到的分类算法是支持向量机，使用的工具是台湾大学林智仁教授等开发的 libsvm。将本文提出的方法 ICIKMDS 与无初始质心优化的普通 k-means 算法(ORD k-means)、随机欠采样(RDS)、SMOTE 算法以及不做任何处理(NODO)做了对比实验，表 3－表 5 分别列出了 3 个语料库上的对比实验结果。

表 3　各方法在 Jingdong_NB_4000 语料库上的实验结果对比

Table 3　Comparison of experimental results of each method in Jingdong_NB_400

k	10		8		6		4	
	$g\text{-}mean$	pr	$g\text{-}mean$	pr	$g\text{-}mean$	pr	$g\text{-}mean$	pr
ICIKMDS	0.827	0.787	0.8394	0.798	0.8326	0.822	0.8572	0.847
RDS	0.823	0.758	0.8464	0.773	0.8533	0.810	0.8689	0.835
SMOTE	0.795	0.617	0.8314	0.680	0.8461	0.748	0.8502	0.761
NODO	0.785	0.449	0.8014	0.483	0.8197	0.598	0.8440	0.692
ORD k-means	0.846	0.766	0.849	0.767	0.8595	0.812	0.8490	0.808

从表 3 可以看出，本文提出的方法在 3 个语料库上的 pr 值都高于其他方法，在语料 Jingdong_NB_4000 上，比 ORD k-means 平均高 2.5%，比随机欠采样平均高 2%，比 SMOTE 算法平均高 10%，比不做任何处理的分类结果平均高 25%；但在整体性能 $g\text{-}means$ 度量上，ORD k-means 和随机欠采样比 ICIKMDS 高 1%左右，ICIKMDS 和 SMOTE 算法的性能没有太大的区别，ICIKMDS 比不做任何处理的方法高 2%。

从表 4 可以看出，在语料 Dangdang_Book_4000 上，本文提出的 ICIKMDS 算法的 pr 值比 ORD k-means 平均高 5%；比随机欠采样平均高 4%，在 k 值为 8 时高 6%；比 SMOTE 算法平均高 27%，在 k=10，8 时最大，约高出 33%；比不做任何处理的方法平均高 46%，在 k=10 时最大，约高 60%。在 $g\text{-}means$ 度量上，ORD k-means 比本文的算法高了 2%，随机欠采样和 SMOTE 算法比本文的算法平均高了 1%，没做任

何处理的方法比 ICIKMDS 平均低了 2%。

表 4 各方法在 Dangdang_Book_4000 语料库上的实验结果对比

Table 4 Comparison of experimental results of each method on Dangdang_Book_4000

k	10		8		6		4	
	g-mean	*pr*	*g-mean*	*pr*	*g-mean*	*pr*	*g-mean*	*pr*
ICIKMDS	0.7571	0.823	0.7928	0.869	0.7920	0.865	0.7992	0.851
RDS	0.7693	0.776	0.7996	0.801	0.8052	0.830	0.8183	0.839
SMOTE	0.7736	0.491	0.7918	0.538	0.7944	0.607	0.8155	0.696
NODO	0.7474	0.227	0.7674	0.327	0.7799	0.410	0.8098	0.583
ORD k-means	0.803	0.785	0.7925	0.808	0.8015	0.800	0.8210	0.822

表 5 各方法在 ChnSentiCorp_htl_ba_4000 语料库上的实验结果对比

Table 5 Comparison of experimental results of each method on ChnSentiCorp_htl_ba_4000

k	10		8		6		4	
	g-mean	*pr*	*g-mean*	*pr*	*g-mean*	*pr*	*g-mean*	*pr*
ICIKMDS	0.8065	0.814	0.7991	0.825	0.7983	0.824	0.8338	0.860
RDS	0.8289	0.776	0.8202	0.768	0.8192	0.773	0.8363	0.815
SMOTE	0.8062	0.563	0.8088	0.602	0.8186	0.659	0.8360	0.729
NODO	0.7647	0.330	0.7804	0.402	0.7983	0.510	0.8320	0.639
ORD k-means	0.836	0.783	0.8245	0.775	0.8445	0.812	0.8405	0.809

