DOI: 10.11992/tis.201706079
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180408.1555.020.html

优化 AUC 两遍学习算法

栾寻, 高尉
（南京大学 计算机软件新技术国家重点实验室, 南京 210023）

摘　要: ROC 曲线下的面积 (简称 AUC) 是机器学习中一种重要的性能评价准则, 广泛应用于类别不平衡学习、代价敏感学习、排序学习等诸多学习任务。由于 AUC 定义于正负样本之间, 传统方法需存储整个数据而不能适用于大数据。为解决大规模问题, 前人已提出优化 AUC 的单遍学习算法, 该算法仅需遍历数据一次, 通过存储一阶与二阶统计量来进行优化 AUC 学习。然而在实际应用中, 处理二阶统计量依然需要很高的存储与计算开销。为此, 本文提出了一种新的优化 AUC 两遍学习算法 TPAUC (two-pass AUC optimization)。该算法的基本思想是遍历数据两遍, 第一遍扫描数据获得正、负样本的均值, 第二遍采用随机梯度下降方法优化 AUC。算法的优点在于通过遍历数据两遍来避免存储和计算二阶统计量, 从而提高算法的效率, 最后本文通过实验说明方法的有效性。

关键词: 机器学习; AUC; ROC; 单遍学习; 在线学习; 排序; 随机梯度下降; 统计量
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)03−0395−04

中文引用格式: 栾寻, 高尉. 优化 AUC 两遍学习算法[J]. 智能系统学报, 2018, 13(3): 395–398.
英文引用格式: LUAN Xun, GAO Wei. Two-pass AUC optimization[J]. CAAI transactions on intelligent systems, 2018, 13(3): 395–398.

Two-pass AUC optimization

LUAN Xun, GAO Wei
(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China)

Abstract: The area under an ROC curve (AUC) has been an important performance index for class-imbalanced learning, cost-sensitive learning, learning to rank, etc. Traditional AUC optimization requires the entire dataset to be stored because AUC is defined as pairs of positive and negative instances. To solve this problem, the one-pass AUC (OPAUC) algorithm was introduced previously to scan the data only once and store the first- and second-order statistics. However, in many real applications, the second-order statistics require high storage and are computationally costly, especially for high-dimensional datasets. We introduce the two-pass AUC (TPAUC) optimization to calculate the mean of positive and negative instances in the first pass and then use the stochastic gradient descent method in the second pass. The new algorithm requires the storage of the first-order statistics but not the second-order statistics; hence, the efficiency is improved. Finally, experiments are used to verify the effectiveness of the proposed algorithm.

Keywords: machine learning; AUC; ROC; one-pass learning; online learning; ranking; stochastic gradient descent; statistics

曲线 ROC 下的面积 (简称 AUC) 是机器学习中一种重要的性能评价准则[1-5], 广泛应用于类别不平衡学习、代价敏感学习、信息检索等诸多学习任务。例如, 在邮件协调过滤或人脸识别中, 某些类别的数据显著多于其他类别, 而类别不平衡性比例[6]可能为 10^6 之多。对 AUC 的研究可以追溯至 20 世纪 70 年代的雷达信号探测分析, 之后 AUC 被用于心理学、医学检测以及机器学习。直观而言, AUC 用于衡量一种学习算法将训练数据中正类数据排在负类数据之前的概率。

由于 AUC 的广泛实际应用, 出现了很多优化

收稿日期: 2017−06−24.　**网络出版日期:** 2018−04−08.
基金项目: 国家自然科学基金青年科学基金项目 (61503179); 江苏省青年基金项目 (BK20150586).
通信作者: 高尉. E-mail: gaow@lamda.nju.edu.cn.

