k-means 和逻辑回归混合策略的不平衡类学习方法

邬长安[1],郑桂荣[1],孙艳歌[1,2],郭华平[1]

[1](信阳师范学院 计算机与信息技术学院,河南 信阳 464000)
[2](北京交通大学 计算机与信息技术学院,北京 100044)
E-mail: hpguo_cm@163.com

摘 要: 不平衡类问题在现实生活中普遍存在,表现为一个类的实例数明显多于另一个类的实例数,其类分布不平衡这一特征导致了传统的分类方法不能很好地处理该类问题. 本文将 *k*-means 和逻辑回归模型相结合,提出一种叫做 ILKL(Imbalanced Learning based on K-means and Logistic Regression) 的算法处理不平衡类问题. 首先,ILKL 使用聚簇方法将多数类划分成一个个子簇,以重新平衡数据集,然后在相对的平衡的数据集上学习逻辑回归模型. UCI 数据集上的实验结果显示,与传统方法相比,本文方法在召回率、*g*-mean 和 *f*-measure 等指标上表现出更好的性能.

关 键 词: 不平衡类; *k*-means; 逻辑回归; 聚方法
中图分类号: TP18 **文献标识码**: A **文 章 编 号**: 1000-1220(2017)09-2119-06

Imbalanced Learning Based on K-means and Logistic Regression Mixed Strategy

WU Chang-an[1],ZHENG Gui-rong[1],SUN Yan-ge[1,2],GUO Hua-ping[1]

[1](College of Computer and Information Technology,Xinyang Normal University,Xinyang 464000,China)
[2](College of Computer and Information Technology,Beijing jiaotong University,Beijing 100044,China)

Abstract: Class-imbalance is very common in real world,which is usually characterized as having more instances of one class than another. Because of imbalanced class distribution,so the traditional classification method doesn't work well on imbalanced class. This paper combines k-means with logistic regression model,and proposes a novel method named ILKL(Imbalanced Learning based on *K*-means and Logistic Regression) for the imbalanced problem. Firstly,ILKL applies clustering method to divide majority class into small clusters to rebalance the dataset for the learning of logistic regression model. The experiments on UCI data sets shows that the proposed method has a significant superiority on measures of recall,*g*-mean and *f*-measure when compared with other state-of-the art methods.
Key words: class-imbalance; *k*-means; logistic regression; cluster algorithms

1 引 言

不平衡类问题也称为稀有类问题. 在二元分类问题中,其表现为一个类的实例(多数类或负类) 明显多于另一个类(少数类或正类) 的实例数[1-3]. 在实际应用中,正确识别少数类实例往往更有价值. 例如,信用欺诈检测中,合法交易远远多于欺诈交易,但正确识别欺诈交易更有意义. 然而,传统的分类方法(如 C4.5、朴素贝叶斯和神经网络等) 假设数据集中各类的实例数目相当,以达到高准确率的分类目的,这导致少数类实例经常被忽略甚至被误分为多数类实例[4,5].

针对以上问题,已经提出了很多处理方法,这些方法主要可归类为如下两类: 基于数据预处理策略和基于算法策略. 主要存在三种数据预处理策略:

1) 通过抽样的方法重新平衡训练数据集[6],如欠抽样[7]、过抽样技术[8];

2) 自主选择更有价值的子训练集学习模型,使用其它实例提高模型的性能[9-11];

3) 根据分类错误代价,加权数据空间[12]. 基于算法处理不平衡类问题的方法是通过调整算法(或相应的目标函数) 使得学习到的模型更倾向于正确分类少数类实例,如两阶段规则学习方法[13] 和单类学习方法. 如 CAO 等[14] 构建了代价敏感神经网络采用粒子群算法优化误分类的代价,从而获取特征子集和结构参数. Alejo 等[15] 在训练集中将 FBP 神经网络和代价函数两种方法相结合用于不平衡类的分类中.

