控制与决策
Control and Decision

一种基于GMM-EM的非平衡数据的概率增强算法

陈刚， 吴振家

引用本文:
陈刚, 吴振家. 一种基于GMM-EM的非平衡数据的概率增强算法[J]. 控制与决策, 2020, 35(3): 763-768.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.0802

文章编号: 1001-0920(2020)03-0763-06 DOI: 10.13195/j.kzyjc.2018.0802

一种基于GMM-EM的非平衡数据的概率增强算法

陈 刚†, 吴振家

(大连海事大学 理学院，辽宁 大连 116026)

摘 要: 非平衡数据的分类问题是机器学习领域的一个重要研究课题. 在一个非平衡数据里,少数类的训练样本明显少于多数类,导致分类结果往往偏向多数类. 针对非平衡数据分类问题,提出一种基于高斯混合模型-均值最大化方法(GMM-EM)的概率增强算法. 首先,通过高斯混合模型(GMM)与均值最大化算法(EM)建立少数类数据的概率密度函数;其次,根据高概率密度的样本生成新样本的能力比低概率密度的样本更强的性质,建立一种基于少数类样本密度函数的过采样算法,该算法保证少数类数据集在平衡前后的概率分布的一致性,从数据集的统计性质使少数类达到平衡;最后,使用决策树分类器对已经达到平衡的数据集进行分类,并且利用评价指标对分类效果进行评判. 通过从UCI和KEEL数据库选出的8组数据集的分类实验,表明了所提出算法比现有算法更有效.

关键词: 分类；非平衡数据；概率密度函数；GMM-EM；概率增强

中图分类号: TP273 **文献标志码:** A

An enhancing probability algorithm for imbalanced datasets based on GMM-EM

CHEN Gang†, WU Zhen-jia

(School of Science，Dalian Maritime University，Dalian 116026，China)

Abstract: The classification of imbalanced datasets has been recognized as a vital issue in the field of machine learning. In an imbalanced dataset, there are obviously fewer training examples of the minority class compared to the majority class so that the result of classification may be biased towards the latter. As a result, the classification performance of whole dataset has a tendency to be poor. Facing on the problem, an enhanced probability algorithm based on the Gaussian mixture model-expectation maximization (GMM-EM) method is proposed for imbalanced datasets. Firstly, the probability density functions (PDFS) of the minority class are obtained by using GMM and EM algorithms. Secondly, because original samples with high probability density have more powerful ability to generate new instances than low probability density samples according to the basic rule of probability theory, an enhanced probability algorithm is given based on PDF of the minority class. The algorithm ensures that the PDFs of the new balanced minority class are in accordance with the original minority class, and makes the minority class balanced in the sense of statistics. Finally, the proposed algorithm and other methods are applied together with a decision tree classifier for assessment. By choosing eight imbalanced datasets from UCI and KEEL repositories, experimental results show that the proposed algorithm is more effective than other methods.

Keywords: classification；imbalanced datasets；probability density functions；GMM-EM；enhancing probability

0 引 言

在现实生活中,研究人员经常会遇到许多非平衡问题,如在用户流失预测[1]、文献分类[2]、邮件检测[3]等. 在这些应用中,标准分类器的分类效果并不理想. 因此,非平衡数据是数据挖掘技术亟待解决的重要问题. 对于解决非平衡问题,可以从数据层面和算法层面进行处理.

数据层方法,也称为预处理方法,常用的解决方案就是对原始数据进行重采样. 这种方法有两种方式: 1) 过采样技术合成少数类样本; 2) 欠采样技术减少多数类样本. 对于过采样技术,目前的算法主要分为3大类: “SMOTE家族”[4-6]、核函数类[7-8]和其他技术[9-10]. 对于“SMOTE家族”, SMOTE[4]是基本算法,它是通过在几个相邻的少数类样本之间进行插值,生成新的少数类样本. 后来提出了E-SMOTE[5]算法是与遗传算法相结合,根据适应度寻找合适的

收稿日期: 2018-06-12；**修回日期:** 2018-11-19.
基金项目: 国家自然科学基金项目(11571056).
责任编委: 王凌.
†通讯作者. E-mail: chengang@dlmu.edu.cn.

