一种改进的少数类样本识别方法

董 璇,蔡立军
(西北工业大学 理学院,陕西 西安 710129)

摘 要:非均衡数据集的分类过程中,产生了向多数类偏斜、少数类识别率较低的问题。为了提高少数类的分类精度,提出了一种 S-SMO-Boost 方法。该方法基于 Adaboost 提升算法迭代过程中错分少数类样本,构造虚拟样本,以加强对易错分样本的训练;其中构造样本利用空间插值方法,即在错分少数类样本周围构造超几何体,在该超几何体内部空间随机插值产生有效虚拟样本。在实际数据集上进行实验验证,结果表明,S-SMO-Boost方法提高了非均衡数据集的分类性能。

关键词:非均衡数据集;超几何体;样本生成;提升算法
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1674-7720(2012)18-0060-03

An improved method on identification of minority class sample

Dong Xuan, Cai Lijun
(Department of Mathematics, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: Analyzing the problem that the classification results is always biased to the majority class in imbalanced data sets. An improved method S-SMO-Boost is proposed. Based on the minorities which are misclassified in the iterative process of Adaboost algorithm, virtual samples are constructed to strengthen the training of minority class samples that are hardly classified .A method called S-SMOTE is used to construct a super geometry based on the minority class samples and its k nearest neighbors. The new virtual samples are generated inside the super geometry. Based on the real data sets, the experiments show that S-SMO-Boost improved the classification performance of imbalanced data sets.

Key words: imbalanced data sets; super geometry; generate samples; boosting algorithm

非均衡数据集的分类问题是模式识别和机器学习的研究热点。所谓非均衡数据集是指数据集合中,某些类的数据样本较多,而其他类数据样本较少[1]。样本较少的为少数类,样本较多的为多数类。非均衡数据集分类问题可应用于风险管理、网络入侵检测、银行预测、医疗诊断等领域。例如,医生疾病诊断中错将癌症病人诊断为正常人,损失会很大。这种情况下少数类样本却是人们更加关注的。针对该特点,传统的分类算法不再适用,有必要寻求好的分类方法使其在类别不均衡条件下,提高对少数类的识别率。

目前,解决非均衡数据集分类问题主要通过两种途径:算法层面方法和数据层面方法。算法层面方法主要是对已有分类算法进行改进或提出新的算法,如李亚军等[2]提出的改进的 Adaboost 算法与 SVM 的组合分类器。数据层面的解决办法有欠抽样方法,随机去掉部分多数类样本使不同类别样本数量均衡,此方法缺点是丢失了多数类的一些重要信息,造成分类性能降低。改进的欠抽样方法有托梅克联系对(Tomek Link)[3]方法、压缩最近邻法(CNN)[4]。简单的过抽样方法随机复制少数类样本的缺点是易导致过学习。Chawla 等[5]提出了 SMOTE(Synthetic Minority Over-sampling Technique)方法,人工合成少数类样本,但是生成样本范围受到极大限制。本文提出了 S-SMO-Boost 方法,利用 Adaboost 提升算法,每次迭代不仅仅增大错分样本权值,还从迭代过程中抽取错分少数类样本,并对该部分样本进行过抽样,过抽样过程采用 SMOTE 的改进方法——空间插值法,增强对错分少数类样本的训练,以训练出一个强分类器,提高分类性能。

1 分类原理

对于含有两个不同类的数据集(多类可化为两类)$S=\{(X_1,y_1),(X_2,y_2),\cdots,(X_m,y_m)\}$,每个数据元组 $X_i(0\leqslant i\leqslant m)$ 用 n 维属性向量 $X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$ 表示,$y_i\in Y=\{1,0\}$

