doi:10.3969/j.issn.1001-4616.2019.01.011

基于概率密度估计的 SMOTE 改进算法研究

李　涛,郑　尚,邹海涛,于化龙

(江苏科技大学计算机学院,江苏 镇江 212003)

[**摘要**]　类别不平衡问题是机器学习与数据挖掘领域中主要关注的问题之一,目前已有多种解决方法,而样本采样技术是其中最为简单有效、同时也是最为常用的一类方法. 本文主要针对 SMOTE(synthetic minority oversampling technique)这一最为流行的采样算法易于受到噪声样本影响及泛化能力差的缺点,提出了一种基于概率密度估计的改进算法. 首先,假定各类样本均服从高斯混合分布,并采用高斯混合模型测得各样本的概率密度,针对各样本在类内与类间所测得概率密度间的排序比较关系来实现噪声信息的过滤. 其次,在过滤后的少数类样本上进行概率密度的重新计算,并根据其特点将其划分为三类:边界样本、安全样本与离群样本. 最后,针对上述三类样本,分别采取不同的策略来进行 SMOTE 采样. 此外,为了进一步提升泛化性能,本文也对 SMOTE 算法的邻域计算规则进行了修正. 通过多个基准的二类不平衡数据集对该算法进行了验证,实验结果表明其是有效且可行的,同时显著优于多种已有的采样算法.

[**关键词**]　类别不平衡,概率密度,样本采样,SMOTE,高斯混合分布
[**中图分类号**]TP181　[**文献标志码**]A　[**文章编号**]1001-4616(2019)01-0065-08

An Improved SMOTE Algorithm Based on Probability Density Estimation

Li Tao,Zheng Shang,Zou Haitao,Yu Hualong

(School of Computer Science,Jiangsu University of Science and Technology,Zhenjiang 212003,China)

Abstract:Class imbalance problem is one of the main problems in the fields of machine learning and data mining. To address this problem,the researchers have proposed lots of methods,in which instance sampling is the simplest,the most effective and the most used approach. As a popular instance sampling algorithm, SMOTE (synthetic minority oversampling technique)tends to be influenced by the noise instances and has poor generalization ability. To deal with this problem, an improved SMOTE algorithm which considers the probability density information is presented in this paper. Firstly, we assume that the instances in each class satisfy Gaussian mixture distribution,hence the Gaussian mixture model is adopted to estimate the probability density of each instance. Then the noisy instances could be removed by comparing rankings of the intra-class and inter-class probability density information. Next,the probability density information would be calculated again on the filtered data set,and then the instances belonging to the minority class could be divided into three groups as below:boundary,safety and outlier. Finally,for the instances in different group,different SMOTE strategies are used to generate the new instances. In addition,to further promote the generalization,the neighborhood calculation rule in SMOTE has also been modified. The experimental results on several binary-class imbalance data sets indicate that the proposed algorithm is effective and feasible. Moreover,it also shows that the proposed algorithm is significantly better than multiple previous algorithms.

Key words:class imbalance,probability density,instance sampling,SMOTE,Gaussian mixture distribution

近年来,类别不平衡问题已逐渐成为了机器学习、模式识别与数据挖掘等领域重点关注的问题之一[1-3]. 事实上,早在 2005 年的 ICDM 国际会议上,该问题便被列为了数据挖掘领域十大挑战性难题之一[4]. 确切地说,对于一个二分类数据集而言,当其中一个类别的样本个数远多于或少于另一个类别样本

收稿日期:2018-08-16.
基金项目:国家自然科学基金(61305058,61572242)、江苏省自然科学基金(BK20130471)、中国博士后特别资助计划项目(2015T80481)、中国博士后科学基金(2013M540404)、江苏省博士后基金(1401037B).
通讯联系人:于化龙,博士,副教授,研究方向:机器学习、数据挖掘. E-mail:yuhualong@just.edu.cn

数的时候,便存在类别不平衡问题. 目前绝大多数传统的分类算法,例如支持向量机、决策树和逻辑回归分类器等,虽然在构造原理上各不相同,但它们却都是建立在样本集均衡分布假设条件之下的,即一旦当数据集出现类别分布不均衡的情况,则会产生严重的后果,使分类边界显著被挤压向少数类区域,从而对少数类的分类精度造成极大影响. 同时,类别不平衡问题在实际应用中也广泛存在,如文本分类[5]、网络入侵检测[6]和软件缺陷检测[7]等.

针对类别不平衡问题,近年来研究人员已开展了大量的研究工作,并提出了诸多行之有效的方法,其大致可分为以下两类:第一类是数据层方法,也可称其为样本采样技术,其主要通过增加或删除样本的方式来修复原始样本集的不平衡分布;而另一类则是算法层方法,主要包括代价敏感学习技术[8]和决策输出补偿技术[9]. 相较于算法层方法,样本采样技术最为突出的优点主要体现在与采用何种分类器无关,而且简单易实现.

