基于支持向量上采样的不平衡数据分类方法

曹 路

(五邑大学信息工程学院 江门 529020) (中山大学数据科学与计算机学院 广州 510006)

摘 要 传统的支持向量机在处理不平衡数据时效果不佳。为了提高少类样本的识别精度,提出了一种基于支持向量的上采样方法。首先根据K近邻的思想清除原始数据集中的噪声;然后用支持向量机对训练集进行学习以获得支持向量,进一步对少类样本的每一个支持向量添加服从一定规律的噪声,增加少数类样本的数目以获得相对平衡的数据集;最后将获得的新数据集用支持向量机学习。实验结果显示,该方法在人工数据集和UCI标准数据集上均是有效的。

关键词 支持向量,采样,不平衡数据,分类

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.017

Imbalanced Data Classification Method Based on Support Vector Over-sampling

CAO Lu

(School of Information Engineering, Wuyi University, Jiangmen 529020, China)

(School of Data Science and Computer Science, Sun Yat-sen University, Guangzhou 510006, China)

Abstract Traditional support vector machine has drawbacks in dealing with imbalanced data. In order to improve the recognition accuracy of the minority class, an over-sampling method based on support vector was proposed. Firstly, K nearest neighbor technology is used to remove the noise from the original data set. Support vector machine learning is then used to obtain the support vector. Noise obeying a certain rule is added to each support vectors of the minority class to increase the number of minority class samples in order to obtain the relative balanced data set. Finally, the support vector machine is learned on the new data set. The experimental results show that the proposed method is effective on both artificial data sets and UCI standard data sets.

Keywords Support vector, Sampling, Imbalanced data, Classification

1 引言

支持向量机(Support Vector Machine,SVM)是基于统计学理论的机器学习方法,具有坚实的理论基础,它克服了传统分类器过学习、局部极值和维数灾难等缺点,已成为目前研究的热点[1]。支持向量机具有坚实的理论基础,主要体现在3个方面:1)最大间隔原理;2)对偶理论;3)核函数的引入。其中最大间隔理论将支持向量机的原始问题最终转化为求解一个凸二次规划问题,由对偶理论顺利引入核函数,用于求解非线性问题的分类。

不平衡数据集在现实生活中广泛存在,如疾病诊断、入侵检测等。在不平衡数据集中,样本数较多的类别称为多类,样本数较少的类别称为少类,其中少类样本的识别是分类的重点。如在入侵检测中,入侵行为的次数一定远远小于正常事件的次数,但如果将一次入侵行为判断为正常事件,可能会遭受严重的损失。支持向量机是一种有监督的学习方法,是基于类分布平衡假设这一条件而设计的,其主要目的是提高整个数据集的分类精度,在处理不平衡数据集时效果不佳。因此,支持向量机的不平衡数据分类问题成了学者们关注的热点。

目前,不平衡数据集分类问题的解决方法主要从算法层面和数据层面两个方面着手。从算法方面处理不平衡数据主要是通过改进现有的算法使分类更偏向于少类,代表性的有代价敏感学习方法[2]、集成算法[3]和提升算法[4]。本文更侧重于数据层面对不平衡数据的分类问题。数据层面的解决方法主要是通过改变训练集中多类和少类的样本数目,减少多类样本数(下采样),增加少类样本数(上采样)或通过混合采样的方法来达到使训练集多类和少类样本大致平衡的目的。具体工作将在第2节说明。

Bishop等人指出在训练集中添加噪声可提高分类器的泛化能力,同时证明了在噪声标准差较小时,在训练集中添加噪声等价于神经网络的正则化,而且正则化系数与噪声标准差有关[5]。Yang等人提出了一种基于高斯分布的虚拟样本生成技术,并且证明了在小样本情况下,利用虚拟样本技术进行学习等价于正则化方法,亦证明对不平衡数据分类问题,利用虚拟样本技术进行学习等价于代价敏感学习[6]。支持向量机是非常依赖于支持向量的分类器,本文根据SVM这一重要特性,提出了一种利用标准正态分布对支持向量上采样的方法。首先利用K近邻的思想对原始数据进行清洗,清除原始数据集中的噪声;然后对少类的每一个支持向量添加服从

