Self-paced Ensemble for Highly Imbalanced Massive Data Classification

刘芷宁

摘要

许多实际应用程序揭示了从不平衡数据学习分类器的困难。随着大数据时代的兴起，越来越多的分类任务使用大量的数据集，但数据集极不平衡且质量低下。在这种情况下，现有的大多数学习方法都存在性能差或计算效率低的问题。为了解决这个问题，我们对类不平衡的本质进行了深入的研究，发现不仅是类之间的不平衡，还有数据本质中嵌入的其他困难，特别是噪声和类重叠，阻碍了我们学习有效的分类器。考虑到这些因素，我们提出了一种新的不平衡分类框架，旨在通过欠采样自步调协调数据硬度来生成强大的集合。大量的实验表明，这个新框架在计算效率很高的同时，即使在高度重叠的类和极端倾斜的分布下，也可以带来健壮的性能。请注意，我们的方法可以很容易地适应大多数现有的学习方法 (例如，C4.5, SVM, GBDT 和神经网络)，以提高它们在不平衡数据上的性能。

关键词：不平衡学习；不平衡分类；集成学习；数据重采样

1 引言

导致机器学习系统在实际应用中表现不佳的原因有很多种：它们可能来自于数据采集过程（如噪声，缺失值）、或者来自于数据集的特性（如类别重叠，庞大的数据规模）、也可能来自于采用的机器学习模型的特性（如模型容量过小/过大）和任务本身（如类别分布不平衡）。我们可以看到在上面描述的这些原因中，“类别不平衡”这一因素本身只占了很小的一部分。单纯的“类别不平衡”并不会对任务的难度有决定性的影响。一个无噪声、无缺失值、类别不重叠的 toy dataset，即使其类别不平衡，分类器也能很容易地得到一个合理的分类边界。反之，对于一个大规模的、复杂的、数据很“脏”的任务（多见于实际的工业应用，也是我们主要考虑的问题），即使其数据集是类别平衡的，要解决这类问题通常也很困难。已有的研究[1-3]已讨论“类别不平衡在分类问题中扮演的角色”，我们认为“类别不平衡”更像是一种“催化剂”，它的存在使得原本就较为困难的分类任务变得加倍困难。

2 相关工作

2.1 传统不平衡学习方法

传统的不平衡学习方法通常只考虑某个或某些因素，且最终表现较为依赖于超参的选择：例如基于距离的重采样方法中考虑的 neighbor 数量会影响方法对噪声的敏感性、cost sensitive 类方法中代价敏感矩阵需要由领域专家设置。

2.2 选取对训练贡献最大的样本/最 informative 的样本

在训练的过程中通过相关信息，指导欠采样策略选取那些对当前集成贡献最大的训练样本。例如 Self-paced Learning[4], (Pool-based) Active Learning[5], Balance Cascade[6]等。

