

复现线上最优成绩

还原A榜成绩：第一步<http://dix.qq.com/teslaml/admin/platform.htm?projectId=10623&flowId=18689>

这一个工作流用于还原A榜成绩。

Extract\_copy ：第五步<http://dix.qq.com/teslaml/admin/platform.htm?projectId=10623&flowId=18698>

这个工作流作用是提取3k训练集、10w测试集、200w测试集的特征，并从10w测试集用挑选候选样本，扩充训练集。

Add200W\_copy：第六步<http://dix.qq.com/teslaml/admin/platform.htm?projectId=10623&flowId=18702>

这个工作流作用是从200w测试集用挑选候选样本，扩充训练集。并得到最终答案。

分布式\_0.7\_copy: 第二步<http://dix.qq.com/teslaml/admin/platform.htm?projectId=10623&flowId=18706>

这个工作流作用是计算相似度规则寻找黑样本。

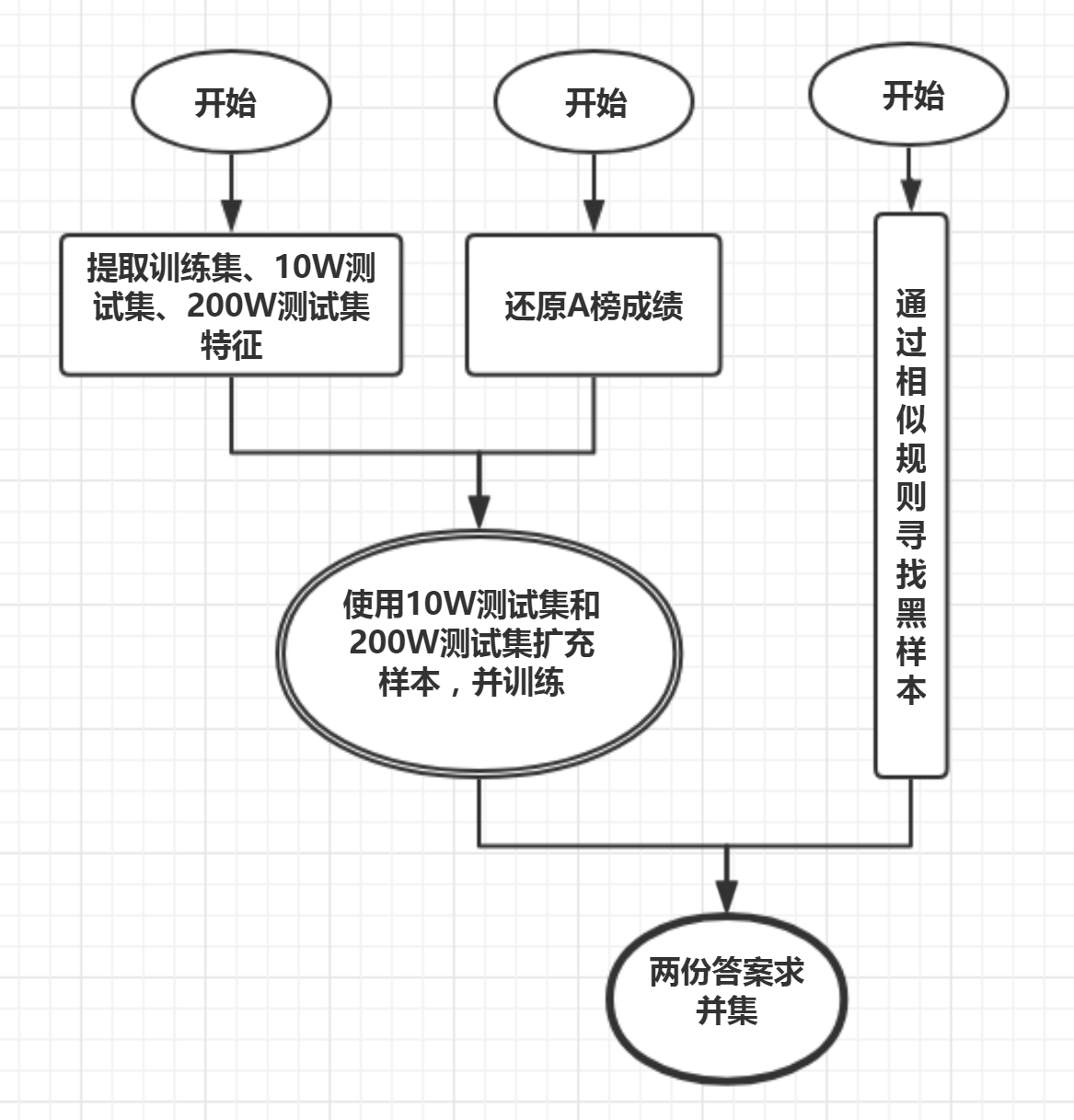
分布式\_0.7\_huitou: 第三步<http://dix.qq.com/teslaml/admin/platform.htm?projectId=10623&flowId=18707>

