基于非平衡数据处理方法的网络在线广告中点击欺诈检测的研究

李 鑫 郭 汉 张 欣 胡方强 帅仁俊
(南京工业大学计算机科学与技术学院 南京 211816)

摘 要 网络在线广告中以套取广告费为目的的点击欺诈检测是机器学习应用的重要内容之一。支持向量机(Support Vector Machine,SVM)是一种优秀的解决二分类和回归问题的机器学习算法,但应用于网络在线广告中的欺诈点击检测时,由于数据集的极端非平衡性,算法性能受到极大的限制。从 FDMA2012 竞赛欺诈发布商检测的真实数据集出发,在详细研究与对比了 3 种非平衡数据处理方法后,选取最佳的混合采样方法对原始数据进行处理,再将其应用于 SVM 分类器。实验结果表明,所提方法能够有效识别实施欺诈点击行为的非法发布商,准确度达到 95% 左右,满足了网络在线广告中点击欺诈检测的要求。

关键词 点击欺诈,支持向量机,非平衡,混合采样
中图法分类号 TP393 **文献标识码** A

Study on Click Fraud Detection in Online Advertising with Imbalanced Data Processing Methods

LI Xin GUO Han ZHANG Xin HU Fang-qiang SHUAI Ren-jun
(College of Computer Science and Technology, Nanjing Tech University, Nanjing 211816, China)

Abstract Click fraud detection in online advertising is one of the most important applications of machine learning. Support vector machine (SVM) is a prominent supervised machine learning algorithm on classification problems with roughly equal distributions datasets. However, when applied to click fraud detection problems, the success of SVM is greatly limited due to the extreme imbalanced distribution of FDMA2012 competition dataset. In this paper, three data preprocess methods, random under-sample (RUS), synthetic minority over-sampling technique (SMOTE) and SMOTE+edited nearest neighbor(ENN), were detailed investigated, followed by SVM classifier to solve the question. Results show that the method combining SMOTE+ENN with SVM achieves accuracy about 95% on minority samples, which basically reaches the requirements of online advertising click fraud detection system.

Keywords Click fraud, SVM, Imbalanced, Mixed-sampling

1 引言

网络在线广告(Online Advertising)是一种以互联网为依托的广告投放形式。根据用户点击广告的数量,运营商按点击数向广告发布商支付广告费用的模式,是网络在线广告中的一种常见的计费方式。某些广告发布商通过伪造虚假用户浏览行为而获取超额利益,从而形成了点击欺诈问题。

由于实施点击欺诈行为的发布商只是个别现象,因此将机器学习应用于点击欺诈检测时,需要解决非平衡数据集(Imbalance Dataset)的处理问题。非平衡数据集也称倾斜数据集,是指在某个数据集合中一类或几类样本的数量较其他样本量稀疏,但通常人们对数据集中少数类的样本更有兴趣[1-2]。对于该类数据,很多分类算法因缺乏少数类数据信息,无法准确地表达出数据集固有的特征,在分类系统中引起决策边界被极大地压缩。虽然最终模型的整体准确率较高,但其并没有有效地检出需要检测的少数类目标样本[2-3]。

SVM 算法是 20 世纪 90 年代由 Vapnik 等人在 VC 维理论和结构风险最小化原则的基础上提出的针对小样本、非线性、高维问题的机器学习算法[4]。SVM 算法的基本模型结构定义为特征空间上寻找间隔最大的线性分类器:

$$f(x)=\mathrm{sgn}[w^{*\mathrm{T}}x+b]=\mathrm{sgn}[\sum_{i=1}^{l}\alpha_i^* y_i(x_i-x)+b] \quad (1)$$

由于拥有良好的实验性能和优良的泛化能力,SVM 算法在模式识别(如文本识别、手写体识别、入侵检测等)领域得到了广泛应用,并取得了良好的效果[5-6]。近年来,国内外很多学者都持续对 SVM 算法进行了改进,使得该算法的应用领域更加广泛[5]。

