文章编号：1004-4353(2018)01-0043-06

基于近邻密度改进的 SVM 不平衡数据集分类算法

刘悦婷

（兰州文理学院 传媒工程学院，甘肃 兰州 730000）

摘要：针对不平衡数据集数据分布不均匀及边界模糊的特点，提出基于近邻密度改进的 SVM(NDSVM)不平衡数据集分类算法. 该算法先计算多数类内每个样本的近邻密度值，然后依据该密度值选出多数类中位于边界区域、靠近边界区域的与少数类数目相等的样本分别与少数类完成 SVM 初始分类，最后用所得的支持向量机和剩余的多数类样本完成初始分类器迭代优化. 人工数据集和 UCI 数据集的实验结果表明，与 WSVM、ALSMOTE-SVM 和基本 SVM 算法相比，本文算法分类效果良好，能有效改进 SVM 算法在分布不均匀及边界模糊数据集上的分类性能.

关键词：支持向量机；不平衡数据集；近邻密度；分布不均匀；边界区域

中图分类号：TP391 **文献标识码**：A

Imbalanced dataset classification algorithm based on NDSVM

LIU Yueting

(*School of Media Engineering*, *Lanzhou University of Arts and Science*, *Lanzhou* 730000, *China*)

Abstract: Aimed at the data of uneven distribution and indistinct boundary in imbalanced dataset, imbalanced dataset classification algorithm based on neighbor density support vector machine (NDSVM) is proposed. In this algorithm, neighbor density value of each sample in the majority is solved firstly. According to the density, the data which on the majority class border or close to the border is equal to the minority samples in quantity, which are selected and the minority class complete SVM initial classification. Then the resulting support vector machine and residual data in the majority class optimize the initial classifier. The simulation results of experiments on the manual and UCI dataset show that compared with WSVM, ALSMOTE-SVM and SVM, NDSVM has better classification performance, which effectively improve the classification performance of SVM algorithm on the uneven distribution and indistinct boundary in imbalanced dataset.

Keywords: support vector machine; imbalanced dataset; neighbor density; uneven distribution; boundary

0 引言

分类是对输入训练样本进行分析、学习后得到决策模型，然后依据该模型预测未知样本，它已成为机器学习领域的一个重要研究方向. 目前，已有众多经典算法可以实现平衡数据的良好分类，如支持向量机法、模糊分类算法、代价敏感学习法和决策树算法等[1]. 但是，在现实中许多应用领域存在明显的不均衡数据，如网络入侵、商业欺诈、文本分类等数据集[2-3]，此类不平衡数据集中少数类数据往往包含重要的信息，因此研究此类少数类信息具有重要意义. 传统分类判决时，分类器总会偏向多数类，把少数类

收稿日期：2017-09-19 **基金项目**：2015 年甘肃省高等学校科研项目(2015B-132)

作者简介：刘悦婷(1979—)，女，副教授，研究方向为电子、自动控制理论.

分到多数类,进而影响了分类效果;因此,如何提高不平衡数据的分类性能已成为众多学者研究的热点[4].目前,不平衡数据分类的方法主要从数据层面和算法层面实现.数据层面完成数据预处理,包括欠采样、过采样和混合采样.欠采样是通过减少多数类样本使数据集均衡,该方法容易造成信息丢失,降低分类器的性能[5-6].过采样是通过复制、插值增加少数类样本使数据集均衡,但会造成过拟合,增加分类器的空间和时间消耗[7-8].混合采样是将欠采样和过采样有效结合从而平衡数据集的分布.