3 本文方法

3.1 本文方法概述

欠采样 + 集成策略进行 boosting-like 的串行训练，最终得到一个加性模型。

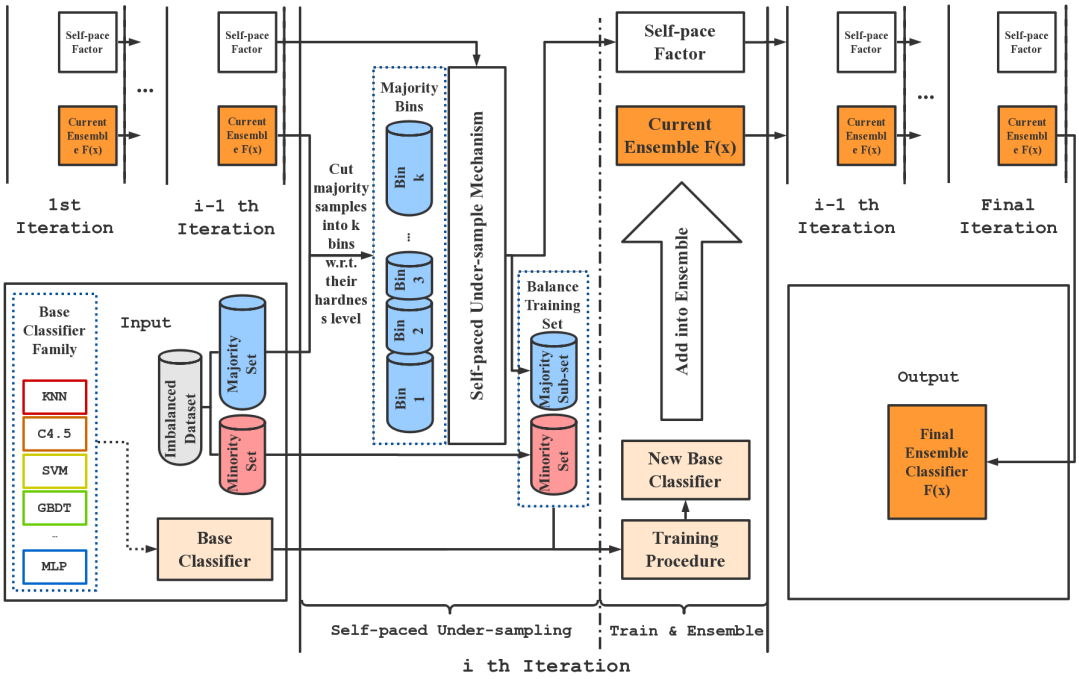

图 1: Self-Paced Ensemble Process.

3.2 分类硬度分布

定义 $\mathcal{H}$ 为一个“分类硬度函数”，任何可分解到单个样本上计算的损失/误差函数都可以被用作分类硬度函数 (e.g., Absolute Error/MSE/Cross-Entropy)。对于一个给定的（trained）模型 $F(\cdot)$，样本 (x, y) 的分类硬度由 $H_x = \mathcal{H}(x, y, F)$ 给出。特别注意分类硬度函数使用当前的模型作为函数的其中一个输入。

3.3 Self-paced Ensemble (SPE)

使用 k 个分桶来近似表示分类硬度分布，并在训练过程中动态更新每个样本的硬度值。

$$B_\ell = \left\{ (x, y) \mid \frac{\ell - 1}{k} \leqslant \mathcal{H}(x, y, F) < \frac{\ell}{k} \right\} w.l.o.g. \mathcal{H}(\cdot) \in [0, 1]$$

3.4 自步因子 α（self-paced factor）

随着训练过程的推进，“简单”样本的数量会快速增长，此时单纯的均衡化采样策略仍然会保留很多这些已被拟合的数据点，导致训练得到的新分类器缺乏 diversity。为了在训练后期增加基学习器的 diversity 并使其关注更加难以分类的样本，作者引入自步因子 α（self-paced factor），它被用来降低那些样本数量过多的 bins 的采样权重。

4 复现细节

4.1 与已有开源代码对比

```
# compute population & hardness contribution of each bin
populations, edges = np.histogram(hardness_c, bins=k_bins)
contributions = np.zeros(k_bins)
index_bins = []
for i_bin in range(k_bins):
    index_bin = ((hardness_c >= edges[i_bin]) & (hardness_c < edges[i_bin + 1]))
    if i_bin == (k_bins - 1):
        index_bin = index_bin | (hardness_c == edges[i_bin + 1])
    index_bins.append(index_bin)
    if populations[i_bin] > 0:
        contributions[i_bin] = hardness_c[index_bin].mean()
```

图 2: 作者的分桶

```
t = np.linspace(0.0, 1.0, k+2)
t = t[1:-1]
sort_index = np.argsort(hardness_c)
index_c = index_c[sort_index]
hardness_c = hardness_c[sort_index]

dist, path = fastdtw(hardness_c, t, dist=lambda a, b: abs(a - b))
edges = set()
for pair in path:
    edges.add(pair[1])
edges = list(edges)
edges.sort()
k_bins = len(edges)
populations = np.zeros(k_bins)
contributions = np.zeros(k_bins)
index_bins = []
```

