基于 CPD-SMOTE 的类不平衡数据分类算法研究

彭如香 杨 涛 孔华锋* 姜国庆 凡友荣
(公安部第三研究所 上海 201204)
(信息网络安全公安部重点实验室 上海 201204)

摘 要 类不平衡现象普遍存在于不同应用领域中,如金融欺诈、网络入侵、垃圾邮件过滤、医学检测,直接采用传统的学习分类算法,分类准确率较低。针对类不平衡情况对分类器的影响,基于传统过采样算法 SMOTE (Synthetic Minority Oversampling Technique)算法处理类不平衡的有效性,致力进一步提升 SMOTE 算法性能,提出一种面向类不平衡数据集分类的改进型 SMOTE 算法——CPD-SMOTE 算法。通过考虑训练集小样本的特征、位置及其周围样本分布,来确定小样本的强相关邻居集,以此作为 SMOTE 最近邻居集,产生新的小样本。实验结果表明,CPD-SMOTE 算法在处理不平衡数据集上相比 SMOTE、Borderline-SMOTE、ADASYN、LN-SMOTE 等算法有所提高。

关键词 SMOTE 类不平衡 分类算法

中图分类号 TP301.6 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.12.048

CLASS IMBALANCE DATA CLASSIFICATION ALGORITHM BASED ON CPD-SMOTE

Peng Ruxiang Yang Tao Kong Huafeng* Jiang Guoqing Fan Yourong
(*Third Research Institute of Ministry of Public Security*, *Shanghai 210204*, *China*)
(*Key Lab of Information Network Security*, *Shanghai 201204*, *China*)

Abstract Class imbalance is a common phenomenon existing in different applications, such as financial fraud, network intrusion, spam filtering and medical detection. If we directly adopt the traditional learning classification algorithm, classification accuracy is low. Aiming at the effect of class imbalance on classifier, this paper proposed an improved SMOTE algorithm, CPD-SMOTE algorithm, which was oriented to the classification of class imbalance datasets. Based on the effectiveness of traditional over-sampling algorithm SMOTE to deal with class imbalance, CPD-SMOTE algorithm was engaged in further improving the performance of SMOTE algorithm. CPD-SMOTE algorithm determined the strong correlation neighborhood set of small samples by considering the characteristics and location of small samples and distribution of their surrounding samples in the training set. It was used as the nearest neighbor set of SMOTE to generate new small samples. Experimental results show that CPD-SMOTE algorithm is better than SMOTE, Borderline-SMOTE, ADASYN and LN-SMOTE in dealing with imbalanced datasets.

Keywords SMOTE Class imbalance Classification algorithm

0 引 言

类不平衡是指属于某一类别的观测样本的数量显著少于其他类别,通常情况下把多数类样本的比例为 100:1、1 000:1,甚至是 10 000:1这种情况下为不平衡数据[1]。类不平衡现象普遍存在着不同应用领域中,如金融欺诈、网络入侵、垃圾邮件过滤、医学检测,直接

收稿日期:2018-08-15。国家重点研发计划课题(2016YFC0800909);公安部科技强警基础工作专项项目(2018GBJC19);上海市科委科研项目(17DZ1101004)。**彭如香**,助理研究员,主研领域:信息安全与数据挖掘。**杨涛**,副研究员。**孔华锋**,研究员。**姜国庆**,助理研究员。**凡友荣**,助理研究员。

采用传统的学习分类算法,分类准确率较低[1-3]。通常采用重采样方法处理类不平衡问题,重采样包括欠采用和过采样两种[1]。相比于传统欠采样方法,SMOTE 算法克服传统随机欠采样导致的数据丢失问题。但是,SMOTE 容易出现过泛化和高方差的问题,进而影响数据分布特征。为了解决这些问题,Borderline-SMOTE、ADASYN、LN-SMOTE 等 SMOTE 改进算法相继被提出[4-6]。这些算法充分考虑小样本的分布,新增的样本不影响小样本的分布。然而这些算法未考虑小样本与大样本的交叉区域,最终形成的数据集将改变原有数据集的分布,而影响分类算法的准确性。

1 相关工作

1.1 SMOTE 算法

SMOTE,即合成少数类过采样技术[1]。该算法由 Chawla 于 2002 年提出,是对随机过采样算法的一种改进。SMOTE 算法的基本思想是对少数类样本进行分析和人工模拟,同时将模拟得到的新样本数据添加到原始数据集当中,使得数据正负样本比例均衡。

SMOTE 算法流程如下:

(1) 对于少数类样本中每一个样本 x,通过计算 x 到该类样本集所有样本的欧式距离,利用 KNN 算法,选出离样本 x 最近的 k 个同类样本点,得到其 K 近邻。

(2) 根据正负样本比例确定采样倍率为 N,对每一个样本 x 分别随机从 K 近邻中选取 N 个样本,假设选择的近邻为 $x_1, x_2, \cdots, x_N$。

(3) 对于样本 x_i 的每一个随机选出的 K 近邻 $\tilde{x}_i$ $(i=1,2,\cdots,N)$,新样本计算公式为:

$$x_{\text{new}} = x + rand(0,1) \times (\tilde{x}_i - x_i)$$

式中:x_i 表示少数类别中的一个样本点;$\tilde{x}_i(i=1,2,\cdots,N)$ 表示从 K 近邻中随机挑选的样本点;$rand(0,1)$ 表示生成 0 ~ 1 之间的随机数。

SMOTE 算法的出现,改进了处理非平衡数据中传统的随机过采样算法,可以有效地对非平衡数据进行纠偏,整体上提高了模型的精度,同时还很大程度上降低了模型的误识率,这是 SMOTE 算法的优点[9]。其缺陷是无法解决非平衡数据的分布问题,容易产生分布边缘化问题,对于边缘的少类样本,对其进行 K 近邻生成样本也位于边缘且会越来越边缘化,这会使得正负样本的边界越来越模糊,加大样本分类的难度。

1.2 类不平衡算法评价指标

根据不同的应用场景,分类器考虑的评价指标也不同,通常基于混淆矩阵进行性能评价。混淆矩阵的定义如表 1 所示[7]:

表 1 混淆矩阵定义

	实际为正类	实际为负类
预测为正类	TP	FP
预测为负类	FN	TN

使用混淆矩阵可以得到如下几个评价指标:

精确度 $Accuracy = (TP + TN)/(TP + FN + FP + TN)$

召回率 $R = TP/(TP + FN)$

准确度 $P = TP/(TP + FP)$

召回率和准确度通常会出现矛盾,这样需要综合考虑,最常见的方法就是 F-Measure(又称为 F-Score)。F-Measure 是 Precision 和 Recall 加权调和平均,其定义如下:

$$F\text{-}Measure = \frac{(\alpha^2+1)P \times R}{\alpha^2(P+R)}$$

另外,可以通过 ROC 曲线评价一个分类器好坏。ROC 曲线是基于样本的真实类别和预测概率来画的,具体来说,ROC 曲线的 x 轴是伪阳性率($FP/(TN + FP)$),y 轴是真阳性率($TP/(TP + FN)$);AUC(Area Under Curve)被定义为 ROC 曲线下的面积。简单地说,AUC 值越大的分类器,正确率越高。

2 CPD-SMOTE 类不平衡数据处理算法

通过上述介绍,SMOET 算法充分考虑小样本的分布,新增的样本不影响小样本的分布。然而该算法未考虑小样本与大样本的交叉区域,最终形成的数据集将改变原有数据集的分布,而影响分类算法的准确性[5-6]。

本文提出一种改进型 SMOET——CPD-SMOTE。CPD-SMOTE 通过考虑训练集小样本 S_i 的特征、位置及其周围样本分布,来确定 S_i 的强相关邻居集 SCN_i。以 SCN_i 作为 SMOTE 最近邻居集,产生新的小样本。这里强相关邻居集 SCN_i 满足:

1) SCN_i 每个元素都为小样本特征向量;

2) 对于 SCN_i 每个元素 C_k,C_k 与 S_i 组成的超矩阵中不包含大样本特征向量。

其算法如下:

算法 1 强相关邻居集确定算法

输入:训练集 $D = \bigcup_{i=1}^{m}(x_i, y_i)$,其中 x_i 为 n 维的特征向量,y_i 为分类标签;s 小样本数;

输出:所有小样本的强相关邻居集 $\{SCN_1, SCN_2, \cdots,$

$SCN_s\}$。

1: $\Gamma = \bigcup_{i=1}^{s} SCN_i$,其中 SCN_i 为空;

2: $\Phi = \bigcup_{i=1}^{s} z_i$,其中 z_i 为小样本特征向量;

3: $\Lambda = \bigcup_{i=1}^{m-s} u_i$,其中 u_i 为大样本特征向量;

4: i = 1;

5: For j in[2,s]

6: 判断 z_i 与 z_j 组成超矩阵中,是否包含 Λ 中的值;