特征. 目前最新设计的CMO-SMOTE[6]算法,对导致数据集不平衡的原因进行了理论分析,并据此生成新的样本. 对于核函数类算法,Kernel-ADASYN[7]算法通过核函数构建自适应的过采样分布来合成少数类样本. PDFOS[8]算法是核函数法与粒子群优化算法相结合的少数类再平衡方法. 对于其他采样技术,OUPS[9]算法是通过减少搜索和选择过程来选择近邻样本,而非随机性选择. ADASYN[10]方法主要针对学习困难程度不同的少数类样本进行过采样,需要学习的程度越困难,所需生成的样本越多;反之,则越少. 对于欠采样技术,该技术通过欠采样方法减少多数类的样本,从而使样本数量与少数类相同. 该方法本身存在一定的不足:容易忽略或剔除有价值的信息,导致分类不准确,该类算法可查阅相关资料获取.

对于算法层面,主要是对分类算法改进以适应非平衡数据,以便在对非平衡数据分类时更有效. 比如HSBagging[11]、MSE-BP[12]、KCBPFELM[13]、BoostedSVM[14]、EFSVM[15]等.

总之,数据层面和算法层面的方法在一定程度上都能够解决非平衡数据的分类问题. 但是,这些方法都存在着一定的不足,譬如:“SMOTE家族”仅考虑几何特性,不涉及样本的统计特征,只是在数量上达到平衡即可,导致新的已平衡的少数类的概率分布往往与原始少数类不同,从而产生质量较差的新样本. 对于kernel-ADASYN算法和PDFOS算法,这两种算法都使用高斯核对少数类的概率密度函数进行估计,虽然在算法上都考虑了少数类样本的概率分布,但这只是基于SVM的一种优化算法,本质上不是一种基于数据集的统计性质的方法. 而欠采样技术容易剔除潜在有价值的信息,因而不采用欠采样技术对非平衡数据平衡化;算法层面对非平衡数据不进行前期的预处理,只是通过改进算法以适用非平衡数据的分类,本质上并没有解决原始数据的非平衡问题.

鉴于此,本文提出一种基于非平衡数据统计性质的概率增强算法. 新算法涉及了少数类原始数据的统计特征,使得算法充分利用了原始数据的统计信息. 同时,遵循新平衡的少数类概率分布与原始少数类一致的原则,生成了高质量少数类样本.

1 高斯混合模型和均值最大化算法

1.1 高斯混合模型

由Chris等[16]在1999年提出的GMM是一个概率统计模型,它可以很好地表示参数空间中数据的空间分布和特征. GMM被定义为高斯密度函数的V-线性组合,即

$$P(\boldsymbol{x}|\boldsymbol{\theta})=\sum_{i=1}^{V}\alpha_i\mathcal{N}_i(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{C}_i),\tag{1}$$

其中:$\mathcal{N}_i(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{C}_i)(i=1,2,\dots,V)$是高斯分布,其均值向量为$\boldsymbol{\mu}_i$,协方差矩阵为$\boldsymbol{C}_i$,$\boldsymbol{x}$代表$p$维数据向量;混合比例$\alpha_i(i=1,2,\dots,V)$满足

$$\sum_{i=1}^{V}\alpha_i=1,\ 0\leqslant\alpha_i\leqslant 1.\tag{2}$$

令$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]^{\mathrm{T}}$是包含有$n$个独立同分布的$p$维样本数据集,$\boldsymbol{Z}=[z_1,z_2,\dots,z_n]^{\mathrm{T}}$表示分别对应于样本观测值的隐含变量,即$z_i$与观测值$\boldsymbol{x}_i$有关,则$z_i$表示$\boldsymbol{x}_i$的隐标签.

