华为－软件所合作项目研究问题

Lijie Xu

2016-12-05

v0.1

（内部科研资料，切勿外传）

**算法稳定性研究问题**

1. **收敛结果的正确性**

算法是否收敛？收敛结果是否正确，如何判定？全局最优还是局部最优？

1. **收敛结果的确定性**

在同样的数据和参数下多次运行，收敛结果是否一致？ 改变数据顺序或一些参数（如样本顺序、分块大小、partition个数等）后，收敛结果是否一致？如果不一致，原因是什么？有哪些除随机数之外的非确定性因素影响一致性？

1. **收敛过程的稳定性**

算法运行过程中的收敛趋势是什么样的（比如每次迭代后的算法结果变化情况，要有图表）？ 如果收敛趋势有波动，那么波动的原因是什么？

1. **收敛结果对比**

同样算法、同样数据在Spark、Flink、Scikit-learn上的收敛结果是否有不同？不同的原因是什么？

1. **稳定性的实证分析**

在Spark、Flink、Scikit-learn的 issues中，有哪些算法稳定性相关的bugs？比如研究LogisticRegression算法，那么去Spark MLlib/ML issues (<https://issues.apache.org/jira/browse/spark/?selectedTab=com.atlassian.jira.jira-projects-plugin:issues-panel>) 中查找LogisticRegression算法相关的Bug/Improvement issues，制作成表格列出来，然后分析是否该bug/issues是否与算法的稳定性相关，如果是，需要简要描述该bug/improvement。Flink的issues在<https://issues.apache.org/jira/browse/FLINK/?selectedTab=com.atlassian.jira.jira-projects-plugin:issues-panel>。Scikit-learn的issues在<https://github.com/scikit-learn/scikit-learn/issues>，及<http://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf> 中的Release history。

**应用可靠性研究问题**

可靠性问题是指应用在运行中出现错误 (比如IOException，OutOfMemory，执行时间过长等failures)。具体的研究问题包括：

1. **可靠性实证分析**

在Spark、Flink 的 issues中，有哪些应用可靠性相关的bugs？比如研究LogisticRegression算法，那么去 Spark MLlib/ML issues 中查找LogisticRegression算法相关的Bug/Improvement issues，制作成表格列出来，然后分析是否该bug/issues是否与可靠性相关，如果是，需要简要描述该bug/improvement。Flink的issues和Scikit-learn的issues类似。

1. **异常数据生成**

什么样的异常数据可以导致应用运行时出现错误？我们的基本想法是“分析程序特征=>利用特征来生成异常数据=>导致应用运行出错”。可以考虑的异常数据特征包括：数据量、数据倾斜、维度过高、分布异常、数据稀疏等。

1. **参数组合测试**

应用在什么样的参数配置下会出现运行时错误？参数包含应用参数（如树的高度）也包含系统参数（如reducer个数），那么这些参数之间是否独立？是否与应用执行时间和资源消耗正相关或负相关？配置参数组合空间很大，如何削减测试空间？

**应用扩展性研究问题**

扩展性研究应用是否能够线性扩展。

与可扩展性相关的系统参数包括：input split size, partition number, partition function等。与可扩展性相关的资源参数包括：Executor number, executor cores/memory等。可扩展性度量指标包括：应用执行时间（如收敛速度），应用资源消耗等。具体问题包括：

1. **应用是否能够线性扩展？**
2. 其他条件不变，只增加数据量，应用执行时间或资源消耗是否线性增加？
3. 数据量不变，增加并行度（减少input split size, 增加partition number）且保证资源够用（executor充足），应用执行时间或资源消耗是否线性减少？

需要对应用的各个jobs/stages的执行时间和资源消耗做细粒度的统计分析。

1. **如果不能达到线性扩展能力（或者线性加速），那么瓶颈在哪里？**需要对应用的各个jobs/stages的执行时间和资源消耗做细粒度的统计分析。

**GC算法研究问题**

每次跑应用，需要统计GC占用的时间（最大、最小、平均），便于开展GC算法方面的研究。