本文应用 SVM 算法来实现网络在线广告的点击欺诈检测,首先需要解决非平衡数据对 SVM 算法的影响,并对算法进行改进,从而提高其处理非平衡数据的性能。

2 数据集

本文采用了 Singapore Management University (SMU)于 2012 年组织的 FDMA2012 竞赛中提供的一份移动广告公司真实点击欺诈检测标准数据集[7]。该数据集由发布商数据集和

本文受国家自然科学基金资助项目(61672279),江苏省重点研发计划项目(BE2015697)资助。

李 鑫 讲师,主要研究方向为机器学习、数据挖掘,E-mail:lixin@njtech.edu.cn;**郭 汉** 硕士生,主要研究方向为机器学习、智能医学信息处理;**张 欣** 硕士生,主要研究方向为机器学习、计算广告;**胡方强** 讲师,主要研究方向为人工智能及嵌入式系统;**帅仁俊** 副教授,主要研究方向为智能建筑、智能医学图像处理。

点击数据集构成(以 CSV 文件格式存储),目的是从正常发布商中检测出实施非法点击行为的欺诈性广告发布商。实验所用的数据统计信息如表 1 所列,其中欺诈发布商和正常发布商的比例分别为 5%和 95%,具有极强的非平衡统计分布特征。

表 1 实验数据统计

点击数据集	广告发布商数据集		
点击总数量	Fraud	OK	总计
5772649	305(5.00%)	5776 (95.00%)	6081

为了将该数据集应用于 SVM 算法,首先需要对提供的原始数据集进行预处理。我们通过分析原始点击数据集中的初始特征,利用 Oentaryo 等[7]的方法为每一个发布商构造了特征向量,最后获得了 118 个不同的预测特征,以及与每个发布商相对应的特征-标签对。

3 非平稳数据集的平衡处理方法

3.1 非平衡数据集对 SVM 的影响

首先,对经过预处理的数据集直接应用 SVM 算法,结果显示对正类样本(欺诈发布商)的检出率为 0。为了进一步说明数据的非平衡性对 SVM 算法的影响,在样本中按照1:1,1:2,1:3,1:4,1:5的比例选取两类数据,并保持其他训练参数一致,利用正类正确率(Positive Accuracy,PACC)和负类正确率(Negative Accuracy,NACC)来衡量分类效果,如图 1 所示。

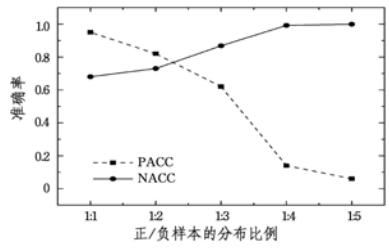

图 1 数据的不平衡性对 SVM 性能的影响

图 1 中,随着两类样本不平衡性的增加,正类样本的检出率逐渐下降,当比例为1:4和1:5时,正类样本的检出率仅分别为 14%和 6.3%。产生这种现象的原因是两类数据分布的不平衡性引起了决策平面的偏倚,从而导致少数类的决策空间被压缩,据此训练得到的 SVM 模型会将少数类错误判别为多数类。

3.2 平衡处理方法

为了解决非平衡数据对 SVM 分类算法的影响,从数据层面出发,针对预处理后的数据进行平衡处理,之后再应用 SVM 算法进行判别,具体实验流程如图 2 所示。主要采用随机采样方法从数据角度来改变非平衡性,包括随机欠采样(Random Under-Sample,RUS)和随机过采样(Random Over-sample,ROS)两种基本方法,以使两类数据达到平衡。

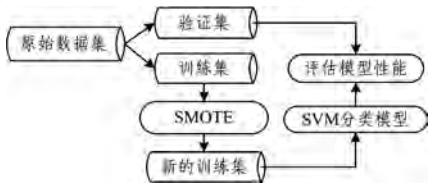

图 2 数据平衡处理实验的流程

3.2.1 随机欠采样方法

随机欠采样方法从多数类中随机选择部分多数类样本数据,再与原有少数类样本进行合并,最后得到新的平衡训练样本集。该方法的缺点是减少了数据集中的有效数据量,导致分类器的分类性能降低,本文第 4 节的实验证实了这一结果。