图 3: 我的分桶

4.2 创新点

DTW (Dynamic time warping) 算法一般是用来度量两个独立时间序列的相似度，尤其是两段序列时长不同情况的方法。DTW 最后求解出两个序列累计匹配的最小距离，对于不等长序列，就有一个时间点对应多个时间点的情况。基于 DTW 一对多且最小化累计距离的思想，预先定义 k 间隔的序列(与该论文的 k 均匀分桶对应)，与所有训练样本的硬度序列进行 DTW 匹配，最后就会获得不均匀的分配，且是 DTW 累计最小距离，从而代替论文作者的 histogram 均匀分桶。

5 实验结果分析

实验在四个数据集：creditcard，covtype，kddcup，payment simulation，其中 kddcup 有 dos 与 prb 和 dos 与 r2l 两种数据集。对比算法包括作者的算法（黄色框），我的算法（红色框）以及其它集成学习方法。

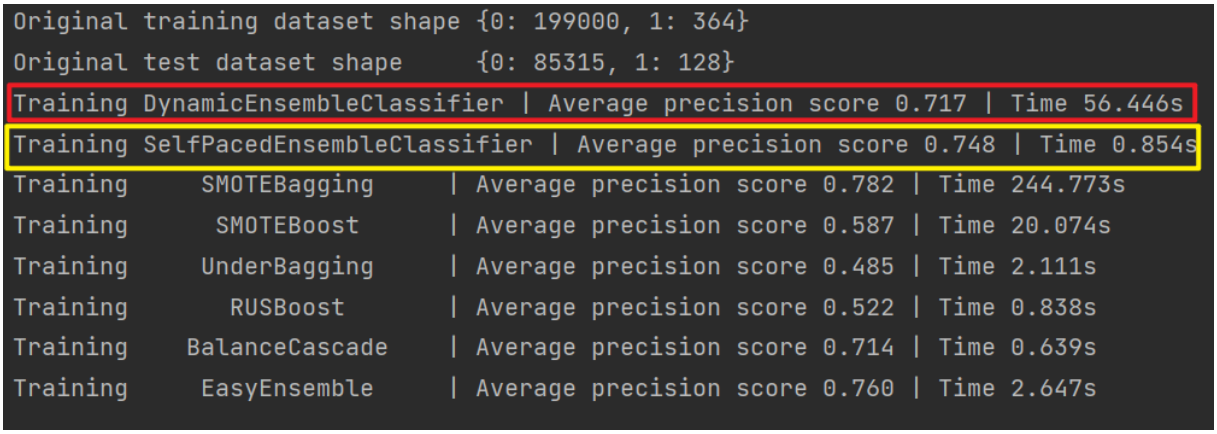

图 4: creditcard，不平衡比 578.88:1

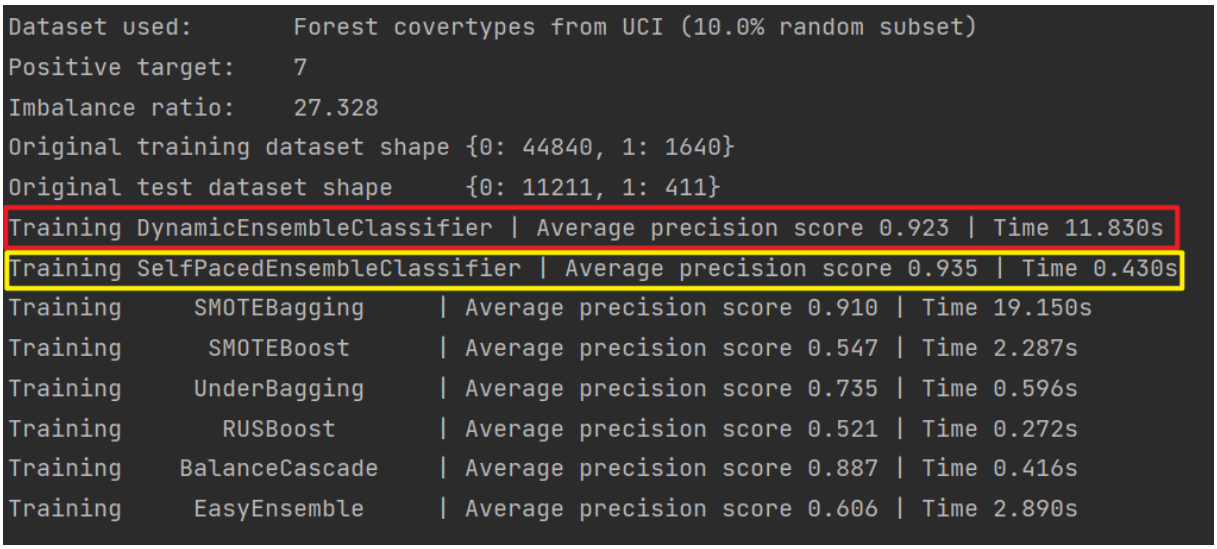

图 5: covtype，不平衡比 27.328:1

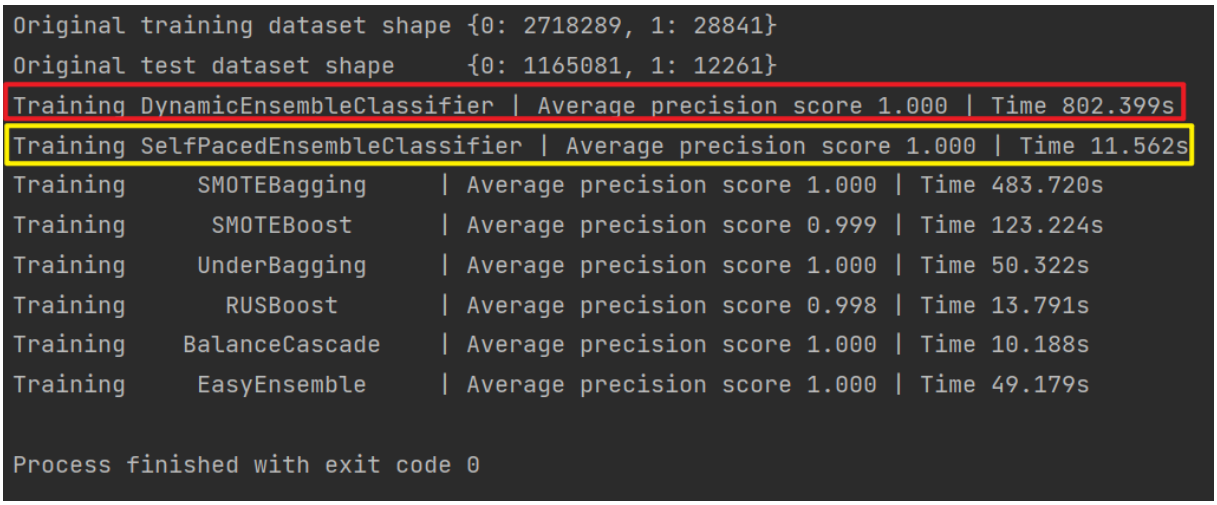

图 6: kddcup 的 dos 与 prb，不平衡比 94.48:1

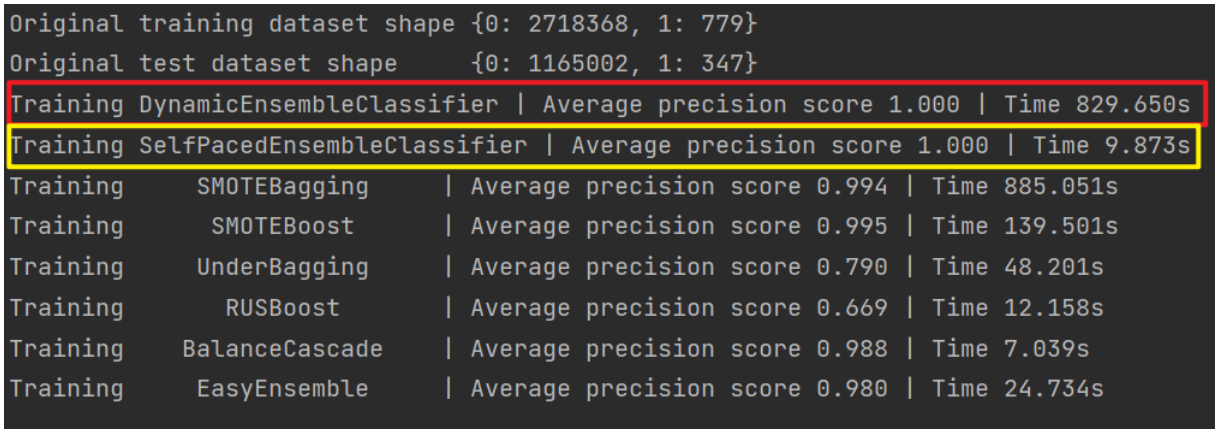

