计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(7):219−223 [doi: 10.15888/j.cnki.csa.006436]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

不平衡数据集中分类超平面参数优化方法①

严晓明

(福建师范大学 数学与信息学院, 福州 350117)
通讯作者: 严晓明, E-mail: yanxm@fjnu.edu.cn

摘　要: 对不平衡数据集 SVM 分类存在着分类结果偏向多数类的情况, 使得分类结果中少数类的 F1-Measure 值偏低. 本文提出一种不改变样本集合的样本数, 并结合样本点总数, 分类过程中的支持向量个数, 少数类和多数类的准确率, 生成权重值对分类超平面参数 *b* 进行优化, 以此提高少数类样本点分类准确率的方法, 并通过实验证明该方法的有效性.

关键词: SVM; 不平衡数据集; 优化; 参数; 分类准确率

引用格式: 严晓明.不平衡数据集中分类超平面参数优化方法.计算机系统应用,2018,27(7):219–223. http://www.c-s-a.org.cn/1003-3254/6436.html

Optimization Method of Classification Hyperplane Parameter under Imbalance Data Set

YAN Xiao-Ming

(College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China)

Abstract: SVM classification result on imbalance data set is partial to majority class. It makes the F1_measure value of minority class inadequate. This paper presents a method which improves the classification accuracy of minority class. The method generates weights to optimize parameters *b* of the classification hyper plane without changing the number of samples, and combines the total number of samples, the number of support vector, the accuracy of minority class and majority class. Finally, the effectiveness of the method is proved by experiments.

Key words: SVM; imbalance data set; optimize; parameter; classification accuracy

传统的 SVM 算法通过分类超平面来判断样本的类别, 在解决不平衡数据的分类问题时, 分类结果会偏向于多数类样本点集合, 使得少数类样本点的分类正确率低, 而多数类分类准确率高.

当前针对不平衡数据集 SVM 分类的改进, 一般集中在数据清洗和算法改进两个方向上. 许多学者都提出了具有代表性的改进方法, 如对于样本的欠采样方法 SMOTE[1], 过采样方法 Tomek links[2]以及它们相应的改进算法[3,4], 都是通过不同方法增加少数类样本或减少多数类样本, 来达到使得不同类别中的样本数量基本相当的目的. 在算法层面上, 代价敏感学习方法[5]对不平衡数据集中少数类和多数类分别设置不同的惩罚参数, 通过调整不同类别的惩罚参数, 提高不平衡数据集的分类效果, Huang[6]改进了代价敏感学习, 通过结合极限学习机来实现动态代价敏感学习; 集成学习方法[7]提出构造不同的弱分类器, 对每个弱分类器设置一个权重并组合成一个强分类器对不平衡数据集进行分类, 在集成学习方法的基础上, Zięba M 等人[8]还结合了主动学习策略对每个弱分类器的代价函数进行改进.

对样本数量的增减, 都会改变使得原始样本数据的分布, 使得分类超平面的位置产生偏差; 而设置不同类别惩罚参数的方法, 对于不同的不平衡数据集中每个类别样本数量和分布情况, 较难对惩罚参数的值进

① 收稿时间: 2017-11-09; 修改时间: 2017-11-29; 采用时间: 2017-12-15; csa 在线出版时间: 2018-06-27

© 中国科学院软件研究所 http://www.c-s-a.org.cn

行预设. 本文提出了一个在保持原数据样本不变的情况下, 应用 SMO 算法解拉格朗日优化方程参数α的同时, 利用不同类别的样本分布特点构造出权重值, 并对超平面方程中的参数b进行优化的方法. 实验结果表明, 分类结果中不平衡数据集少数类的分类正确率更高, 相应的 F1-Measure 指标也得到了改善, 并且对于各种类别样本分布情况的不平衡数据集, 有着较好的适应能力.

1 样本数量不平衡对分类超平面的影响

SVM 算法得到一个间隔边界$2/\|w\|$最大的分类超平面$y=w\cdot x+b$, 目标函数为:

$$\begin{aligned}&\min\ \tfrac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{N}\varepsilon_i\\&\text{s.t.}\ \ 1-\varepsilon_i-y_i(wx_i+b)\le 0,\ \ 0\le\varepsilon_i\end{aligned}\tag{1}$$

其中, ε_i为每个样本点的松弛变量, C为惩罚参数.