3.2.2 随机过采样方法

随机过采样方法选取少数类样本为处理对象,采用复制少数类样本的策略来增加少数类样本的数量[8],从而提升样本的平衡性。但简单的过采样方法可能导致无用信息不断增加,产生模型的过拟合问题。

不同于随机过采样中简单地对样本进行复制,本文使用的合成少数类过采样技术(Synthetic Minority Oversampling Technique,SMOTE)是人工合成新的少数类样本,这能有效地避免模型过拟合的问题[8]。SMOTE 算法对少数类中的一个样本 Pub_{ij} 计算其 K 近邻,并设置采样倍率 N,每一个随机选出的近邻 Pub_{ij} 分别与原样本 Pub_{ij} 按照式(2)构建新的样本 Pub_{new},rand(0,1)为 0 与 1 之间的随机数。

$$Pub_{\text{new}}=Pub_i+\text{rand}(0,1)\times(Pub_{ij}-Pub_i) \tag{2}$$

3.2.3 混合采样方法

本文采用的数据集中少数类与多数类样本数量的比例为1:17,如果仅使用 SMOTE 算法使两类样本达到平衡,可能同样会产生过分泛化的问题。因此,在 SMOTE 算法的基础上,利用 Guatavo 等[9]提出的混合采样方法,将欠采样与过采样方法相结合,即将 SMOTE 算法与最近邻规则(Edited Nearest Neighbor,ENN)相结合,对于其中一种类型样本,当与其最近的 3 个近邻样本中属于相异类型的样本数超过 2 个时,删除这些相异类型样本。例如,对于某个负类样本,在与其最近的 3 个近邻样本中,当包含 2 个或 2 个以上正类样本时,删除这些正类样本;当包含 2 个负类样本和 1 个正类样本,或 3 个均为负类样本时,样本保持不变。从多数类样本和少数类样本两方面进行处理,以避免上述两种方案所带来的负面影响。

4 实验与结果分析

本文进行一系列实验来验证所提方法对非平衡数据进行处理时的有效性。为避免抽样的偶然性与模型过拟合,每次采用 10 折交叉验证法进行实验,并将平均值作为最后的分类结果。构建 SVM 分类器时选用高斯径向基函数(RBF),其他选用默认参数。实验中,数据预处理过程对特征向量值的计算、非平衡数据的处理、SVM 模型训练分析等均利用 Python 完成。从几百万原始点击数据出发,实现了对超过 3000 个广告发布商的点击欺诈检测。

4.1 评价指标

衡量非平衡数据的分类性能时,通常使用混淆矩阵作为性能评价指标。常见的混淆矩阵结构如表 2 所列。

表 2 分类预测混淆矩阵

实际类	预测类	
	Fraud	OK
Fraud	TP	FN
OK	FP	TN

根据混淆矩阵,可以使用预测准确率(Accuracy)、正类正确率(PACC)、负类正确率(NACC)、G_mean 值以及 ROC 曲线等多个指标来评价系统分类的性能。其中:

$$Accuracy=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{NP}+\text{FP}+\text{TN}}$$

$$PACC=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$NACC=\frac{TN}{FP+FN}$$

$$G\_mean=\sqrt{PACC\times NACC}$$

4.2　实验结果

首先利用随机欠采样方法对非平衡数据进行处理，实验结果如表 3 所列。

表 3　RUS+SVM 的实验结果

实验编号	*Accuracy*	*PACC*	*NACC*	*G_mean*
1	0.6026	0.9688	0.5822	0.7510
2	0.6059	0.9688	0.5857	0.7533
3	0.6002	0.9688	0.5796	0.7493
4	0.5820	0.9688	0.5605	0.7369
5	0.5977	0.9688	0.5770	0.7477

利用随机过采样方法，分类器对正类的预测准确率达到了 95%以上，完全达到了目标效果；但其对负类的预测准确率低于 60%，这意味着存在较大数量的误判问题，使得正常发布商被检测为欺诈发布商。其主要原因是，由于未考虑样本的分布情况，在对多数类样本进行采样时具有较大的随机性，同时采样的样本数量远少于原样本数量，因此会造成一些信息缺失，未被采样的多数类样本可能携带有更加重要的信息[10]。