7: 若不包含,则将 z_j 加入到 SCN_i 集合中;

8: Fori in[2,s - 1]

9: For j in[1,i - 1]

10: 判断 SCN_j 中是否包含 z_i

11: 若包含,则将 z_i 加入到 SCN_j 集合中;

12: For j in[i + 1,s]

13: 判断 z_i 与 z_j 组成超矩阵中,是否包含 Λ 中的值;

14: 若不包含,则将 z_j 加入到 SCN_i 集合中;

15: i = s;

16: For j in[1,s - 1]

17: 判断 SCN_j 中是否包含 z_i

18: 若包含,则将 z_i 加入到 SCN_j 集合中;

19: 若不包含,则将 z_j 加入到 SCN_i 集合中;

20: 计算非空强相关邻居集中元素数量平均值 k;

21: 对于所有为空的强相关邻居集 SCN_i

22: 从大样本集 Λ 中随机挑选 k 个 z_i(小样本)的最近邻居 Λ'

23: 对于 Λ'中的所有 u_j

24: 取 α 为 n 维向量,元素取值为[0,0.5]之间的随机值;

25: $z_{new} = z_i + (u_j - z_i) \times \alpha$

26: 将 z_{new}加入到 SCN_i 中

下面通过二维图简单说明强相关邻居节点选择方法,如图 1 所示。

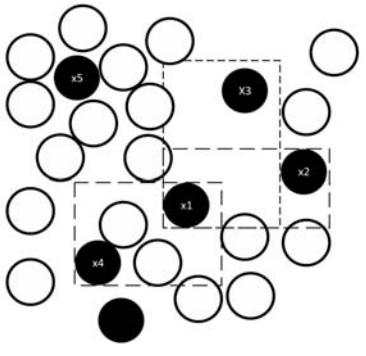

图 1　小样本 x1 的强相关邻居节点选择示意图

图 1 中,实心圆表示小样本节点,空心圆表示大样本节点。通过上述强相关邻居集确定算法,判断 x2 是否是 x1 的强相关节点的依据是由 x2 与 x1 组成的超矩阵中,是否包含大样本节点确定的。从图中看出 x2 满足该条件,故可将 x2 加入 x1 的强相关邻居集 SCN_1 中。同理,x3 也为 SCN_1 的元素,而 x4 将不会被选到 SCN_1 中。虽然 x4 到 x1 的距离小于 x3,但是 x4 与 x1 之间存在大样本节点。若不考虑大样本节点的存在,采用传统的 SMOTE,在 x4 与 x1 连线上随机产生新的节点,将影响大样本的分布,进而影响整个样本空间的分布。采用本文中的提出强相关邻居集确定算法,x4 将不会被纳入到 SCN_1。然后,将 SCN_1 作为 SMOTE 算法中邻居集产生新的小样本节点仍服从原来的分布。对于 x5,根据上述判断条件,将不存在符合条件的邻居,所得到的强相关邻居集为空。考虑到这种情况,本算法充分考虑节点邻居数量均衡的特点,首先计算非空强相关邻居集中元素数量平均值 k,然后从大样本集中随机挑选 k 个 x5 最近邻居,其次从 x5 与每个邻居的连线不超过一半(x5 为起点)上随机挑选一个点作为 x5 的强相关邻居。至此,每个小样本的强相关邻居集都已确定。接下来将得到的强相关邻居集作为 SMOTE 算法中邻居集,完成新样本的生成,其算法过程如下:

算法 2　CPD-SMOTE 算法

输入:训练集 $D = \bigcup_{i=1}^{m} (x_i, y_i)$,其中 X_i 为 n 维的特征向量,y_i 为分类标签;N:待生成的小样本数量;

输出:新生成的小样本集 S;$\Phi = \bigcup_{i=1}^{s} z_i$,其中 z_i 为小样本特征向量。

1: 设 α 为 n 维向量,元素取值为[0,1]之间的随机值;

2: 使用上述强相关邻居集确定算法,得出强相关邻居集 $\{SCN_1, SCN_2, \cdots, SCN_s\}$

3: For i 从 1 到 N

4: 随机挑选小样本 Φ 中的点 z_j:

5: 对于 SCN_i 中元素 z_j

6: $z_{new} = z_i + (z_j - z_i) \times \alpha$

7: 将这些新的点加入 S 中

3　实验设置与分析

3.1　实验数据

本文实验采用来自 Kaggle 上的两个数据集,paySim 和 CreditCard Fraud Detection。

数据集 paySim 一种移动金融交易数据数据集,该数据集包括总样本数 6 362 620 个,其中正常交易样本数为 6 354 407,欺诈交易样本数为 8 237 个,比例为 772:1。paySim 数据集字段表如表 2 所示。