多元高斯混合模型由式(1)定义,$\mathcal{N}_i(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{C}_i)(i=1,2,\dots,V)$是少数类数据集的$p$维高斯密度,其参数是均值向量$\boldsymbol{\mu}_i$、协方差矩阵为$\boldsymbol{C}_i$、每个分量密度是$p$个变量的高斯函数,其形式如下:

$$f(\boldsymbol{x}|\boldsymbol{\theta}_i)=\frac{1}{(2\pi)^{\frac{p}{2}}}\frac{1}{|\boldsymbol{C}_i|^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{C}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\Big\}=\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i,\boldsymbol{C}_i).\tag{3}$$

其中:$\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{\mu}\in\boldsymbol{R}^P,\boldsymbol{C}\in\boldsymbol{R}^{P\times P})$,期望值为$E(\boldsymbol{x})=\boldsymbol{\mu},\boldsymbol{\mu}=[\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\dots,\boldsymbol{\mu}_V]$;协方差阵为$E[(\boldsymbol{x}-\bar{\boldsymbol{x}})(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathrm{T}}]=\boldsymbol{C},\boldsymbol{C}=[\boldsymbol{C}_1,\boldsymbol{C}_2,\dots,\boldsymbol{C}_V]$.

对于整个少数类数据集$\boldsymbol{X}$,其联合概率密度函数为

$$P(\boldsymbol{x}|\alpha,\boldsymbol{\mu},\boldsymbol{C})=\prod_{j=1}^{n}\Big\{\sum_{i=1}^{V}\alpha_i\mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{C}_i)\Big\}.\tag{4}$$

因此,上述问题的解决方法是对参数$\boldsymbol{\theta}=(\alpha_1,\alpha_2,\dots,\alpha_V,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\dots,\boldsymbol{\mu}_V,\boldsymbol{C}_1,\boldsymbol{C}_2,\dots,\boldsymbol{C}_V)$的最大似然估计.

1.2 均值最大化算法

EM算法是一种基于极大似然估计的优化算法,用于含有隐变量的概率模型参数的极大似然估计或者极大后验概率估计. 该算法可分为两步:第1步称为E步,它基于初始值的参数或之前的迭代值来计算似然函数的期望;第2步是M步,它将似然函数最大化,并转换为可获得的新参数值. EM算法是反复迭代的过程,直到上述两个步骤收敛.

对于一个给定的GMM,如果能够区分每个观测值$\boldsymbol{x}_k$是属于哪个高斯类,然后确定参数$(\alpha_1,\alpha_2,\dots,\alpha_V,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\dots,\boldsymbol{\theta}_V)$,其中$\boldsymbol{\theta}_i=(\boldsymbol{\mu}_i,\boldsymbol{C}_i)$,则分类算法很容易实现;相反,如果知道这些参数$(\alpha_1,\alpha_2,\dots,\alpha_V,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\dots,\boldsymbol{\theta}_V)$,则确定每个数据$\boldsymbol{x}_k$是最有可能属于

哪一高斯类. 然而,以上两种情况都是未知的,它们可以被看作是隐含的变量,所以EM算法可以应用于不完整的数据来估计这些参数.

当GMM的高斯类数为V时,样本大小为n,则E步的期望是

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{l=1}^{V}\sum_{i=1}^{n}\gamma(Z_{il})\log\alpha_l + \sum_{l=1}^{V}\sum_{i=1}^{n}\gamma(Z_{il})\log\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}_l). \tag{5}$$

其中:$\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}_l)$是第$l$个高斯类的概率分布;$\gamma(Z_{il})$是第$l$个高斯类的后验概率,其计算方法是

$$\gamma(Z_{il}) = \frac{\alpha_l\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}_l)}{\sum\limits_{j=1}^{V}\alpha_j\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\theta}_j)}. \tag{6}$$

为了计算$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$,可利用如下公式重新估计M步中GMM的每个参数:

$$\boldsymbol{\alpha}_l^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma(Z_{il}), \tag{7}$$

$$\boldsymbol{\mu}_l^{(t+1)} = \frac{1}{n\alpha_l^{(t+1)}}\sum_{i=1}^{n}\boldsymbol{x}_i\gamma(Z_{il}), \tag{8}$$

$$\boldsymbol{C}_l^{(t+1)} = \frac{1}{n\alpha_l^{(t+1)}}\sum_{i=1}^{n}\gamma(Z_{il})(\boldsymbol{x}_i-\boldsymbol{\mu}_l^{(t+1)})(\boldsymbol{x}_i-\boldsymbol{\mu}_l^{(t+1)})^{\mathrm{T}}. \tag{9}$$

这两个步骤形成了迭代关系,直到EM算法收敛.