接下来使用 SMOTE 算法进行数据平衡处理。由于 SMOTE 算法中的近邻数（K 值）以及采样倍率（*ratio*）对最终构建的 SVM 分类器的性能有一定的影响，为获得更好的分类效果，首先对 K 值和采样倍率对模型分类准确性的影响进行研究。图 3 与图 4 分别是 K 值和倍率（*ratio*=1 表示两类数据样本数量为1∶1）对 SVM 分类器预测结果中正类正确率 *PACC* 和负类正确率 *NACC* 的影响。可以发现，随着采样率的变化，不同 K 值下的 *PACC* 与 *NACC* 分别经历了一个由低到高（图 3）和由高到低（图 4）的阶跃型变化趋势，整体走势大致相同。当倍率为 0.5 时，*PACC* 接近于 1，同时 *NACC* 也较高，且此时的 *G_mean* 值也较大，说明此时分类器对两种样本的检测均达到了较高水平。当采样率继续增大时，*PACC* 基本保持不变，而 *NACC* 出现下降，分类器的性能明显恶化。进一步比较采样倍率为 0.5 时 K 值的变化对 3 种指标的影响，综合考虑 *NACC*，*PACC* 和 *G_mean* 3 个参数，可以发现 K 的最佳取值为 5。因此，在随后的实验中将 *SMOTE* 算法中的 K 值和采样倍率 *ratio* 分别设定为 5 和 0.5。

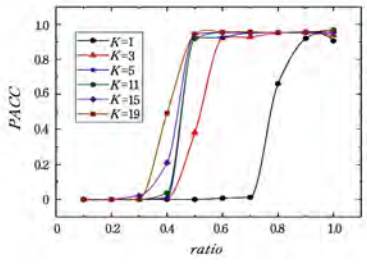

图 3　K 值和采样率 *ratio* 对 SVM 分类器 *PACC* 的影响

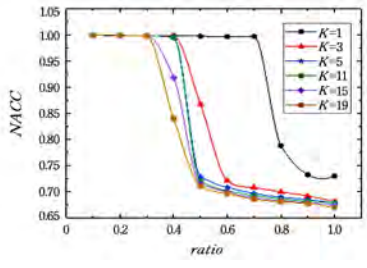

图 4　K 值和采样率 *ratio* 对 SVM 分类器 *NACC* 的影响

确定 K 值和采用倍率后，应用 SMOTE 算法将处理后的数据送入 SVM 分类器进行分类实验，结果如表 4 所列。对比表 3 和表 4 可以发现，*PACC* 虽略有下降，但仍维持在 95%以上，同时 *NACC* 由 58%提升至 67%左右，分类器的性能有了较为显著的提升。

表 4　SMOTE+SVM 的实验结果

实验编号	*Accuracy*	*PACC*	*NACC*	*G_mean*
1	0.6981	0.9538	0.6819	0.8011
2	0.7075	0.9538	0.6752	0.8025
3	0.6955	0.9538	0.6687	0.7993
4	0.6950	0.9538	0.6746	0.8020
5	0.6921	0.9569	0.6561	0.7923

最后，将 SMOTE+ENN 混合采样方法处理后的数据作为新的数据集用于训练 SVM 分类器，仍选用高斯径向基函数（RBF）作为 SVM 的核函数，其他选用默认参数，实验结果如表 5 所列。表 5 显示，采用混合采样方法处理数据后，正类样本的正确率 *PACC* 和负类样本的正确率 *NACC* 较单独 *SMOTE* 方法又有了一定的提升，*PACC* 升高至与 *RUS* 方法相同，同时 *NACC* 继续提升至 70%左右。

表 5　SMOTE+ENN+SVM 的实验结果

实验编号	*Accuracy*	*PACC*	*NACC*	*G_mean*
1	0.7049	0.9688	0.6902	0.8177
2	0.7148	0.9688	0.7006	0.8238
3	0.7106	0.9688	0.6963	0.8213
4	0.7049	0.9688	0.6902	0.8177
5	0.7082	0.9688	0.6936	0.8197