为 X_i 的类标号，即数据元组所属类别。对数据集进行分类即将数据集分为训练集和检验集，分类原理如图 1 所示。

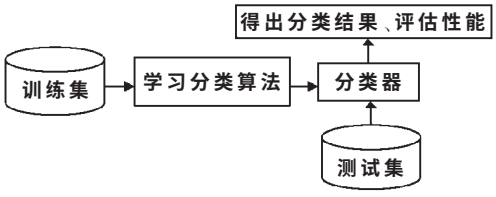

图 1 分类原理流程图

对于两类非均衡类数据集 $S=P\cup N$，$|P|=n_p$，$|N|=n_N$，设 $n_N=\lambda n_p$，λ 为非均衡率，$\lambda>1$ 且 $\lambda\in Q^+$。图 2 为一个非均衡数据集样本分布情况。

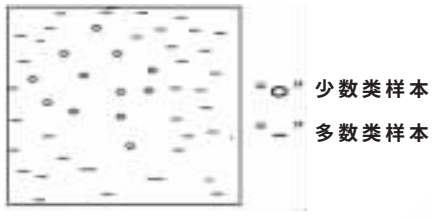

图 2 非均衡数据集样本分布情况

训练集为数据集中抽取的样本，由于训练集中两类样本数量同样相差极大，训练分类算法得到的分类器分类结果向多数类偏斜，故对测试集进行分类时，少数类样本识别率较低。本文提出 S-SMO-Boost 方法。

2 S-SMO-Boost 方法

Adaboost 是一种流行的提升算法，利用样本的权值来确定训练集的抽样分布。令 $S=\{(X_j,y_j)|j=1,2,\cdots,m\}$ 表示包含 m 个训练样本的集合，初始时所有样本权值均为 $1/m$，根据训练样本的抽样分布来抽取样本，得到新的样本集，由该训练集训练一个分类器，对原数据集样本进行分类。每一轮迭代结束时更新训练样本的权值，增加错误分类样本权值，减少正确分类样本权值。使分类器在随后迭代中关注那些很难分类的样本。算法中基分类器 C_i 的重要性依赖于它的错误率。组合分类器的最终结果通过取每个基分类器预测的加权平均得到。但是，由于少数类样本与多数类样本数量相差较大，即使增大错分少数类样本的权值，抽取的少数类样本仍然很少。利用 S-SMO-Boost 方法，在每次迭代中，记录错分少数类样本，利用空间插值法（S-SMOTE）在其周围产生相似的少数类样本，加入训练集中，进入下次迭代，训练分类器，有针对性地加强了对错分少数类样本的训练，提高了少数类样本的识别率。

S-SMO-Boost 方法具体算法如下：

输入：训练集 $S=\{(x_j,y_j)|j=1,2,\cdots,m\}$，其中 $y_j\in(1,0)$，k 表示迭代次数。

(1)初始化权值 $w=\{w_j=1/m|j=1,2,\cdots,m\}$。

(2)for $i=1$ to k do

(3)根据 w，通过对 S 进行抽样（有放回），产生训练集 S_i。用 S_i 训练基分类器 C_i，用 C_i 对原训练集 S 中所有样本分类。

(4)计算加权误差 $\varepsilon_i=\frac{1}{m}\left[\sum_{j=1}^{m}w_j\delta(C_i(x_j)\neq y_j)\right]$。

(5)if $\varepsilon_i>0.5$，重设样本权值，$w=\{w_j=1/m|j=1,2,\cdots,m\}$，返回步骤(4)。

end if

(6)对于 x_j，if $y_j=1$，而 $C_i(x_j)=0$，则记录 x_j。

(7)对其利用空间插值方法，产生虚拟样本 $syn'_i\in P_{creat}$。

(8)$a_i=\frac{1}{2}\ln\frac{1-\varepsilon_i}{\varepsilon_i}$。

(9)根据 $w_j^{i+1}=\frac{w_j^{(i)}}{Z_i}\times\begin{cases}e^{-a_i}, C_i(x_j)=y_j\\ e^{a_i}, C_i(x_j)\neq y_j\end{cases}$ 更新样本的权值。