SVM 算法要找的两条间隔边界是两类别样点中所有间距最小的样本点之间的最大距离. 在不平衡数据集下生成分类超平面时, 之所以分类结果会偏向多数类样本集, 本质上是因为对所有类别的样本点都使用相同的惩罚系数C.

SVM 算法对每个样本点加入松弛变量ε_i, 在求解分类超平面的过程中比较该点的松弛变量ε_i到该样本点所属类别的间隔分界面距离的大小. 如果样本点到分界面的距离小于该样本点的ε_i, 则表示分界面对于该样本点是可正确分类的. 加入了松弛变量后, 目标函数多了一项对相应的ε_i的值进行累加再乘上惩罚系数C. 在优化的过程中目标函数关注的是所有样本点的松弛变量之和的C倍最小. 在公式 (1) 的约束条件中惩罚参数C是对于所有样本点的松弛变量ε_i的约束, 分类正确的样本x_i'也必须满足这样的约束. 而分类正确的样本点x_i'在两条间隔边界之外, 即$y_i(wx_i+b)\ge 1-\varepsilon_i$, 这些样本点并不是要找的支持向量, 而且对应的$\varepsilon_i$值比较大, 也在目标公式 (1) 中进行了累加. 由于不平衡数据集两个类别的样本点数量上的差异以及相同的惩罚系数C, 使得松弛变量ε_i对少数类的样本点惩罚量的累加相对于多数类样本点偏少, 导致分类超平面向少数类靠近, 即分类结果偏向于多数类样本集.

在上文提到对不平衡数据集代价敏感的 SVM 算法中, 将惩罚参数设置为C^+与C^-, 分别表示对于少数类样本点和多数类样本点松弛变量ε_i^+和ε_i^-的约束. 在应用拉格朗日方法转换成对偶问题时, 参数w和b的求解变成了对α的求解, 即实际上C^+与C^-最后都变成了对α值的约束: $0\leqslant\alpha_i\leqslant C^+$和$0\leqslant\alpha_i\leqslant C^-$, 如公式 (2)[5,9]:

$$\begin{aligned}\max\quad & W(\alpha)=-\tfrac{1}{2}\Big(\sum_{i,j=1}^{N}\alpha_i y_i\alpha_j y_j x_i*x_j\Big)+\sum_{i=1}^{N}\alpha_i\\\text{s.t.}\quad & 0\le\alpha_i\le C^+,\ 0\le\alpha_i\le C^-,\ \sum_{i=1}^{N}\alpha_i y_i=0\end{aligned}\tag{2}$$

用公式 (2) 的方式对不平衡数据集进行优化时, 在算法运行前, 要人为设置两个参数C^+与C^-, 通过少数类样本集设置的惩罚参数大于多数类样本集的惩罚参数, 即C^+与C^-. 首先, C^+要比C^-预设值大多少, 是一个人工经验的问题, 由于样本的分布和数量上的区别, 预设值的大小不容易确定; 其次, 这两个参数最后转换成条件$0\leqslant\alpha_i\leqslant C^+$和$0\leqslant\alpha_i\leqslant C^-$, 又由于$C^+>C^-$, 实际上是最后去判断$\alpha_i$与$C^-$的大小, 公式可以合并成$0\leqslant\alpha_i\leqslant C^-$, 即用 SMO 算法求解时, 对不同类别的惩罚系数的作用被弱化成了对较小的那个惩罚系数C^-的约束, 相应地对分类效果的作用也弱化了.