这个工作流作用是计算相似度规则寻找黑样本。

分布式\_order\_same\_dot:第四步

<http://dix.qq.com/teslaml/admin/platform.htm?projectId=10623&flowId=18709>

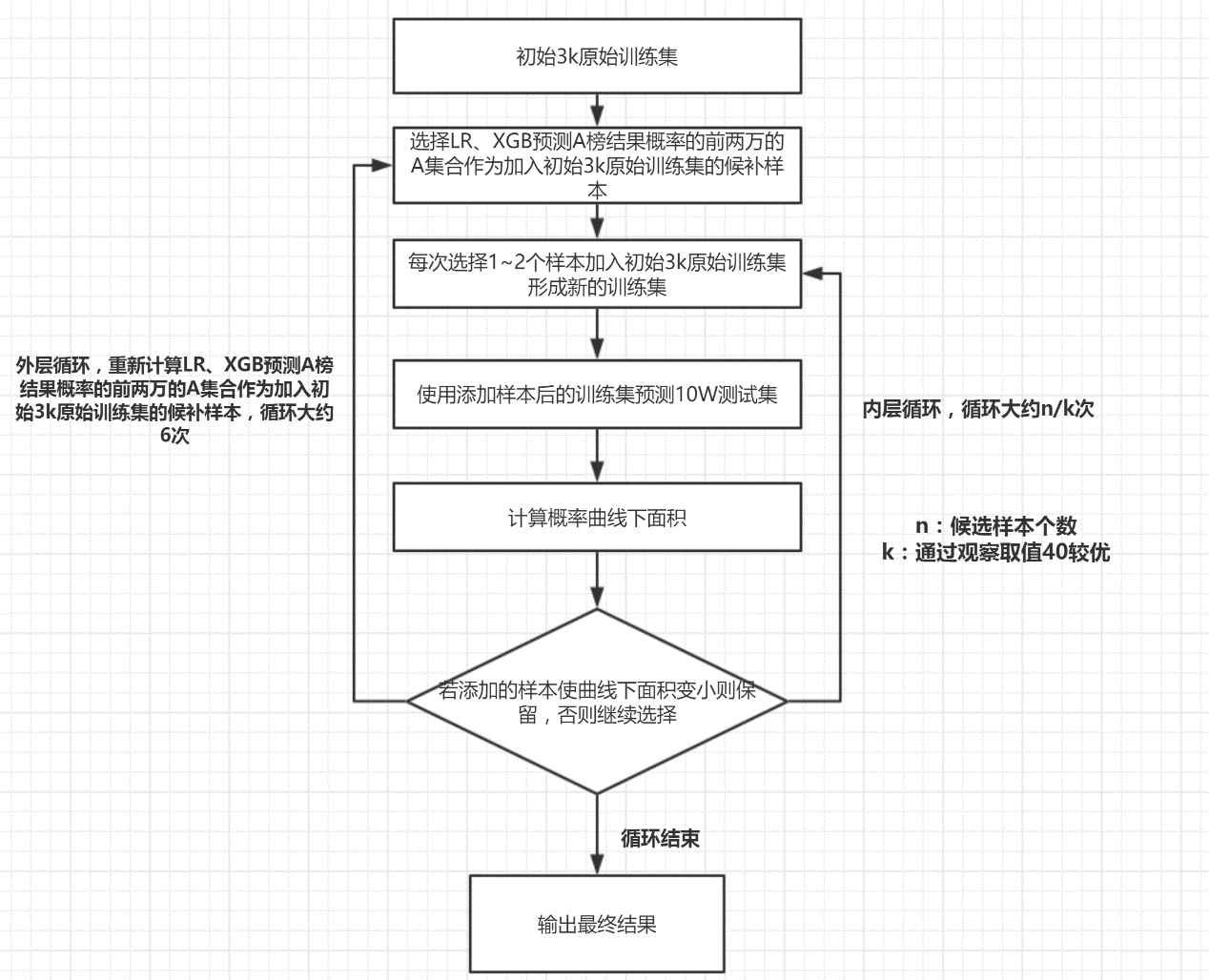
这个工作流作用是计算相似度规则寻找黑样本。



算法说明：

1)分别提取3K训练集、10W测试集、200W测试集特征。

2)扩充训练集思路及还原A榜成绩。

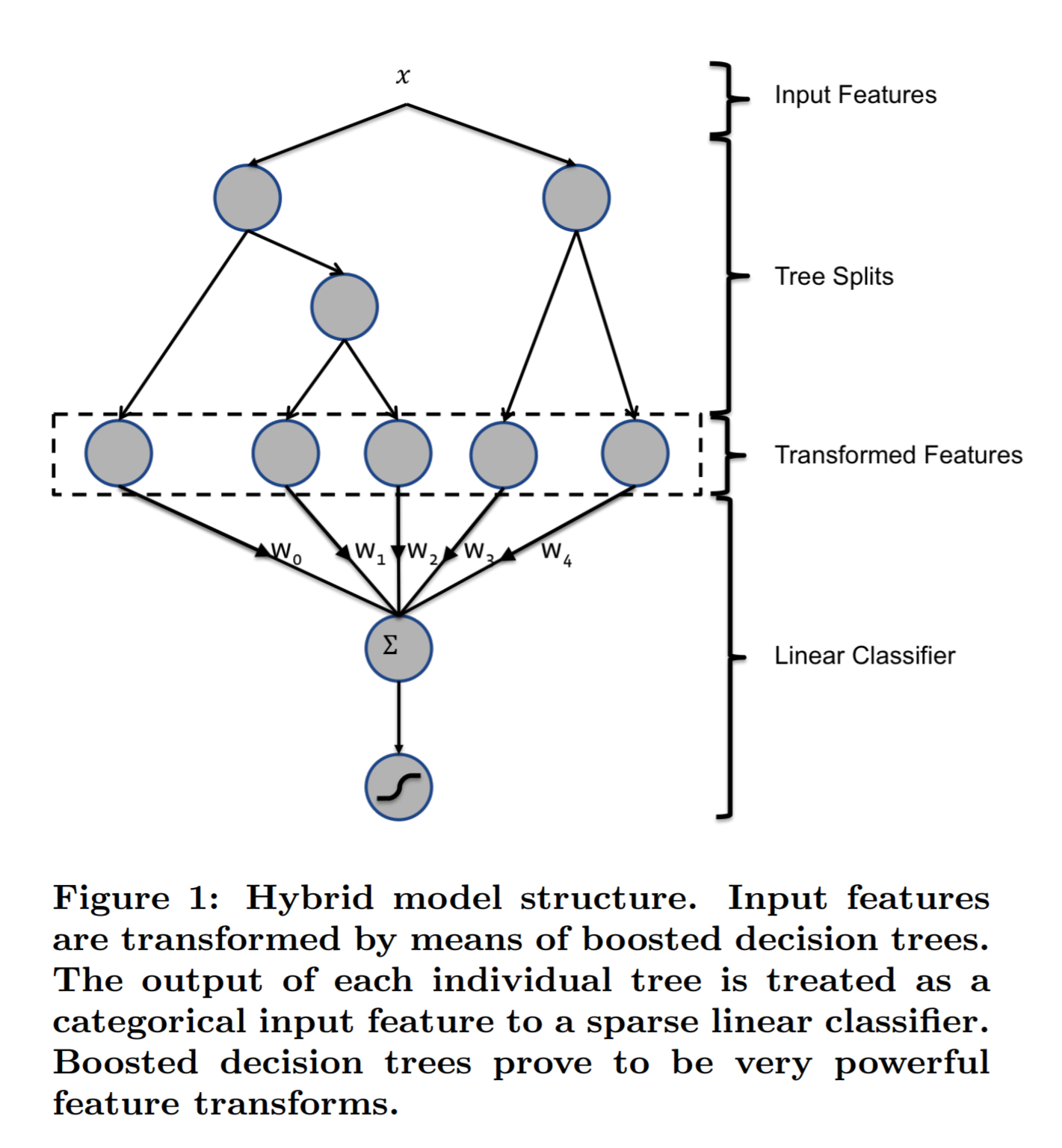


注：



为什么选择LR和XGB的A集合作为扩充训练集的候选样本？

思路来源于本次赛题做模型融合不太方便，而在比赛中做差异性较大的模型对成绩有至关重要的提升，因此参考论文kdd2014（paper title），使用了XGB构造新特征的方法。对于本赛题，先用已有特征训练XGB模型，然后利用XGB模型学习到的树来构造新特征，最后把这些新特征加入原有特征（需要离散）一起训练模型。