ТЕСТИРОВАНИЕ ВОСПРОИЗВОДИМОСТИ ВЫЧИСЛЕНИЙ В ОБЛАЧНОЙ И РЕАЛЬНОЙ КЛАСТЕРНЫХ СРЕДАХ

Т.А. БАРДАДЫМ, С.П. ОСИПЕНКО
Институт кибернетики имени В.М. Глушкова
НАН Украины, Киев, Украина
tamara.bardadym@gmail.com, baston888@gmail.com

Аннотация. *Описаны результаты численного эксперимента по исследованию воспроизводимости биомедицинских вычислений, проведенных с помощью облачного сервиса OpenStack и реального кластера СКИТ-4.5.*
Ключевые слова: *облачные технологии, воспроизводимые вычисления, платформа кластерная.*

Создание новых технологий обработки данных побуждает исследователей к рациональному выбору сред и средств проведения вычислений. При этом следует заранее осознавать преимущества тех или иных вариантов и возможности переноса вычислений из одной среды в другую. Особое внимание при этом следует обращать на воспроизводимость вычислений независимо от того, где именно проведены расчеты. Ранее авторы получили опыт использования контейнерной технологии Docker для обеспечения воспроизводимости при использовании данных The Cancer Genome Atlas (TCGA) [1]. Следующим шагом является изучение возможностей параллельных вычислений и выбора подходящих технологических подходов. Действительно, поиск информативных признаков на массиве данных генной экспрессии большой размерности на обычном компьютере иногда длился более десяти часов. Для использования параллельных вычислений надо иметь возможность обеспечить все условия, которые способны поддерживать воспроизводимость таких расчетов на разных платформах. В сравниваемых в этой работе тестовых вычислительных средах используются современные технологии, которые позволяют достичь этой цели. Одна из сред построена с помощью программной системы с открытым кодом OpenStack [2], которая широко используется в мире для создания публичных и частных облачных сервисов, в том

числе научных. Созданная среда представляет собой виртуальный кластер, то есть совокупность виртуальных машин, объединенных в виртуальную сеть, на которых установлено соответствующее программное обеспечение, включая менеджер ресурсов Slurm. Кластер состоит из 4-х вычислительных и одного управляющего узла.

Для численных экспериментов была выбрана задача кросс-валидации модели линейной классификации, построенной с использованием негладкой оптимизации (с применением модуля NonSmoothLC, который был разработан Ю.П. Лаптиным [3] на основе методов негладкой оптимизации [4-5]). Эта задача была выбрана потому, что ее решение требует привлечения большого количества вычислительных ресурсов, а также потому, что позволяет применить модель масштабирования, где отдельные подзадачи могут выполняться независимо друг от друга.

Кросс-валидация классификационной модели была проведена на биомедицинских данных по экспрессии генов. В модель вошли 249 наиболее информативных показателей генной экспрессии, предварительно отобранных из множества 20 тыс. Количество наблюдений равно 152. Расчеты осуществлялись в программной среде R (версия 4.1.1) со всеми необходимыми для проведения кросс-валидации пакетами (включая NonSmoothLC), расположенными в контейнеризированом приложении Singularity. Контейнеризация избавила от необходимости инсталляции среды R и ее библиотек непосредственно на кластере (в случае использования реального кластера для обычного пользователя это было бы невозможно). Кроме того, контейнеризация обеспечила одно из условий портабельности вычислений между виртуальными и реальным кластерными средами. Данные и программный код кросс-валидации были расположены вне контейнера, что позволило использовать адаптированную к кластерным вычислениям версию кода.

Для обеспечения масштабирования вычислений была использована технология MPI. Она применялась для распределения задач между параллельными процессами (контейнерами с R) с использованием языка Python. Приложение на Python запускалось как несколько параллельных процессов, которые, в свою очередь, запускали экземпляры контейнеров для выполнения вычислений, передавая им в качестве параметров индексы наборов данных, на

которых необходимо провести кросс-валидации и другую необходимую информацию. Вычисления проводились с помощью контейнеризированных приложений Singularity. Результаты анализа хранились в виде файлов данных R. Агрегация полученных результатов кросс-валидации проводилась отдельным не параллельным процессом.