AUC 学习方法，如支持向量机方法[7-8]、集成学习 boosting 算法[9-10]，以及梯度下降算法[11]。这些方法需要存储整个训练数据集，算法在运行时需要扫描数据多遍，因此难以解决大规模学习任务。在理论方面，AGARWAL 和 ROTH[12]给出了优化 AUC 可学习性的充分条件和必要条件，而 GAO 和 ZHOU[13]则根据稳定性给出了可学习性的充要条件。

针对大规模 AUC 优化学习，ZHAO 等[14]于 2011 年提出优化 AUC 的在线学习算法，该方法借助于辅助存储器，随机采取正样本与负样本。而辅助存储器的大小与数据规模密切相关，因此很难应用于大规模数据或不断增加的数据。为此，GAO 等[3]于 2013 年提出优化 AUC 的单遍学习方法，该算法仅需遍历数据一次，通过存储一阶与二阶统计量优化 AUC 学习。

在实际应用中，存储与计算二阶统计量依旧需要较高的存储与计算开销。因此，本文提出了一种新的优化 AUC 两遍学习算法 TPAUC (two-pass AUC optimization)。该算法遍历数据两遍：第一遍统计正负样本均值，第二遍通过随机梯度方法进行优化 AUC 学习。新算法只需计算与存储一阶统计量，而不需要存储二阶统计量，从而有效地提高效率，最后本文通过实验验证了该算法的有效性。

1　TPAUC 学习方法

设示例空间$X \subset R^d$和Y分别表示样本的输入空间和输出空间，本文关注二分类问题，于是有$Y=\{+1,-1\}$。假设 D 表示空间$X\times Y$上潜在的联合分布。假设训练数据集为

$$\boldsymbol{S}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$$

其中每个训练样本是根据分布D独立同分布采样所得。进一步假设分类器$f: X\to R$为一个实值函数。给定样本 $\boldsymbol{S}$ 和函数f，AUC(f,$\boldsymbol{S}$)定义为

$$\sum_{i\neq j}\frac{I[f(x_i)>f(x_j)]+\frac{1}{2}I[f(x_i)=f(x_j)]}{T_{S^+}T_{S^-}} \tag{1}$$

式中：$I[\cdot]$为指示函数，如果判定为真，其返回值为 1，否则为 0；T_{S^+}和T_{S^+}分别表示训练集中正、负类样本的样本数。

直接优化 AUC 往往等价于 NP 难问题，从而导致计算不可行。在实际应用中，一种可行的方法是对优化表达式 (1) 进行一种替代损失函数：

$$L(f,S)=\sum_{i\neq j}\frac{l(f(x_i)-f(x_j))I[y_i>y_j]}{T_{S^+}T_{S^-}} \tag{2}$$

式中 $l: R\to R^+$是一个连续的凸函数，常用的函数包括指数损失函数、Hinge 损失函数、Logistic 损失函数等。由于损失函数定义于一对正样本和负样本之间，该替代函数又被称为“成对替代损失函数 (pairwise surrogate loss)”。

借鉴于优化 AUC 单遍学习算法[3]，本文采用最小二乘损失函数，即在式 (2) 中有

$$l(f(x_i)-f(x_j))=(1-f(x_i)+f(x_j))^2$$

为简洁起见，不妨假设样本总数为T，其中正样本数为T_+，负样本数为T_-，以及设优化函数为

$$L(\boldsymbol{w})=\sum_{i\neq j}\frac{l(\boldsymbol{w})I[y_i>y_j]}{T_{S^+}T_{S^-}}$$

式中$l(\boldsymbol{w})=(1-\langle \boldsymbol{w},\boldsymbol{x}_i\rangle+\langle \boldsymbol{w},\boldsymbol{x}_j\rangle)^2$，经过计算整理可得

$$\begin{aligned}L(\boldsymbol{w})=1&+\frac{1}{T_+}\sum_{i:y_i=1}\langle \boldsymbol{w},\boldsymbol{x}_i\rangle^2+\frac{1}{T_-}\sum_{i:y_i=-1}\langle \boldsymbol{w},\boldsymbol{x}_i\rangle^2-\\&\frac{1}{T_+T_-}\sum_{i:y_i=1}\langle \boldsymbol{w},\boldsymbol{x}_i\rangle\sum_{i:y_i=-1}\langle \boldsymbol{w},\boldsymbol{x}_i\rangle-\\&\frac{2}{T_+}\sum_{i:y_i=1}\langle \boldsymbol{w},\boldsymbol{x}_i\rangle+\frac{2}{T_-}\sum_{i:y_i=-1}\langle \boldsymbol{w},\boldsymbol{x}_i\rangle\end{aligned}$$

