500 переменных), а также на данных спектрометрической съемки в инфракрасном диапазоне (1955731 образцов, 148 переменных), произведенной вдоль русла одной из рек Норвегии. Испытания производились на арендованном у компании Amazon кластере Hadoop, состоящем из 10 вычислительных узлов. Алгоритм показал приемлемое ускорение, как в синтетическом тесте, так и на реальных данных.

Библиографический список

1. Dean J., Ghemawat S. Simplified data processing on large clusters // Operating Systems Design and Implementation. – 2004. – P. 137–149.

2. Apache Hadoop. [Электронный ресурс] Режим доступа – http://hadoop.apache.org.

3. Amazon Elastic MapReduce. [Электронный ресурс] Режим доступа – http://aws.amazon.com/elasticmapreduce/.

4. Дрейпер Н. Прикладной регрессионный анализ. – М.: Финансы и статистика, 1987. – 717 с.

5. Apache Commons Math. [Электронный ресурс] Режим доступа – http://commons.apache.org/math/.

Распределенные алгоритмы построения интервальной регрессии

В.Д. Пятков, С.И. Жилин
АлтГУ, Барнаул

Метод построения регрессионной зависимости по экспериментальным данным при интервальной ошибке в выходной переменной [1] довольно широко используется в практике эмпирического моделирования и для краткости именуется *интервальной регрессией* (ИР). Суть метода сводится к оцениванию множества допустимых значений параметров регрессии, совместных как с используемой моделью регрессии, так и с набором ограничений, вытекающих из интервального характера ошибки наблюдения выходной переменной.

При построении наиболее употребимой линейной по параметрам регрессии задача оценивания множества допустимых значений параметров сводится к решению нескольких задач линейного программирования. Поэтому вычислительная сложность алгоритма построения ИР определяется эффективностью методов линейного программирования: при увеличении числа наблюдений время построения модели, как минимум, растет полиномиально. При обработке большого объема данных распределение процесса построения модели ИР на несколько

вычислительных узлов может оказаться единственным способом решить задачу за обозримое время.

Настоящая работа направлена на поиск способов распределенного решения задачи построения ИР. Изучены следующие возможности распределенного решения задачи построения ИР.

1. *Использование параллельных методов линейного программирования.* В частности, при использовании симплекс-метода ускорение может быть достигнуто за счет распараллеливания обхода многогранника, являющегося множеством решений задачи построения ИР.

2. *Параллельное решение независимых задач оптимизации, возникающих при построении ИР.* Каждый из вычислительных узлов занимается решением только своего подмножества задач оптимизации, а итоговая модель получается агрегацией решений частных задач на одном из вычислителей.

3. *Распараллеливание по данным с применением концепции* MapReduce [2, 3]. Входной набор наблюдений дробится на порции, для каждой из которых на отдельном вычислительном узле решается подзадача построения модели ИР. Далее модели, построенные на каждом вычислителе, агрегируются в общую результирующую модель (соответствующую полной совокупности наблюдений). За счет отсеивания на каждом из вычислительных узлов неинформативных наблюдений, размерность итоговой задачи оптимизации существенно сокращается.

Результаты численных экспериментов свидетельствуют о целесообразности использования приемов 2 и 3 при обработке наборов данных с количеством наблюдений, достаточным для того, чтобы сформировать для вычислительных узлов подзадачи, время решения которых превышает накладные расходы на распараллеливание. Выигрыш же от использования параллельного симплекс-метода оказывается незначительным.

Библиографический список

1. Оскорбин Н.М., Максимов А.В., Жилин С.И. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия Алтайского госуниверситета. – 1998. – №1 (5). – С. 37–40.

2. Ерохин Г.Н., Камышников А.И., Оскорбин Н.М. Обработка больших баз данных методами линейного программирования // Управление, математическое моделирование и оптимизация на базе ПЭВМ: Меж-вуз. сб. науч. работ. – Барнаул: АГУ, 1993. – С. 143–147.

3. Dean J., Ghemawat S. Simplified data processing on large clusters // Operating Systems Design and Implementation. – 2004. – P. 137–149.