SVM在非平衡数据集中的应用

黄秀丽,王 蔚
(南京师范大学 教育科学学院,江苏 南京 210097)

摘 要:在一个数据集中,至少有一个类别相对与其他类别有很少的样本,则这样数据集可以称为高度倾斜的或者是非平衡的数据集。非平衡数据在现实中普遍存在。在非平衡数据分类中,传统机器学习算法的分类表现受到了阻碍。支持向量机(SVM)基于结构风险最小化原则,是近几年发展起来的机器学习方法。分析了SVM在非平衡数据集中的应用情况,同时提出了几种SVM运用于非平衡数据集中的主要改进方法,这些方法对于非平衡数据的分类有很好的分类效果。

关键词:非平衡数据;SVM;机器学习
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2009)06-0190-04

Application of SVM in Imbalances Dataset

HUANG Xiu-li, WANG Wei
(School of Education, Nanjing Normal University, Nanjing 210097, China)

Abstract: A training dataset is called imbalance if at least one of the classes are represented by significantly less number of instances than the others. The class imbalance problem occurs when there is significantly less number of observations of the target concept. Various real-world classification tasks suffer from this phenomenon. The class imbalance problem has been known to hinder the learning performance of classification algorithms. The support vector machine theory is based on the minimization principle to structure risk. Support vector machine is an algorithm of machine learning that has developed during these years. Summarizes the state of the application of SVM in imbalances data. Then introduce some algorithms improved to get good performance.

Key words: imbalances data; SVM; machine learning

0 引 言

在一个数据集中,至少有一个类别相对与其他类别有很少的样本,则这样数据集可以称为高度倾斜的或者是非平衡的数据集(Imbalances dataset)。对于一个标准的两类分类问题,样本较多的类被称为正类,样本较少的类别被称为负类。现实生活中,导致非平衡类别的一方面是有意义的数据普遍较少;另外一个原因是由于对一些特定类别的样本的搜索限制。实际生活中的非平衡问题有:识别欺诈信用卡交易,文本分类,蛋白质数据库分类,以及从卫星图像中探测特定的目标。通常,把正类样本误判为负类的损失远大于把负类样本误判为正类的损失。AAAI分别在2000年和2003年举办两届研讨会,专题讨论非平衡学习问题,这两届研讨会可以看作是这一问题引起全面关注的标志。

在非平衡数据分类中,传统机器学习算法主要考虑的是各类学习样本数量大致平衡的情形,其评价标准主要是基于精度的,得到的数据边界将会严重的向目标类倾斜。结果,负误识(False negative)的比率就特别高。支持向量机(SVM)是Vapnik等人提出的一类新型机器学习方法,以统计学习理论为基础,具有严格的理论和数学基础[1]。不同于神经网络、决策树等传统算法基于经验最小化原则,SVM基于结构风险最小化原则即同时考虑经验风险和置信范围,获得了良好的泛化性能。文献[2]对非平衡数据的学习做了一个全面而又系统地分析,采用35个来自不同领域的现实非平衡数据集,对11种学习算法进行了研究。提出目前大多数关于非平衡数据集的研究集中在决策树上,未来的研究可以多考虑SVM,神经网络等的改进方法。

实验研究表明,相对而言,SVM分类器对数据的非平衡性更不敏感[3]。因此,人们对SVM提出了各种改进方法以更好地处理非平衡数据。

收稿日期:2008-09-28;**修回日期**:2008-12-12
基金项目:全国教育科学"十五"规划教育部重点基金项目(DCA050056);江苏省教育科学"十一五"规划项目(D/2006/01/096)
作者简介:黄秀丽(1986-),女,硕士研究生,研究方向为机器学习与数据挖掘;王 蔚,博士,教授,研究方向为机器学习与数据挖掘、信号处理等。