构造的新特征向量是取值0/1的，向量的每个元素对应于XGB模型中树的叶子结点。当一个样本点通过某棵树最终落在这棵树的一个叶子结点上，那么在新特征向量中这个叶子结点对应的元素值为1，而这棵树的其他叶子结点对应的元素值为0。新特征向量的长度等于GBDT模型里所有树包含的叶子结点数之和。

举例说明。下面的图中的两棵树是XGB模型中的，第一棵树有3个叶子结点，而第二棵树有2个叶子节点。对于一个输入样本点x，如果它在第一棵树最后落在其中的第二个叶子结点，而在第二棵树里最后落在其中的第一个叶子结点。那么通过XGB获得的新特征向量为[0, 1, 0, 1, 0]，其中向量中的前三位对应第一棵树的3个叶子结点，后两位对应第二棵树的2个叶子结点。

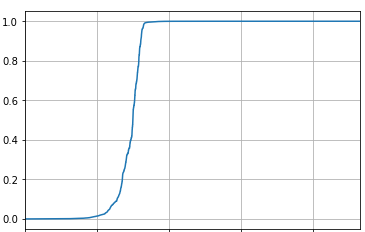
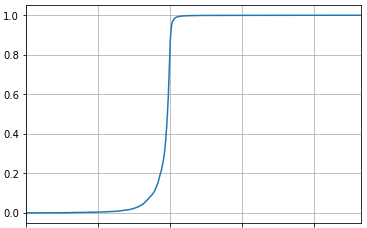
为什么使用预测概率曲线下面积作为选择样本的指标？

