##### QG工作室数据挖掘小组实验报告

|  |  |  |
| --- | --- | --- |
| 实习生：李炜乐 | 导师：张润鹏 | 日期：2019 年 7 月23日 |

|  |
| --- |
| **实验名称：随机森林算法** |
| **已完成内容：**   1. **实现了ID3,C4.5,CART分类决策树的随机森林的算法** 2. **实现了CART决策树回归森林算法** |
| **未完成内容：暂无** |
| **未完成原因：暂无** |
| **需要帮助：暂无** |

|  |  |
| --- | --- |
| **实验总结** | |
| **知识点总结：**   1. **随机森林包括行取样、列取样，建树三个步骤。其中行取样就是去数据集的子集，列取样为随机选取k个属性。** 2. **ID3和C4.5和CART分类随机森林算哒相似的工作原理是通过行取样和列取样来取数据集，通过决策树建树算法实现多颗决策树，并且储存到列表中，然后进行对测试集的测试，通过每一棵决策树的最优值选取最多的取值作为结果。** 3. **ID3和C4.5所使用数据集为lenses数据集，CART数据集为adult数据集。ID3的正确率为82.3%左右， C4.5正确率为78.2%左右，CART正确率约为81.5%左右。ID3运行时间为3.15s，C4.5运行时间为4.35s，CART运行时间为112.2s（因为数据集较多）。** 4. **CART回归森林算法跟其他森林的相似之处是行取样和列取样都相同，对于建树的过程，采取了使用CART树回归，预测结果为全部树的预测的均值。** 5. **CART回归森林的数据集属于深林火灾数据集，该数据集在CART上的拟合效果并不是很好，所以在CART回归森林的拟合效果也不是很出众。** 6. **CART的数据集为森林火灾数据集，预测出来均误差比为-33（预测值减实际值除于实际），运行时间为143.3s左右** | |
| **遇到问题**：对于回归森林并不能很好的预测出很好的预测值，预测值于实际值相差较大 | **解决过程：对于其他数据集拟合性能挺好的，这个数据集的结果较多为0值，预测可能一定程度会影响了数据的预测** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **导师评价** | | | | |
| **实验分数** | **知识掌握情况** | **代码编写能力** | **建议** | **评价日期** |
|  |  |  |  |  |