在语料 ChnSentiCorp_htl_ba_4000 上，本文提出的 ICIKMDS 算法的 *pr* 值比 ORD k-means 平均高 4%；比随机欠采样平均高 5%，其中在 $k=8$ 时最大，要高 6%；比 SMOTE 算法平均高 19%，在 $k=10$ 时最大，约高出 25%；比不做任何处理的方法平均高 36%，在 $k=10$ 时最大，高了 48%。在 *g-means* 度量上，ORD k-means、随机欠采样和 SMOTE 算法比本文提出的算法均高 1%～2%，没做任何处理的方法比本文方法低。

因此，本文提出的算法在小类的分类准确度上有了很大的提高，同时在整体性能上与其他算法的差距不大，这主要是因为本文算法通过欠采样找到了大类中最具代表性的样本，但同时是以降低对大类样本的准确率来提高对小类样本分类的准确率。由于现实中一般要求对小类样本分类的准确率较高，但同时整体性能也不能太差，因此所提算法具有使用价值。

接下来将 3 个数据集上的数据整合到一起，取它们的平均值来测试本文算法的效果。将 *pr* 度量和 *g-mean* 度量放在一个表中，结果如表 6 所列。

表 6 3 个数据集上各种方法的平均 *pr* 值实验结果对比

Table 6 Comparison of average *pr* values of various methods on three data sets

k	10		8		6		4	
	g-mean	*pr*	*g-mean*	*pr*	*g-mean*	*pr*	*g-mean*	*pr*
ICIKMDS	0.797	0.808	0.810	0.831	0.808	0.837	0.83	0.856
RDS	0.807	0.770	0.822	0.780	0.826	0.804	0.841	0.829
SMOTE	0.792	0.557	0.811	0.607	0.820	0.671	0.834	0.729
NODO	0.765	0.335	0.783	0.404	0.799	0.508	0.829	0.638
ORD k-means	0.828	0.778	0.822	0.783	0.835	0.808	0.837	0.813

从表 6 可以看出，本文提出的算法的 *pr* 值比其他算法都高，在整体性能上本文提出的算法比 ORD k-means 和随机欠采样表现得略差，但比其他算法好。

图 1、图 2 直观地显示了表 6 中的结果。

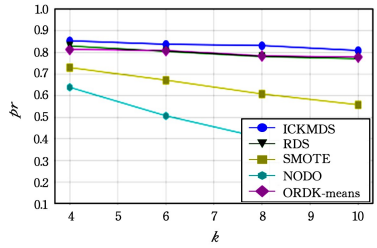

图 1 3 个数据集上各种方法的平均 *pr* 值对比

Fig. 1 Average *pr* value comparison of various methods on three datasets

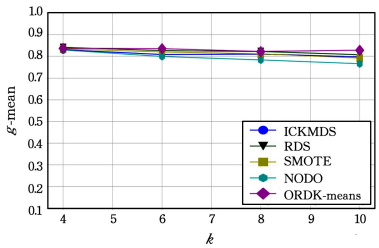

图 2 3 个数据集上各种方法的平均 *g-mean* 值对比

Fig. 2 Average *g-mean* value comparison of various methods on three datasets

从图 1 和图 2 可以看出，本文算法在 *pr* 性能度量上明显好于其他算法，而且比较稳定；而在整体性能上的表现与其他算法的区别不大，且好于不做任何处理的方法。因此，本文提出的算法在小类样本的分类上表现出了较好的效果，且整体性能几乎未受影响，具有较好的实用性。

结束语 针对实际应用中数据集的不均衡问题，本文用聚类的思想在多类样本上欠采样，同时对聚类容易陷入局部最优的问题给予改进，提出了初始质心的求解过程，这样就能使数据集在分布上达到均衡，从而提高分类器的性能，特别是小类样本的分类准确率。但是，本文算法增加了聚类的时间复杂度，而且对文本分类的性能并不优于其他算法；同时，由于本文提出的方法只是在大类样本集中采用欠采样的技术来解决数据集的分布不均衡问题，在小类样本集中采用过采样技术来较大程度地提高分类器的整体性能是未来的主要工作。

参 考 文 献

[1] ZHAI Y, YANG B R, QU W. Overview of Imbalanced Data Mining [J]. Computer Science, 2010, 37(10): 27-32. (in Chinese)
翟云，杨炳儒，曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学，2010，37(10)：27-32.