算法层面是对分类算法本身进行操作,包括对传统算法的改进、众多算法的集成等.改进算法主要是通过调整分类边界、改变概率密度等措施修改算法在数据集上的偏置,使得决策面偏向于少数类,提高少数类的分类性能[9].文献[10]提出了先由 SVM 确定近邻数目,再由 KNN 算法完成分类的方法;文献[11]提出了 WSVM 算法,该算法按照聚类权重定性选择出对分类决策面起大小作用的多数样本,然后将选取的多数类样本与少数类完成 SVM 训练.不平衡数据集按照不同特点可分为以下 3 类[12]:①两类数据数量差别很大,类分布比较均匀;②两类数据数量相当,但类分布差别较大,如一类比较集中,一类比较分散;③两类数据数量和类分布差别都很大.传统分类方法适用于研究第一种情况,不适用于后面两种情况.文献[13]提出了基于实例重要性的不平衡数据分类问题,但忽略了类内不均匀对分类精度的影响.本文综合文献[12-13]的思想,提出基于近邻密度的平衡法,按照多数类每个样本的近邻密度选择出对分类决策面起大小作用的样本,将训练集按照样本作用大小分别与相同数目的少数类相结合进行重新组织训练,并通过实验验证了本文算法的优越性.

1 SVM 算法

基于统计学习理论的 SVM 是建立在结构风险最小化原理上的,其目的是寻求最优分类面.对于原始特征空间中不可分的问题,SVM 通过核函数将低准线性不可分样本映射到更高维的特征空间中,从而将线性问题转化为求解线性约束的二次规划问题.

给定样本 (x_i,y_i), $i=1,2,\cdots,n$, $x_i\in\mathbf{R}^n$, $y_i\in\{-1,+1\}$. 基本 SVM 算法利用最大间隔寻找决策分类超平面,其决策判别函数为

$$f(x)=\boldsymbol{w}^{\mathrm{T}}x+b, \tag{1}$$

式中 $\boldsymbol{w}\in\mathbf{R}^n$ 为权向量,$b\in\mathbf{R}$ 为阈值.构造 SVM 优化模型[3] 为

$$\min_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i,$$

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,\ \xi_i\geqslant 0,\ i=1,2,\cdots,n. \tag{2}$$

式中:ξ_i 为松弛变量,反映样本被错分的程度;C 为惩罚常数,控制对错分样本的惩罚程度.为求解式(2)二次规划问题,构造 Lagrange 函数:

$$L_{(\boldsymbol{w},b,\xi_i)}=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)-1+\xi_i)-\sum_{i=1}^{n}\beta_i\xi_i, \tag{3}$$

式中 α_i 和 β_i 为 Lagrange 算子.利用式(3)分别对 $\boldsymbol{w},b,\xi_i$ 求偏导并置零,得式(2)的对偶问题:

$$\max_{\alpha}-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j(x_i\cdot x_j)\alpha_i\alpha_j+\sum_{j=1}^{n}\alpha_j,$$

$$\text{s.t. } \sum_{i=1}^{n}\alpha_iy_i=0,\ 0\leqslant\alpha_i\leqslant C. \tag{4}$$

根据 KTT 条件,得

$$\boldsymbol{w}^*=\sum_{i=1}^{n}x_iy_i\alpha_i^*,\ b^*=y_j-\sum_{i=1}^{n}\alpha_i^*y_i(x_i\cdot x_j), \tag{5}$$

将式(5)代入式(2),求得判别函数为

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x,x_i)+b^*\right), \tag{6}$$

式中 $K(x,x_i)$ 为核函数. 将高维特征空间两个训练样本的内积$(x\cdot x_i)$ 用输入空间样本的核函数代替，得

$$K(x,x_i)=(x\cdot x_i). \tag{7}$$

2　基于近邻密度改进的 SVM 算法(NDSVM)