设正、负样例的协方差矩阵分别为

$$\boldsymbol{S}_T^+=\frac{1}{T_+}\sum_{i:y_i=1}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}},\ \boldsymbol{S}_T^-=\frac{1}{T_-}\sum_{i:y_i=-1}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$$

以及设正样例与负样例的均值分别为

$$\boldsymbol{c}_T^+=\frac{1}{T_+}\sum_{i:y_i=1}\boldsymbol{x}_i,\ \boldsymbol{c}_T^-=\frac{1}{T_-}\sum_{i:y_i=-1}\boldsymbol{x}_i$$

因此表达式$L(\boldsymbol{w})$可以进一步化简、分解为

$$L(\boldsymbol{w})=\sum_t L_t(\boldsymbol{w})/T \tag{3}$$

当$y_t=1$时，有

$$L_t(\boldsymbol{w})=1+\langle \boldsymbol{w},\boldsymbol{x}_t-\boldsymbol{c}_T^+\rangle^2/T_+ +\langle \boldsymbol{w},\boldsymbol{c}_T^+-\boldsymbol{c}_T^-\rangle^2-2\langle \boldsymbol{w},\boldsymbol{c}_T^+-\boldsymbol{c}_T^-\rangle$$

当$y_t=-1$时，有

$$L_t(\boldsymbol{w})=1+\langle \boldsymbol{w},\boldsymbol{x}_t-\boldsymbol{c}_T^-\rangle^2/T_- +\langle \boldsymbol{w},\boldsymbol{c}_T^--\boldsymbol{c}_T^+\rangle^2-2\langle \boldsymbol{w},\boldsymbol{c}_T^--\boldsymbol{c}_T^+\rangle$$

考虑在损失函数中加入正则项，以防止模型过拟合。本文采用随机梯度下降方法[15-19]，因此

$$\boldsymbol{w}_t=\boldsymbol{w}_t-\eta\nabla L_t(\boldsymbol{w}_{t-1})$$

只需得到关于$\boldsymbol{w}_{t-1}$的梯度表达式，而梯度只需对式 (3) 中$L_t(\boldsymbol{w})$表达式直接求导可得。

因此，本文的思想是不需存储协方差矩阵$\boldsymbol{S}_T^+$和$\boldsymbol{S}_T^-$，需利用均值$\boldsymbol{c}_T^+$和$\boldsymbol{c}_T^-$，从而即可进行优化 AUC 学习。本方法的核心只需要数据的一阶统计量，而不需要二阶统计量，从而将算法空间复杂度降至$O(d)$。同时，该公式中$\boldsymbol{c}_T^+$和$\boldsymbol{c}_T^-$是整个样本空间中正例和负例的均值，在第 1 次遍历数据时不可知，因此需要遍历数据两遍。

本文方法的基本流程可以分为两步：第 1 步遍历数据，统计正样本和负样本均值$\boldsymbol{c}_T^+$和$\boldsymbol{c}_T^-$；第 2 步遍历将利用数据的均值计算得到梯度，然后利用随机梯度下降法更新$\boldsymbol{w}$而完成优化 AUC 的学习，并在实验中取得很好的效果。

2　实验验证

本文将在标准真实数据集和高维数据集实验验证所提方法的有效性, 其中 8 个标准数据集分别为 diabetes、fourclass、german、splice、usps、letter、magic04、a9a。数据集中样本数量从 768~32 561 不等, 样本维度的范围从 8~256。所有数据集的特征都被规范到[-1, 1], 多分类问题被转变为两分类问题, 随机将类别划分成两类。