样本采样技术又可细分为:欠采样技术和过采样技术. 欠采样的基本思想是通过删除部分多数类样本,从而使数据集达到平衡. 这种做法的弊端在于会导致部分分类信息缺失,从而降低最终的分类性能. 而过采样的基本思想则是通过插入新样本的方法来增加少数类样本,可有效保留样本的原始分类信息. 但是其也存在自身的缺点,即容易产生过适应现象. SMOTE 算法[10]有效修正了这一缺陷,然而 SMOTE 算法在合成新少数类样本过程中,也通常会忽视噪声和离群样本的影响,使其影响范围进一步扩大,同时也未考虑样本分布信息的作用,进而导致最终所建立的分类模型无法达到最优.

近年来,很多研究人员也已经对 SMOTE 算法进行了改进,如 BSO(borderline-SMOTE)算法[11]认为边界区域才是对分类面位置起支撑作用的关键,故仅采用处于边界区域的少数类样本来合成新样本,其共有 2 个不同的版本,BSO1 在合成新样本时采用的是同类 K 近邻,而 BSO2 算法则采用所有训练样本共同计算 K 近邻. 上述算法的缺陷在:当训练样本的类别不平衡比率极大时,可能会造成选取的少数类边界样本极度不准确,进而影响采样的效果. SN-SMOTE(surrounding neighborhood-based SMOTE)算法[12]仅仅重新定义了邻域的计算方式,在一定程度上提升了模型的鲁棒性,但仍未考虑去除噪声及离群样本的影响. SL-SMOTE(safe-level-based SMOTE)算法[13]则考虑了噪声对 SMOTE 算法的影响,其采用同类样本在参照样本 K 近邻中所占的比例来划分噪声样本与安全样本,进而在生成新样本时令合成的新样本更靠近少数类安全区域,降低噪声传播的风险. 同样,该算法也会在类别不平衡比率较大时,误判噪声和安全样本. SMOTE-IPF(SMOTE-iterative partitioning filter)[14]也是一种改进的 SMOTE 算法,它虽然也能在一定程度上有效地缓解噪声的影响,可是在合成新样本时,仍未考虑样本的先验分布信息,进而对不同样本进行区别对待.

本文针对上述这些算法所存在的不足,提出了一种改进的 SMOTE 算法:基于概率密度估计的 SMOTE 算法(PDE-SMOTE,probability density estimation-based SMOTE). 该算法首先分别在每个类别上利用高斯混合模型来拟合该类样本的概率密度. 然后,根据每个样本在类内和类间概率密度的排序比较关系找到并删除噪声点. 进而,在去噪后的训练样本集上重新计算各样本的概率密度. 最后,再根据各少数类样本的概率密度分布特点将其划分为三类:边界样本、安全样本和离群样本,从而针对每类样本的特点分配个性化的 SMOTE 参数,以达到最优的采样效果. 特别地,为了提升数据分布的泛化性,本文也继承了 SN-SMOTE 算法的邻域计算模式. 通过 12 个基准的类别不平衡数据集对 PDE-SMOTE 算法的有效性与可行性进行了实验验证,结果表明该算法要显著优于多种已有的采样算法.

1 方法

1.1 高斯混合模型

作为一种经典的概率分布模型,高斯混合模型(gaussian mixture model,GMM)已在诸多实际应用领域得到广泛的应用,如图像处理[15]、概率密度拟合[16]等. 本文主要关注高斯混合模型的概率密度拟合能力,用以计算各样本的概率密度,以探索样本集的真实分布,并找出其中的噪声及离群样本. 从理论上而言,可以通过增加模型个数的方式,令高斯混合模型拟合任意概率分布. 换言之,任意的概率分布都可以采用多个高斯分布函数去联合加权近似.

众所周知,一维高斯函数可表示如下:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

式中,μ 和 σ 分别表示均值与标准差. 而多维高斯函数则可表示如下:

$$N(x;\mu,\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)\right]. \tag{2}$$

式中,μ 和 $\boldsymbol{\Sigma}$ 分别表示均值和协方差矩阵,d 则为样本的维度. 而所谓高斯混合模型,则是采用若干个高斯模型的加权组合而成,其中,模型的个数需预先设定. 在该模型中,某个样本 x_i 的概率密度可通过下式计算得出:

$$p(x_i)=\sum_{j=1}^{M}w_jN_j(x_i;\mu_j,\boldsymbol{\Sigma}_j). \tag{3}$$

式中,M 表示高斯模型个数,w_j 则表示第 j 个模型所占权重,其限定条件为:

$$\sum_{j=1}^{M}w_j=1. \tag{4}$$

在采用高斯混合模型对未知样本分布进行估计时,最为重要的步骤便是“参数估计”,而最为常用的方法则是最大似然估计法(EM 法). 其本质是令样本在估计的概率密度模型上的概率值达到最大. 鉴于概率密度值通常较小,连乘容易造成浮点数下溢,故在计算中通常将其转化为对数运算,其对应计算公式如下:

$$l(X|\boldsymbol{\Theta})=\log\prod_{i=1}^{n}p(x_i)=\sum_{i=1}^{n}\log\sum_{j=1}^{M}w_jN_j(x_i;\mu_j,\boldsymbol{\Sigma}_j). \tag{5}$$

