【小夕精选】如何优雅而时髦的解决不均衡分类问题

微调 夕小瑶的卖萌屋 2018-10-28

之前小夕因项目需要研究了一小阵子的不均衡（文本）分类问题，不过没有研究的太过深入，也没有总结出一套成体系的处理思路。正好今天发现数据挖掘大佬「微调」在知乎上写了一个言简意赅又很具有实际操作价值的回答，于是搬过来分享给大家啦～相关方向的小伙伴记得点击文末**阅读原文**关注「微调」大佬哦。

模型如何评价

先谈谈这种极端的类别不平衡的评估问题，我们一般用的指标有（前两个是全局评估，最后一个是点评估）：

- ROC曲线下的面积（AUC_ROC）
- mean Average Precesion（mAP），指的是在不同召回下的最大精确度的平均值
- Precision@Rank k。假设共有 *n*个点，假设其中*k*个点是少数样本时的Precision。这个评估方法在推荐系统中也常常会用。

选择哪个评估标准需要取决于具体问题。而在上线前怎么确定你的模型已经达标?这个需要AB test，每个公司都有不同的标准，很难一概而论。**重点是新系统至少要比现有系统在某方面有了提升，而全新模型至少应该符合从业者的基本预期**。也要认识到大部分情况下上线的模型都不可能是完美的，我的个人建议是可以利用「已有的监督模型+人工」做主动学习（active learning）。比如先上线一个不完美的模型，每次将模型预测中最不确定的部分（预测值在临界点附近的样本）交给人工验证，并重新训练逐步提高模型预测的精准度。

如何解决问题

至于如何处理数据不平衡的问题，最传统的思路还是使用过采样和欠采样等。相关资料大家看的比较多的是08年的Survey Paper [1]，比较新和前沿的做法可以参考[2]，可以至少读一下Related Works部分了解一下这些年来常用的非平衡数据处理方法有哪些。比较科普的文章可以参考我的回答：

微调：欠采样（undersampling）和过采样（oversampling）会对模型带来怎样的影响?
https://www.zhihu.com/question/269698662/answer/352279936

里面也介绍了一些常用的工具。去年其实也写过一篇类似的文章，可以参考：

如何处理数据中的「类别不平衡」？
https://zhuanlan.zhihu.com/p/32940093

如果上述方法表现依然不好，还有几个方法可供尝试：

1. **有监督的集成学习**：可以先用采样的方法建立*k*个平衡的训练集，每个训练集上单独训练一个分类器，并对*k*个分类器结果取平均。一般在这种情况下，每个平衡训练集上都需要使用比较简单的分类器，如逻辑回归。**其实在实际使用中，这种方法不一定会比集成树模型更好**，可能还不如使用xgboost。但在复杂问题上多尝试一些手段是好的，说不定有奇效。
2. **无监督的异常检测**：异常检测指的是从数据中找到那些异常值，比如你案例中的“广告”。无监督的异常检测一般依赖于对于数据的假设，比如广告和正常的文章内容很不相同，那么一种假设是广告和正常文章间的欧式距离很大。**无监督异常检测最大优势就是在不需要数据标签，如果在对数据假设正确时效果甚至可以比监督学习更好，尤其是当获取标签成本很高时**。具体的科普文章可以参考我的回答：

   微调：数据挖掘中常见的『异常检测』算法有哪些?
   https://www.zhihu.com/question/280696035/answer/417091151

3. **半监督异常集成学习：**如果把1和2的思路结合起来，你可以试试半监督的方法，具体做法可以参考[3]。简单而言，你可以现在原始数据集上使用多个无监督异常方法来抽取数据的表示，并和原始的数据结合作为新的特征空间。在新的特征空间上使用集成树模型，比如xgboost，来进行监督学习。**无监督异常检测的目的是提高原始数据的表达，监督集成树的目的是降低数据不平衡对于最终预测结果的影响。这个方法还可以和我上面提到的主动学习结合起来，进一步提升系统的性能。**当然，这个方法最大的问题是**运算开销比较大**，需要进行深度优化。
4. **高维数据上的半监督异常检测：**考虑到文本文件在转化后往往维度很高，可以尝试一下最近的一篇KDD文章[4]，主要是找到高维数据在低维空间上的表示，以帮助基于距离的异常检测方法。

总结来看，我建议从以下顺序尝试：

- 直接在数据上尝试有监督的集成学习（方法1）
- 直接在数据上使用多种无监督学习，观察哪一类算法的效果更好（方法2）
- 结合以上两点(方法3)
- 如果以上方法都不管用，尝试方法4
- 使用方法1, 3，4时，可以加入主动学习
- **如果以上方法均不奏效，最靠谱的还是找更多人做数据标注，毕竟数据为王。**从效果上看往往是「监督学习>>半监督学习>无监督」，能用监督就不要依赖无监督。

数据挖掘项目的本质就是试错，所以很难有确定的答案。抛开准确率不谈，另外的重要因素包括系统的效率和耦合度。前者指的是运算开销，后者指的是设计与维护开销，这些在设计方案时都要考虑到。**最终上线的版本不一定是最强力的那个，往往是最适合的那个。**

参考文献

[1] He, H. and Garcia, E.A., 2008. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9), pp.1263-1284.

[2] Roy, A., Cruz, R.M., Sabourin, R. and Cavalcanti, G.D., 2018. A study on combining dynamic selection and data preprocessing for imbalance learning. *Neurocomputing*, *286*, pp.179-192.

[3] Zhao, Y.; Hryniewicki, M.K. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Rio, Brazil, 8–13 July 2018.

[4] Pang, G., Cao, L., Chen, L. and Liu, H., 2018. Learning Representations of Ultrahigh-dimensional Data for Random Distance-based Outlier Detection. *arXiv preprint arXiv:1806.04808*.