1　非平衡数据中 SVM 的应用

1.1　支持向量机概述

支持向量机(SVM)算法的目的在于找到一个最优超平面,使得边界数据垂直于超平面的距离最大。以两类分类为例,假设$((x_1,y_1)\cdots(x_n,y_n))$是训练数据集,$x_i$是代表数据,$y_i\in(-1,+1)$是分类标签。通过该数据集,SVM 建立一个在高维特征空间的线性区分。通过最大化边界距离来区分这两个类。对应如下的二次规划函数:

$$\min\frac{1}{2}w\cdot w^{\mathrm{T}}+C\sum_{i=1}^{N}\xi_i \tag{1}$$

$$\text{s.t.}\begin{cases}\forall\, iy_i(w\cdot\Phi(x_i)+b)\geq 1-\xi_i\\ \forall\, i\xi_i\geq 0\end{cases} \tag{2}$$

其中 ξ_i 是松弛变量,刻画训练样本集被错分的程度。通过拉格朗日函数转换成对偶问题。

$$\max W(a)\equiv\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j}a_ia_jy_iy_jK(x_ix_j) \tag{3}$$

$$\text{s.t.}\begin{cases}\forall\, i0\leq a_i\leq C\\ \sum_{1}^{n}a_iy_i=0\end{cases} \tag{4}$$

$\Phi(\cdot)$是从输入空间到特征空间的映射,$K(x_i,x_j)=(\Phi(x_i),\Phi(x_j))$是 kernel 矩阵,对应于$a_i\neq 0$的样本点被称为支持向量。SVM 本质上是一种二次规划问题。算法的实现上,考虑到内存和运算量等问题,人们提出了块算法、分解算法和序贯最小最优化算法(SMO)。

1.2　SVM 应用于非平衡数据集

近来关于非平衡类问题的研究集中以下几个方面。第一是代价敏感性训练[4]。正类误分的代价比负类误分的代价高。这个方法需要对分类错误的代价指定一个合适的值。第二是对于原始数据集重新采样。处理非平衡数据集最常用的方法就是采样,采样的基本思想就是通过改变训练数据的分布来消除或减少数据的非平衡。第三是用一种基于识别(单类分类)的监督学习替代基于判别(二类分类)的监督学习[5,6]。

支持向量机(SVM)作为一种表现优异的分类算法,将它应用于非平衡数据集受到了大家的广泛关注,在文献[7]中,对非平衡数据中 SVM 的表现进行了研究。对于 SVM 的改进主要表现在以下几个方面。

1.2.1　代价敏感性

和传统算法一样,SVM 不具有代价敏感性,不能直接用于代价敏感学习[8],因此需要对 SVM 进行一定的改进。1997 年,Edgar Osuna 等人就提到用改进的 C - SVM 解决非平衡样本集,并首次使用两个正则化参数来分别控制两类的错误惩罚,即 C^+ 和 C^-,给出的二次优化形式为:

$$\min\frac{1}{2}\|w\|^2+C^+\Big(\sum_{iy_i=+1}\xi_i\Big)+C^-\Big(\sum_{iy_i=-1}\xi_i\Big) \tag{5}$$

1999 年 Veropoulos 等对 C - SVM 支持向量机进行改造,对不同的类设置不同的惩罚参数 C 即对正类和反类赋予不同的代价,作为 SVM 的惩罚因子。最优问题变为:

$$\min_{a}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_ih_ja_ia_jK(x_i\cdot x_j)-\sum_{j=1}^{l}a_j=e^{\mathrm{T}}a+\frac{1}{2}a^{\mathrm{T}}Q^{\mathrm{T}}a \tag{6}$$

$$\text{约束条件:}\begin{cases}\sum_{i=1}^{l}y_ia_i=0\ \ 0\leq a_i\leq C^{+1}\text{if } y_i=+1\\ 0\leq a_i\leq C\text{ if } y_i=-1\end{cases} \tag{7}$$

然而由于式中存在$\sum_{i=1}^{l}y_ia_i=0$约束条件,因此增加一个类的 a_i 值,会不可避免地增加另外一类的 a_i 值来满足上式等式的约束条件,这样对非平衡数据分类就不能取得很好的效果[9]。2003 年 Raskutti 等人进一步改进了 Veropoulos 的代价选取方式[10]。