式中,$\boldsymbol{\Theta}=(\theta_1,\cdots,\theta_M)^{\mathrm{T}}$,$\theta_j=(w_j,\mu_j,\boldsymbol{\Sigma}_j)$.

上述公式的极值可采用 EM 法求得:首先,假设已知各高斯模型的参数(通常随机给定初始值),采用其估计每个高斯模型的权重;然后,基于估计的权重,再去更新各高斯模型的参数. 迭代重复上述两个步骤,直到达到稳定状态为止. 采用上述算法,且在模型个数设置合理时,可以无限逼近样本集的真实分布.

1.2　基于概率密度的样本去噪策略

接下来,考虑如何利用样本的概率密度信息进行去噪. 众所周知,对于分类问题而言,判定一个样本是噪声样本的标准较为容易确定,即观察其是否分布在本类内较为稀疏,同时在异类内又较为稠密的区域,若是,则判定其为噪声点,若否,则判定其为非噪声点. 基于此标准,不难根据高斯混合模型所反馈的概率密度信息来找出原始训练集中的那些噪声样本. 然而,鉴于在类别不平衡数据中,不同类在样本规模上差异往往较大,从而造成概率密度估计值处于不同量纲而无法直接比较的问题,故本文摒弃了直接采用类内-类间概率密度值大小比较的方法,转而采用类内-类间概率密度排序的策略来找到并移除噪声样本. 这种做法的好处在于概率密度的量纲可能会随着样本规模与分布形态而发生改变,而样本间的概率密度排序关系却不会受到这些因素的影响.

基于上述思想,不难看出:若要找到并移除噪声,首先需要从各类中选取一定比例概率密度极小的样本,然后将其置于异类的高斯混合模型中计算其概率密度,并观察其在异类样本中的概率密度排序,若其概率密度大于一定的排序比例,则可判定其为噪声样本,进而进行移除. 上述 2 种比例可以视为 2 个阈值参数,分别用 α 与 β 来加以表示,其中,α 表示类内稀疏样本的比率,即备选的噪声样本比率;而 β 则用来表示异类稠密区域的边界,进而确定哪些样本应该被认定为噪声点. 由此可见,α 与 β 的取值范围均应在[0,1]之间,且不宜为其分配过大值.

下面给出基于 GMM 模型的样本去噪算法流程:

算法 1　基于 GMM 的样本去噪算法

输入:原始样本集 $\boldsymbol{\Theta}$,比例参数 α、β,高斯模型个数 M
输出:去噪后的样本集 $\boldsymbol{\Phi}$

算法流程:

1.将 $\boldsymbol{\Theta}$ 中的各类样本进行分离,并在每类样本上分别训练 GMM 模型,用以评估类内样本的概率密

度,GMM 模型参数为 M;

2.对每类中的样本按照概率密度从小到大的顺序进行排列,并按照排序选取排在前类内样本个数×α 个样本,将其作为备选噪声样本;

3.将一类的备选噪声样本置于异类的 GMM 模型中,并计算得到这些样本在异类分布中的概率密度;

4.按照概率密度从小到大的顺序对异类样本进行排序,并选取排在异类样本个数×β 的样本,以其所对应的概率密度作为截断值,并逐一比较本类备选噪声样本在异类中的概率密度与截断值的大小关系,若大于截断值,则视其为噪声并移除,若否,则判定其为非噪声样本;

5.得到去噪后的样本集 Φ.

1.3 基于概率密度的样本区域划分及个性化采样策略

如前文所述,除噪声以外,传统的 SMOTE 算法及其改进算法也通常忽略了样本先验分布信息的作用,而对全部样本等同对待,进而为其分配相同的参数,即使考虑到了样本先验分布信息的作用,也往往采用的是 K 近邻的判别策略,会受到类别不平衡比率的影响而造成判别结果的不准确与不稳定. 针对这一问题,本文仍然采取了基于 GMM 的概率密度估计信息来描述样本的先验分布,这种方法的优点在于摒弃了样本的近邻关系,因而提升了描述的准确性与鲁棒性.