与以上做法不同,本文将 *k*-means 和逻辑回归同时应用于不平衡类分类问题中,提出了一种叫做 ILKL(Imbalanced Learning based on *K*-means and Logistic Regression) 的不平衡类学习算法. 该方法使用 *k*-means 聚簇方法将多数类划分成一个个子簇,分别将每个子簇关联新的类标号,将这些小簇视为一个个少数类实例,然后在重新平衡的数据集下建立逻辑回归模型进行分类. 该做法基于如下原因: *k*-means 有利于增强数据集中少数类实例的线性可分性(逻辑回归为线性分类模型) . 实验显示,与传统逻辑回归、欠抽样逻辑回归以及过

收稿日期: 2016-07-18 收修改稿日期: 2016-09-02 基金项目: 国家自然科学基金项目(61501393) 资助; 河南省科技厅科技计划项目(162102210310) 资助; 河南省教育厅科技研究重点项目(15A520026) 资助; 信阳师范学院研究生科研创新基金重点项目(2015KYJJ39) 资助.
作者简介: **邬长安**,男,1959 年生,硕士,教授,CCF 高级会员,研究方向为数字图像处理、模式识别、数据挖掘; **郑桂荣**,女,1989 年生,硕士研究生,研究方向为数据挖掘; **孙艳歌**,女,1981 年生,博士研究生,讲师,CCF 会员,研究方向为机器学习、计算机视觉; **郭华平**,男,1982 年生,博士,讲师,CCF 会员,研究方向为数据挖掘、人工智能.

抽样逻辑回归相比,ILKL 在召回率、g-mean 和 f-measure 上都表现出明显优势.

论文剩余部分组织如下: 第 2 节首先介绍聚簇方法和逻辑回归模型,然后将二者有机结合用于不平衡类分类问题中; 第 3 节讨论参数学习; 第 4 节分析实验结果; 第 5 节总结.

表 1 ILKL 的算法流程

Table 1 Flowchart of ILKL algorithm

学习阶段:
输入:
D: 训练集
K: 划分的簇数
L: 逻辑回归学习方法
输出: 学习到的逻辑回归模型
过程:
1. $D_{maj}=\phi$; //负类(多数类)实例集合
2. $D_{min}=\phi$; //正类(少数类)实例集合
3. **for** each $\mathbf{x}_i \in$ D **do** //$\mathbf{x}_i$ 表示类标号未知样本
4. **if**(y_i = maj) //y_i 是多数类
5. $D_{maj}=D_{maj}\cup\{\mathbf{x}_i\}$; //将类标号未知样本放入多数类实例集中
6. **else**
7. $D_{min}=D_{min}\cup\{\mathbf{x}_i\}$; //将类标号未知样本放入少数类实例集中
8. **end for**
9. S = k-means(D_{maj}, K); //使用 k-means 划分多数类为 k 个小簇, S 表示 k 个簇的集合
10. **for** each $S_j \in S$ **do** //对于每个子簇 S_j
11. **for** each $\mathbf{x}_i \in S_j$ **do**
12. $y_i=j$; //重新设置子簇的每个为 j
13. **end if**
14. **end for**
15. **for** each $\mathbf{x}_i \in D_{min}$ **do**
16. $y_i=K+1$; //重新设置少数类实例的类标号为 $K+1$
17. **end for**
18. $D=D_{min}$;
20. **for** each $S_i \in S$ DO
21. $D=D\cup S_i$ //合并处理后的实例集合
22. **end for**
23. $\boldsymbol{M}$ = L(D); //使用逻辑回归学习分类模型
24. **return** $\boldsymbol{M}$

预测
输入:
$\boldsymbol{M}$: 学习到的逻辑回归模型
$\mathbf{x}$: 未知的,待预测的实例
K: 簇的数目
输出:
样本 $\boldsymbol{x}$ 的预测类标号
过程:
1. $P=M(\mathbf{x})$; //每个类的分布
2. $p_{maj}=\max(P,1,K)$ //待预测类为多数类,则其具有高置信度
3. $p_{min}=P_K+1$ //待预测类为多数类,则其具有高置信度
4. p = normalize($[p_{maj}, p_{min}]$); //规范化后验概率
5. **retun** p;

2 基于 k-means 和逻辑回归混合策略的不平衡类学习算法

2.1 算 法

现有聚类算法有划分方法[17] 和层次方法[18] 等,本文采用的是 k-means 划分方法,其算法思想是: 首先从聚类对象中随机选出 k 个对象作为类簇的质心,对剩余每个对象,根据它们分别到这 k 个质心的距离,将它们划分到最相似的簇(k-means 用欧式距离来量化相似度). 然后重新计算质心位置,不断反复上述过程,直到准则函数收敛为止[16]. 这些簇满足以下条件: (1) 每簇都有一个中心质点,(2) 在 k-means 算法中,每个对象都能找到属于自己的簇[17], 每一个质点都是它周围所有点坐标的算术平均值.