2　基于GMM-EM的概率增强算法的提出

根据上述分析,本文利用GMM模型及EM算法,针对非平衡数据提出一种基于GMM-EM的概率增强算法,其具体内容如下.

step 1: 基于GMM-EM获取少数类的概率分布. 首先,利用GMM获取高斯型少数类数据集;其次,为了得到少数类的概率分布,使用极大似然估计方法对高斯混合模型的参数进行估计. 所处理的数据集具有隐含变量,因此选用EM优化算法来估计混合分布的参数.

step 1.1: 初始化. 由于对参数初始值的依赖,EM算法通常收敛于局部极小值. 为了提高EM算法的性能,使用如下方法作为每个参数的初始值:

$$\alpha_1^0 = \alpha_2^0 = \ldots = \alpha_V^0 = \frac{1}{V};$$

$$\boldsymbol{\mu}_l^0 = \frac{1}{n\alpha_l^0}\sum_{i=1}^{n}\boldsymbol{x}_i\gamma(Z_{il}),\ l=1,2,\ldots,V;$$

$$\boldsymbol{C}_l^0 = \frac{1}{n\alpha_l^0}\sum_{i=1}^{n}\gamma(Z_{il})(\boldsymbol{x}_i-\boldsymbol{\mu}_l^0)(\boldsymbol{x}_i-\boldsymbol{\mu}_l^0)^{\mathrm{T}},$$

$$l=1,2,\ldots,V.$$

step 1.2: E步. 该步主要是对后验概率的估计,后验概率$\gamma(Z_{il})$的计算方法由式(6)给出.

step 1.3: M步. 利用式(7)～(9),得到均值向量和协方差阵.

step 2: 确定每个少数类样本的概率密度权重. 通过使用GMM-EM对参数$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{C}_i)$进行估计,从而获得少数类的概率密度函数. 因此,每个少数类样本的概率密度权重可由下式计算得到:

$$\boldsymbol{\omega}_{\boldsymbol{x}_i} = \frac{\boldsymbol{f}_i}{\sum\limits_{i=1}^{n}\boldsymbol{f}_i},\ i=1,2,\ldots,n. \tag{10}$$

其中:$\boldsymbol{f}_i$为少数类样本$\boldsymbol{x}_i$的概率密度,n为少数类样本量.

step 3: 确定少数类样本的过采样数量. 根据高概率密度的样本生成新样本的能力比低概率密度的样本更强的性质,在step 2已经得到少数类每个样本的概率密度权重,使用下式计算每个少数类样本的过采样数量:

$$\boldsymbol{S}_{\boldsymbol{x}_i} = [\boldsymbol{\omega}_{\boldsymbol{x}_i}\ \ \Delta],\ i=1,2,\ldots,n, \tag{11}$$

使得$\sum\limits_{i=1}^{n}\boldsymbol{S}_{\boldsymbol{x}_i} = \Delta$,其中[ ]为取整函数,$\Delta = N_{\max} - N_{\min}$表示多数类与少数类样本数量之差. 生成新样本的每个少数类样本的过采样数量可根据式(11)计算得到,具体情形有以下3种:

1) 若$\boldsymbol{S}_{\boldsymbol{x}_i} = 0$,则令$\boldsymbol{S}_{\boldsymbol{x}_i} = 1$,使得原始少数类样本按照概率密度权重降序生成新样本,直到新的少数类样本总数量等于Δ;

2) 若$\sum\limits_{i=1}^{n}\boldsymbol{S}_{\boldsymbol{x}_i} > \Delta$,则剔除由低概率密度样本所生成的新样本直到新的少数类样本,总数量等于Δ;

3) 若$\sum\limits_{i=1}^{n}\boldsymbol{S}_{\boldsymbol{x}_i} < \Delta$,则由高概率密度样本再次按概率权重降序生成新样本直到新的少数类样本,总数量等于Δ.