到稿日期:2016-08-20 返修日期:2016-10-31 本文受广东省特色创新类项目(2015KTSCX143),广东省青年创新人才项目(2015KQNCX172),江门市科技计划项目(江科[2016]189号,江科[2015]138号),五邑大学青年基金(2013zk07,2015zk11)资助。

曹 路(1983—),女,博士生,讲师,主要研究领域为模式识别、机器学习,E-mail:caolu20001742@163.com。

一定规律的噪声，使生成的样本的均值和方差的期望与原始少类样本的期望和方差接近，通过这种对支持向量上采样的方法增加少数类样本的数目以获得相对平衡的数据集；最后将获得的新数据集用支持向量机学习。实验结果显示，本文所提方法是可行的。

2 相关知识

2.1 不平衡数据分类数据层面的处理

在数据层面，最简单的上采样和下采样的方法是随机复制少类样本和随机减少多类样本，其虽能达到使数据平衡的目的，但实际效果并不理想[7]。人们更倾向于采用各种启发式的算法从数据层面对不平衡数据集进行处理。

在上采样方面，SMOTE 算法（Synthetic Minority Over-sampling Technique）是最为典型的上采样算法，其基本思想是通过 K 近邻的方法寻找给定少类样本的 k 个最近邻，通过在给定样本点和 k 个最近邻之间插值构造人工样本，以达到增加少类样本数的目的[8]。SMOTE 算法能在一定程度上改善不平衡数据分类的性能，但因未考虑邻近样本点的分布，易引起噪声，不仅容易过拟合还会增加算法复杂度。BSMOTE（Borderline Synthetic Minority Over-sampling Technique）是基于 SMOTE 的改进方法，该算法只考虑少类边界样本的线性插值，以保证生成样本的有效性[9]。Gao 等人提出了一种基于核密度估计的上采样方法，该方法先通过核密度估计获得少类样本的概率密度函数，再根据概率密度函数对少类样本进行上采样，获得比较符合原始少类样本分布的合成样本[10]。文献[11]提出了两种概率抽样方法 RACOG（Rapidly Converging Gibbs）和 wRACOG（wrapper-based Rapidly Converging Gibbs），这两种方法均是通过联合概率分布的数据属性和吉布斯抽样生成新的少数类样本，其中，RACOG 根据预定义滞后的马尔科夫链选取新的样本，wRACOG 选取分类器在概率上最容易错分的样本为新的样本。文献[12]提出了一种基于马氏距离的上采样方法 MDO（Mahalanobis Distance-based Over-sampling），该方法通过保留少数类样本的协方差结构，并通过概率等高线生成新的样本，减少不同类别重叠的风险。

在下采样方面，OSS（One Side Selection）算法是一种较为普遍的下采样方法，其将多类样本分为噪声样本、边界样本、冗余样本和安全样本，根据 Tomek Links 技术去掉噪声点和边界点以减少少类样本数目[13]。文献[14]提出了一种基于聚类的下采样方法，该方法将所有的训练样本分为若干个簇，然后在聚类中选择有代表性的数据作为训练样本，以提高少数类的分类精度。文献[15]提出了一种利用多层感知器动态采样数据的方法 DyS（Dynamic Sampling Method）来解决多类不平衡问题。对每一个训练样本，该算法计算其可能被选中的概率，当概率大于某一个随机数时，即认为该样本是具有代表性的点，否则不被用于训练过程。文献[16]提出了一种基于 BP 神经网络的单边动态下采样技术 ODU（One-sided Dynamic Under-sampling），利用 BP 神经网络通过不断迭代训练的方法确定多类中用于训练的样本以平衡数据集。文献[17]提出了一种基于多元敏感性的下采样技术。该方法对多类样本进行聚类获得样本的分布信息，以提高采样的多样性；然后通过随机敏感策略从每一个簇中挑选多类和少类，最后利用不断的聚类和迭代来获得一个相对平衡的数据集。