逻辑回归是一种经典的概率统计分类模型[20],大多数无约束最优化技术都可以应用到逻辑回归的求解过程中. 对于两类问题,逻辑回归定义如下:

$$p(y=j\mid\mathbf{x}_i)=\frac{\exp(w_i^T\mathbf{x}_i+b_i)}{1+\sum_{l=1}^{M-1}\exp(w_l^T\mathbf{x}_i+b_l)},j=1,\cdots,M-1$$

$$p(y=M\mid\mathbf{x}_i)=\frac{1}{1+\sum_{l=1}^{M-1}\exp(w_l^T\mathbf{x}_i+b_l)} \quad (1)$$

定义一个数据集 $D=\{(\mathbf{x}_i,y_i)\mid i=1,2,\cdots,N\}$,其中 $y_i\in\{1,2,\cdots,M\}$,y_i 是实例 $\boldsymbol{x}_i$ 的类标号,其对数似然函数为:

$$L(\mathbf{w})=-\ln p(\mathbf{y}\mid\mathbf{w})=\sum_{j=1}^{M}\sum_{\mathbf{x}_i\in C_j}\ln p(y=j\mid\mathbf{x}_i) \quad (2)$$

目标函数(2) 作为代价函数来估计模型的参数,本文就是用这个方法来处理不平衡类问题. 本文采用 k-means 和逻辑回归模型相结合的学习算法—ILKL,在达到高准确率分类的同时也规避了将少数类误分为多数类或直接忽略少数类的风险. 算法伪代码如表 1 所示,下面给出算法的具体描述.

该算法首先用 k-means 对多数类聚簇处理,然后在重平衡的数据集上学习逻辑回归. 其具体操作可归为以下两个阶段: 在学习阶段,ILKL 首先划分训练集 D 为多数类实例集 D_{maj}和少数类实例集 D_{min}(第 1 ~ 8 行). 然后,ILKL 使用聚类方法 k-means 划分 D_{maj}为 k 个簇(第 9 行),并为每个簇 S_j 关联一个类标号 j(设置簇中的每个实例的类标号为 j)(第 10 ~ 14 行). 同理,ILKL 标记少数类实例用 D_{min}. ILKL 结合上述标记的实例使用相应的学习方法 M 使训练集重新平衡,便于学习逻辑回归模型 M (第 18-23 行),然后返回 M; 在预测阶段,ILKL 使用学习到的模型预测实例属于每个类标号的置信度,并设置最高置信度(非关联到少数类的置信度) 为实例属于多数类的置信度. 然后 ILKL 将其规范化,使可信度总和等于 1,并返回其规范分布. ILKL 学习过程的时间复杂度取决于学习簇和逻辑回归算法本身. 在预测阶段,ILKL 将预测结果映射到多数类和少数类上.

与传统的分类算法不同,论文从数据的角度研究不平衡类分类问题,该算法首先使用 k-means 将多数类数据集划分成一个个小簇并关联新的类标号,以期望在重新平衡的训练集上学习逻辑回归模型. 实验结果表明本文算法可以提高逻辑回归模型在不平衡类分类问题中的范化性能,结果分析详见 4.3.

2.2 讨 论

不同于之前的不平衡数据集预处理方法,ILKL 使用聚类方法对多数类进行划分,在类分布均匀的训练数据集上学习逻辑回归模型,使其达到线性可分的目的,实验显示这种方法比重采样技术更有效. 不平衡类问题还可细分为类内不平衡,

和类间不平衡类内不平衡是指一个类的实例数明显多于另一个类,类内的不平衡使得不平衡类问题研究更加复杂化[19]. 为了解决这一难题,本文采用 k-means 和逻辑回归混合的策略,首先在源数据集上用 k-means 将多数类数据集划分成多类小簇(如图 1 所示,整个图视为一个不平衡数据集,五角星代表多数类实例,四角星代表少数类实例) ,并将每个簇关联一个新的类标号,然后,在重平衡的数据集上学习分类模型.

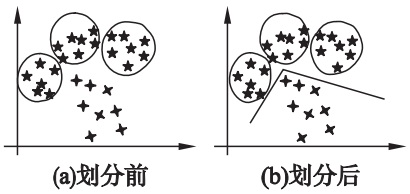

图 1　多数类划分前后,学习到的分类器

Fig. 1　Classifier learned on data set (a) before partition majority class (b) after partition majority class

如图 1 给出了在划分多数类前后,所学习到的分类模型. 从图 1(b) 显示,划分后学习到的线性模型能正确识别所有少数类实例. 这样就可以增强不平衡数据集的线性可分性,有利于提升逻辑回归在不平衡数据集上的泛化性能,大大降低少数类被误分或忽略的概率,更进一步提高了分类准确率.