2.1　问题分析

图 1 为密度不均匀数据集的示意图，其中 a、b、c 分别代表 3 个簇所在的范围，由图 1 知密度关系 $a>b>c$. 传统的分类算法很难找到统一的邻域半径来发现 a、b、c 这 3 个类别，若邻域半径取值较大，a 和 b 容易被认定为一个类；反之 c 被视为噪声. 要解决该问题，简单的方法是人为设定邻域半径值，然而在数据集密度分布情况未知的情况下，人为设定邻域半径存在较大困难，因此难以对分布不均匀数据集进行正确分类. 基于此，本文算法考虑通过计算每个样本近邻密度来判断样本所在的区域. 样本密度以该样本为中心的 k 近邻的平均距离的倒数来衡量，样本的 k 近邻密度越大，说明样本近邻之间越密集，样本所在区域越接近簇类中心；反之，样本的 k 近邻密度越小，样本近邻之间越稀疏，样本所在区域越接近类边界. 由于多数类的边界区域数据的密度值最小，很容易发生错分，因而其对分类结果“作用最大”；靠近边界的样本是单例数据或有助于克服噪声的影响数据，因而其对分类结果“作用较大”；其余区域的样本对分类结果“作用较小”. 因此，本文采用平均距离的倒数标识样本的密度信息，从而判定样本所在的区域.

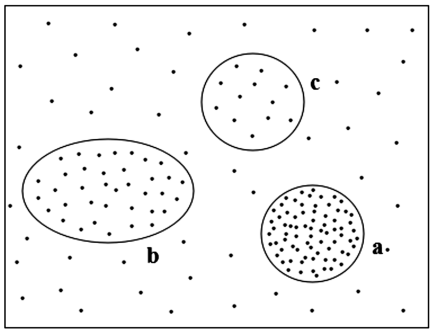

图 1　不均匀分布数据集

2.2　近邻密度的定义

给定包含 q 个数据点的 m 维数据集 Z，$z=(z_1,z_2,\cdots,z_m)\in Z$，$y=(y_1,y_2,\cdots,y_m)\in Z$，则 z 与 y 之间的欧氏距离为

$$d(z,y)=\sqrt{\sum_{i=1}^{m}(z_i-y_i)^2}. \tag{8}$$

设有对象 z，集合 D，且 $z\notin D$，用式(8) 计算 z 到集合 D 中任一点 d_j 的距离，表示为 $d(z,d_j)$. 将所有 $d(z,d_j)$ 按照降序排列，得到与 z 近邻的 k 个对象 $d_1,d_2,\cdots,d_k$，表示为

$$Ne(z)=\{d_j\mid d_j\in D,j=1,2,\cdots,k\}. \tag{9}$$

对任意 $z_i\in Z$，$d_j\in Ne$，z_i 的近邻密度 ρ_i 定义为

$$\rho_i=\frac{1}{\sum_{j=1}^{k}d(z_i,d_j)/k}. \tag{10}$$

由式(10) 知，ρ_i 越大，对象 z_i 的近邻越密集；反之，对象 z_i 的近邻越稀疏.

2.3　NDSVM 算法流程

本文提出的 NDSVM 算法首先计算多数类中每个样本的近邻密度，然后再根据密度值的大小分别选出多数类边界区域、靠近边界区域的样本，并且所选样本数目与少数类数目相同，保证样本的均衡性. 由于对分类结果作用的大小依次为多数类的边界区域的样本、靠近边界区域的样本、剩余区域的样本，因此本文先从多数类样本中选取与少数类数目相等的密度值最小、次最小的两部分样本，再将选取样本分别与少数类样本完成 SVM 初始分类，以此保证训练样本数量的平衡性，最后用剩余的多数类样本对

初始分类器进行迭代优化. NDSVM 算法的具体流程为:

Step 1 变量初始化. P 表示少数类样本集合,p 表示少数类样本总数;N 表示多数类样本集合,n 表示多数类样本总数;N_{order} 表示多数类样本按密度值降序排列的集合;$N_{\text{order\_behind}}$ 是 N_{order} 集合中最后 p 个样本组成的集合;$N_{\text{order\_behindf}}$ 是 N_{order} 集合中次最后 p 个样本组成的集合;$N_{\text{order\_other}}$ 是 N_{order} 集合中剩余样本组成的集合.

Step 2 从训练样本中分离出多数类样本,用集合 N 表示.

