详解xgboost算法的样本不平衡问题

原创 石头 机器学习算法那些事 2019-01-18

XGBoost官方文档对参数scale_pos_weight的定义：

scale_pos_weight [default=1]

- Control the balance of positive and negative weights, useful for unbalanced classes.
- A typical value to consider sum(negative instances)/sum(positive instances)

翻译：

调节正负样本权重的平衡，常用来处理不平衡的正负样本数据。

典型值算法：

scale_pos_weight = 负样本总数/正样本总数。若训练负样本总数是500，正样本总数100，那么设置scale_pos_weigh为 5。

scale_pos_weight的含义：

字面意思是正样本权重尺度，是一个关于调节正样本权重的变量。 好吧，让我们深入XGBoost的源码分析：

源码：

```
for (omp_ulong i = n - remainder; i < n; ++i)
{
    auto y = info.labels_[i];
    bst_float p = Loss::PredTransform(preds_h[i]);
    bst_float w = info.GetWeight(i);
    w += y * ((param_.scale_pos_weight * w) - w);          # 权重更新方程
    gpair[i] = GradientPair(Loss::FirstOrderGradient(p, y) * w,
            Loss::SecondOrderGradient(p, y) * w);
}
```

我们重点关注权重更新方程那一行，可得如下结论：

当scale_pos_weight > 1时，即训练集的负样本总数大于正样本总数，y表示类别，由权重更新方程可知，正类对应的权重w增加，负类对应的权重不变 。同理可知scale_pos_weight < 1时的权重更新 。

讲到这里，是不是觉得参数 scale_pos_weight的内容已经讲完了？其实不是，xgboost官方文档对定义又加了如下补充：

For common cases such as ads click through log, the dataset is extremely imbalanced. This can affect the training of XGBoost model, and there are two ways to improve it.

- If you care only about the overall performance metric (AUC) of your prediction
  - Balance the positive and negative weights via `scale_pos_weight`
  - Use AUC for evaluation
- If you care about predicting the right probability
  - In such a case, you cannot re-balance the dataset
  - Set parameter `max_delta_step` to a finite number (say 1) to help convergence

翻译：

常见的情况比如广告点击日志，数据集是极其不平衡的，导致训练的xgboost模型受到影响，有两种方法可以改善：

1）如果仅仅关注预测问题的AUC指标，那么你可以调节 scale_pos_weight参数来帮助训练数据不平衡带来的收敛问题 。

2）如果关注预测概率的准确性问题，那么你就不能调节scale_pos_weight参数来改变样本权重的方法帮助收敛，可通过设置参数 max_delta_step 为一个有限的值来帮助收敛 。

说一说我对这两种方法的理解：

如果仅仅关注AUC指标，那么我们可以调节参数scale_pos_weight，因为改变scale_pos_weight并没有改变AUC值，假设我们预测四个人是正样本的概率分别是：0.2,0.4,0.6,0.8，调节该参数后预测四个人是正样本的概率分别是0.3,0.6,0.7,0.9，根据AUC的计算公式，调节参数前后的AUC并没有改变，但是样本的预测概率却发生改变了。因此，调节参数 scale_pos_weight改变了预测概率，并没有改变AUC指标 。

也可以通过贝叶斯的角度去理解第二种方法中不能改变权重的原因：

[例] 训练数据集有1000个正常人，100个癌症病人，假设不改变权重，利用贝叶斯思想预测测试样本是正常人的先验概率大约是0.9，癌症病人的先验概率大约是0.1 。若大大增加癌症病人的权重，有可能导致测试样本是正常人和癌症病人的先验概率都是0.5，这完全不符合现实情况 。因此，改变权重导致了预测概率的不可信问题 。

参考

https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html

https://xgboost.readthedocs.io/en/latest/parameter.html

https://blog.csdn.net/h4565445654/article/details/72257538

推荐阅读

XGBoost算法原理小结

XGBoost参数调优小结

浅谈频率学派和贝叶斯学派