3　参数学习

本文将 ILKL 用于不平衡类分类问题中,先对多数类聚簇,以期所有类实例数相当. 算法 1 中存在一个输入参数 k (聚簇数目) ,本文选择 3 个数据集(letter、sick 和 vowel) 作为代表数据集,数据集详细信息如表 2 所示. 图 2 显示参数 k 的

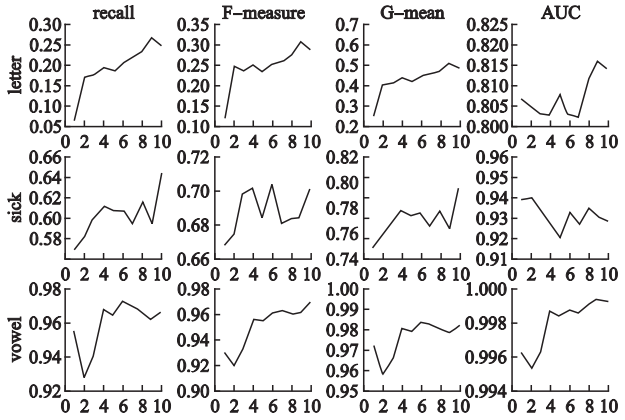

图 2　参数 k 对 ILKL 算法性能的影响

Fig. 2　Impact of k on ILKL

取值对 ILKL 算法的 recall、g-mean、f-measure 和 AUC 的影响. 大图中,从上到下每行分别显示 letter、sick 和 vowel 三个数据集的结果,从左往右每列分别对应不同的聚簇数目下 recall、g-mean、f-measure 和 AUC 分别在三个数据集上的实验结果. 每幅小图中的横坐标代表 k 的取值从 1 增加到 10(步长为 2) ,纵横坐标分别代表在三个数据集上相应的实验指标值. 图 2 显示,每对应一个 k 值需要在一个数据集上执行 100 次实验,获得 100 个结果,取这 100 个实验结果的平均值作为最终结果,平均结果体现为图 2 中的点坐标. 另外需要注意的是,由于相应的标准差结果显示不清晰,因此在 letter 这个数据集上做了 10 次验证实验.

如图 2 所示,随着聚簇数目 k 值的增加可发现如下结果规律: (1) 在 letter 数据集上 recall、f-measure 和 g-means 的曲线快速上升; (2) 在 sick 和 vowel 这两个数据集上曲线的变化不稳定; (3) AUC 曲线的振动很大. 结果(1) 和(2) 表明 ILKL 通过划分多少类实例能有效提高逻辑回归在不平衡类问题上的学习性能,结果(3) 中 AUC 指标值表明 ILKL 不能提高逻辑回归在不平衡问题上的学习能力. 逻辑回归是一个判别模型尽管它能对一个待预测的实例输出一组连续的数值,但 AUC 还是更适合做生成模型 (如贝叶斯) 的评估指标,所以实验中,未比较算法 AUC 性能. 根据图 2 结果,设置聚簇数目 k 为多数类与小数类数目的比率(下取整) ,若比率超过 10,则设置聚簇数目 k = 10.

4　实验结果与分析

4.1　数据集和实验设置

我们从机器学习库 UCI 里随机选取 12 个数据集[1] 并对对每个数据集采用一次十折交叉验证来分析算法的性能. 数据集 sick、breast-cancer、credit-g、waveform-5000、colic. ORIG、diabetes,and breast-w 等均为两类不平衡数据集,其它两类不平衡数据集来源于多数类数据集,将这两类中的一类作为正类另一类作为负类[21],百分比只占 0. 0205 (高度不平衡) 到 0. 3448 (轻微不平衡) 的称为少数类. 数据集的详细信息如表 2 所示,其中 Per,Exs,Attrs 和 Cls 分别表示少数类百分比,数据集大小,属性个数和个别类数目,本文全部实验都是在 weka 平台中完成的.