Под воспроизводимостью в контексте задачи кросс-валидации понимается возможность получить одни и те же результаты как при изменении параметров масштабирования, так и при проведении вычислений на разных платформах. Проверка была выполнена следующим образом. Результатом кросс-валидации являются показатели: надежности полученной модели (коэффициент F1 Соренсена) [7], величина промежутка между классами и коэффициенты полученной модели (249 коэф. модели + 1 свободный член = 250). Всего 252 переменные, которые проверялись на воспроизводимость для 36 (вариантов данных) * 36 (вариантов количества параллельных процессов) * 5 (серий) = 6480 вариантов вычислений в случае СКИТ-4.5 и 36 * 4 * 5 = 720 вариантов для OpenStack. Фактически проверялась матрица результатов кросс-валидации размером 252 колонок на 7200 строк. Процедура анализа этой матрицы заключалась в подсчете стандартных отклонений для каждого из 252 показателей по каждому из 36-ти вариантов данных, использованных при кросс-валидации (как для каждой среды отдельно, так и для двух сред вместе). В случае абсолютного совпадения результатов суммарное стандартное отклонение равнялось бы нулю. В реальности мы получили среднее значение стандартного отклонения для переменных равным 3.0e-07. Максимальное стандартное отклонение было равно 7.5e-05, то есть на воспроизводимость вычислений не повлияло ни их масштабирование, ни перенос расчетов на другую кластерную среду.

Примерно одинаковыми были также показатели нагрузки и использования памяти на этих разновидностях кластеров. Впрочем, следует отметить, что здесь на результат могли повлиять особенности задачи кросс-валидации, которая позволяет применить модель масштабирования, где отдельные подзадачи могут выполняться независимо друг от друга.

Выводы. Проведенное исследование продемонстрировало возможность развертывания на приватном сервере среды OpenStack и создания виртуального кластера Slurm. Тестовые вычисления, проведенные на реальных данных с помощью виртуального кластера и с помощью реального кластера Института кибернетики СКИТ-4.5 продемонстрировали обеспечение портабельности и воспроизводимости результатов. Подробная информация о создании виртуального кластера и сравнении вычислений на ней с вычислениями на реальном кластере СКИТ-4.5 опубликована в [6].

Исследование выполнено при поддержке Национальной академии наук Украины (тема ВФ.115.41).

Литература

1. Bardadym T.O., Gorbachuk V.M., Novoselova N.A., Osypenko S.P., Skobtsov V.Yu., Intelligent analytical system as a tool to ensure the reproducibility of biomedical calculations // Artificial Intelligence. – 2020. – № 3. – P. 65–78.
2. Sefraoui O., Aissaoui M., Eleuldj M. OpenStack: toward an open-source solution for cloud computing. International Journal of Computer Applications. 2012. **55 (**3). P. 38–42.
3. Zhuravlev Y.I., Laptin Y.P. et al. Linear classifiers and selection of informative features // Pattern Recognition and Image Analysis. – 2017. – Т. 27. – № 3. – С. 426–432.
4. Шор Н. З., Журбенко Н. Г. Метод минимизации, использующий операцию растяжения пространства в направлении разности двух последовательных градиентов. Кибернетика. – 1971. – № 3. – С. 51–59.
5. Шор Н.З. Методы минимизации недифференцируемых функций и их приложения. – К.: Наук. думка, 1979. – 199 с.
6. Бардадим Т.О., Лефтеров О.В., Осипенко С.П. Досвід тестового розгортання OpenStack і порівняння віртуального та реального кластерних середовищ // Кібернетика та комп’ютерні технології. – 2021. – № 3. – С. 74 – 85. https://doi.org/10.34229/2707-451X.21.3.0
7. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter. 1948. 5. P. 1–34.