表 2 字段含义表

序号	名称	值	含义
1.	step	1	时间跳隔,按每小时一个跳隔,该字段取值 1 ~ 743,为一个月的数据
2.	type	PAYMENT	数据交易类型,包括五种: CASH_OUT:现金转出 PAYMENT:付账 CASH_IN:现金转入 TRANSFER:转账 DEBIT:借贷
3.	amount	1 060.31	交易金额
4.	nameOrig	C429214117	交易起始账户
5.	oldbalanceOrg	1 089.0	交易起始账户初始额
6.	newbalanceOrig	28.69	交易起始账户交易后额度
7.	nameDest	M1591654462	交易结束账户
8.	oldbalanceDest	0.0	交易结束账户初始额
9.	newbalanceDest	0.0	交易结束账户交易后额度
10.	isFraud	0	是否欺诈标识
11.	isFlaggedFraud	0	检测出是否欺诈标识

数据集 CreditCard Fraud Detection 为信用卡交易数据。该数据集包含两天内发生的交易,其中 284 807 笔交易中有 492 笔被盗刷。数据集为不平衡数据集,类(被盗刷)占所有交易的 0.172%。它只包含作为 PCA 转换结果的数字输入变量。该数据集属性除了时间和金额保留原始值,其他属性采用 PCA 转换进行了变换,转换特征 V1,V2,…,V28,特征"Class"为响应变量,如果发生被盗刷,则取值 1,否则为 0。

通过上述分析,这两类数据均为不平衡数据集。

3.2 实验分析

paySim 数据集和 CreditCard Fraud Detection 数据集都给出了目标列,为监督学习的应用场景。是否是欺诈交易或信用卡是否被盗刷是一个二元分类问题,本文选用基分类算法是支持向量机 SVM 算法作为基分类器[7-8]。

为了验证 CPD-SMOTE 算法与其他几种类不平衡处理算法 SMOTE、Borderline-SMOTE、ADASYN、LN-SMOTE 的性能,本文构建的分类算法模型首先分别通过这些算法将数据进行类平衡处理,使得数据正负样本比例均衡。然后分别采用 SVM 算法对数据分类,并采用 10 倍交叉验证方法验证模型的性能。这里 SVM 算法采用 skik-learn 算法集中,表 3 - 表 5 分别给出了不同算法在两个数据集上评测指标均值,包括召回率 *R*、F-Measure 与 AUC 均值。实验结果表明,相比其他类平衡处理算法,CPD-SMOTE 算法效果明显,学习性能更优。

表 3 基分类器为 SVM 下的不同算法的召回率 R 均值

Dataset	PaySim	CreditCard FraudDetection
SOMTE	0.672	0.654
Borderline-SMOTE	0.721	0.798
ADASYN	0.831	0.787
LN-SMOTE	0.801	0.856
CPD-SOMTE	0.912	0.891

表 4 基分类器为 SVM 下的不同算法的 F-Measure 均值

Dataset	PaySim	CreditCard FraudDetection
SOMTE	0.783	0.763
Borderline-SMOTE	0.632	0.687
ADASYN	0.871	0.892
LN-SMOTE	0.862	0.897
CPD-SOMTE	0.932	0.962

表 5 基分类器为 SVM 下的不同算法的 AUC 均值

Dataset	PaySim	CreditCard FraudDetection
SOMTE	0.654	0.677
Borderline-SMOTE	0.763	0.765
ADASYN	0.782	0.763
LN-SMOTE	0.867	0.872
CPD-SOMTE	0.901	0.912

4 结 语

针对类不平衡情况对分类器的影响,本文在 SMOTE 算法的基础上,提出了 CPD-SMOTE 算法。该算法通过考虑训练集小样本的特征、位置及其周围样本分布,来确定小样本的强相关邻居集,以此作为 SMOTE 最近邻居集,产生新的小样本,实现数据正负样本比例均衡,并采用 SVM 算法对数据进行分类预测。在相同的基分类器 SVM 情况下,相比其他 4 种算法(SOMTE、Borderline-SMOTE、ADASYN、LN-SMOTE),CPD-SMOTE 算法在处理不平衡问题时能提升分类性能。

参 考 文 献

[1] 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.