step 4: 产生新样本. 为了保证新生成的样本在原样本点附近而不至于与相邻原样本所生成的新样本点产生交叉,定义$\boldsymbol{x}_j$为与少数类样本$\boldsymbol{x}_i$距离最近的少数类样本点,其度量采用欧氏距离$D(\boldsymbol{x}_j,\boldsymbol{x}_i)(j\neq i)$,即

$$\boldsymbol{x}_j = \min_{j\neq i}^{n} D(\boldsymbol{x}_j,\boldsymbol{x}_i),\ i,j=1,2,\ldots,n. \tag{12}$$

新生成的样本$\boldsymbol{x}_i^k$可由下式计算得到:

$$\boldsymbol{x}_i^k = \boldsymbol{x}_i \pm \frac{1}{2}rD(\boldsymbol{x}_j,\boldsymbol{x}_i),$$

$$r\in(0,1),\ k=1,2,\ldots,\boldsymbol{S}_{\boldsymbol{x}_i}. \tag{13}$$

其中: r为$(0,1)$之间的随机数; $\boldsymbol{x}_i^k=[x_i^{1(k)},x_i^{2(k)},\dots,x_i^{p(k)}]$表示少数类第$i$个原始样本生成第$k$个新样本.

step 5: 分类及评估. 为了验证本文算法的有效性, 将决策树作为基分类器, 对本文算法与其他方法进行比较(SMOTE, Borderline1SMOTE, ADASYN, Borderline2SMOTE). 此外, 评价指标(AUC、Acc、Sen、Spe、Pre和F_1值)用于评价本文算法与其他算法之间的分类性能. 本文算法伪代码如下所示.

算法1 基于GMM-EM的概率增强算法.

输入: N个训练样本数据集$D=\{\boldsymbol{x}_i,y_i\},i=1,2,\dots,N$, 其中$\boldsymbol{x}_i$代表第$i^{\text{th}}$样本, 且$y_i$是$\boldsymbol{x}_i$的类别标签;

输出: 产生新的少数类样本.

1. 1) 反复迭代直至收敛.
2. for each i, l do
3. set $\gamma(Z_{il})$;
4. end for
5. 更新以下参数: $\boldsymbol{\alpha}_l^{(t+1)},\boldsymbol{\mu}_l^{(t+1)},\boldsymbol{C}_l^{(t+1)}$.
6. 2) 产生新样本.
7. $\boldsymbol{f}_i$表示每一个p维少数类样本的概率密度值.
8. $\Delta=N_{\max}-N_{\min}$定义为新样本生成数.
9. n代表少数类样本数.
10. for $i=1$ to n do
11. 计算$\boldsymbol{\omega}_{\boldsymbol{x}_i}$,
12. 计算$\boldsymbol{S}_{\boldsymbol{x}_i}$,
13. if $\boldsymbol{S}_{\boldsymbol{x}_i}=0$ then $\boldsymbol{S}_{\boldsymbol{x}_i}=1$.
14. else if $\sum_{i=1}^{n}\boldsymbol{S}_{\boldsymbol{x}_i}>\Delta$ then 剔除由低概率密度样本产生的新样本直至数量等于Δ.
15. else if $\sum_{i=1}^{n}\boldsymbol{S}_{\boldsymbol{x}_i}<\Delta$ then 再次由高概率密度样本产生新样本直至数量等于Δ.
16. end if
17. end for
18. while 数据非平衡 do
19. for $i=1$ to n do
20. 计算$\boldsymbol{x}_j$;
21. 计算$\boldsymbol{x}_i^k$;
22. end for
23. end while

3 算例分析

利用不同行业、不同程度的不平衡比例和不同的特性, 从UCI和KEEL机器学习库中, 选择8组数据集进行实验. 实验的主要目的是将所提出的新算法与其他预处理方法进行比较. 同时, 5种过采样方法(SMOTE, Borderline1SMOTE, Borderline2SMOTE, ADASYN和新算法)将与决策树一起用于评估. 为了使训练更加准确, 实验采用了6次交叉验证的方法.