TPAUC 算法的学习率参数η和正则化参数λ范围都为$\{2^{-10},2^{-9},2^{-8},\cdots,2,4\}$。首先将数据集划分为训练集和测试集, 参数的选择通过在训练集上进行五折交叉验证来确定。选定参数后, 再在测试集上进行 5 遍五折交叉验证, 将这 25 次的结果取平均值作为最终的测试结果。

本文比较了如下 5 种算法:

1) OPAUC: 优化 AUC 单遍学习算法[3]。

2) OAMseq: 优化 AUC 的在线学习算法[14]。

3) OAMgra: 优化 AUC 的在线学习算法[14]。

4) Online Uni-Exp: 优化加权单变量指数损失函数[20]。

5) Online Uni-Squ: 优化加权单变量平方损失函数[20]。

实验结果如表 1 所示, 不难发现, 本文提出的优化 AUC 两遍学习方法 TPAUC 性能与 OPAUC 相当, 但明显优于 OAMseq、OAMgra、Online Uni-Exp 以及 Online Uni-Squ。

本文选用 8 个高维稀疏数据集, 分别为 real-sim、rcv、rcv1v2、sector、sector.lvr、news20、ecml2012、news20.b。数据集中样本数量从 9 619~456 886 不等。特征维度的范围为 20 985~1 355 191。实验设置与标准数据集相似, 实验结果如表 2 所示。可以发现, TPAUC 算法在高维稀疏数据上与其他算法的效果具有可比性或性能更优。

表 1　TPAUC 在低维数据集上性能比较

Table 1　Comparisons of TPAUC on low-dim. datasets

数据集	diabetes	fourlcass	german	splice	usps	letter	magic04	a9a
TPAUC	0.841 1	0.830 9	0.798 1	0.915 1	0.971 3	0.811 5	0.831 9	0.900 5
OPAUC	0.830 9	0.831 0	0.797 8	0.923 1	0.962 0	0.811 4	0.838 3	0.900 2
OAMseq	0.825 4	0.830 6	0.774 7	0.859 4	0.931 0	0.754 9	0.823 8	0.842 0
OAMgra	0.829 5	0.829 5	0.772 3	0.886 4	0.934 8	0.760 3	0.825 9	0.857 1
Online Uni-Exp	0.821 5	0.828 1	0.790 8	0.893 1	0.953 8	0.811 3	0.835 3	0.900 5
Online Uni-Squ	0.825 8	0.829 2	0.789 9	0.915 3	0.956 3	0.805 3	0.834 4	0.894 9

表 2　TPAUC 在高维数据集上性能比较

Table 2　Comparisons of TPAUC on high-dim. datasets

数据集	real-sim	rcv	rcv1v2	sector.lvr	sector	news20	ecml2012	news20.b
TPAUC	0.975 3	0.990 3	0.976 5	0.996 6	0.923 7	0.891 0	0.962 0	0.640 1
OPAUC	0.974 5	0.980 2	0.963 3	0.996 5	0.929 6	0.884 0	0.963 0	0.640 6
OAMseq	0.984 0	0.988 5	0.968 6	0.996 5	0.916 3	0.854 3	0.920 0	0.631 4
OAMgra	0.976 2	0.985 2	0.960 4	0.995 5	0.904 3	0.834 6	0.965 7	0.635 1
Online Uni-Exp	0.991 4	0.990 7	0.982 2	0.996 9	0.921 5	0.888 0	0.982 0	0.634 7
Online Uni-Squ	0.992 0	0.991 8	0.981 8	0.966 9	0.920 3	0.887 8	0.953 0	0.623 7

3　结束语

ROC 曲线下的面积 (简称 AUC) 是机器学习中一种重要的性能评价准则, 由于 AUC 定义于正负样本之间, 传统方法需存储整个数据而不能适用于大数据。为此 Gao 等提出优化 AUC 的单遍学习算法, 该算法仅需遍历数据一次, 通过存储一阶与二阶统计量来进行优化 AUC 学习。本文致力于减少二阶统计量的计算与存储开销, 提出一种新的优化 AUC 两遍学习算法 TPAUC。新提出的算法只需计算与存储一阶统计量, 而不需要存储二阶统计量, 从而有效地提高效率, 最后本文通过实验验证了该算法的有效性。

参考文献:

[1] HSIEH F, TURNBULL B W. Nonparametric and semiparametric estimation of the receiver operating characteristic curve[J]. The annals of statistics, 1996, 24(1): 25–40.