当要求不同样本尺寸的两类有类似属性的边界时(即各自的支持向量数与其总样本数的比值相等时,或者是两类的错误率接近时),Hong - Gunn Chew 等人详细分析了 C - SVM 算法中因类别大小非平衡而造成对分类精度影响的原因,并提出了相应的解决方法。通过一系列证明得出以下结论:

$$\frac{C^+}{C^-}=\frac{N_-}{N_+} \tag{8}$$

其中,N_- 是少数类样本数,N_+ 是多数类样本数。针对非平衡样本集,SVM 算法使用不同的正则化参数(C^+ 和 C^-)来分别控制两类错分样本的惩罚程度时,使用这一结论,可以得到很好的效果[11]。

1.2.2　采样

基本采样包括 under - sampling 和 over - sampling。under - sampling 是通过减少多数类样本来平衡两类样本,这样就可能忽略潜在的有用的多数类样本。而 over - sampling 是通过复制少数类样本来完成的。引入了额外的训练集,基本采样包括 under - sampling 和 over - sampling。under - sampling 是通过减少多数类样本来平衡两类样本,这样就可能忽略潜在的有用的多数类样本。而 over - sampling 是通过复制少数类样本来完成的。引入了额外的训练集,更严重的是可能会引起数据的过度拟合问题。此外,过采样方法需要更多新建实例的内存空间和基于学习算法的内存空间(例如:在 kernel 分类算法中的扩展 kernel 矩阵)。

高级采样方法则将 under－sampling 和 over－sampling 巧妙结合起来，SMOTE(Synthetic Minority Over－Sampling)引入新的非重复的人造少数样本[12]。这种方法增加了通用性，而不是像精确复制样本一样会引起过度拟合。SMOTE 是一种过取样的方法，与随机过取样方法的最大不同在于，它能生成一些新的样本然后添加到新的训练样本集。首先，计算样本集中每个样本的 k 个近邻(一般取 5 个近邻)。对于每个样本，每次从它的 k 个近邻中随机选取一个，在它们所连成的线段上随机选取一个点作为新产生的样本点。这样的过程重复 n 次。最终生成数量为原来样本集数目的 $n+1$ 倍的新的样本集。在文献[7]中，对非平衡数据中 SVM 的表现进行了研究。利用文献[13]的 SMOTE 算法来 Over－Sampling 数据，并且用不同的分类错误代价来训练 SVM。

1.2.3 单类分类

单类支持向量机(one－class SVM)方法是对支持向量机分类器的扩展，其基本思想是通过估计目标类样本在特征空间中的密度分布，从而对未知样本做出“是”或“非”的评估，从某种意义上来讲，这种对样本性质的评价方法更加类似目标检测的过程，即根据单类样本的信息，希望找到类似的目标，而对非目标的其它背景作为一个整体来考虑，避免了复杂的样本采样来对背景信息作完备的描述[14]。

在支持向量机研究领域，B. Scholkopf 等人提出了单类支持向量机[15]。one－ class SVM 方法最早被应用在函数概率密度估计的领域，one－ class SVM 实现方法存在两种途径，一种是构造与原点分离的超平面来实现，一种是构造超球体的方法实现。两种方法在核函数选择径向基函数的情况下是等价的。单类分类被认为适合用在数据出现严重不平衡时。文献[14]采用径向基核函数按第一种方法实现单类支持向量机，并在一种非平衡数据集遥感图像目标的检测中取得了很好的效果。文献[6]中提出了基于球体的单向模糊 SVM，首先获得了多数类的最小超球体，然后利用中心和直径给出多数类的模糊实例成员，减少了噪声和多余实例的影响。