其中 Z_i 是一个正规因子，确保 $\sum_{j=1}^{k}w_j^{i+1}=1$。

(10)将 P_{creat} 归入训练集 S。

(11)end for

(12)输出 $C^*(x)=\arg\max_y\sum_{i=1}^{k}a_i\delta(C_i(x)=y)$。

上述算法中，步骤(6)~(7)对迭代过程中错分的少数类利用空间插值法进行过抽样，在其邻域空间内产生类似有效少数类样本；步骤(10)将虚拟样本加入训练集，同样加大了少数类被抽到的概率，加强了对易错分少数类样本的训练，降低了数据集的非均衡程度，有效提高了非均衡数据集的分类性能，其有效性将在实验中得到验证。

3 空间插值方法

SMOTE 方法是一种过抽样方法，该方法在少数类 p 与其同类 k（$k=5$）近邻 p' 之间的连线上人工合成少数类样本 $p_{new}=p+\text{random}(0,1)\times(p'-p)$，其中 random(0,1) 是介于 0 与 1 之间的随机数。根据需要循环以上过程生成更多新的虚拟样本，避免了过度拟合问题。

其缺点是：(1)没有针对性，对于分类正确的少数类样本，同样产生虚拟样本，增加了分类成本；(2)生成样本仅介于少数类样本之间的连线上，限制了生成范围，不符合真实数据分布情况，有一定局限性。

空间插值法 S-SMOTE(Space Synthetic Minority Over-sampling Technique)对 SMOTE 方法中少数类样本的生成范围进行了改进。设数据集中少数类样本集为 $P=\{(p_1,1),(p_2,1),\cdots,(p_i,1)\}(0\leqslant i\leqslant n_p)$，$p_i$ 为 P 中样本元组；多数类样本集为 $N=\{(n_1,0),(n_2,0),\cdots,(n_j,0)\}(0\leqslant j\leqslant n_N)$，1 和 0 分别为少数类和多数类标号。$n_N=\lambda n_p(\lambda>1$ 且 $\lambda\in Q^+)$。

空间插值法的基本思想如下：

(1)对少数类样本 p_i，利用欧式空间距离公式求其 k（$k=5$）近邻。

(2)利用该少数类及其 k 近邻构造超几何体（三维空间中为四面体），在该超几何体内随机插值，产生虚拟少

数类样本，相比 SMOTE 方法，生成样本范围变大。对于存在多数类近邻的少数类，更容易被错分，故在分类过程中贡献较大，因此构造部分边界虚拟少数类样本。图3表示利用空间插值法在超几何体内随机产生虚拟少数类样本。

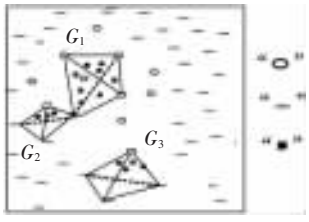

图 3 用空间插值法在超几何体内产生虚拟少数类样本图

产生虚拟样本具体步骤：

(1)求少数类样本 p_i 的 k 近邻，设 k_1 为其 k 近邻中多数类样本个数。

(2)若 $k_1=0$，则从 k 近邻中随机取 4 个设为 p_{i_1}、p_{i_2}、p_{i_3}、p_{i_4}构成一个四面体，如图 4(左)所示。据式(1)~式(3)构造四面体内部样本，$syn'_{inew}=(syn_{i1},syn_{i2},\cdots,syn_{in})\in P_{creat}$

$$a_i=p_{i_4}+r_1(p_{i_3}-p_{i_4}) \tag{1}$$

$$b_i=p_{i_2}+r_2(a_i-p_{i_2}) \tag{2}$$

$$syn_i=p_{i_1}+r_3(b_i-p_{i_1}) \tag{3}$$

其中，$r_1,r_2,r_3\in(0,1)$。

(3)若 $k_1=1$，则将该多数类样本看作噪声去除，从剩余近邻中选取 3 个与 p_i 构成超几何体产生虚拟样本 $syn'_i\in P_{creat}$。