综合比较 3 个实验结果可以发现，虽然较 SMOTE 算法对整体样本的预测正确率提升有限，但采用混合采样方法处理非平衡数据后构建的 SVM 分类模型对整体样本的预测正确率比另两种方法要高。总体来说，混合抽样方法可以在一定程度上避免因为欠抽样造成的重要信息丢失以及过抽样造成的过拟合现象。

上述实验证明，与单纯采用 SVM 方法相比，RUS+SVM，SMOTE+SVM 和 SMOTE+ENN+SVM 3 种方法对欺诈发布商的检测性能均有较大程度的改善。为更加直观地分析改善程度，利用 ROC 曲线和 *AUC* 对 3 种方法作进一步的比较，ROC 曲线如图 5 所示。

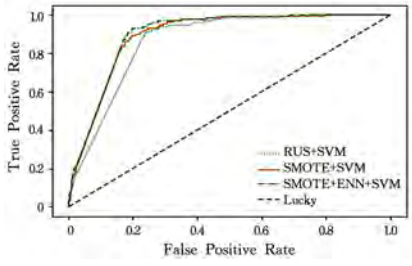

图 5　ROC 曲线对比图

ROC 曲线越往左上角凸，即越靠近（0，1）点，说明分类器的性能越好。图 5 中，SMOTE+ENN+SVM 方法的 ROC 曲线最靠近左上角（0，1），而且完全覆盖 SMOTE+SVM 与 RUS+SVM 的 ROC 曲线。另一方面，3 种方法的 *AUC* 值分别为 0.88，0.85 和 0.80，同样表明 SMOTE+ENN+SVM 方法的处理效果最佳，其对目标数据集中欺诈发布商的检测获得了较高的准确度（95%左右），基本满足网络在线广告中对

发布商点击欺诈检测问题的要求。

结束语 非平衡数据集中少数类别样本的检测是实际工作中经常会遇到的问题，针对非平衡数据的分类检测也一直是异常检测研究的重要内容。通常，少数类数据代表危害较大的恶意或非法行为，对该类别样本的检测必须要有较高的准确度。本文分别利用 RUS，SMOTE 及 SOME+ENN 3 种方法，结合 SVM 算法对网络在线广告欺诈点击问题进行研究。实验结果表明，使用 SOME+ENN 混合采样方法，结合 SVM 算法对网络在线广告欺诈点击检测最为有效，其对少数实施点击欺诈的非法广告发布商的检测率达 95%左右。

参 考 文 献

[1] ZHANG S，SADAOUI S，MOUHOUB M. An Empirical Analysis of Imbalanced Data Classification[J]. Computer & Information Science，2015，8(1)：151-162.

[2] 尹留志. 关于非平衡数据特征问题的研究[D]. 合肥：中国科学技术大学，2014.

[3] JIAN C，GAO J，AO Y. A new sampling method for classifying imbalanced data based on support vector machine ensemble[J]. Neurocomputing，2016，193(C)：115-122.

[4] VAPNIK V N. The nature of statistical learning theory [M]. New York：Springer Verlag，1995.

[5] 崔建明. 基于 SVM 算法的文本分类技术研究[J]. 计算机仿真，2013，30(2)：299-302.

[6] 董亚楠，刘学军，李斌. 一种基于用户行为特征选择的点击欺诈检测方法[J]. 计算机科学，2016，43(10)：145-149.

[7] OENTARYO R，LIM E P，FINEGOLD M，et al. Detecting click fraud in online advertising：a data mining approach [J]. Journal of Machine Learning Research，2014，15(1)：99-140.

[8] CHAWLA NV，BOWYER KW，HALL LO，et al. SMOTE：synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research，2011，16(1)：321-357.

[9] GUSTAVO E A，BATISTA P A，RONALDO C，et al. A study of the behavior of several methods for balancing machine learning training data[J]. SIGKDD Explorations，2004，6(1)：20-29.

[10] 于化龙，高尚，赵靖，等. 基于过采样技术和随机森林的不平衡微阵列数据分类方法研究[J]. 计算机科学，2012，39(5)：190-194.