2.2 不平衡数据集支持向量机分类性能的影响

为了测试不平衡数据集对支持向量机的影响，利用均匀分布产生两类样本点，如图 1 所示，在横轴上服从(0,1)的均匀分布，在纵轴上分布服从(0,1)和(−1,0)的均匀分布。在样本数均衡的情况下，理想情况下的线性分类面应该是通过横轴的直线($y=0$)。从图 1(a)中可以看到，当两类样本点均为 1000 时，通过支持向量机所获得的分类面比较接近于理想分类面。图 1(b)－图 1(d)中多类和少类的数目分别为1000∶200，1000∶50 和 1000∶10。从图中可以看到，随着不平衡率的变化，在多类样本数目不变而少类样本数不断减少的情况下，分类面会远离理想分类面而偏向于少类样本，这会导致支持向量机对少类样本产生较大的测试误差，不利于少类样本的识别。

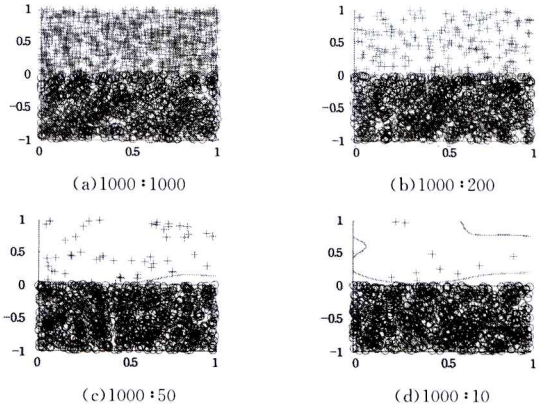


图 1 不同不平衡率下分类面的变化

3 基于支持向量的上采样技术

支持向量机是通过最大化分类间隔实现的，距离最大化间隔决策边界最近点的集合是支持向量，即决策面实际是由支持向量决定的。在考虑支持向量机下的不平衡数据分类时，要给予少类的支持向量更多关注以提高少类样本的识别能力。本文提出了基于支持向量的上采样方法，具体流程如下。

Step 1 将原始数据集分为训练集 T 和交叉验证集 V 两部分。

Step 2 利用 K 近邻的思想过滤训练集 T 中的噪声样本。方法如下：需要给定样本的 K 个近邻($K=5$)，超过 4/5 为相反类别，可以认为该样本为噪声，需在训练集中滤除。去噪后可获得新的训练集 T'。

在本算法中，清除噪声样本是有必要的，清除噪声样本不仅可以在一定程序上减少训练样本，提高分类效率，更为重要的是为下一步对支持向量的上采样进行预处理，因为如果少类中的噪声样本为支持向量，将其上采样后，将在训练集中引入更多噪声而影响分类性能。

Step 3 获得训练集 T' 中的多类 T_{maj} 和少类 T_{min}，利用支持向量机对训练集 T' 分类，以获得支持向量。

Step 4 对支持向量进行上采样。

Step 4.1 计算 T' 中少类样本的个数 n、少类样本每个特征的均值 μ_i 和方差 σ_i^2，$i=0,\cdots,m$。m 为原始样本的特征个数，每一维的属性为 a_i，$i=0,\cdots,m$。

Step 4.2 对每个支持向量的每个属性按式(1)生成合成样本。

$$a_i^{new}=a_i-\Delta_i\cdot\frac{\sigma_i}{\sqrt{n}} \tag{1}$$

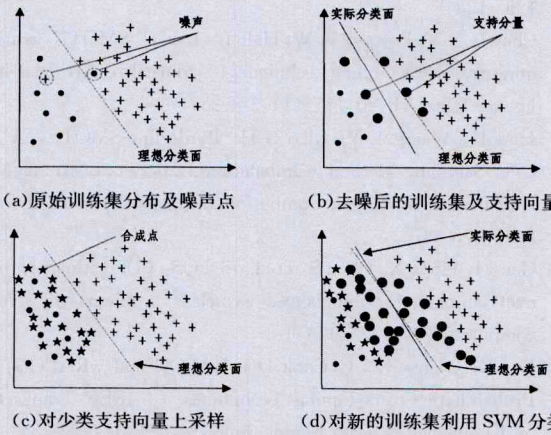
其中，a_i 为少类样本每一维的属性，Δ_i 为服从标准正态分布的随机变量，n 为少类样本的个数，σ_i 为少类样本每个特征的标准差。根据中心极限定理可知，当 $n\to\infty$ 时，生成的样本的均值和方差的期望与原始少类样本的期望和方差接近。