(4)若 $2\leqslant k_1\leqslant 4$，则从其 k 近邻中随机选取 3 个与 p_i 构成超几何体，p_i 为顶点，如图 4(右)所示，$syn'_i=p_i+r_3\times\frac{1}{2}(b_i-p_i)\in P_{creat}$，集中于 p_i 为顶点等比缩小的几何体内。

(5)若 $k_1=5$，则该少数类被认为是噪声，不再产生虚拟样本。

(6)记录 $|P_{creat}|$，循环此过程直至产生所需虚拟样本数量。

步骤(2)中，少数类 p_i 的 k 近邻均为少数类，构造超几何体 G_1(如图 3 所示)，G_1 内随机虚拟样本以该少数类样本及其 k 近邻为近邻，则根据 K-NN 算法思想，近邻多为少数类的样本属于少数类的概率很大，故可归入少数类样本集 S 中。步骤(3)、(4)中，k 近邻中存在多数类样本时，当产生的虚拟样本靠近多数类时，根据最近邻思想，则其可能属于多数类，故构造超几何体 G_2、G_3(如图 3 所示)，将生成样本范围控制在以 p_i 为顶点的等比缩小的几何体内，使生成样本更靠近少数类样本，提高了生成样本的质量，避免噪声产生。所以空间插值的方法能生成有效的虚拟少数类样本。

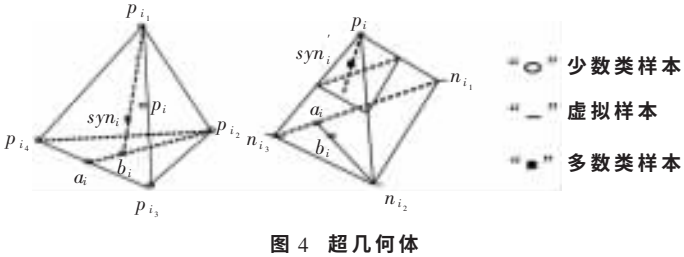

图 4 超几何体

4 实验结果及其分析

4.1 数据集

实验采用 Haberma、Pima Indians、Wisconsin-breast-cancer、Machine 4 个公开数据集，它们均选自 UCI 机器学习数据库[6]。4 个数据集除 Machine 外均为两类非均衡数据集，对 Machine 数据集选取类标为“3”的一类作为少数类，将其余类别合并为多数类。表 1 为 4 个数据集的基本信息。

表 1 数据集描述

数据集	Pima	Wisc	Mach	Habe
样本总数	768	699	207	306
少数类样本数	268	241	29	81
多数类样本数	500	458	178	225
属性个数	8	9	7	3
少数类比/%	34.8	34.4	14.0	26.5

4.2 实验结果

仿真实验在 CPU 双核 2.80 GHz、内存 2 GB 的 PC 机上进行，将 S-SMO-Boost 方法分别与单独使用 S-SMOTE、SMOTE、J48 方法进行比较，均以 J48 为基分类器，并采用十折交叉验证法。其中，S-SMOTE 方法处理数据时，循环其构造样本的过程直至 $|P_{creat}|=n_N-n_P=(\lambda-1)n_P$。则 P_{creat}、P 与剩余多数类样本集合并为均衡数据集训练分类器。

分类性能评价准则采用几何均值 $G-mean$ 值与少数类的 $F-value$ 值：

$$G-mean=\sqrt{(TP/(TP+TN))\times(TN/(TN+FP))}$$

$$F-value=\frac{(1+\beta^2)\times Recall\times Precision}{\beta^2\times Recall\times Precision}$$

其中，TP 与 TN 分别表示正确分类的少数类与多数类数量，FP 与 FN 分别表示错分为少数类与多数类的样本数量。$G-mean$ 值中 $TP/(TP+TN)$ 指少数类精确度，$TN/(TN+FP)$ 指多数类精确度，只有两者的值都大时，几何均值才会大，因此几何均值能合理地评价非均衡数据集的整体分类性能。$F-value$ 值中 $Recall=TP/(TP+FN)$ 与 $Precision=TP/(TP+FP)$ 分别表示少数类查全率和查准率，两者值都大时 $F-value$ 值才会大，因此 $F-value$ 值能正确反映少数类的分类性能。