通过阅读赛题描述，得知测试集中包含在训练集中未出现的新的攻击手段，即训练样本假设空间不存在的假设。但未出现的新的攻击手段毕竟只是少量样本，所以可以通过添加少量测试样本以扩充训练集（直接对添加的样本赋予黑样本标记，0）。这样就可以使概率曲线面积迅速下降,我们的目标是



式中， 为概率曲线下面积， 为添加候选样本。

(1)未添加样本的概率曲线下面积 (2)添加样本的概率曲线下面积

至此，还原A榜成绩思路完成，(仅保留A榜测试集pairs(ID,LABEL),这套特征弃用)

2.重新对3k训练集、10W测试集、200W测试集使用另外一套特征进行提取，扩充训练集思路同上，通过观察，用于比赛的200W测试集和3k训练集样本分布差异较大，扩充样本变得极为困难，我们队伍大约扩充了十个样本（扩充思路同上），成绩大约有0.5的提升。

3.规则说明

通过在初赛中积累的经验，我们发现有部分机器轨迹存在模仿人类轨迹的现象（即人类几乎不可能画出两个一模一样的轨迹）。即某样本在训练集中为白样本，标签为1。但在测试集中这个样本变为了黑样本，标签为0。对于这类现象是违反训练集假设空间中，因此对于此类标签，我们对轨迹完全相同的样本有