2 参数 b 与分类超平面的关系

对公式 (1) 求解w和b, 可得:

$$\begin{cases}w=\displaystyle\sum_{i=1}^{N}\alpha_i y_i x_i\\[2ex] b=-\dfrac{\max_{i:y_i=-1}w^{\mathrm{T}}x_i+\min_{i:y_i=1}w^{\mathrm{T}}x_i}{2}\end{cases}\tag{3}$$

上式中的α_i是拉格朗日乘子. 在求解α_i时, 存在着三种情况: 当样本点在 SVM 分类超平面的两条间隔为$2/\|w\|$分界面之外时, α_i的值为 0; 当样本点位于间隔边界上时, $0\leqslant\alpha_i\leqslant C$; 而最后一种情况$\alpha_i=C$对应的样本点位于间隔边界之间. 在应用 SVM 对不平衡数据集进行分类时, 大部分的多数类的样本点位于间隔边界之外, 即这些样本点对应的α_i的值为 0.

在求得w和b后, SVM 的分类超平面就可以确定, 即$y=w\cdot x+b$. 两条与分类超平面间隔为$1/\|w\|$的分界面也随之确定下来. 此时, 要提高少数类样本点的分类准确率, 有以下几种方法: 第一是使间隔分界面变大, 即增大$1/\|w\|$. 当间隔分界面增大的时候, 多数类的支持向量, 即满足$0\leqslant\alpha_i\leqslant C$条件的样本点个数也大量增加, 由于多数类的样本数量要明显大于少数类的样本, 因此并不会改善松弛变量ε_i对少数类样本点惩罚量的累加值, 这种方法收到的效果不佳; 第二种方法是改变分类超平面的斜率, 即w. 由于两类别样点中所有间距

© 中国科学院软件研究所 http://www.c-s-a.org.cn

最小的样本点之间的最大距离已经确定, 再去改变了斜率w, 分类超平面并没有对少数类分类准确率的提高产生太大的作用; 第三种方法是改变$y=w\cdot x+b$中参数b的值. 改变参数b的值, SVM 分类超平面的截距发生变化, 使得超平面向多数类方向移动, 显而易见, 可以增加少数类样本的分类准确率. 此时由于不平衡数据集中两类样本点数量相差较大以及两类样本点的当前分类的正确率已经产生了差别, 并且如果两类样本点的数量相差越大, 支持向量的数量越多, 分类超平面就应当向多数类的方向移动的距离越大, 从这样的思路出发, 本文对分类超平面的参数b的值增加一个权重, 使得改进了参数b后的分类超平面更靠近多数类样本点, 即更有利于少数类样本集合.

下面用一个例子说明参数b的值对于分类结果的影响. 从[1, 5]区间上的随机选取 220 个随机 2 维样本点的数据集, 其中多数类样本点 210 个, 用实心圆点表示, 少数类样本点用加号表示, 在图 1 中, 实线为分类超平面, 两条虚线为间隔为$2/\|w\|$分界面, 样本点外另加三角形框表示错分样本, 另加上一个圆形框表示支持向量. 对该不平衡数据集采用传统 SVM 分类算法进行分类, 惩罚系数C的值取 5, 得到的结果如图 1 所示.

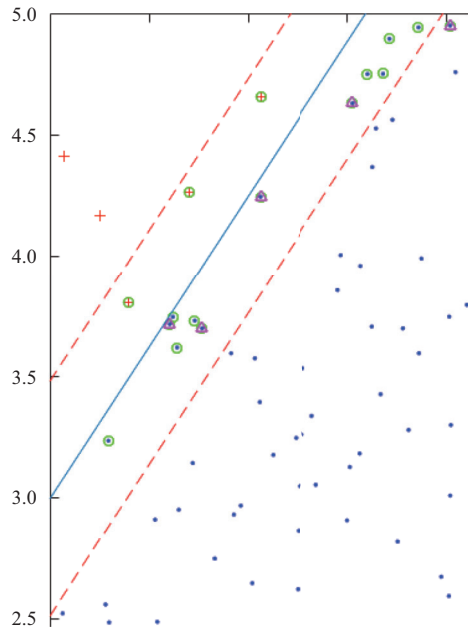

图 1 不平衡数据集分类结果局部放大

SVM 分类最终得到的结果是一个分类超平面$y=w\cdot x+b$, w和b对分类超平面位置起了决定性的作用. 而对于不平衡数据集而言, 少数类由于样本数量较少, 一旦出现错分样本, 会对该类样本的正确率产生比多数类出现的错分更大的影响. 因此有必要对α_i进行计算的过程中对参数b的生成再进行优化.