运用 SVM 对非平衡数据集进行分类的过程中，经常会综合运用以上的几个方面。文献[16]提出：两类非平衡数据问题(IDP)中，基于判别(discriminative)的二类 SVM 通过人工平衡数据集，往往因为少数类实例的缺乏使结果受到影响。基于识别(recognition)的单类 SVM 仅仅通过实例较多的数据集进行训练，但是区分度不高。所以通过对两种方法的集成提高了对 IDP 问题的解决性能。在文献[8]中，分别实现了基于过取样的代价敏感 SVM，基于欠取样的代价敏感 SVM，以及基于 SMOTE 的代价敏感 SVM。并且指出基于欠取样的代价敏感 SVM 是一种很好的算法，但是在数据集严重不平衡时，该方法是严重失效的。文献[17]中提出了一种通过改变 kernel 矩阵来调整类边界的方法，该方法在非平衡数据集中取得了很好的效果。文献[11]提出在数据严重不平衡的情况下，通过只训练少数类的方法，来识别多数类，并取得了一定的成效。

2 几种非平衡数据分类的 SVM 算法

文献[18]通过对多个支持向量机集成的办法解决两类样本非平衡问题。该系统首先将多数类样本分为 K 个子集，每个子集的训练样本数量与少数类样本数量基本平衡，然后复制 K 个少数类样本集分别和 K 个多数类样本子集合并生成 K 个新的训练样本集。下一步便对这 K 个新的训练样本集分别进行训练，从而得到 K 个支持向量机，最后用多分类器组合的方法将 K 个支持向量机集成，或者用 K 个支持向量机进行组合预测分类。运用数据预处理方法进行非平衡数据集分类是先对训练数据集进行预处理，然后用处理过的训练数据集训练分类器。

不同于一些传统方法，文献[6]中提出了用主动学习策略来处理非平衡类的问题。根据等式 5 可知，只有支持向量对解决效果有影响。问题就集中在怎样在数据集中选择最有意义的实例，也就是最接近超平面的实例。SVM 主动学习的基本方式是：从已有的训练集中学习 SVM，选择最接近超平面的实例，然后在新选择的实例加入到训练集重新训练。

在选择最接近超平面的实例的步骤中，每次都要搜索整个数据集。可以通过迭代方法查询小池数据来代替查询整个数据集。迭代过程中，每次在数据集 X_N 中随机选择 L 个实例，然后在随机数据集 $X_L, L << N$ 中，选择最有意义的实例 x_i, x_i 在 X_N 的前 $p\%$ 最有意义实例中的概率为 $(1-\eta)$。可证 L 中至少有一个实例在前 $p\%$ 的概率为 $1-(1-p\%)^L$。可知：

$$1-(1-p\%)^L = 1-\eta \text{ 则 } L = \log\eta/\log(1-p\%) \tag{9}$$

因为 L 独立于 N，所以不受整个数据集规模的影响。结果显示在构造平衡数据，主动学习可以作为一个更有效的重采样的替换物。主动学习避免了在欠采样中丢失信息的风险以及在过采样中的负担。主动学习已经有一些学者提出[14]。作为一种简单的方法，系统的研究显示它对非平衡数据有很好的效果。

3 结束语

非平衡数据集日益受到大家的关注,传统算法在该问题上有很大的局限。关于神经网络、决策树、SVM 等的各种改进方法被提出。文中分析了 SVM 在非平衡数据集中的应用情况,提出了 SVM 几种主要的改进方法。各种方法有不同的适用领域,同时正确选择分类结果的度量也很重要。也可以对不同的 SVM 运用集成和组合等方式,提高对非平衡数据整体的分类表现。

参考文献:

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer,2000:138 - 167.

[2] Van Hulse j J,Khoshgoftaar T M,Napolitano A. Experimental Perspectives on Learning from Imbalanced Data[C]//In Proceedings of the 24th International Conference on Machine Learning. New York:ACM,2007:143 - 146.

[3] Jo T,Japkow Icz N,Stephens. The class imbalance problem: a systematic study[J]. Intelligent Data Analysis,2002,6(5):203 - 231.

[4] Domingos P. Metacost:A general method for making classifiers cost - sensitive[C]//In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:Elsevier Science Inc,1999:230 - 232.

[5] Han P Hui, Mao Binghuan, Lv Hairong, et al. One - Sided Fuzzy SVM Based on Sphere for Imbalanced Data Sets Learning[C]//In proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Washington. IEEE Computer Society,2007:166 - 170.