3.1 数据集的选择

在实验中, 选择8组不同的数据集, 如表1所示. 对于每一个数据集, 显示了数据集名称、样本数量、样本的属性个数和不平衡率(IR). 其中, IR定义为多数类样本与少数类样本的数量之比, 其计算公式为

$$\text{IR}=\frac{\text{多数类样本数量}}{\text{少数类样本数量}}. \tag{14}$$

表1 非平衡数据集的基本情况

数据集	样本量	属性数	IR
Wine(1-2)	130	13	1.20
Wine(2-3)	119	13	1.48
Glass(1-3)	87	9	4.12
Glass(1-5)	83	9	5.38
Pima_indians	768	8	1.87
Banknote	1 372	4	1.25
New_thyroid	185	5	4.29
Yeast1	1 484	8	2.46

3.2 参数估计结果

将样本的均值与方差作为EM算法的初值, 对少数类高斯混合模型的参数进行估计. 根据参数的初始值或者上一步迭代的模型参数来计算出后验概率, 即隐变量的期望值; 再对后验概率构造似然函数并最大化, 从而获得新的参数值. 当上一步迭代的参数值与当前参数值之差小于给定值时, 停止迭代, 最终结果作为每个参数的估计值. 这样, 每个数据集参数估计便计算出来, 结果如表2所示.

3.3 算法比较

在获得上述8组实验数据集少数类的概率分布后, 利用算法1的step 2 ~ step 4生成新样本得到平衡后的少数类数据集. 一般情况下, 当数据集存在类别不平衡时, 错误率并不是一个合适的评价指标. 因此, 表3的混淆矩阵用于计算AUC、Acc、Sen、Spe、Pre和F_1值, 使用它们作为性能评估指标.

ROC曲线上的一组点是通过调整阈值得到的, 曲线越凸且靠近左上方, 代表对应的分类器泛化能力越强. 在ROC曲线上, X轴代表假阳率

$$\text{FPR}=\frac{\text{FP}}{\text{TN}+\text{FP}}, \tag{15}$$

Y轴代表真阳率

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}}. \tag{16}$$

AUC是ROC曲线下到X轴以及直线$x=1$所围成的面积, 即ROC曲线的积分, 以定量的方式表示该

表 2　非平衡数据集的基本情况

数据集	参数
Wine(1-2)	$\boldsymbol{\mu}=[12.977, 3.497, 2.341, 16.834, 106.339, 2.695, 2.907, 0.273, 1.843, 5.037, 1.054, 3.056, 1115.712]$ $\boldsymbol{\sigma}=[0.788, 0.739, 0.067, 1.247, 10.410, 0.185, 0.520, 0.004, 0.514, 2.078, 0.025, 0.086, 219.635]$
Wine(2-3)	$\boldsymbol{\mu}=[13.860, 3.233, 2.355, 20.713, 99.312, 1.664, 0.620, 0.411, 1.316, 6.600, 0.644, 1.673, 629.900]$ $\boldsymbol{\sigma}=[0.446, 1.162, 0.141, 2.235, 10.410, 0.261, 0.012, 0.017, 0.270, 2.078, 0.113, 0.193, 113.892]$
Glass(1-3)	$\boldsymbol{\mu}=[1.519, 14.770, 3.574, 1.274, 72.405, 0.579, 8.794, 0.008, 0.057]$ $\boldsymbol{\sigma}=[0.003, 0.251, 0.017, 0.129, 0.497, 0.223, 0.289, 0.035, 0.105]$
Glass(1-5)	$\boldsymbol{\mu}=[1.519, 11.030, 1.582, 1.936, 72.366, 4.541, 12.500, 0.188, 0.061]$ $\boldsymbol{\sigma}=[0.003, 0.739, 2.090, 0.460, 1.232, 2.055, 1.197, 0.584, 0.149]$
Pima_indians	$\boldsymbol{\mu}=[5.797, 141.257, 70.825, 22.164, 100.336, 35.143, 0.678, 37.067]$ $\boldsymbol{\sigma}=[3.734, 31.880, 21.452, 17.650, 138.430, 7.249, 0.124, 10.948]$
Banknote	$\boldsymbol{\mu}=[-1.521, 0.432, 2.028, -1.117]$ $\boldsymbol{\sigma}=[1.841, 5.308, 5.779, 2.772]$
New_thyroid	$\boldsymbol{\mu}=[95.286, 16.957, 4.585, 1.211, -0.045]$ $\boldsymbol{\sigma}=[18.493, 3.752, 5.029, 0.187, 0.048]$
Yeast1	$\boldsymbol{\mu}=[0.471, 0.481, 0.530, 0.188, 0.503, 0.000, 0.447, 0.337]$ $\boldsymbol{\sigma}=[0.011, 0.011, 0.011, 0.022, 0.042, 0.000, 0.010, 0.111]$