对于图 1 中的不平衡数据集而言, 生成的分类超平面如果能向多数类方向移动, 即图 1 中实线向下移动, 在本例中相当于减小b, 就会提高少数量样本点的分类准确率.

3 参数 b 针对少类样本集的优化

本节提出一种直接在 SMO 算法求解α时, 不改变样本点的分布及数量, 结合两类样本分类正确率, 每类样本点和支持向量数量, 对参数b增加权重值的方法. 目的是使得最终的分类超平面的位置与传统的 SVM 得到的分类超平面相比, 向多数类样本方向移动, 从而增加少数类样本的分类正确率.

具体的优化算法步骤如下:

Step1. 对所有的乘子α_i进行顺序扫描, 直到得到第一个不满足 KKT 条件的α'为止.

Step2. 在除α'以外的满足 KKT 条件乘子中, 找出使得函数y对输入x_i的预测值与真实输出类标记y_i之差值最大的α'', 即背离 KKT 条件最严重的α'', 进行更新.

Step3. 计算更新了α'和α''后的少数类分类正确率和多数类分类正确率:

$$Acc^{+}=\frac{TP}{NUM\_P},\ Acc^{-}=\frac{TN}{NUM\_N} \tag{4}$$

其中, TP表示少数类分类正确的个数, NUM_P表示少数类总样本数; TN表示多数类分类正确的个数, NUM_N表示多数类总样本数. sup_pos和sup_neg分别为少数类和多数类的支持向量个数, 生成一个对分类超平面参数b的权重, 如公式 (5):

$$\begin{aligned}\chi&=\log\left((sup\_pos+sup\_neg)\cdot\frac{NUM\_N}{NUM\_P}\right)\\ C\_b&=\frac{Acc^{-}}{Acc^{+}}\cdot\chi\end{aligned} \tag{5}$$

Step4. 在对b进行更新时, 如果该样本点是少数类的样本点, 则以一定的概率乘上C_b.

Step5. 如果达不到迭代次数, 或者更新的过程中α_i发生了变化, 则跳到步骤 1 继续执行, 否则结束循环.

Step3 中的中χ是一个常数, 结合了不平衡数据集中的每个类别的支持向量个数和样本数, 应用这个常数对两类样本分类正确率的值Acc^{-}/Acc^{+}放大, 最后对参数b的值按步骤 4 进行更新. 当Acc^{+}为 0 时, Acc^{-}/Acc^{+}的值用 1 代替. 显然C_b是一个大于 1 的值.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

增加了权重C_b后, 在对少数类样本点计算时, 使参数b乘上C_b这个大于 1 常数, 并且如果两类样本点的数量相差越大, 支持向量的数量越多, 该权重值就越大, 使得改进后的分类超平面更靠近多数类样本点, 即更有利于少数类样本集合.

在 SMO 算法的求解过程中, 迭代更新后, 会使得求得的分类超平面向多数类方向移动, 对不平衡数据集问题, 会使得少数类的分类正确率提高, F1-Measure 指标得到改善, 并且由于仅在迭代过程中增加了若干条计算语句和一个分析是否为少数类样本的判断语句, 算法的时间复杂度没有发生变化.

4 实验与结果分析

下面设置了两组数据在 Matlab 2016a 中来验证本文算法 (以下用 SVM_Improved 表示), 一组数据为上文中的人工数据集, 另一组为 UCI[10]公共数据集中的 6 个不平衡数据集; 实验环境的计算机配置为: CPU 为 core i5, 内存 4 G, 操作系统为 Windows10.

4.1 人工数据集

图 2 和本文第 2 节中的图 1 分别是采用传统 SVM 和 SVM_Improved 得到的分类超平面, 惩罚系数C都为 5.0.