根据类别不平衡问题的特点及对分类算法与各类样本先验分布信息的依赖关系进行分析,本文建议将少数类样本划分为 3 种不同的类别,并对隶属于每种类别的样本采用个性化的策略来进行采样. 三类样本分别为:安全样本、边界样本及离群样本. 其中,安全样本通常位于类内的高密度区域,边界样本则位于两类的交叠区域,通常在同类与异类中的概率密度均相对较低;而离群样本则位于整个特征空间的稀疏区域,其往往表现为无论在同类还是异类中,均有很低的概率密度.

接下来,探讨如何通过 GMM 模型对上述三类样本进行区分. 鉴于上述三类样本之间没有一个清晰的划分界限,故仍采用设置阈值的策略来对其进行区分. 三类样本所对应的阈值仍然各自对应于一个比例值,分别为 λ_1、λ_2 与 λ_3,用以表示三类样本在训练集少数类样本中所占有的比例,这 3 个比例值可由用户根据实际情况或经验来进行人为设置,约束条件为:

$$\lambda_1+\lambda_2+\lambda_3=1. \tag{6}$$

根据经验,训练集中的安全样本处于高密度区域,故 λ_1 的取值往往应远大于另外两个阈值.

在进行 3 种不同类型的样本划分时,应首先根据少数类 GMM 模型所反馈的概率密度信息,确定对应于 λ_1 比例的高概率密度样本,并将其划入安全区域. 然后,将剩余的样本放入多数类 GMM 模型,对应参数 λ_2,选取其中概率密度较大的样本,并将其划入边界区域. 而剩余的样本,则划归为离群样本.

为了进一步提升采样结果的泛化性,同时降低采样的风险,本文采用了两种策略来对其加以保障. 其一是借鉴了 SN-SMOTE 算法[12]的邻域计算策略,其二是为不同类型的少数类样本分配不同的邻域参数.

不同于 SMOTE 算法的邻域计算规则,SN-SMOTE 在计算样本邻域时,不再是简单选取距离其最近的 K 个近邻,而是采用了如下的过程:对于目前的主样本,首先在少数类样本中找到其最近邻样本,作为其一近邻,其二近邻则是少数类中距主样本和其一近邻连线中心点最近的样本,其三近邻则是距离主样本点与其一、二两近邻所构成的三角形质心点最近的少数类样本,以此类推,可以找到其全部 K 近邻. 上述策略可以最大限度地保证采样结果的泛化性. 而在生成新样本时,SN-SMOTE 仍继承了 SMOTE 算法的策略,即在主样本与其某个随机近邻的连线上随机合成一个新样本:

$$x_{\text{new}}=x+r\times(x'-x). \tag{7}$$

式中,x,x'与 x_{new}分别代表主样本,主样本的一个随机近邻样本及新和成的样本,而 r 则表示一个(0,1)之间的随机数,用以保证新样本出现在主样本与其近邻样本的连线之上.

针对 3 种类型样本的特点,所设计的个性化参数设置如下:

(1)安全样本:如主样本处于安全区域,则证明其处于同类高密度区域,邻域参数 K 宜取一个相对较大值,本文选取 $K=5$;

(2)边界样本:如主样本处于边界区域,则表明其处于概率密度相对稀疏的区域,而这一区域又较为重要,对最终分类模型的生成起支撑作用,故希望新合成的样本仍位于此区域,故应取一个适中的邻域参数,本文选取 $K=3$;

(3)离群样本:如主样本处于此区域,则表示其处于密度极低区域,希望生成的新样本能远离这一区域,故令邻域参数 $K=1$,同时,随机数 r 的取值范围变为(0.5,1),用以保证新生成样本更靠近近邻样本,而非主样本.

下面给出基于 GMM 模型的样本区域划分及个性化采样算法流程:

算法 2　基于 GMM 的样本区域划分及个性化采样算法

输入:去噪后的样本集 Φ,比例参数 λ_1,λ_2 与 λ_3,高斯模型个数 M
输出:采样后的样本集 Ω

算法流程:

1.将 Φ 中的多数类及少数类样本进行分离,并在每类样本上分别训练 GMM 模型,用以评估类内样本的概率密度,GMM 模型参数为 M;

2.对少数类中的样本按照概率密度从大到小的顺序进行排列,并按照排序选取排在前类内样本个数×λ_1 个样本,将其标注为安全样本;

3.将少数类中剩余样本置于多数类的 GMM 模型中,并计算得到这些样本在异类分布中的概率密度,按照从大到小的顺序进行排序;

4.选取排在前类内样本个数×λ_2 个样本,将其标注为边界样本,剩余样本则标记为离群样本;

5.随机选取少数类中的一个样本作为主样本,观察其样本类型,并选取与其样本类型相对应的邻域参数进行计算,生成新的样本,重复这一过程直至两类样本数量相同为止;

6.得采样后的样本集 Ω.