Step 4.3　重复上一步骤直至生成所需样本个数。

Step 5　生成的样本、少类样本 T_{min} 和多类样本 T_{maj} 一起构成新的训练集，以获得相对平衡的数据集。

Step 6　用分类器对获得的新的数据集进行分类，并通过交叉验证的测试集进行性能评估。

通过图 2 可以直观地了解基于支持向量的上采样算法的原理。



(a)原始训练集分布及噪声点　(b)去噪后的训练集及支持向量

(c)对少类支持向量上采样　(d)对新的训练集利用 SVM 分类

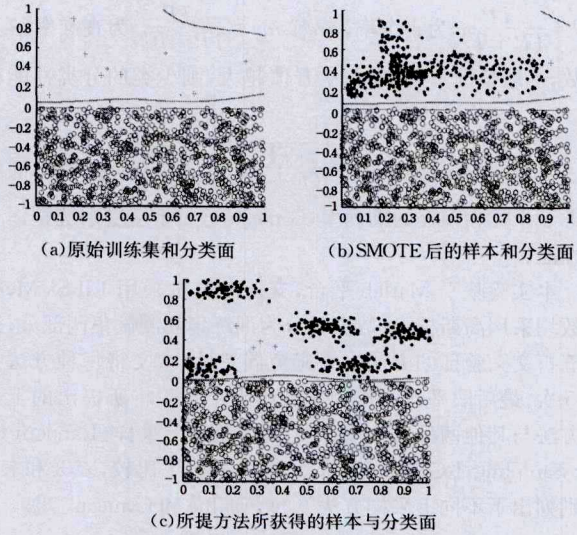
图 2　基于支持向量的上采样算法图示

原始训练集数据分布如图 2(a)所示，其中圆形为少类，“+”为多类，根据 K 近邻的思想，找到多类和少类中的噪声点并进行滤除。图 2(b)为滤除掉噪声以后的训练集用支持向量进行分类，图中的圈加和圈点分别为多类和少类的支持向量，虚线为理想分类面，实线为利用支持向量机所获得的分类面。支持向量机基于最大间隔理论，在数据高度不平衡时所获得的分类面与理想分类面有一定差距。图 2(c)为对少类支持向量进行上采样后所获得的新的训练集的示意图。从图中可以看到，对支持向量进行上采样，不仅可以增加少类样本的数目，而且可以增加少类的边界样本点。图 2(d)为上采样后的样本、少类样本和多类样本构成新的数据集后，用支持向量机分类后的效果，虚线为理想分类面，实线为实际分类面，圆圈所标注的均为支持向量。可以看到，由于对原训练集中的支持向量进行了上采样，增加了少类的边界样本点，使分类面不再向少类倾斜，实际分类面与理想分类面靠近。

4　实验

4.1　人工数据集

为了验证基于支持向量的上采样方法的可行性，采用人工数据集进行验证。实验中的人造样本服从均匀分布，理想情况下的线性分类面是通过横轴的直线。在本实验中，原始训练集中多类和少类的数目比为 1000∶10，多类样本点用圆圈表示，少类样本点用加号表示，实线为通过 SVM 获得的分类面，虚线为理想分类面。为获得相对平衡的数据集，经过上采样后的少类样本数设为 500，实心圆点为上采样后的合成点。图 3 为不同方式下的分类效果图，本实验所用数据集均采用高斯核的 SVM 分类。