表 3　混淆矩阵

真实值	预测值	
	少数类	多数类
少数类	TP	FN
多数类	FP	TN

ROC曲线对应的分类器的泛化能力,其计算公式如下:

$$\mathrm{AUC}=\frac{1+\mathrm{TPR}-\mathrm{FPR}}{2}. \tag{17}$$

准确率(Acc)是正确分类的样本数占总样本数的比值,其计算公式如下:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{TN}}. \tag{18}$$

灵敏度(Sen)用来计算少数类能被正确分类的比例. 此外,灵敏度也称召回率或真阳率,其计算公式如下:

$$\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{19}$$

特效性(Spe)被用来计算多数类能被正确分类的比例,其计算公式如下:

$$\mathrm{Spe}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \tag{20}$$

查准率(Pre)是被正确预测为少数类样本的数量占所有被预测为少数类样本数量的比例,其计算公式如下:

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{21}$$

此外,F值也是非平衡数据分类问题常用的性能指标. 它是Sen和Pre的结合,是数据非平衡问题有效度量. F值依赖β因子,该参数取值是从0到正无穷,且用来控制Sen和Pre. 当$\beta=0$时,F值退化为Sen,相反,当$\beta=\infty$时,F值退化为Pre.

$$F=\frac{(1+\beta^2)\cdot\mathrm{Sen}\cdot\mathrm{Pre}}{\mathrm{Pre}+\beta^2\cdot\mathrm{Sen}}. \tag{22}$$

在实验中,设定$\beta=1$来计算F值,其公式如下:

$$F_1=\frac{2\cdot\mathrm{Sen}\cdot\mathrm{Pre}}{\mathrm{Pre}+\mathrm{Sen}}. \tag{23}$$

用本文算法与其他5种算法在AUC、Acc、Sen、Spe、Pre和F_1值评价指标中进行比较. 限于篇幅,省略了各个评价指标的结果并用表4进行汇总. 表4统计了上述不同评估指标中每种方法的获胜次数,结果显示了本文算法优于其他方法.

表 4　获胜次数比较结果

	获胜次数						
	AUC	Acc	Sen	Spe	Pre	F_1	Total
ORIGINAL	0	0	0	0	1	1	2
SMOTE	1	0	0	0	0	0	1
SMOTE_BL1	0	2	1	4	3	1	11
SMOTE_BL2	0	0	1	0	1	0	2
ADASYN	0	1	1	1	2	0	5
本文算法	7	5	6	5	6	6	35

3.4　算法时间复杂度探讨

在GMM-EM中,迭代次数设为1 000次,则时间复杂度为$O(1)$;E步和M步时间复杂度分别为$O(n^2)$和$O(n)$,n代表训练样本数,时间复杂度为$O(n^2+n+1)$. ORIGINAL算法计算每个少数类样本的概率密度权重及过采样数量,此时间复杂度为$O(x)$,x代表少数类的样本数. 新设计的避免新样本出现交叉和重

叠现象的方法,其时间复杂度为$O(sx)$,s代表生成新样本数. 因此,总的时间复杂度为$O(n^2+n+1+x+sx)$. SMOTE算法为$O(sx)$,SMOTE-Borderline1算法为$O(x+kx+1)$,SMOTE-Borderline2算法为$O(x+kx+1)$,ADASYN算法为$O(kx+sx+1)$,k代表近邻数.

由上可见,时间复杂度最大为本文算法,最小为SMOTE-Borderline系列. 本文算法虽然所耗时间相对较长,但是根据表4各算法比较结果来看,其有效性是可以接受的.