1.4　本文方法

结合算法 1 与算法 2,不难将本文所提出的 PDE-SMOTE 算法流程描述如下:

算法 3　PDE-SMOTE 算法

输入:原始不平衡样本集 Θ,比例参数 α、β、λ_1,λ_2 与 λ_3,高斯模型个数 M
输出:采样后的样本集 Ω

算法流程:

1.调用算法 1 得到去噪后的不平衡样本集 Φ;

2.在去噪后的样本集 Φ 上调用算法 2 得到采样后的平衡样本集 Ω.

2　实验结果与讨论

2.1　数据集

为验证本文所提出算法的有效性与可行性,采用了 12 个基准的二类不平衡数据集对其性能进行了测试. 这些数据集分别采集自 Keel 数据库[17]及 UCI 数据库[18],它们具有不同的特征数、样本数及类别不平衡比率,有关这些数据集的详细信息可参见表 1.

表 1　本文所用数据集

Table 1　The data sets used in this paper

数据集	特征数	样本数	不平衡比率	数据集	特征数	样本数	不平衡比率
glass1	9	214	2.06	abalone9_18	8	731	16.40
new_thyroid1	5	215	5.14	haberman	3	306	2.78
vehicle1	18	846	2.90	wisconsin	9	683	1.86
pima	8	768	1.87	ILPD	10	583	2.50
yeast_2_vs_4	8	514	9.08	seeds2v13	7	210	2.00
ecoli1	7	336	3.36	vowel0	13	988	10.00

2.2　实验设置

为了证明本文所提出的 PDE-SMOTE 算法具有更优的性能,本文选用了 8 种对比算法,第 1 种是不对原

始数据集进行任何采样处理的 ORI 算法,其也可以被视为基准算法,用于确定采样过程的必要性,其他几种比较算法分别包括:随机过采样 ROS(random oversampling)、SMOTE[10]、BSO1[11]、BSO2[11]、SL-SMOTE[13]、SN-SMOTE[12]和 SMOTE-IPF[14]. 分类器则采用最为常用的支持向量机(SVM,support vector machine)分类器,核函数为高斯径向基函数,其最优参数组合通过格搜索(Grid Search)的方式确定.

实验的硬件环境为:Intel Core i5-2467M 处理器;CPU 主频 1.60GHz;内存 4GB;操作系统为 Windows 7;编程环境 Matlab2014a.

PDE-SMOTE 算法的有关参数,根据实验结果所反馈的经验统一设置如下:高斯混合模型参数 M 统一设置为 5,α 与 β 均取为 0.1,λ_1 取值为 0.8,λ_2 与 λ_3 均取值为 0.1. 而对于其他比较算法中所特有的参数,均根据对应文献中的最优设置来进行分配.

此外,众所周知,对于不平衡分类任务而言,整体分类精度不再是一个准确的性能度量指标,故本文采用 F-measure 及 G-mean 这两个常用的性能测度来评价各种比较算法的性能[19-20]. 此外,为了真实地反应各种比较算法的性能,本文采用 20 次随机十折交叉验证的方式给出最终结果,结果以(均值±标准差)的形式呈现.

2.3　结果与讨论

表 2 与表 3 分别给出了各种比较算法在 12 个二类不平衡数据集上的实验结果. 其中,表 2 对应 F-measure 测度,而表 3 对应 G-mean 测度. 在每个数据集上,最优结果以粗体标记的形式给出.

表 2　各种比较算法在 12 个数据集上的 F-measure 测度值

Table 2　F-measure results of various comparison algorithms on the 12 used data sets