(a)原始训练集和分类面　(b)SMOTE 后的样本和分类面

(c)所提方法所获得的样本与分类面

图 3　不同方式下的分类效果图

图 3(a)所示为原始样本和通过 SVM 所获得的分类面。SMOTE 是一种最典型的上采样方法，图 3(b)所示为将少类样本进行 SMOTE 上采样后所获得的训练集与分类面，可以看到，相较于原始不平衡数据集，SMOTE 上采样能在一定程度上提高分类性能。由于 SMOTE 是少类样本之间的线性插值，采样具有一定的盲目性，因此即使能降低多类和少类的不平衡率，其效果并不令人满意。图 3(c)为采用基于支持向量的上采样的方法所获得的样本和分类面，其中星形为原始少类样本经过 SVM 所获得的支持向量。从图中可以看到，采用支持向量上采样后，生成了对分类影响较大的边界样本，使最终获得的分类面非常贴近理想分类面，相较于 SMOTE 算法，所提方法对少类的预测效果有了较大提升。

4.2　UCI 数据集

为进一步评估所提方法的性能，选择 5 组 UCI 标准数据集进行测试。为简化实验，将多类样本转化为二值类处理。对多个类别的数据集，其中一类为少数类，剩余的合并为多数类。数据集的总体描述如表 1 所列。

表 1　数据集

训练集	样本数	属性	多类	少类	不平衡率
Germen	1000	2	700	300	2.33
Glass7	214	9	185	29	6.4
Vowel	990	13	900	90	10.0
Yeast	1332	8	1248	84	14.9
Letter	200000	16	19266	734	26.25

传统的分类器以分类精度为评价标准。对于不平衡数据分类，以分类精度来衡量存在不合理性。如对于高度不平衡数据，即使将少类样本全部判定为多类，由于多类样本数目较多，也会得到较高的分类精度。目前，不平衡问题分类的评价标准有 F-value 和 G-mean 等，均是建立在混淆矩阵的基础之上，如表 2 所列。

表 2　混淆矩阵

	预测为正(少)类	预测为负(多)类
正(少)类样本	TP 正确分类的正类数	FN 错误分类的负类数
负(多)类样本	FP 错误分类的正类数	TN 正确分类的负类数

F-value 定义为 $F=\dfrac{(1+\beta^2)\times recall\times precision}{\beta^2\times recall+precision}$，其中 *re-*

$call=\frac{TP}{TP+TN}$为查全率，$precision=\frac{TP}{TP+FN}$为查准率，$\beta$为参数，一般情况下可取为1。$F$值越大，则少类的分类性能越好。

G-mean定义为$F=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}$。由于同时考虑了少类和多类的准确率，G-mean可用于衡量系统整体分类性能。

本实验基于Matlab平台，支持向量机采用LIBSVM，核函数均采用高斯核，将数据集分为训练集和测试集两部分，采用五折交叉验证的方法。为避免偶然性，本文将每种方法执行5次，最后取平均值作为实验结果。实验将所提出的上采样方法与其他两种常用上采样算法随机上采样(Random Over-Sampling，ROS)和SMOTE算法进行了比较，表3和表4分别列出了不同上采样方法下的F-value和G-mean。

表3　数据集在不同方法下的F-value

数据集	SVM	ROS	SMOTE	所提方法
Germen	0.561	0.560	0.617	0.672
Glass7	0.815	0.827	0.854	0.886
Vowel	0.827	0.842	0.897	0.911
Yeast	0.632	0.637	0.702	0.781
Letter	0.543	0.574	0.689	0.765

表4　数据集在不同方法下的G-mean

数据集	SVM	ROS	SMOTE	所提方法
Germen	0.703	0.709	0.718	0.745
Glass7	0.813	0.842	0.885	0.913
Vowel	0.843	0.867	0.891	0.925
Yeast	0.663	0.669	0.743	0.865
Letter	0.554	0.561	0.713	0.796

从表3和表4中可以看出，由于随机上采样方法具有一定的随意性，在利用支持向量机进行分类时并不能带来明显的效果；SMOTE上采样方法在F-value和G-mean上相较于随机上采样有所提升。在不平衡率较低的数据集Germen和Glass7上，所提方法在F-value和G-mean上均有不同程度的提高；在不平衡较高的数据集Letter上，基于支持向量的上采样方法亦有不错的表现。G-mean和F-value的提高说明所提方法不仅可以提高不平衡数据整体的分类性能，同时能改善少类的分类性能。