表 2　实验数据集

Table 2　Description of experimental data sets

	数据集	Per	Exs	Attrs	Cls
d1	letter	0. 0367	20000	17	26
d2	sick	0. 0612	3772	30	2
d3	anneal. ORIG	0. 0746	898	39	6
d4	balance-scale	0. 0786	625	5	3
d5	vowel	0. 0909	990	11	11
d6	vehicle	0. 1552	846	19	4
d7	breast-cancer	0. 2572	286	10	2
d8	credit-g	0. 3000	300	21	2
d9	waveform-5000	0. 3306	5000	41	3
d10	colic. ORIG	0. 3370	368	28	2
d11	diabetes	0. 3490	768	9	2
d12	breast-w	0. 3448	699	10	2

为了评估 ILKL 的性能,本文把 ILKL 分别与 LR-US、LR-SMOTE 和 LR 相比较,其中 LR 表示直接用传统逻辑回归处理不平衡类分类问题,LR-US(LR-SMOTE) 表示先对数据集进行欠抽样(过抽样) 技术处理后再运行 LR. 此外,设置聚簇数目 k 为多数类与小数类数目的比率(下取整) ,若比率超过

[1] http: //www. ics. uci,edu/mlearn/MLRepository. html

10,则设置聚簇数目 $k=10$.

4.2 不平衡类分类问题评估标准

评价指标是评估一个算法是否有效的重要手段,传统方法中最常用的就是精度指标. 对于二元分类问题,少数类记为正类,多数类记为负类,表 3 概括了分类模型正确和不正确的实例数目的混淆矩阵,因此精度可定义为:

表 3 混淆矩阵

Table 3 Confusion marix

	预测的正类(+)	预测的负类(-)
实际正类(+)	*TP*	*FN*
实际负类(-)	*FP*	*TN*

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

然而,在不平衡分类问题中仅用精度作为度量指标是不够的,还需要其他评价指标如: 召回率、f-measure 和 g-mean 等作为评估算法性能好坏的重要指标. 定义如下:

$$recall=\frac{TP}{TP+FN}\quad precision=\frac{TP}{TP+FP}$$
$$f-measure=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision}$$
$$g-mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TF}{TF+FN}} \tag{4}$$

其中 β 是调整精度的相对重要性和召回率的一个系数(通常 $\beta=1$). 从方程(4)可以看出 *f*-measure 和召回率结合来衡量的有效性分类的加权比例,召回率和精度由问题本身确定,*f*-measure 代表召回率和精度之间的调和平均数. *f*-measure 和 *g*-mean 是评估算法性能的两个重要指标. 具体来说,*g*-mean 是以类的准确分类为基础,*g*-mean 值越大说明算法性能越好.

本文使用召回率、*f*-measure 和 *g*-mean 来评估算法在不平衡数据集上的分类性能. 仅仅通过准确率是不能够充分评估一个算法的好坏,一个好的算法需要在提高召回率、f-measure 和 *g*-mean 的同时不降低准确率.

4.3 实验结果评估分析

实验结果显示 LR-US、LR-SMOTE 和 LR 分类效率显然不及 ILKL. 表 4-表 6(见下页)显示在准确率、召回率、*g*-mean、*f*-measure 四个评估指标中,至少有三个指标(召回率、*g*-mean 和 *f*-measure)能说明本文算法更有效. 表中各个实心圆(空心圆)表明 ILKL 显著优于(略优于)其他算法,本文使用统计意义下置信度为 95% 的 t-测试比较算法的分类性能. 表最后一行显示各个算法在 12 个数据集上的平均水平.