Data-set	ORI	ROS	SMOTE	BSO1	BSO2	SL-SMOTE	SN-SMOTE	SMOTE-IPF	PDE-SMOTE
glass1	0.592 9±0.048 2	0.607 2±0.013 6	0.633 8±0.020 8	0.625 7±0.013 9	0.634 8±0.015 1	0.622 4±0.033 7	0.635 7±0.006 2	0.637 6±0.022 3	0.644 1±0.008 0
new_thyroid1	0.905 9±0.036 5	0.941 8±0.014 0	0.893 8±0.020 7	0.901 3±0.021 5	0.860 6±0.035 9	0.865 9±0.040 2	0.903 4±0.033 9	0.906 5±0.014 8	0.921 8±0.006 6
vehicle1	0.440 4±0.021 4	0.655 5±0.010 2	0.662 9±0.001 2	0.673 6±0.005 3	0.622 5±0.006 7	0.659 6±0.012 2	0.664 8±0.013 5	0.658 5±0.001 8	0.672 1±0.015 8
pima	0.617 1±0.009 1	0.631 3±0.007 0	0.645 1±0.006 5	0.634 8±0.008 5	0.637 7±0.004 0	0.645 9±0.020 0	0.641 4±0.001 7	0.627 2±0.009 4	0.663 1±0.002 2
ecoli1	0.750 9±0.018 0	0.778 1±0.014 7	0.753 8±0.004 6	0.757 0±0.002 0	0.758 0±0.021 2	0.774 7±0.008 6	0.786 7±0.012 3	0.757 1±0.006 1	0.789 1±0.016 0
yeast_2_vs_4	0.744 4±0.031 1	0.731 6±0.021 4	0.728 7±0.014 0	0.703 3±0.012 9	0.703 3±0.012 9	0.791 4±0.042 8	0.768 1±0.024 5	0.745 9±0.016 6	0.783 5±0.021 0
abalone9_18	0.138 0±0.030 1	0.404 5±0.027 6	0.408 8±0.013 8	0.447 1±0.013 1	0.101 6±0.053 6	0.391 6±0.009 1	0.389 3±0.030 7	0.446 0±0.017 2	0.416 0±0.027 5
haberman	0.225 5±0.016 3	0.466 8±0.012 7	0.458 7±0.011 5	0.464 6±0.017 5	0.301 2±0.028 8	0.455 9±0.025 3	0.441 2±0.011 5	0.460 5±0.009 5	0.478 7±0.025 4
wisconsin	0.947 3±0.000 9	0.955 4±0.001 9	0.950 3±0.000 2	0.953 3±0.002 6	0.953 1±0.000 2	0.956 7±0.001 4	0.955 4±0.000 3	0.951 3±0.002 7	0.962 5±0.001 9
ILPD	0.010 0±0.005 4	0.557 6±0.005 1	0.524 6±0.002 8	0.539 4±0.011 1	0.445 4±0.019 9	0.532 2±0.002 8	0.451 2±0.025 2	0.549 3±0.006 3	0.556 5±0.004 9
seeds2v13	0.955 4±0.001 4	0.955 4±0.002 9	0.949 0±0.007 1	0.950 1±0.003 4	0.955 4±0.002 9	0.964 6±0.006 2	0.951 4±0.002 9	0.955 4±0.006 5	0.959 4±0.006 5
vowel0	0.978 9±0.008 6	1.000±0.000 0	1.000±0.000 0	1.000±0.000 0	0.949 5±0.034 7	0.978 9±0.007 0	0.964 5±0.007 2	0.974 7±0.002 5	1.000±0.000 0

表 3　各种比较算法在 12 个数据集上的 G-mean 测度值

Table 3　G-mean results of various comparison algorithms on the 12 used data sets

Data-set	ORI	ROS	SMOTE	BSO1	BSO2	SL-SMOTE	SN-SMOTE	SMOTE-IPF	PDE-SMOTE
glass1	0.663 9±0.035 4	0.681 9±0.014 5	0.709 1±0.017 4	0.708 8±0.011 5	0.723 0±0.012 4	0.720 0±0.032 2	0.711 9±0.007 2	0.718 5±0.023 0	0.720 9±0.012 7
new_thyroid1	0.917 7±0.031 2	0.950 7±0.013 9	0.903 8±0.003 4	0.919 3±0.002 4	0.910 9±0.027 8	0.881 0±0.037 5	0.927 3±0.026 0	0.981 2±0.001 0	0.957 6±0.013 7
vehicle1	0.553 6±0.016 5	0.792 4±0.009 2	0.801 7±0.001 0	0.814 0±0.003 9	0.752 1±0.005 5	0.785 2±0.010 9	0.784 6±0.012 3	0.795 9±0.001 1	0.809 4±0.014 1
pima	0.693 3±0.007 6	0.729 0±0.006 4	0.724 2±0.004 2	0.740 0±0.006 9	0.717 1±0.004 5	0.671 1±0.014 8	0.720 1±0.002 7	0.709 4±0.009 0	0.739 4±0.001 7
ecoli1	0.809 0±0.020 4	0.855 8±0.008 5	0.865 5±0.006 7	0.864 9±0.007 3	0.853 7±0.019 8	0.865 0±0.009 4	0.857 2±0.013 1	0.874 7±0.004 7	0.888 6±0.005 3
yeast_2_vs_4	0.832 9±0.022 9	0.889 4±0.014 1	0.886 8±0.013 0	0.892 3±0.014 3	0.892 3±0.014 3	0.892 4±0.033 8	0.872 7±0.011 5	0.909 8±0.014 8	0.899 9±0.005 9
abalone9_18	0.215 6±0.040 6	0.753 4±0.023 6	0.751 0±0.003 1	0.799 2±0.018 8	0.152 4±0.086 0	0.652 4±0.007 7	0.701 8±0.027 1	0.756 3±0.015 8	0.770 1±0.029 2
haberman	0.371 8±0.017 7	0.588 3±0.014 9	0.609 8±0.014 2	0.613 6±0.019 3	0.446 3±0.023 1	0.624 4±0.036 5	0.599 8±0.007 2	0.601 9±0.014 8	0.635 8±0.022 7
wisconsin	0.969 7±0.000 9	0.969 5±0.001 4	0.967 5±0.000 5	0.970 6±0.001 3	0.967 5±0.000 5	0.967 7±0.001 2	0.968 6±0.000 8	0.967 4±0.001 5	0.972 2±0.002 1
ILPD	0.032 7±0.016 2	0.631 2±0.006 5	0.632 5±0.003 4	0.648 1±0.013 3	0.585 6±0.016 6	0.632 5±0.006 2	0.632 3±0.022 6	0.667 7±0.006 2	0.668 8±0.002 6
seeds2v13	0.964 4±0.002 6	0.968 3±0.000 8	0.964 7±0.004 0	0.962 5±0.001 8	0.958 3±0.000 8	0.969 6±0.003 3	0.968 3±0.000 8	0.968 3±0.000 7	0.970 3±0.000 7
vowel0	0.979 9±0.008 2	1.000±0.000 0	1.000±0.000 0	1.000±0.000 0	0.952 9±0.030 4	0.979 9±0.006 6	0.997 8±0.000 3	0.998 3±0.000 3	1.000±0.000 0