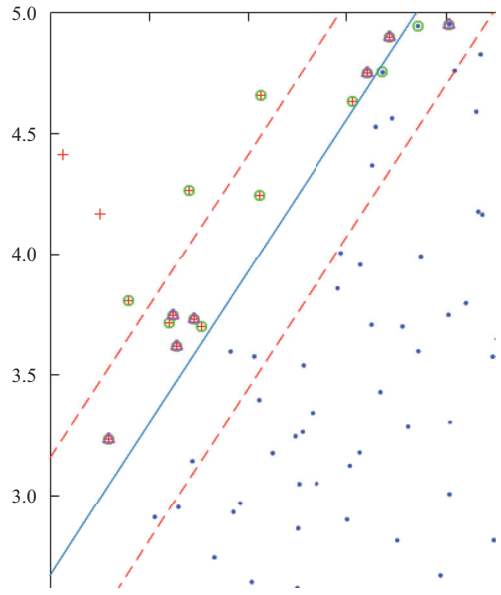

图 2 本文算法生成的分类超平面放大图

第 2 节图 1 中的少数类分类正确率为 50%, 多数类正确率为 100%, 即少数类分对 5 个, 分错 5 个; 多数类分对 210 个, 没有错分样本点; 支持向量一共 16 个, 支持向量中少数类和多数类各一半, 少数类的 F1-Measure 为 0.67. 图 2 中的少数类正确率为 90%, 多数类正确率为 97.14%, 少数类分对 9 个, 分错 1 个, 多数类分对 204 个, 分错 6 个. 支持向量的情况和图 1 中相同, 而少数类 F1-Measure 为 0.93. 图 1 中的分类超平面方程为: $y = 1.3 * x + 1.7$, 而图 2 的为$y = 1.3 * x + 1.4$, 两个算法的实际分类面间隔都为 0.48. 从该数据集的实验上可以看到, 采用 SVM_Improved 算法, 参数b的值变化后, 分类超平面更靠近多数类样本集, 使得少数类样本点的分类性能得到较大的提升.

4.2 UCI 中的不平衡数据集

从 UCI 中抽取 6 个不同的不平衡数据集, 分别为 heart disease, balance scale, yeast, abalone, haberman, ecoli. 如表 1 所示.

表 1 UCI 不平衡数据集

数据集	样本总数	目标类别	少数类:多数类
heart disease	303	Class1	55:248
balance scale	625	Balance	49:576
yeast	1484	ME1 至 ME3	258:1226
abalone	4177	Class19	32:4145
haberman	306	2	81:225
ecoli	336	om&omL	25:311

在这 6 个 UCI 数据中, 不平衡数据集里有多个类别的, 将其中的一个或若干个类别合并设置为少数类, 即表 1 中为目标类别列, 而将其余类别的样本合并设置为多数类. 如数据集 yeast 中, 将类别标签为 ME1, ME2 和 ME3 这三个类别的 44,51,163 个样本合并成少数类, 而将剩余的标签为 CYT 等 7 个类别共 1226 个样本组成多数类. 每个数据集少数类和多数类样本数的对比为表 1 中的最后一列. 这四个数据集中, heart disease 选择的是 Cleveland 数据库; abalone 数据集的多数类与少数类样本数相差最大, 达到 129 倍, 其它五个数据集的多数类与少数类样本数相差 5 至 15 倍之间.

对这 6 个数据集的实验结果如表 2 所示, 其中 Pr、Re 和 F1_M 分别表示 Precision(查准率), Recall(召回率) 和 F1-Measure. 算法 SVM_1 为对少数类和多数类分别设置不同的惩罚参数的代价敏感学习方法. 在实验中, 惩罚参数C都为 5, 算法 SVM_1 中对少数的惩罚参数C^+为 5, 对多数类的惩罚参数C^-为 3.

从表 2 的数据中可以看出: 对于两个类别样本数量不同的多个不平衡数据集中, SVM_Improved 算法的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

少数类样本 F1_measure 的值都有不同程度的提升. 特别地对于 haberman 数据集, 由于属性数只有 4 个, 并且这些属性值为整数又较接近, 即两类样本点在分类超平面附近有较多的分布, 本文算法对于少数类的分类正确的样本数较 SVM_1 算法多了 11, 虽然此时的多数类样本点的分类正确的数量有一定的下降, 但是最后少数类 F1_measure 的值提升较大; 对于 ecoli 数据集, 样本属性值的特点和 haberman 数据集类似, 属性数增加到 8 个, 少数类分类正确的样本较 SVM_1 增加了 13% 左右, 和 haberman 数据集的结果接近.