[2] ELKAN C. The foundations of cost-sensitive learning[C]// Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, WA, 2001: 973–978.

[3] GAO Wei, JIN Rong, ZHU Shenghuo, et al. One-pass AUC optimization[C]//Proceedings of the 30th International Conference on Machine Learning. Atlanta, GA, 2013: 906–914.

[4] HAND D J. Measuring classifier performance: a coherent alternative to the area under the ROC curve[J]. Machine learning, 2009, 77(1): 103–123.

[5] EGAN J P. Signal detection theory and ROC analysis, series in cognition and perception[M]. New York: Academic Press, 1975.

[6] WU Jianxin, BRUBAKER S C, MULLIN M D, et al. Fast asymmetric learning for cascade face detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(3): 369–382.

[7] BREFELD U, SCHEFFER T. AUC maximizing support vector learning[C]//Proceedings of ICML 2005 Workshop on ROC Analysis in Machine Learning. Bonn, Germany, 2005.

[8] JOACHIMS T. A support vector method for multivariate performance measures[C]//Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 377–384.

[9] FREUND Y, IYER R, SCHAPIRE R, et al. An efficient boosting algorithm for combining preferences[J]. Journal of machine learning research, 2003, 4: 933–969.

[10] RUDIN C, SCHAPIRE R E. Margin-based ranking and an equivalence between AdaBoost and RankBoost[J]. Journal of machine learning research, 2009, 10: 2193–2232.

[11] HERSCHTAL A, RASKUTTI B. Optimising area under the ROC curve using gradient descent[C]//Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada, 2004.

[12] AGARWAL S, ROTH D. Learnability of bipartite ranking functions[C]//Proceedings of the 18th Annual Conference on Learning Theory. Bertinoro, Italy, 2005: 16–31.

[13] GAO Wei, ZHOU Zhihua. Uniform convergence, stability and learnability for ranking problems[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 1337–1343.

[14] ZHAO Peilin, HOI S C H, JIN Rong, et al. Online AUC maximization[C]//Proceedings of the 28th International Conference on Machine Learning. Bellevue, WA, 2011: 233–240.

[15] GAO Wei, WANG Lu, JIN Rong, et al. One-pass AUC optimization[J]. Artificial intelligence, 2016, 236: 1–29.

[16] SHALEV-SHWARTZ S, SINGER Y, SREBRO N, et al. Pegasos: primal estimated sub-gradient solver for SVM[J]. Mathematical programming, 2011, 127(1): 3–30.

[17] CESA-BIANCHI N, LUGOSI G. Prediction, learning, and games[M]. New York: Cambridge University Press, 2006.

[18] HAZAN E, KALAI A, KALE S, et al. Logarithmic regret algorithms for online convex optimization[C]//Proceedings of the 19th Annual Conference on Learning Theory. Pittsburgh, PA, 2006: 499–513.

[19] DEVROYE L, GYÖRFI L, LUGOSI G. A probabilistic theory of pattern recognition[M]. New York: Springer, 1996.

[20] KOTŁOWSKI W, DEMBCZYŃSKI K, HÜLLERMEIER E. Bipartite ranking through minimization of univariate loss[C]//Proceedings of the 28th International Conference on Machine Learning. Bellevue, WA, 2011: 1113–1120.

作者简介：

栾寻，男，1994 年生，硕士研究生，主要研究方向为大规模机器学习、推荐系统。