[6] Ertekin S,Huang Jian,Bottou L,et al,Learning on the Border:Active Learning in Imbalanced Data Classification[C]// Proceedings of the sixteenth ACM conference on information and knowledge management. New York: ACM,2007:127 - 136.

[7] Akbani R,Kwek S,Japkowicz N. Applying support vector machines to imbalanced datasets[C]//Proc. of European Conference on Machine Learning. Berlin:Springer,2004:39 - 50.

[8] 李　刚.代价敏感的支持向量机监督学习研究[D].南京:南京师范大学,2007.

[9] Veropoulos K, Campbellc, Cristianinin. Controlling the sensitivity of support vector machine[C]//In:Dean T. Proc of the Sixteenth International Joint Conference on Artificial Intelligence (IJCA I99), Workshop ML3, Stockholm: [s. n.], 1999:55 - 60.

[10] Raskutti B,Kowalczyk A. Extreme balancing for svms:a case study[C]//Workshop on Learning from Imbalanced Datasets II,ICML. [s.l.]:AAAI Press,2003:178 - 181.

[11] Chew Hong - Gunn,Crisp D J,Bogner R E ,et al. Target detection in radar imagery using support vector machine with training size biasing[C]//Sundararajan N. Proceeding of the sixth International Conference on Control, Automation, Robotics and Vision. Singapore:[s. n.],2000:80 - 85.

[12] 张　琦,吴　斌,王　柏.非平衡数据训练方法概述[J].计算机科学,2005(8):181 - 183.

[13] Chawla N V,Bowyer K W,Hall L O,et al. Smote:Synthetic minority over - sampling technique[J]. Journal of Artificial Intelligence Research (JAIR),2002,16:321 - 357.

[14] 王凯峰,秦前清.基于单类 SVM 的遥感图像目标检测[J].计算机工程与应用,2005(3):63 - 65.

[15] Scholkopf B,Platt J C,Shawe - Taylor J,et al. Estimating the support of a high - dimensional distribution[J]. Neural Computation,2001,13(7):1443 - 1472.

[16] Wu P Gang,Chang P Y. Aligning Boundary in Kernel Space for Learning Imbalanced Dataset[C]//In Proceedings of the Fourth IEEE International ConferenceWashington: IEEE Computer Society,2004:265 - 272.

[17] Li P Peng,Chan PKap Luk,Fang Wen. Hybrid Kernel Machine Ensemble for Imbalanced Data Sets[C]//In Proceedings of the 18th International Conference on Pattern Recognition. Washington:IEEE Computer Society,2006:1108 - 1111.

[18] Yan R, Liu Y, Jin R, et al. On predicting rare classes with SVM ensembles in scene classification[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hong Kong:[s. n.],2003:21 - 24.

(上接第 189 页)

[3] 段松云. 无人机着陆数学模型研究——三轮着地滑行[J]. 系统仿真学报, 2004,16(6):1296 - 1299.

[4] 顾宏斌. 飞机地面运行的动力学模型[J]. 航空学报, 2001,22(2):163 - 167.

[5] Jr Doyle G R. A Review of Computer Simulation for aircraft - surface dynamics[J]. Journal of aircraft, 1986,23(4):257 - 265.

[6] 邹美英. 飞机防滑刹车系统新型控制律设计与仿真研究[D]. 西安:西北工业大学, 2006.

[7] 徐冬苓. 飞机防滑刹车系统的建模与仿真研究[J]. 测控技术, 2005,23(11):66 - 68.

[8] Lindsley N J, Talekar N B. A new tire model for aircraft landing gear dynamics[C]//2000 International ADAMS User Conference. Michigan,USA:[s. n.],2000.

[9] Bakker E, Nybory L, Pacejka H B. Tyre Modelling for Use in Vehicle Dynamics Studies[P]. USA:SAE870421,1987.

[10] 王纪森,何长安.飞机轮胎与跑道间结合系数模型的研究[J]. 西北工业大学学报, 2000,18(4):569 - 571.

[11] 高泽迥,林　宏,赵世春,等. 飞机地面载荷若干问题的探讨[J].航空学报, 1994,15(1):8 - 16.