[2] VISA S, RALESCU A. Issues in Mining Imbalanced Data Sets-A Review Paper[C]// Sixteen Midwest Artificial Intelligence & Cognitive Science Conference. 2005: 67-73.

[3] XIE N N. Text classification algorithm based on unbalanced data

set [D]. Chongqing: Chongqing University, 2013. (in Chinese)
谢娜娜. 基于不均衡数据集的文本分类算法研究[D]. 重庆：重庆大学，2013.

[4] YOU M Y, CHEN Y, LI G Z. New feature selection algorithm in unbalanced problem Im-IG [J]. Journal of Shandong University (Engineering Science), 2010, 40(5): 123-128. (in Chinese)
尤鸣宇，陈燕，李国正. 不均衡问题中的特征选择新算法 Im-IG [J]. 山东大学学报(工学版)，2010，40(5)：123-128.

[5] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[M]// Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. Berlin: Springer, 2005: 878-887.

[6] FAN X N. Data imbalance classification problem[D]. Hefei: University of Science and Technology of China, 2011. (in Chinese)
范先念. 数据不平衡分类问题研究[D]. 合肥：中国科学技术大学，2011.

[7] NÚÑEZ H, GONZALEZ-ABRIL L, ANGULO C. Improving SVM Classification on Imbalanced Datasets by Introducing a New Bias[J]. Journal of Classfication, 2017, 34(3): 427-443.

[8] ABDELLATIF S, BEN HASSINE M A, BEN YAHIA S, et al. ARCID: A New Approach to Deal with Imbalanced Datasets Classification[C]// International Conference on Current Trends in Theory and Practice of Informatics. 2018: 569-580.

[9] ZHANG Y, FU P P, ZHANG Y T. Large scale data classification based on hierarchical clustering and resampling[J]. Journal of Computer Applications, 2013, 33(10): 2801-2803. (in Chinese)
张永，浮盼盼，张玉婷. 基于分层聚类及重采样的大规模数据分类[J]. 计算机应用，2013，33(10)：2801-2803.

[10] CHAWLA N, BOWYER K, HALL L, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[11] LONGADGE R, DONGRE S. Class Imbalance Problem in Data Mining: Review[C]// International Journal of Computer Science and Network, 2013, 2(1).

[12] GUO Y W, LIU X X. Study on the Method of Information Gain Feature Selection in Chinese Text Classification [J]. Computer Engineering and Applications. 2012, 48(27): 119-122. (in Chinese)
郭亚维，刘晓霞. 中文文本分类中信息增益特征选择方法的研究[J]. 计算机工程与应用. 2012，48(27)：119-122.

[13] YIN L Z, GE Y, XIAO K L, et al. Feature selection for high-dimensional unbalanced data[J]. Neurocomputing, 2013, 105(1): 3-11.

[14] BAI P, ZHANG X B, ZHANG B, et al. Support vector machine theory and engineering application examples [M]. Xi'an: Xi'an University of Electronic Science and Technology Press, 2008: 28-30. (in Chinese)
白鹏，张喜斌，张斌，等. 支持向量机理论及工程应用实例[M]. 西安：西安电子科技大学出版社，2008：28-30.

[15] AKBANI R, KWEK S, JAPKOWICZ N. Applying Support Vector Machines to Imbalanced Datasets[C]// European Conference on Machine Learning. Springer Berlin Heidelberg, 2004.

[16] DIAO C X. W-SVM Model for Unbalanced Data Set Classification [D]. Hefei: Hefei University of Technology, 2012. (in Chinese)
刁翠霞. 面向不均衡数据集分类的 W-SVM 模型[D]. 合肥：合肥工业大学，2012.

[17] GUO H Y, VIKTOR H L. Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1): 30-39.