机器学习 作业四

Hugo Zhang

1

随机森林算法如下:

Algorithm Random Forest Algorithm

Require:
训练数据集$D = (x_i, y_i)_{i=1}^N, x_i \in \mathcal{X} \subseteq R^d, y_i \in \mathcal{Y} \subseteq R, i = 1, 2, ..., N$;
选取特征个数k;
基分类器个数T

Ensure:
集成分类器$f(x)$

1: **for** each $t \in [1, T]$ **do**
2: 从D中有放回的随机抽取N个样本得到D_t
3: 从训练数据的d个特征中随机选择k个特征作为决策树的划分特征
4: 从D_t使用决策树算法在选中的k个维度上学得基分类器$f_t(x)$
5: **end for**
6: **return** 集成分类器$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{t=1}^{T} I(f_t(x) = y)$;

2

Bagging算法是将多个弱分类器并行训练，每个分类器的训练数据是通过从原数据自助采样得到的，最后按照投票方法得到集成分类器。随机森林是使用决策树作为基分类器的Bagging方法，其中还加入了一定的随机性，即对每个基分类器，随机选择k个特征而非全部特征进行训练。与Bagging方法相比，随机属性的选择添加了一定的随机扰动，有助于增加基分类器的多样性差异，进而可以提升其泛化能力；与此同时，由于特征数量减少，因此训练速度加快。一般而言，使用足够多的基分类器，我们期望随机森林方法能够比Bagging方法取得更好的效果。

使用sklearn对两个算法在分类任务上进行实验，随机森林使用默认超参数，即100个estimator，max feature取sqrt特征数，不限制max depth且min samples split取2；Bagging同样使用默认超参数，基分类器为决策树，使用10个estimator。使用sklearn内置的数据集，前四个是小数据集，最后用了一个稍微大一点的数据集，数据集的性质和最终结果如下（随机使用4/5训练，使用1/5测试，进行5次实验并取test accuracy均值）：

数据集	样本数	特征数	随机森林	Bagging
Iris	150	4	97.3%	96.0%
Digits	1797	64	94.7%	93.1%
Wine	178	13	96.7%	91.7%
Breast Cancer	569	30	95.4%	94.6%
Olivetti Faces	400	4096	76.2%	73.8%

这样的数据是符合预期的，随机森林引入了更多的随机性、提高训练速度从而可以使用更多的基分类器，理应取得更好的效果。