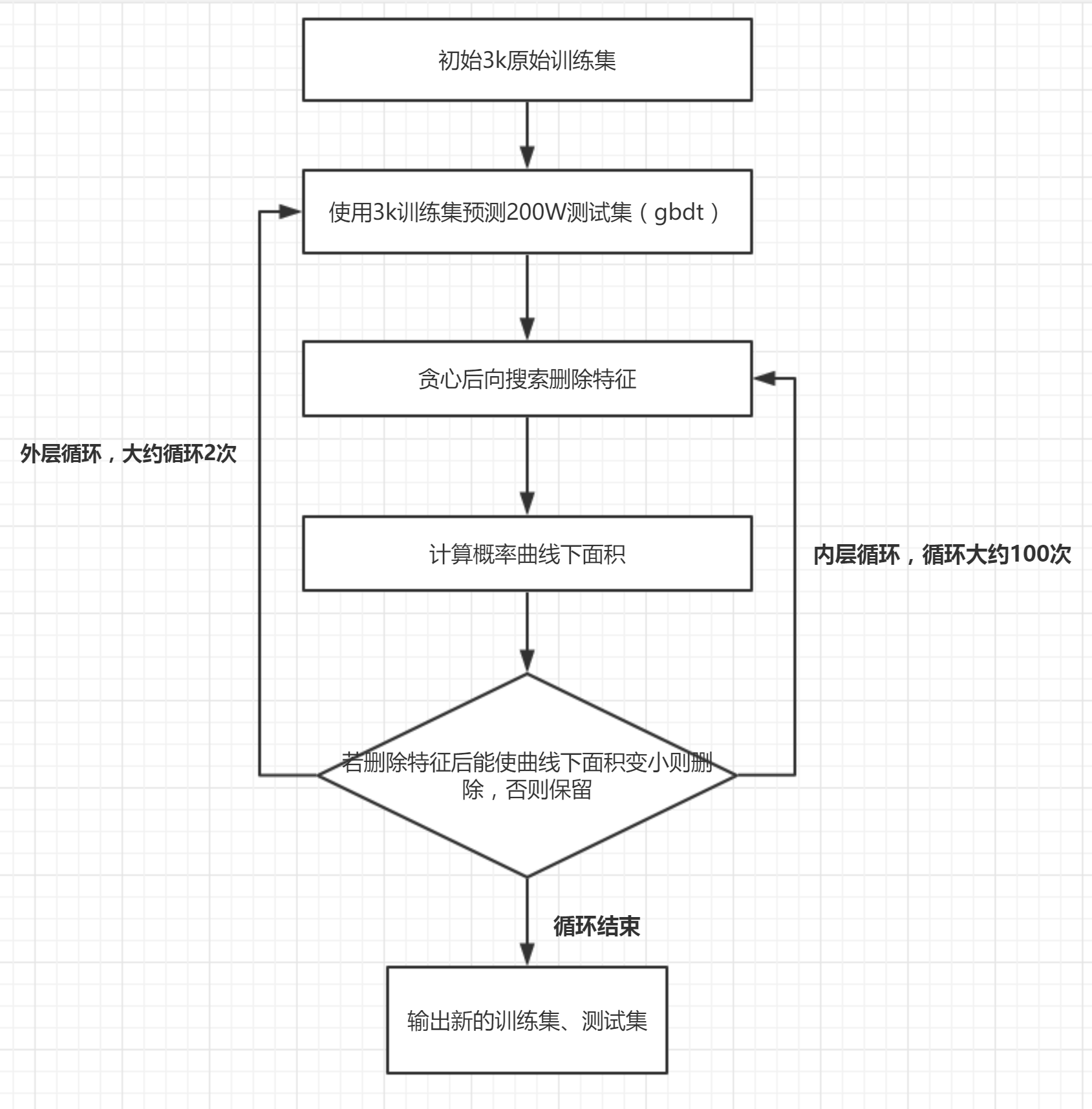
式中d为距离计算公式， 为设定阈值，M为测试集。

对于机器模仿人类轨迹，还存在模仿人类轨迹的同时，在轨迹尾部加上一点干扰。处理这个模仿现象，我们采用截断较长轨迹的尾部计算相似距离。

失败的尝试：

特征选择：

在比赛的最后阶段，我们希望通过特征选择的方式提升分数。



大致思路：采用后向搜索，遍历训练集特征，每次在候选特征中尝试删除1~2个特征,并记录特征index(这个过程大约重复一百次)，然后重复整个过程直至循环结束。



但不幸的是，特征选择没有提升分数。

参考文献

He X, Pan J, Jin O, et al. Practical Lessons from Predicting Clicks on Ads at Facebook[J]. 2014(12):1-9.