图 7: kddcup 的 dos 与 r2l，不平衡比 3448.82:1

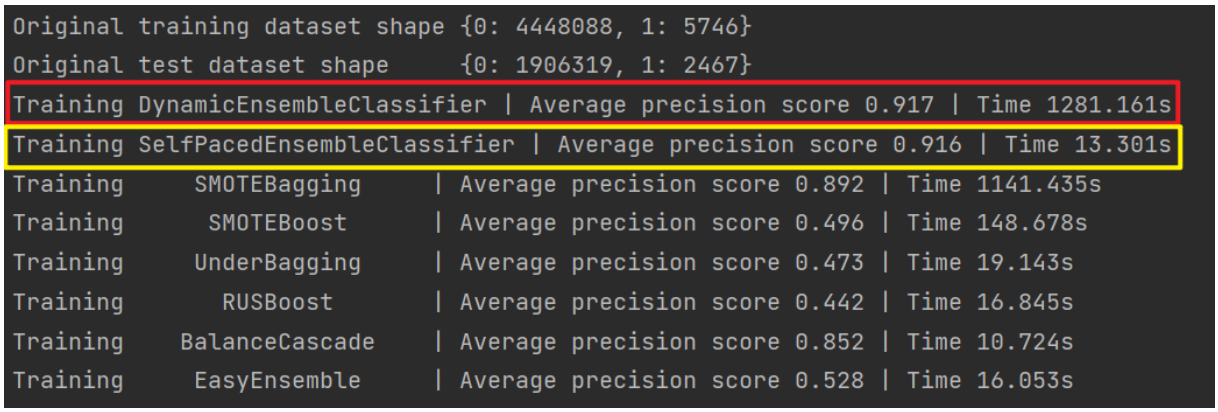

图 8: payment simulation，不平衡比 773.70:1

从结果来看，基于 DTW 的分桶在 Average precision 指标上，与作者的简单分桶相比，基本上没有提升，反而时间复杂度更高了。但从最后的 payment simulation 数据集的微小提升可以思考，也许好的不均匀分桶会有效果。

6 总结与展望

使用“分类硬度分布”的概念来归纳影响不平衡学习的复杂因素，从而指导数据欠采样训练，想法和效果都很好。只是对作者的均匀分桶，我觉得可以有更好的处理，因为样本的硬度并不一定在范围内是均匀分布的。基于 DTW 的聚类分桶可以实现不均匀分桶，从实现效果来看并没有提升，并且时间复杂度更高了不少。有可能是单纯 DTW 聚类本身的缺陷，还可以尝试其它聚类的方法，但聚类一般时间复杂度都很高，所以要想一个快速聚类的方法是一个难点，我想不均匀分桶从逻辑上更合理一点，只能说值得探索。

参考文献

[1] GARCÍA V, SÁNCHEZ J, MOLLINEDA R. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets[C]//Iberoamerican congress on pattern recognition. 2007: 397-406.

[2] NAPIERAŁA K, STEFANOWSKI J, WILK S. Learning from imbalanced data in presence of noisy and borderline examples[C]//International conference on rough sets and current trends in computing. 2010: 158-167.

[3] PRATI R C, BATISTA G E, MONARD M C. Learning with class skews and small disjuncts[C]//

Brazilian Symposium on Artificial Intelligence. 2004: 296-306.

[4] KUMAR M, PACKER B, KOLLER D. Self-paced learning for latent variable models[J]. Advances in neural information processing systems, 2010, 23.

[5] SETTLES B. Active learning literature survey[J]., 2009.

[6] LIU X Y, WU J, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 39(2): 539-550.