Step 3 从集合 N 中任选样本点 n_i,用式(10)计算样本的近邻密度. 依次类推,计算集合 N 中所有样本的近邻密度,以密度值降序排列集合 N 中的所有样本,得到集合 N_{order};

Step 4 判断 p 和 n 的关系. 若 $p < n \leqslant 2p$,则认为训练样本是平衡样本,用传统 SVM 训练样本,得到分类结果;若 $n > 2p$,则认为是不平衡样本,转入 Step 5;

Step 5 集合 P 和 $N_{\text{order\_behind}}$ 组成的两类平衡集合为 M_{PN1},用 M_{PN1} 训练 SVM,得到支持向量机 S_{PN1},多数类支持向量的个数为 n_{neg1},少数类支持向量的个数为 n_{pos1};

Step 6 集合 P 和 $N_{\text{order\_behindf}}$ 组成的两类平衡集合为 M_{PN2},用 M_{PN2} 训练 SVM,得到支持向量机 S_{PN2},多数类支持向量的个数为 n_{neg2},少数类支持向量的个数为 n_{pos2};

Step 7 在不影响分类精度的同时,使用支持向量集取代训练样本集进行训练,这样可以缩短训练时间. 由支持向量机 S_{PN1}、S_{PN2} 和 $N_{\text{order\_other}}$ 组成集合 M_{PN3},从 M_{PN3} 中提取全部支持向量的个数($n_{\text{pos1}} + n_{\text{pos2}} + n_{\text{neg1}} + n_{\text{neg2}}$),提取与支持向量相同数目的多数类,完成 SVM 分类器的迭代训练,并完成支持向量的更新. 当满足式(11)的 $T < 0.9$ 时,返回到 Step 4;当满足 $T \geqslant 0.9$ 时,迭代训练停止,输出分类结果.

$$T = \text{多数类支持向量总个数} / \text{少数类支持向量总个数}. \tag{11}$$

3 实验分析

3.1 评价指标

由于传统分类器的性能指标存在较大缺陷,本文采用了如下的评价指标:

1) G-mean. G-mean 为综合评价少数类和多数类两类样本的分类性能的指标,若分类器分类偏向于某一类,则会影响另一类的分类正确率,其计算公式为 $G\text{-mean} = \sqrt{Sensitivity \times Specificity}$.

2) F-measure. F-measure 是查全率和查准率两个评价方式的综合,能有效反映分类器对少数类样本分类性能的敏感程度,其计算公式为 $F\text{-measure} = \dfrac{2 \times Sensitivity \times Precision}{Sensitivity + Precision}$.

3) AUC. AUC 能全面地描述分类器在不同判决阈值时的性能,其计算公式为

$$AUC = (1 + \frac{TP}{FP + FN} - \frac{FP}{TN + FP})/2.$$

其中:不均衡数据集中少数类别的样本为正类 P,多数类别的样本为负类 N;TP 表示实际为正类而被预测为正类的样本个数;FN 表示实际为正类而被预测为负类的样本个数;FP 表示实际为负类而被预测为正类的样本个数;TN 表示实际为负类而被预测为负类的样本个数;查全率 $Sensitivity = TP/(TP + FN)$,为少数类样本的正确率;特异度 $Specificity = TN/(TN + FP)$,为多数类样本的正确率;查准率 $Precision = TP/(TP + FP)$,为被正确分类的正类样本占被分为正类的全部样本比值.

3.2 NDSVM 算法的实验验证与比较

为验证本文算法的可行性,用 Matlab 2014a 编写程序,选取 Dataset 人工数据集、UCI 数据集为实验对象进行测试,将测试结果与文献[8]的 ALSMOTE-SVM 算法、文献[11]的 WSVM 算法和 SVM 算法的测试结果进行比较. Dataset 是包含 1018 个数据点的不均匀人造数据集,如图 2 所示. 从 UCI 库中

选取不平衡性较轻的 Iris、Glass Identification 数据集和不平衡性高的 Spectf Heart、Ecoli 数据集完成实验，如表 1 所示. SVM 分类器的参数设置为：选取高斯函数为核函数，核宽度＝1，惩罚因子 C＝1 000，初始 α＝0.2，多数类中近邻 k＝10，实验迭代运行 20 次. 4 种算法在 5 个数据集上运行得到的 G-mean、F-measure 结果如表 2 所示. 在人工数据集 Dataset、Glass Identification 和 Spectf Heart 上运行 4 种算法得到的 AUC 变化曲线分别如图 3— 图 5 所示.