通过对比实验发现,与其它三种方法相比,尽管 ILKL 在准确率这一评估指标上不占优势,但其结果仍在在可接受范

表 4 ILKL、LR-US、LR-SMOTE 和 LR 的召回率及标准差

Table 4 Recall and standard error of ILKL, LR-US, LR-SMOTE and LR

dataset	ILKL	LR-US	LR-SMOTE	LR
d1	0.9292 ±0.0314	0.7562 ±0.0313 •	0.1608 ±0.0400 •	0.0763 ±0.0296 •
d2	0.8614 ±0.0574	0.8875 ±0.0743	0.7096 ±0.0880 •	0.6317 ±0.0910 •
d3	1.0000 ±0.0000	1.0000 ±0.0000	0.8976 ±0.1188 •	0.8976 ±0.1188 •
d4	0.7900 ±0.2234	0.4850 ±0.2082 •	0.0000 ±0.0000 •	0.0000 ±0.0000 •
d5	0.9889 ±0.0351	0.9889 ±0.0351	0.9667 ±0.0750	0.9778 ±0.0468
d6	0.9562 ±0.0624	0.9547 ±0.0369	0.9600 ±0.0516	0.9550 ±0.0497
d7	0.6264 ±0.2058	0.5931 ±0.1419	0.4778 ±0.1811	0.3597 ±0.1723 •
d8	0.6967 ±0.0895	0.7100 ±0.0802	0.5933 ±0.1004 •	0.4900 ±0.0802 •
d9	0.9292 ±0.0209	0.9341 ±0.0200	0.9244 ±0.0207	0.8603 ±0.0316 •
d10	0.6867 ±0.1595	0.7263 ±0.1343	0.6712 ±0.1230	0.4782 ±0.1501 •
d11	0.7284 ±0.1052	0.7044 ±0.1154	0.7346 ±0.1057	0.5698 ±0.1290 •
d12	0.9630 ±0.0448	0.9587 ±0.0389	0.9630 ±0.0448	0.9505 ±0.0541
average	0.8460	0.8110	0.6768	0.6039

表 5 ILKL、LR-US、LR-SMOTE 和 LR 的 f-measure 值及标准差

Table 5 F-measure and standard error of ILKL, LR-US, LR-SMOTE and LR

dataset	ILKL	LR-US	LR-SMOTE	LR
dataset	ILKL	LR-US	LR-SMOTE	LR
d1	0.2992 ±0.0207	0.1824 ±0.0123 •	0.2531 ±0.0575	0.1374 ±0.0510 •
d2	0.6795 ±0.0578	0.4547 ±0.0399 •	0.7154 ±0.0843	0.7036 ±0.0756
d3	0.8557 ±0.0884	0.7410 ±0.1537	0.8251 ±0.1083	0.8205 ±0.1093
d4	0.2940 ±0.0738	0.1245 ±0.0556 •	0.0000 ±0.0000 •	0.0000 ±0.0000 •
d5	0.9836 ±0.0265	0.7081 ±0.0823 •	0.9234 ±0.0597 •	0.9420 ±0.0383 •
d6	0.9563 ±0.0624	0.9547 ±0.0369	0.9500 ±0.0516	0.9550 ±0.0497
d7	0.4885 ±0.1194	0.4689 ±0.1365	0.4260 ±0.1506	0.4044 ±0.1783
d8	0.5722 ±0.0554	0.5970 ±0.0512	0.5748 ±0.0839	0.5409 ±0.0717 •
d9	0.8741 ±0.0172	0.8438 ±0.0222 •	0.8452 ±0.0221 •	0.8335 ±0.0247 •
d10	0.5733 ±0.0988	0.6162 ±0.0830	0.5769 ±0.0657	0.5038 ±0.1220
d11	0.6629 ±0.0620	0.6571 ±0.0633	0.6703 ±0.0603	0.6301 ±0.0893
d12	0.9550 ±0.0282	0.9510 ±0.0317	0.9550 ±0.0282	0.9501 ±0.0317
average	0.6829	0.6083	0.6429	0.6184

表 6　ILKL、LR-US、LR-SMOTE 和 LR 的 *g*-mean 值及标准差

Table 6　*G*-mean and standard error of ILKL, LR-US, LR-SMOTE and LR

dataset	ILKL	LR-US	LR-SMOTE	LR
dataset	ILKL	LR-US	LR-SMOTE	LR
d1	0.8812 ±0.0162	0.7531 ±0.0191 •	0.3972 ±0.0517 •	0.2707 ±0.0571 •
d2	0.8859 ±0.0332	0.8766 ±0.0362	0.8333 ±0.0533 •	0.7888 ±0.0567 •
d3	0.9854 ±0.0101	0.9666 ±0.0263	0.9348 ±0.0663 •	0.9341 ±0.0659 •
d4	0.7318 ±0.1051	0.4582 ±0.1133 •	0.0000 ±0.0000 •	0.0000 ±0.0000 •
d5	0.9932 ±0.0178	0.9519 ±0.0254 •	0.9758 ±0.0370	0.9836 ±0.0232
d6	0.9611 ±0.0316	0.9576 ±0.0176	0.9657 ±0.0281	0.9671 ±0.0273
d7	0.6034 ±0.1109	0.5838 ±0.1261	0.5586 ±0.1397	0.5320 ±0.1480
d8	0.6895 ±0.0471	0.7106 ±0.0437	0.6869 ±0.0677	0.6487 ±0.0569 •
d9	0.9157 ±0.0128	0.8969 ±0.0166 •	0.8965 ±0.0165 •	0.8794 ±0.0192 •
d10	0.6658 ±0.0875	0.6985 ±0.0694	0.6662 ±0.0606	0.6097 ±0.1010
d11	0.7360 ±0.0515	0.7317 ±0.0528	0.7433 ±0.0503	0.7027 ±0.0749
d12	0.9668 ±0.0219	0.9637 ±0.0235	0.9668 ±0.0219	0.9615 ±0.0268
average	0.8346	0.7941	0.7188	0.6898