(上接第 331 页)

通过仿真结果得出，对于基于概率型的动态帧时隙 ALOHA 算法及其改进算法，其系统效率即系统吞吐量不会超过 37%。当标签数等于帧时隙数时系统有最大的吞吐量 36.8%；当标签相对少量时，随着标签的增加，使用动态帧时隙 ALOHA 算法会使系统效率达到最大值；但当标签超过一定数量继续增加时，动态帧时隙 ALOHA 算法已不适应该情况，使用本文的动态帧时隙 ALOHA 改进算法来识别标签能维持较高的系统效率和吞吐量。

结束语 本文在标签估计中提出了动态因子均值估计算法，并对该标签估计算法进行了 MATLAB 仿真。本文算法能够对未识别标签进行准确地估计，使估计的误差率维持在 5%左右，为下一步准确地分组提供依据。最后结合该标签估计算法对动态帧时隙 ALOHA 算法进行改进。对动态帧时隙 ALOHA 改进算法通过 MATLAB 进行识别标签数量从 0 增加到 1000 的仿真。仿真实验证明：1)基于动态因子均值标签估计算法的动态帧时隙 ALOHA 算法能够保证帧时隙的较高的系统利用率且稳定在 30%以上。2)基于动态因子均值标签估计算法的动态帧时隙 ALOHA 改进算法能够使用较少的帧时隙数完成标签的识别任务，比传统动态帧时隙 ALOHA 算法所用时隙数减少了 45%左右。

本文重点是对标签估计算法做出改进，下一步可以在此基础上重点对标签分组法进行改进，使基于概率型的 ALOHA 算法的帧时隙资源利用率更加逼近理论值，使系统吞吐率更加稳定。

参 考 文 献

[1] 单剑锋，陈明，谢建兵，等. 基于 ALOHA 算法的 RFID 防碰撞技术研究[J]. 南京邮电大学学报(自然科学版)，2013，33(1)：56-61.

[2] 潘雪峰，曹加恒，PAN X F，等. 一种改进的动态帧时隙 ALOHA 算法[J]. 微电子学与计算机，2016，33(6)：95-99.

[3] GITAKRISHNAN R，SATISH U. Fitted dynamic framed slotted ALOHA anti-collision algorithm in RFID systems[C]// Proceedings of the International Conference on Information Technology and Multimedia. 2012：1-6.

[4] 潘思丞，王慧琴，张小红. 静态环境中分组 ALOHA 防碰撞算法研究[J]. 计算机工程与应用，2016，52(20)：114-117.

[5] 杨帆，徐焕良，谢俊，等，基于双空闲因子的 RFID 防碰撞算法研究[J]. 计算机工程与科学，2016，38(7)：1440-1446.

[6] XU Y，CHEN Y. An improved dynamic framed slotted ALOHA Anti-collision algorithm based on estimation method for RFID systems[C]// Poceedings of the IEEE International Conference on RFID. 2015：1-8.

[7] 刘金艳，冯全源. 无线射频识别多标签防碰撞算法综述[J]. 计算机集成制造系统，2014，20(2)：440-451.

[8] 卢迪，李绅龙，许成舜. CHI 标签估计下自适应帧长调整 DFSA 算法[J]. 哈尔滨理工大学学报，2015，20(1)：56-60.

[9] CHONG S K，LAI N S. Dynamic framed slotted ALOHA algorithm for RFID systems with enhanced tag estimation technique [C]// proceedings of the IEEE International Conference on Rfid-Technologies and Applications. 2013 .

[10] SCHOUTE F C. Dynamic Frame Length ALOHA [J]. Mobile Communications，1983，31(4)：565-568.

[11] VOGT H. Efficient Object Identification with Passive RFID Tags[C]// 2002 IEEE International Conference on Proceedings of the Systems，Man and Cybernetics. 2002.

[12] 庞宇，彭琦，林金朝，等. 基于分组动态帧时隙的射频识别防碰撞算法[J]. 物理学报，2013，62(14)：488-495.

[13] 钱东昊，张琨，张磊. 基于标签识别码分组的防碰撞算法研究[J]. 计算机应用与软件，2015，32(7)：252-254.