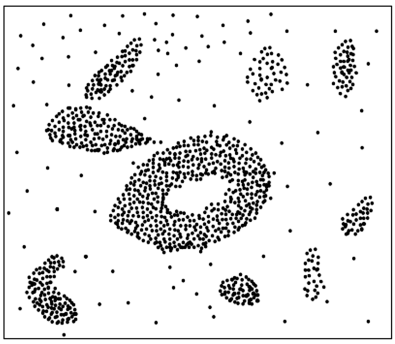

图 2　人工数据集 Dataset

表 1　实验数据集信息

数据集	样本总数	属性	类别数	处理后样本分布		
				训练集(正，负)	训练样本的比例/%	测试集(正，负)
Dataset	1 018	12	9	735(52，683)	7.07	283(48，235)
Iris	150	4	3	94(40，54)	42.55	56(15，41)
Glass Identification	214	10	6	132(25，107)	18.94	82(20，62)
Spectf Heart	267	22	2	89(9，80)	10.11	178(37，141)
Ecoli	336	8	3	104(25，79)	24.04	232(98，134)

表 2　4 种算法的实验结果

数据集	SVM 算法		ALSMOTE-SVM 算法		WSVM 算法		NDSVM 算法	
	G-mean	F-measure	G-mean	F-measure	G-mean	F-measure	G-mean	F-measure
Dataset	83.8	82.6	91.5	89.8	93.9	93.3	97.2	96.8
Iris	94.5	94.0	97.6	97.0	97.9	97.2	97.7	97.2
Class Identification	93.7	93.5	95.6	94.8	95.2	94.6	97.3	96.6
Spectf Heart	75.9	73.3	90.2	89.5	91.3	90.2	96.5	96.2
Ecoli	81.3	79.7	91.6	90.1	93.4	93.1	97.8	97.1

由表 2 可知，在各数据集中，除了 WSVM 算法的 Iris 数据值(97.9)外，NDSVM 算法的 F-measure、G-mean 性能值均高于其他各算法的性能值，且 SVM 分类器的性能值相对最低. 这表明，本文提出的 NDSVM 算法的性能优于其他 3 种算法，这是因为本文算法通过近邻密度将多数类样本进行排序，能够保证每次参与分类器训练的多数类与少数类的个数平衡，同时还考虑了类边界的样本信息，因此本文提出的 NDSVM 算法的性能有了较大提高.