围之内. 因此,在不平衡类分类问题中,准确率不是一个理想的评价指标,故本节没有给出准确率的比较结果. 为了更直观的显示 ILKL 的有效性,图 3-图 5 用四个数字表示四种算法的序[22]: 在一个数据集中,表现最佳的算法序为 1.0,次好的算法序为 2.0,以此类推. 若两种算法结果都最佳,则这两个算法序均为 1.5.

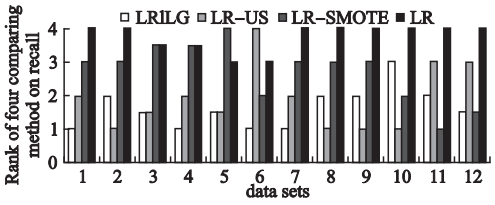

图 3　ILKL、LR-US、LR-SMOTE 和 LR 召回率上的序

Fig. 3　Ranks of ILKL, LR-US, LR-SMOTE and LR on recall

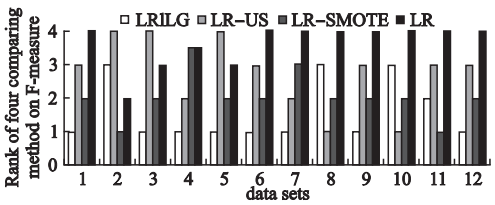

图 4　ILKL、LR-US、LR-SMOTE 和 LR 在 F-measure 上的序

Fig. 4　Ranks of ILKL, LR-US, LR-SMOTE and LR on F-measure

表 4 汇总了四种算法在召回率上的实验结果,对应的图 3 显示四个算法的序. 从表 4 知,12 个数据集中,有 9 个数据集的实验结果显示 ILKL 明显优于 LR,且 ILKL 的平均召回率为 0.2421,远远高于 LR(召回率在[0,1]之间) . 这些实验结果表明,ILKL 适合不平衡类问题,可以提高逻辑回归在正类(及不平衡类) 上的学习性能. 同样地,ILKL 还优于 LR-US 和 LR-SMOTE. 特别地,ILKL 分别在 2 个和 5 个数据集上明显优于 LR-US 和 LR-SMOTE. 此外,从图 3 可以看出在 12 个数据集上 ILKL、LR-US、LR-SMOTE 和 LR 的平均序分别为 1.6、1.9、2.7 和 3.8.

表 5 显示了 ILKL、LR-US、LR-SMOTE 和 LR 在 f-measure 上的性能,图 4 对应四种算法在 12 个数据集上的序. 从表 5 知,相比于 LR-US、LR-SMOTE 和 LR,ILKL 显示了更好的性能. 尤其,ILKL 分别在 5 个、3 个、5 个数据集上分别优于 LR-US、LR-SMOTE 和 LR. 图 4 显示在 10 个、8 个、12 个数据集上 ILKL 分别优于 LR-US、LR-SMOTE 和 LR. 此外,ILKL 在 7 个数据集的 f-measure 值都领先.

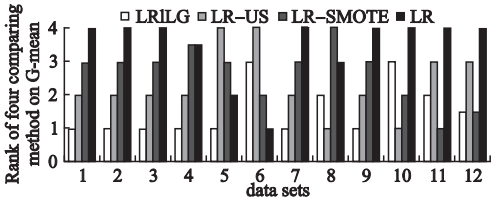

图 5　ILKL、LR-US、LR-SMOTE 和 LR 在 G-mean 上的序

Fig. 5　Ranks of ILKL, LR-US, LR-SMOTE and LR on G-mean

与 f-measure 类似,表 6 和图 5 分别显示 ILKL、LR-US、LR-SMOTE 和 LR 四种算法的 G-mean 值和序. 表 6 显示 ILKL 分别在 4 个、5 个、6 个数据集上的实验结果 分别优于 LR-US、LR-SMOTE 和 LR,图 5 显示四个算法相比,ILKL 在 10 个、9 个、11 个数据集上的序比另外三个算法高. ILKL 的平均序最高 1.5,紧接着 LR-US(2.3) 、LR-SMOTE(2.7) 、LR (3.5) .