结束语　针对不平衡数据分类问题，提出了一种基于支持向量的上采样方法。该方法利用支持向量机对支持向量比较敏感的这一特性，按照一定规则在支持向量的周围添加一定数量的噪声用于平衡数据集，同时使生成的样本符合原始数据的统计特性。人工数据集和标准UCI数据集的实验结果均显示，所提方法是有效的。由于上采样的方法增加了样本数量，在一定程度上会影响分类器的复杂度，因此提出既有良好效果又有较快收敛速度的算法是下一步的研究内容；同时，将所提方法用于多类别不平衡数据集，亦值得深入研究。

参考文献

[1] Vapnik V N. 统计学习理论[M]. 许建华，张学工，译. 北京：电子工业出版社，2004

[2] Castro C L, Braga A P. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data[J]. IEEE Transactions on Neural Networks & Learning Systems, 2013, 24(6): 888-899

[3] Li Y, Liu Z D, Zhang H J. Review on Ensemble Algorithms for Imbalanced Data Classification [J]. Application Research of Computers, 2014, 31(5): 1288-1291(in Chinese)
李勇，刘战东，张海军. 不平衡数据的集分类方法综述[J]. 计算机应用研究，2014，31(5)：1288-1291

[4] Galar M, FernaáNdez A, Barrenechea E, et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches[J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2012, 42(4): 463-484

[5] Bishop C. Training with Noise is Equivalent to Tikhonov Regularization[J]. Neural Computation, 1995, 7(1): 108-116

[6] Yang J, Yu X, Xie Z Q, et al. A novel virtual sample generation method based on Gaussian distribution [J]. Knowledge-Based Systems, 2011, 24(6): 740-748

[7] He H, Garcia E A. Learning from Imbalanced Data [J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9): 1263-1284

[8] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321-357

[9] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[M]// Advances in Intelligent Computing. Springer Berlin Heidelberg, 2005: 878-887

[10] Gao M, Hong X, Chen S, et al. PDFOS: PDF estimation based over-sampling for imbalanced two-class problems [J]. Neurocomputing, 2012, 138(11): 1-8

[11] Das B, Krishnan N C, Cook D J. RACOG and wRACOG: Two Probabilistic Oversampling Techniques[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(1): 222-234

[12] Abdi L, Hashemi S. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques [J]. Soft Computing, 2014, 19(12): 3369-3385

[13] Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection [C] // Proceedings of the Fourteenth International Conference on Machine Learning. 2000: 179-186

[14] Yen S J, Lee Y S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications, 2009, 36(3): 5718-5727

[15] Lin M, Tang K, Yao X. Dynamic sampling approach to training neural networks for multiclass imbalance classification[J]. IEEE Transactions on Neural Networks & Learning Systems, 2013, 24(4): 647-660

[16] Fan Q, Wang Z, Gao D. One-sided Dynamic Undersampling No-Propagation Neural Networks for imbalance problem[J]. Engineering Applications of Artificial Intelligence, 2016, 53(c): 62-73

[17] Ng W W, Hu J, Yeung D S, et al. Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems[J]. IEEE Transactions on Cybernetics, 2014, 45(11): 2402-2412

[18] Zhang X S, Luo Q. Unbalanced Data Classification Algorithm Based on Clustering Ensemble Under-sampling [J]. Computer Science, 2015, 42(11): 63-66(in Chinese)
张枭山，罗强. 一种基于聚类融合欠抽样的不平衡数据分类方法[J]. 计算机科学，2015，42(11)：63-66

[19] Cao L, Wang P. Imbalanced Data Classification Based on SMOTE Sampling and the Support Vector Machine [J]. Journal of wuyi university(Natural Science Edition), 2015, 29(4): 27-31(in Chinese)
曹路，王鹏. 基于SMOTE采样和支持向量机的不平衡数据分类[J]. 五邑大学学报(自然科学版)，2015，29(4)：27-31