(下转第 268 页)

受到隐含层节点数的影响;在隐含层传输函数选择上,建议使用 tansig 作为传输函数;在原始输入和参考输入中噪声线性相关时,神经网络噪声抵消系统对信噪比小于 0 dB 的输入信号有显著的噪声抵消效果;两路输入中噪声非线性相关时,系统对信噪比大于 0 dB 的输入信号有提高信噪比的效果。噪声抵消系统中 BP 神经网络采用优化后的隐含层节点数公式,具有很强的降噪能力,可用于低信噪比下噪声抵消。

参 考 文 献

[1] Widrow B, Glover J R, Mccool J M, et al. Adaptive noise cancelling: principles and applications[J]. Proceedings of the IEEE, 2005, 63(12):1692-1716.

[2] 邹进, 曹茜红, 韩迎春,等. 基于自适应噪声抵消的微弱振动信号提取方法[J]. 探测与控制学报, 2015(5):47-50.

[3] Jamel T M, Mohamed H A. Noise canceller using a new modified adaptive step size LMS algorithm[J]. Wseas Transactions on Signal Processing, 2014(10):637-644.

[4] Jagadesh T, Mahalakshmi P. A novel pipelined adaptive RLS filter for ECG noise cancellation[J]. Research Journal of Applied Sciences Engineering & Technology, 2015, 11(5):501-506.

[5] Bai L, Yin Q. A modified NLMS algorithm for adaptive noise cancellation[C]//IEEE International Conference on Acoustics Speech and Signal Processing. IEEE, 2010: 3726-3729.

[6] Sun G, Xue G, Li C. The ezperiment of RBF network used in active noise controlling in tracked vehicle cabins[C]//8th International Symposium on Test and Measurement. 2009: 505-508.

[7] Zhang M. Application of BP neural network in acoustic wave measurement system[J]. Modern Physics Letters B, 2017, 31:1740052.

[8] Liu S, Pan J, Yang M H. Learning recursive filters for low-level vision via a hybrid neural network[C]//European Conference on Computer Vision. Springer, Cham, 2016: 560-576.

[9] Miry M H, Miry A H, Khleaf H K. Adaptive noise cancellation for speech employing fuzzy and neural network[C]//International Conference on Energy, Power and Control. IEEE, 2010:289-296.

[10] 李安平, 刘国荣. 一类非线性系统的自组织模糊神经网络控制[J]. 电机与控制学报, 2016, 20(12):82-91.

[11] 李晓艳. 基于神经网络的自适应噪声抵消的研究[D]. 武汉:武汉理工大学, 2010.

[12] 蒋威, 张东阳. 基于量子神经网络的无线电引信干扰对消[J]. 微计算机信息, 2011(5):199-201.

[13] 周伟, 吴晗平, 吴晶,等. 紫外目标探测弱信号处理方法研究[J]. 红外技术, 2012(9):508-514.

[14] Cai Z F, Jian Z, Chen L D. Harmonic analysis approach using enhanced adaline neural network[J]. Journal of Zhejiang University, 2009, 43(1):166-171.

[15] Dixit S, Nagaria D. Neural network implementation of least-mean-square adaptive noise cancellation[C]//International Conference on Issues and Challenges in Intelligent Computing Techniques. IEEE, 2014:134-139.

[16] 彭耿, 黄知涛, 陆凤波等. 中频通信信号信噪比的快速盲估计[J]. 电子与信息学报, 2010, 32(1):102-106.

[17] Zhao B, Lu H, Chen S, et al. Convolutional neural networks for time series classification[J]. Journal of Systems Engineering and Electronics, 2017, 28(1):162-169.

[18] Zhang L, Wang F, Sun T, et al. A constrained optimization method based on BP neural network[J]. Neural Computing & Applications, 2018, 29(2):413-421.

(上接第 262 页)

[2] García V, Sánchez J S, Mollineda R A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2012, 25(1): 13-21.

[3] Wang S, Yao X. Using Class Imbalance Learning for Software Defect Prediction[J]. IEEE Transactions on Reliability, 2013, 62(2):434-443.

[4] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[C]//International Conference on Advances in Intelligent Computing. Springer-Verlag, 2005:878-887.

[5] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//IEEE International Joint Conference on Neural Networks. IEEE, 2008:1322-1328.

[6] Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data[C]//Computational Intelligence and Data Mining. IEEE,2011:104-111.

[7] Breiman L. Random forests[J]. Machine Learning,2001,45(1):5-32.

[8] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests[J]. Pattern Recognition, 2011, 44(2):330-349.

[9] 尹华. 面向高维和不平衡数据分类的集成学习算法研究[D]. 武汉:武汉大学,2012.