机器学习实验-Ensemble Learning

2012011307 黄必胜

一 实验目的

基于两种不同分类器（SVM, Decision Tree），比较两种不同的 Ensemble Learning 算法（Bagging, AdaBoosting.M1）的效果。

二 实验工具

编程语言 ：python
调用库 ：sklearn
数据集 ：WEBSPAM-UK2006

三 实验过程与结果分析

实验的完成情况如下，

- 完成了 SVM，Decision Tree 与 Bagging、AdaBoosting 的两两组合。
- 引入其他两种的分类器，KNN 和 Naïve Baynes
- 对数据进行归一化处理
- 对于处理数据的不均衡性，尝试着在训练时只取数量相等的正例和负例

实验参数如下

- 测试样例比例为 20%
- 交叉检验次数为 5
- Bagging 迭代次数为 10
- AdaBoosting 迭代次数为 10

结果如下：

Bagging 和 AdaBoosting 算法过程中都训练了多个分类器，最终综合各个分类器的结果进行输出。为了分析 Ensemble Learning 是否产生了作用，可以通过比较迭代过程中每一个 Hypothesis 的平均正确率和最后输出的正确率。表中结果为交叉检验后的平均结果。

Ensemble 策略	分类器	迭代平均正确率	最终正确率	提高比例
Bagging	SVM	72.0%	72.2%	0.3%
Bagging	DTree	84.2%	88.8%	5.5%
Bagging	KNN	72.8%	74.8%	2.7%

Bagging	SVM + Trun	60.4%	72.9%	20.7%
Bagging	DTree+Trun	81.9%	85.4%	4.3%
AdaBoosting	SVM	72.0%	72.1%	0.1%
AdaBoosting	DTree	84.0%	88.5%	5.4%
AdaBoosting	Baynes	75.1%	75.4%	0.4%
AdaBoosting	SVM + Norm	90.0%	90.2%	0.2%
AdaBoosting	Dtree + Norm	77.3%	86.6%	12.0%

注：表中 Norm 表示对数据进行归一化处理。Trun 表示截取训练集使得正例负例数量相等。

结果分析：

观察图表，总结后可以得出以下结论，

1. 分类正确率效果较好的策略组合为 SVM+Norm，分类正确率为 90.0%，但 Ensemble Learning 的提升效果不明显(正确率提升 0.2%)。
2. 在给定的数据集上，Bagging 策略和 AdaBoosting 策略对于 SVM 分类器的提升效果都不明显(Bagging 0.3%, AdaBoosting 0.1%)，但是对 Decision Tree 都有很大的提升作用(Bagging 5.5%，AdaBoosting 5.4%)。分析原因，可能是 SVM 分类器较为稳定，而 Decision Tree 可能较容易出现过拟合现象，相对而言较不稳定，而 Ensemble Learning 可以更好地提升不稳定分类器的性能。
3. 归一化操作，对于 SVM 分类器而言，能够十分明显地提高分类器的正确率，但是对于 Decision Tree 分类器没有那么明显的提升作用。
4. 对于数据的不均衡性，尝试训练中取数量相等的正例和负例。该尝试的分类效果不佳，迭代过程中的正确率就相比之前就降低了。分析原因可能是截取之后数据量变小，训练样例变少而影响了性能，以及可能增加了不稳定性，使得 Ensemble Learning 的效果很明显（这可以由 Bagging+SVM+Trun 策略中正确率提升了 20.0%推测）；而且对于正负例不均衡的测试样本，有偏的估计可能才能取得更好的分类效果。