各种场景的处理方法

July 1, 2020

1 不均衡样本集的处理 [1]

1.1 场景描述

在训练二分类模型时，经常会遇到正负样本不均衡的问题，例如医疗诊断、网络入侵检测、信用卡反诈骗等。对于很多分类算法，如果直接采用不均衡的样本集来进行训练学习，会存在一些问题。例如，如果正负样本比例达到 1:99，则分类器简单地将所有样本都判为负样本就能达到 99% 的正确率，显然这并不是我们想要的，我们想让分类器在正样本和负样本上都有足够的准确率和召回率。

1.2 问题

对于二分类问题，当训练集中正负样本非常不均衡时，如何处理数据以更好地训练分类模型?

1.3 处理方法

1.3.1 基于数据的方法

主要是对数据进行重采样，使原本不均衡的样本变得均衡。

直接的随机采样虽然可以使样本集变得均衡，但会带来一些问题：过采样对少数类样本进行了多次复制，扩大了数据规模，增加了模型训练的复杂度，同时也容易造成过拟合；欠采样会丢弃一些样本，可能会损失部分有用信息，造成模型只学到了整体模式的一部分。

为了解决上述问题，通常在过采样时并不是简单的复制样本，而是采用一些方法生成新的样本，这样可以降低过拟合的风险。

在实际应用中，具体的采样操作可能并不总是如上述几个算法一样，但基本思路很多时候还是一致的。例如，基于聚类的采样方法，利用数据的类簇信息来指导过采样/欠采样操作；经常用到的数据扩充 (data augmentation) 方法也是一种过采样，对少数类样本进行一些噪音扰动或变换（如图像数据集中对图片进行裁剪、翻转、旋转、加光照等）以构造出新的样本；而 Hard Negative Mining 则是一种欠采样，把比较难的样本抽出来用于迭代分类器。

1.3.2 基于算法的方法

在样本不均衡时，也可以通过改变模型训练时的目标函数（如代价敏感学习中不同类别有不同的权重）来矫正这种不平衡性；当样本数目极其不均衡时，也可以将问题转化为 one-class learning / anomalydetection。本节主要关注采样，不再细述这些方法（我们会在其它章节的陆续推送相关知识点）。

References

[1] "Hulu 机器学习问题与解答系列 | 第四弹：不均衡样本集的处理." [Online]. Available: https://mp.weixin.qq.com/s?__biz=MzA5NzQyNTcxMA==&mid=2656430319&idx=1&sn=2e22a6371e30929b4aeacb33565e5184&scene=19#wechat_redirect