从表 2 与表 3 的实验结果,可以得出如下结果:

(1)对比基线算法 ORI,各种采样算法的性能都有或多或少的提升. 这也直接说明了对于不平衡数据而言,对其进行采样以达到样本分布重平衡的策略是有效且可行的.

(2)在 new_thyroid1、ILPD 及 vowel0 等 3 个数据集上,即使采用最原始的 ROS 算法,其分类性能也能

达到最优或较优. 根据前人研究经验,不难得出这几个数据集的分布较为简单,数据集中几乎不包含噪声且两类样本间具有较大间隔的结论[21]. 事实上,在此类数据集上,若采用哪些复杂的采样算法,不但无助于提升模型的质量,还会增加建模的时间复杂度.

(3)在大多数数据集上,各种 SMOTE 的改进算法都或多过少地展现出了比传统 SMOTE 算法更优的性能. 这一结果也间接地证明了前文对 SMOTE 算法缺点分析的正确性.

(4)对比其他几种 SMOTE 的改进算法,本文算法取得了明显更优的分类结果. 确切地说,本文所提出的 PDE-SMOTE 算法在 7 个数据集上,取得了最优的 F-measure 结果,在 6 个数据集上,获得了最优的 G-mean 结果. 实际上,对比 SN-SMOTE 算法,本文算法增加了去噪的环节,同时考虑了样本先验分布信息的作用,而相对于 SL-SMOTE 及 SMOTE-IPF 算法,本文算法又改变了邻域的计算规则,使采样结果更具泛化性,同时其所采用的去噪策略也要更加准确. 这也解释了为何本文所提出的 PDE-SMOTE 算法要明显优于其他的比较算法.

2.4　显著性检测

此外,为了检验本文算法与其他几种算法相比,其结果是否在统计学上具有显著优势,本文也采用了 Friedman 非参数假设检验及 1×N Holm 后置假设检验法对各类算法的性能进行了统计学比较. 有关这一统计学过程的详细知识可参见文献[22],而相关的统计学测试软件可从以下网址获取:http://sci2s.ugr.es/sicidm/. 表 4 给出了各种比较算法在 12 个数据集上的平均排序(average ranking,AR)和调整后的 p 值(adjusted p-value,APV)结果.

表 4　各类算法的显著性比较,显著性水平 $\alpha=0.05$

Table 4　Statistical significance comparison of various algorithms, the significance level is 0.05

Algorithm	AR_F	APV_F	AR_G	APV_G
PDE-SMOTE	1.86	—	1.46	—
ORI	9.54	8.58×10^{-7}	9.67	1.34×10^{-8}
ROS	7.56	0.001 0	6.50	0.001 4
SMOTE	5.65	0.009 6	6.58	0.001 2
BSO1	4.54	0.024 7	5.21	0.016 8
BSO2	9.32	5.57×10^{-6}	8.75	6.51×10^{-7}
SL-SMOTE	5.17	0.032 8	5.21	0.016 8
SN-SMOTE	6.01	0.012 8	6.09	0.003 2
SMOTE-IPF	4.53	0.026 7	4.46	0.026 7

从表 4 的结果可以看出:本文算法无论在 F-measure,还是 G-mean 测度上,均取得了最小的平均排序值,且与其他各类算法相比,在 $\alpha=0.05$ 显著水平下,均是显著更优的,进而表明该算法在统计学上显著优于其他比较算法.