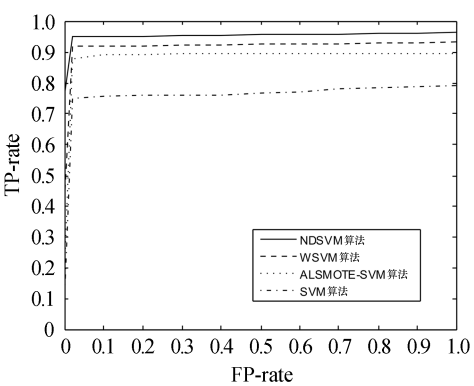

图 3　4 种算法在 Dataset 人工数据集上运行所对应的 AUC 曲线

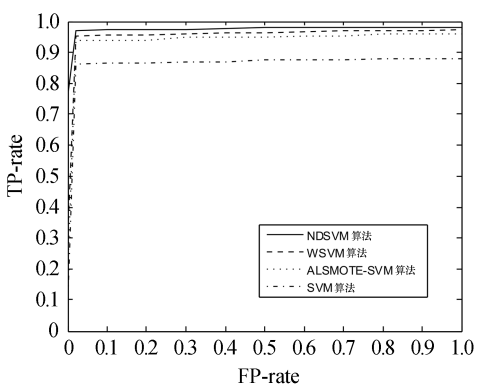

图 4　4 种算法在 Glass Identification 数据集上运行所对应的 AUC 曲线

AUC 值,即 ROC 曲线与横轴围成区域的面积值可以反应分类器的性能. AUC 曲线越接近(0,1)点,且其与横轴围成的面积越大,分类器效果越好. 由图 3—图 5 可以看出,在各数据集中,NDSVM 算法得到的 AUC 值均高于其他 3 种算法的 AUC 值,且 SVM 算法的 AUC 值均为最低. 这说明 NDSVM 算法的性能优于其他 3 种算法,具有可行性.

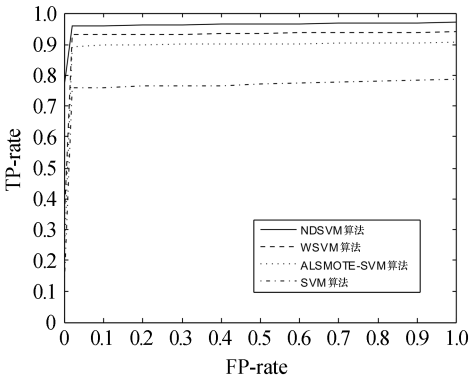

图 5 4 种算法在 Spectf Heart 数据集上运行所对应的 AUC 曲线

4 结论

本文针对传统分类方法对不均匀分布、边界信息模糊的不平衡数据集识别性能较低的问题,提出了一种基于近邻密度改进的 SVM 算法,即 NDSVM 算法. 实验结果显示,NDSVM 算法的性能优于 SVM、ALSMOTE-SVM、WSVM 算法,对于不平衡数据集具有良好的有效性和可行性. 如何更好地协调相关参数的取值及降低时间复杂度是今后需要进一步研究的目标.

参考文献:

[1] Jason V H, Taghi K. Knowledge discovery from imbalanced and noisy data[J]. Data & Knowledge Engineering, 2009,68:1513-1542.
[2] 翟云,杨炳儒,曲武. 不平衡类数据挖掘研究综述[J]. 计算机科学,2010,37(10):27-32.
[3] 张静静. 基于不平衡数据集的支持向量机模型与算法研究[D]. 北京:中国农业大学,2015.
[4] 李勇,刘战东,张海军. 不平衡数据的集成分类算法综述[J]. 计算机应用研究,2014,31(5):1287-1291.
[5] 程险峰,李军,李雄飞. 一种基于欠采样的不平衡数据分类算法[J]. 计算机工程,2011,37(13):147-149.
[6] 李荣陆,胡运发. 基于密度的 KNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展,2004,41(4):539-545.
[7] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: syn-thetic minority over-sampling technique[J]. Journal of Artifi-cial Intelligence Research, 2002,16:321-357.
[8] 张永,李卓然,刘小丹. 基于主动学习 SMOTE 的非均衡数据分类[J]. 计算机应用与软件,2012,29(3):91-93.
[9] 孟军. 不平衡数据集分类算法的研究[D]. 江苏:南京理工大学,2014.
[10] Wang Chinheng, Lee Lamhong, Rajkumar R, et al. Ahybrid text classification approach with low dependency on parameter by integrating *K*-nearest neighbor and support vector machine[J]. Expert Systems with Applications, 2012,39:11880-11888.
[11] 王超学,张涛,马春森. 基于聚类权重分阶段的 SVM 解不平衡数据集分类[J]. 计算机工程与应用,2015,51(21):133-137.
[12] 刘万里,刘三阳,薛贞霞. 不平衡支持向量机的平衡方法[J]. 模式识别与人工智能,2008,21(2):136-141.
[13] 杨扬,李善平. 基于实例重要性的 SVM 解不平衡数据分类[J]. 模式识别与人工智能,2009,22(6):913-918.
[14] Lin Y, Lee Y, Wahba G. Support vector machines for classification in non standard situations[J]. Machine Learning, 2002,46(1/2/3):191-201.