5　总　结

本文提出了基于 *k*-means 的数据预处理的逻辑回归学习方法 ILKL(Imbalanced Learning based on K-means and Logistic Regression) 提高逻辑回归在不平衡类中的分类性能. 不同于传统的逻辑回归方法,该方法采用 k-means 聚类进行了数据预处理,将多数类实例划分为一个个子簇,然后用 ILKL 学习到的模型在重平衡的数据集上进行分类. 实验结果表明,本文数据预处理方法比传统逻辑回归,欠抽样逻辑回归,过抽样逻辑回归在召回率、g-mean、f-measure 和准确率上效果更优. 由于逻辑回归的易扩展性,因此本文未来工作结合不平衡具

体分布情况,在传统逻辑回归基础上修改其目标函数,使其新的算法具有更好的泛化性能.

References:

[1] He H, Garcia E A, Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[2] He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications [C]. Wiley-IEEE, 2013.

[3] Yao P, Wang Z, Jiang H, et al. Fault diagnosis method based on cs-boosting for unbalanced training data [J]. Journal of Vibration, Measurement & Diagnosis, 2013, 33(1): 111-115.

[4] Liu X Y, Wu J X, Zhou Z H. Learning imbalanced multiclass data with optimal dichotomy weight [C]. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2013: 478-487.

[5] Martin P D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation [J]. Journal of Machine Learning Technologies, 2011, 2(1): 37-63.

[6] Liu X Y, Wu J X, Zhou Z H. Exploratory under sampling for class imbalance learning [C]. In Proceedings of IEEE International Conference on Data Mining, 2006: 965-969.

[7] Varassin C G, Plastino A, Leitão H C D G , et al. Undersampling strategy based on clustering to improve the performance of splice site classification in human genes in database and expert systems applications [C]. In Proceeding of 24th IEEE International Workshop on DEXA, 2013: 85-89.

[8] 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.

[9] Zhu J, Hovy E H. Active learning for word sense disambiguation with methods for addressing the class imbalance problem [C]. EMNLP-CoNLL, 2007, 7: 783-790.

[10] Ertekin S. Adaptive oversampling for imbalanced data classification [C]. Information Sciences and Systems, 2013: 261-269.

[11] Mi Y. Imbalanced classification based on active learning smote [C]. Research Journal of Applied Sciences, 2013, 5(3): 944-949.

[12] Wang B X, Japkowicz N. Boosting support vector machines for imbalanced data sets [J]. Knowledge and Information Systems, 2010, 25(1): 1-20.

[13] Joshi M V, Agarwal R C, Kumar V. Mining needles in a haystack: classifying rare classes via two-phase rule induction [C]. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2001: 91-102.

[14] Cao P, Zhao D, Zaiane O R. A pso-based cost-sensitive neural network for imbalanced data classification [C]. In Trends and Applications in Knowledge Discovery and Data Mining, 2013: 452-263.

[15] Alejo R, Garcia V, Sotoca J M, et al. Improving the performance of the rbf neural networks trained with imbalanced samples [C]. Computational and Ambient Intelligence, 2007: 162-169.

[16] Nanda S J, Panda G. A survey on nature inspired metaheuristic algorithms for partitional clustering [J]. Swarm and Evolutionary Computation, 2014, 9(16): 1-18.

[17] Arthur D, Vassilvitskii S. k-means + +: the advantages of carefull seeding [C]. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007: 1027-1035.

[18] Reddy C K, Vinzamuri B. A survey of partitional and hierarchical clustering algorithms [C]. Data Dlustering: Algorithms and Applications, 2013: 87-110.

[19] Sun Y M, Kamel M S, Andrew K C. Cost-sensitive boosting for classification of imbalanced data [J]. Patter Recognition, 2007, 40(12): 3358-3378.

[20] Murphy K P. Machine learning: a probabilistic perspective [C]. MIT Press, Cambridge, Massachusetts London, England, 2012.

[21] Hulse J V, Khoshgoftaar T M, Napolitano A. Experimental perspectives on learning from imbalanced data [C]. In Proceedings of the Twenty-Fourth International Conference, 2007: 935-942.

[22] Demsar J. Statistical comparisons of classifiers over multiple data sets [J]. Journal of Machine Learning Research, 2006, 7(1): 1-30.