机器学习实验 2 报告

计 71 庞博予 2017011315

一、实验设计

1．预处理

实验的第一步是将不定长的评论文本转化为可以用来训练的定长向量。这一步中常用的方法是 tf-idf 以及 word2vec。这里使用了 word2vec 模型，对训练集中的所有文本进行了训练，得到较常见词汇（出现次数多于一定阈值）的词向量。将文本中的每个词的词向量做平均得到一个文本的平均向量。

2．集成学习

实验中实现了两种集成学习方式，一种是 bagging、一种是 adaboost。

由于数据集规模较大、文本向量长度较长，故在原始算法上稍作改进以提高可行性和运行效率。

Bagging 方法中，每一轮从全体训练集中随机抽取一定比例的数据作为训练集训练，并将训练得到的模型缓存，在测试时用每一轮训练出来的模型分别预测并将结果取均值。

Adaboost 方法中，本应将全体训练集按权重作为实际训练集进行每一轮的训练，但是由于其训练复杂度过大以及速度极慢，故采用按权重抽样的方法获得与理论上的训练集在期望意义上相同分布的实际训练集。采用这个抽样训练集进行训练后用全体训练集测试该模型，并根据预测结果正确与否进行权重调整。重复上述过程若干轮。

集成学习中用到的基础分类器分别是 SVM 和决策树，两者均是用 python 的 sklearn 库中默认的实现。需要注意的是，在 bagging 方法中，可以使用 SVR 或 SVC 两种不同形式的 SVM 来训练模型，而在 adaboost 中由于需要判断每一个数据是否预测正确，在不对算法做出修改的情况下只能使用 SVC 进行训练，若使用 SVR 则会因为预测结果为实数而无法判断是否预测准确。

二、实验结果与性能

轮数*训练规模	Bagging+SVM	Bagging+决策树	Adaboost+决策树
30*10000	0.98967	1.00810	
50*20000	0.97749	1.00149	
20*50000	-	1.01975	1.01492
30*60000	-	1.00684	1.00849
50*50000	-	0.99452	0.99524
50*80000	-	0.99623	1.031

在上述几种不同的参数情况下相对 RMSE 最小的算法是 Bagging+SVM 算法。运行速度上，以 SVM 为基础分类器的算法组合运行速度远远慢于同等条件的决策树算法。且决策树的运行速度受训练规模影响较小，而 SVM 运行速度受训练规模影响很大。

注意到 adaboost+决策树的组合中在训练规模较小时没有给出运行结果，因为在训练规模较小时哪怕决策树在该样本上拟合的很好，在全体数据集上的预测正确率低的令人扼腕，其错误率很难低于 50%，基本几次迭代后就会因为无法训练出可以被接受的下一个模型而自动结束。这样结束的算法误差都是极大的。这也是决策树的天然缺点之一，容易过拟合。

而其中 adaboost+svm 的组合由于每次要对全体训练集做测试，时间复杂度过高，而且由于 SVR 并不能适应 adaboost 算法且 SVC 在该问题上表现性能*极差*，故虽然通过参数运行

该组配置，在此并不做性能展示。

目前可以跑出的 RMSE 最小的算法参数是 Bagging 结合 SVM 在 50 轮次随机抽样每次抽样 20000 条数据，其 RMSE 为 0.97749，截止到写报告时在 kaggle 上排名第 37/61。

三、分析与讨论

1．Bagging 与 adaboost

Bagging 与 adaboost 的天然区别之一在于 adaboost 需要对每一次迭代中的模型性能进行评价，然后根据不同数据在在该次迭代模型中的表现调整数据的权值；而 bagging 只需要每次抽样即可。在多分类问题中，adaboost 模型便需要做出适应性调整：使用分类器而不是回归器来训练模型或者修改评价算法不再用预测是否正确作为调整依据而是以“预测结果是否与真实答案够接近”作为调整标准，而后者又需要引入新的参数来控制这个“够接近”的标准。

此外，在多分类问题中 adaboost 本身的正确率限制需要做出修改，原算法不接受预测正确率<50%的模型，而五分类问题中理论上该阈值就要修改为 20%，但这样显然约束力大打折扣，实际效果也未必好。

1．SVM 与决策树

决策树相比 SVM 而言功能更加单一化，即基本上只能用于分类而不能用于回归。而本问题中评分虽然是离散的类似于分类问题，但本质是一个连续的评分空间，评分之间存在偏序关系，这时候决策树只能支持的分类模型就显得力不从心。

2．不同的预处理方法

实验中的词嵌入部分一开始采用了 128 维词向量空间，后改为 256 维词向量空间。虽然同为 word2vec 模型仍然有很大局限性，但是可以看到将维度从 128 提升到 256 后正确率有了小幅的提升。