3　结束语

针对 SMOTE 算法的采样结果易于受噪声影响,未考虑样本先验分布信息的作用及易于产生过适应现象等诸多缺点,本文结合高斯混合模型提出了一种基于概率密度估计的 SMOTE 改进算法:PDE-SMOTE. 该算法首先采用高斯混合模型估计各训练样本的概率密度,然后根据其反馈值来进行去噪,进而再在去噪后的训练样本上重新训练高斯混合模型,利用各样本的概率密度进行样本所处区域的划分. 此外,为了提升采样结果的泛化性,本文也借鉴了 SN-SMOTE 算法的邻域计算规则来对各区域样本进行个性化的采样. 大量实验结果表明了该算法是有效及可行的.

鉴于该算法需两次调用高斯混合模型进行建模,且需对训练样本进行划分,故不可避免地会大幅增加训练的时间开销,这将是未来工作待改进之处. 此外,目前该模型仅可用于解决二类不平衡问题,如何将其扩展到多类不平衡学习领域,也是未来待解决的问题之一. 最后,需要注意的是,本文算法不适宜用于解决少数类样本极度稀少的类别不平衡问题,因为在此场景下,采用高斯混合模型对少数类所进行的概率密度估计将变得极为不准确,故会影响后续处理的精确性,进而影响到最终的采样及分类效果.

[参考文献]

[1] GANGANWAR V. An overview of classification algorithms for imbalanced datasets[J]. International journal of emerging technology and advanced engineering,2012,2(4):42-47.

[2] SUN Y,WONG A K C,KAMEL M S. Classification of imbalanced data:a review[J]. International journal of pattern recognition and artificial intelligence,2009,23(4):687-719.

[3] HULSE V J,KHOSHGOFTAAR T M,NAPOLITANO A. An exploration of learning when data is noisy and imbalanced[J]. Intelligent data analysis,2011,15(2):215-236.

[4] YANG Q,WU X. 10 challenging problems in data mining research[J]. International journal of information technology and decision making,2006,5(4):597-604.

[5] LIU Y,HAN T L,SUN A. Imbalanced text classification:A term weighting approach[J]. Expert systems with applications,2009,36(1):690-701.

[6] THOMAS C. Improving intrusion detection for imbalanced network traffic[J]. Security and communication networks,2013,6(3):309-324.

[7] WANG S,YAO X. Using class imbalance learning for software defect prediction[J]. IEEE transactions on reliability,2013,62(2):434-443.

[8] BATUWITA R,PALADE V. FSVM-CIL:fuzzy support vector machines for class imbalance learning[J]. IEEE transactions on fuzzy systems,2010,18(3):558-571.

[9] YU H L,MU C,SUN C Y,et al. Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data[J]. Knowledge-based systems,2015,76(1):67-78.

[10] 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.

[11] HAN H,WANG W Y,MAO B H. Borderline-SMOTE:A new over-sampling method in imbalanced data sets learning[C]// International Conference of Intelligent Computing. USA:ICIC,2005:878-887.

[12] GARCIA V,SáNCHEZ J S,MARTíN-FéLEZ R,et al. Surrounding neighborhood-based SMOTE for learning from imbalanced data sets[J]. Progress in artificial intelligence,2012,1(4):347-362.

[13] BUNKHUMPORNPAT C,SINAPIROMSARAN K,LURSINSAP C. Safe-level-SMOTE:safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem[C]//Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Germany:Springer-Verlag,2009:475-482.

[14] SáEZ J A,LUENGO J,STEFANOWSKI J,et al. SMOTE-IPF:addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information sciences,2015,291(5):184-203.

[15] 向日华,王润生. 一种基于高斯混合模型的距离图像分割算法[J]. 软件学报,2003,14(7):1250-1257.

[16] 吴福仙,温卫东. 极大似然最大熵概率密度估计及其优化解法[J]. 南京航空航天大学学报(自然科学版),2017,49(1):110-116.

[17] ALCALA F J,FEMANDEZ A,LUENGO J,et al. KEEL data-mining software tool:data set repository,integration of algorithms and experimental analysis framework[J]. Journal of multiple-valued logic and soft computing,2011,17(2/3):255-287.

[18] BLAKE C,KEOGH E,MERZ C J. UCI repository of machine learning databases[EB/OL]. http://www.ics.uci.edu/mlearn/MLRepository.html,1998.

[19] HE H,GARCIA E A. Learning from imbalanced data[J]. IEEE transactions on knowledge & data engineering,2009,21(9):1263-1284.

[20] GUO H X,LI Y,SHANG J,et al. Learning from class-imbalanced data:Review of methods and applications[J]. Expert systems with applications,2016,73:220-239.

[21] LóPEZ V,FERNáNDEZ A,GARCíA S,et al. An insight into classification with imbalanced data:Empirical results and current trends on using data intrinsic characteristics[J]. Information sciences,2013,250:113-141,2013.

[22] DEMSAR J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of machine learning research,2006,7:1-30.

[责任编辑:黄 敏]