图 5 表示分别用四种方法对 4 个数据集分类时得到的少数类 $F-value$ 值。同种方法得到的 $F-value$ 值点用线连起可清晰显示，利用 S-SMO-Boost 方法得到的 $F-value$ 值相比其他方法均有一定程度的提高。

表 2 对不同方法，分别比较了 4 个数据集的 $G-mean$ 值，由实验结果可知，直接用 J48 进行分类得到的值最

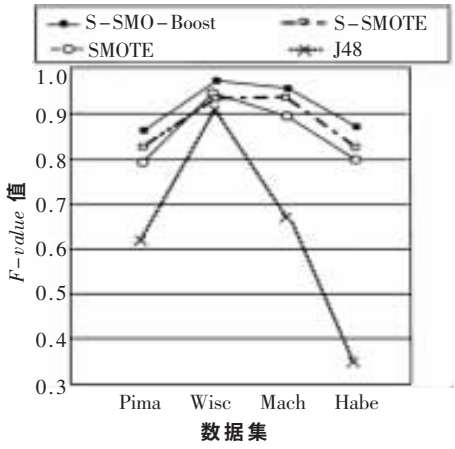

图 5 数据集少数类 *F-value* 值比较

表 2 数据集 *G-mean* 值比较

方法	Pima	Wisc	Mach	Habe
J48	0.697 1	0.949 0	0.774 4	0.508 0
SMOTE	0.801 0	0.959 5	0.966 2	0.814 9
S-SMOTE	0.827 1	0.967 0	0.968 0	0.839 2
S-SMO-Boost	0.846 6	0.978 9	0.979 4	0.865 3

小，因为数据集严重不均衡。相比 SMOTE 方法，S-SMOTE 在少数类邻域空间内插值产生有效虚拟样本，并加强靠近边界少数类样本的训练，故分类性能相对较好。S-SMO-Boost 将空间插值法融入提升算法，在迭代过程中利用错分样本产生虚拟样本，增强对错分少数类样本的训练，且增大错分样本的权值，加大迭代中作训练集的概率，并将弱分类器组合成强分类器。由表 2 知，用 S-SMO-Boost 方法得到的 *G-mean* 值最大，提高了非均衡数据集的整体分类性能。

为了解决非均衡数据集中少数类识别率较低的问题，本文提出了 S-SMO-Boost 方法，利用空间插值方法，产生有效虚拟样本，并将其与提升算法融合，加强对错分少数类样本的训练。经实验验证，该方法提高了少数类识别率和数据集整体分类性能。

参考文献

[1] WEISS G.Mining with rarity：an unifying framework[J]. Sigkdd Explorations，2004，6(7)：7-19.

[2] 李亚军，刘晓霞，陈平.改进的 AdaBoost 算法与 SVM 的组合分类器[J].计算机工程与应用，2008，44(32)：140-142.

[3] TOMEK I.Two modi-cations of CNN[J].IEEE Transactions on Systems Man and Communications，1976，SMC-6：769-772.

[4] MANNILA，LIU，MOTODA.Adavances in instance selection for instance-based leaning algorithms[J].Data Mining and Knowledge Discovery，2002(6)：153-172.

[5] CHAWLA N，BOWYER K，HALL L，et al.SMOTE：synthetic minority over-sampling echnique[J].Journal of Artificial Intelligence Research，2002(16)：321-357.

[6] BLAKE C，MERZ C.UCI repository of machine learning databases[DB/OL].1998.http://archive.ics.uci.edu/ml/.

(收稿日期：2012-05-17)

作者简介：

董璇，女，1988年生，硕士研究生，主要研究方向：非均衡数据集的分类方法。

蔡立军，男，1963年生，副教授，硕士研究生导师，主要研究方向：飞行器的控制、制导与仿真，微分对策，分散控制。