表 2 实验结果对比 (单位: %)

数据集	SVM			SVM_1			SVM_Improved		
	Pr	Re	F1_M	Pr	Re	F1_M	Pr	Re	F1_M
heart-disease	77.59	81.82	79.65	80.36	81.82	81.08	79.66	85.45	82.46
balance scale	60.61	81.63	69.57	61.76	85.71	71.79	61.11	89.80	72.73
yeast	52.56	75.58	62.00	52.49	77.52	62.60	54.16	78.29	64.03
abalone	77.27	53.13	62.96	64.29	56.25	60.00	59.52	78.13	67.57
haberman	33.3	38.27	35.63	41.18	43.21	42.17	43.40	56.57	49.20
ecoli	76.92	80.00	78.43	75.00	84.00	79.25	75.00	96.00	84.21

对于 abalone 数据集, 样本相差 129 倍时, TP 增加从 17 个样本增加到 25 个样本, 少数类分类正确的数量提高的同时多数类识别错误的样本数也较 SVM 算法增加了 12 个样本, F1-Measure 的值增加了近 5%, 该数据集的 F1-Measure 指标的提升也与 haberman, ecoli 这样的数据集接近. heart-disease, balance scale, yeast 三个数集的 TP 分别较 SVM 算法增加了 2,4,7 个样本, 即少数类样本分类正确数量增加了 2% 至 5%, 多数类正确率基本不变.

这是由于在分类超平面参数b优化后, 少数类的样本点正确率大幅提升的结果. 由以上几个 UCI 数据集以及人工数据集实验还可以得出以下结论: 对于在分类超平面附近有相对较多的少数类样本点的数据集, 如数据集 haberman, ecoli, abalone, 本文算法可以使得少数类的分类精度得到较大的改善, F1-Measure 值有较大的改进.

5 结语

本文提出一种改进不平衡数据集少数类样本分类精确度的 SVM_Improved 方法, 在求解α的过程中, 结合了不平衡数据集中的每个类别的支持向量个数和样本总数以及多数类和少数类样本的正确率比生成一个参数C_b, 对 SVM 的分类超平面参数b进行优化. 实验结果表明, 该方法改善了不平衡数据集的少数类 F1-Measure 指标, 特别在分类超平面附近有较多的少数类支持向量的数据集, 少数类样本点的正确率有较大改进.

参考文献

1 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.

2 Kubat M, Matwin S. Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the Fourteenth International Conference on Machine Learning. Nashville, TN, USA. 1997. 179–186.

3 刘胥影, 吴建鑫, 周志华. 一种基于级联模型的类别不平衡数据分类方法. 南京大学学报 (自然科学), 2006, 42(2): 148–155.

4 Wang KJ, Adrian AM, Chen KH, *et al*. A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan. Computer Methods and Programs in Biomedicine, 2015, 119(2): 63–76. [doi: 10.1016/j.cmpb.2015.03.003]

5 Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines. Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, Sweden. 1999. 55–60.

6 Huang YW. Dynamic cost-sensitive ensemble classification based on extreme learning machine for mining imbalanced massive data streams. International Journal of u- and e-Service, Science and Technology, 2015, 8(1): 333–346. [doi: 10.14257/ijunesst]

7 Schapire RE. A brief introduction to boosting. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Stockholm, Sweden. 1999. 1401–1406.

8 Zięba M, Tomczak JM. Boosted SVM with active learning strategy for imbalanced data. Soft Computing, 2015, 19(12): 3357–3368. [doi: 10.1007/s00500-014-1407-5]

9 吴敏. 支持向量机分类算法的若干研究[硕士学位论文]. 南京: 南京邮电大学, 2014.

10 UCI. Welcome to the UC Irvine Machine Learning Repository! http://archive.ics.uci.edu/ml.

© 中国科学院软件研究所 http://www.c-s-a.org.cn