4 结 论

本文基于非平衡数据的统计特性提出了一种基于GMM-EM的概率增强算法. 该算法根据少数类样本的统计特征进行再平衡,并通过算例验证了所提出算法的有效性. 其优点是充分利用了数据的统计特征生成新样本,避免了在生成数据时出现交叉重叠的情况. 本文算法虽然获得了较好的分类效果,但还可以从以下3个方面展开研究: 1) FG分类非平衡问题是本文的研究对象,今后可推广到多分类非平衡问题; 2) 本文算法针对的是非平衡比例小于10的数据集,今后可扩展到非平衡比例大于10的分类问题; 3) 本文并没有对EM算法初值问题展开详细的讨论,这也是今后需要进一步研究的问题.

参考文献(References)

[1] Bing Zhu, Bart Baesens, Seppe K L M vanden Brouke. An empirical comparison of techniques for the class imbalance problem in churn prediction[J]. Information Sciences, 2017, 408(10): 84-99.

[2] Laza R, Pavon R, Reboiro-Jato M, et al. Evaluating the effect of unbalanced data in biomedical document classification[J]. Journal of Integrative Bioinformatics, 2011, 8(3): 177-189.

[3] Dai H. Class imbalance learning via a fuzzy total margin based support vector machine[J]. Appl. Softcomputing, 2015, 31(6): 172-184.

[4] Chawla N, Bowyer K, Hall L, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(3): 321-357.

[5] Deepa T, Punithavalli M Dr. An E-SMOTE technique for feature selection in high-dimensional imbalanced dataset[C]. The 3rd International Conference on Electronics Computer Technology. Kanyakumari: IEEE, 2011: 322-324.

[6] Zhou Changsheng, Liu Bin, Wang Shihai. CMO-SMOTE: Misclassification cost minimization oriented synthetic minority oversampling technique for imbalanced learning[C]. The 8th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou: IEEE, 2016: 353-358.

[7] Bo Tang, Haibo He. Kernel ADASYN: Kernel based adaptive synthetic data generation for imbalanced learning[C]. 2015 IEEE Congress on Evolutionary Computation (CEC). Sendai: IEEE, 2015: 664-671.

[8] Ming Gao, Xia Hong, Sheng Chen, et al. Probability density function estimation based over-sampling for imbalanced two-class problems[C]. The 2012 International Joint Conference on Neural Networks (IJCNN). Brisbane: IEEE, 2012: 10-15.

[9] William A Rivera, Amit Goel, Peter Kincaid J. OUPS: A combined approach using SMOTE and propensity score matching[C]. The 13rd International Conference on Machine Learning and Application. Detroit: IEEE, 2014: 424-427.

[10] Haibo He, Yang Bai, Edwardo A Garcia, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. 2008 IEEE International Joint Conference on Neural Networks. Hong Kong: IEEE, 2008: 1322-1328.

[11] Yang L, Yiu-ming C, Yuan Y T. Hybrid sampling with bagging for class imbalance learning[C]. The Pacific-Asia Conference on Knowledge Discovery and Data Mining. Switzerland: Springer, Cham, 2016: 14-26.

[12] Alejo R, García V, Pacheco-Sánchez J H. An efficient over-sampling approach based on mean square error back-propagation for dealing with the multi-class mbalance problem[J]. Neural Process Lett, 2015, 42(3): 603-617.

[13] Shi-Xiong X, Fan-Rong M, Bing L, et al. A kernel clustering-based possibilistic fuzzy extreme learning machine for class imbalance learning[J]. Cognitive Computation, 2015, 7(1): 74-85.

[14] Maciej Zieb, Jakub M T. Boosted SVM with active learning strategy for imbalanced data[J]. Soft Computing, 2015, 19(12): 3357-3368.

[15] Fan Qi, Wang Zhe, Li Dongdong, et al. Entropy-based fuzzy support vector machine for imbalanced datasets[J]. Knowledge-Based Systems, 2017, 115(1): 87-99.

[16] Chris S, Grimson W E L. Adaptive background mixture models for real-time tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Fort Collins: IEEE, 1999: 246-252.

作者简介

陈 刚(1964−), 男, 教授, 从事数据挖掘、信息提取等研究, E-mail: chengang@dlmu.edu.cn;

吴振家(1993−), 男, 硕士生, 从事机器学习和数据挖掘的研究, E-mail: statistic573@163.com.

(责任编辑: 孙艺红)