большой выбор самых разных моделей искусственных нейронных сетей. Для решения данной задачи, предположительно, лучше всего подходят такие модели как CNN (Сверточные нейронные сети) и LSTM (Долгая краткосрочная память). Их часто применяю в задачах классификации, обладают достаточной скоростью обработки и имеют высокую точность, что полностью удовлетворяет нашим потребностям. Но эти алгоритмы являются сложными по построению, поэтому как альтернативу возможно использовать Наивный байесовский классификатор, который имеет более простую архитектуру, но при этом тоже достаточно хорошо справится с нашей задачей.

Нейросетевое моделирование для автоматической настройки задач в системах обработки больших данных

Бобряков А.С.
Научный руководитель — доцент, к.ф-м.н. Осипова В.А.
МАИ, Москва

В настоящее время с увеличением объема хранимых данных возникает потребность эти данные обрабатывать для получения статистик и новой информации, необходимой для дальнейшего анализа. Для этого применяют системы обработки больших данных: Hadoop, Spark и другие.

Основной проблемой использования таких систем является их предварительная настройка. Для каждой выполняемой задачи необходимо минимизировать целевые затраты: время выполнения, используемую память, другие целевые критерии, присущие конкретной области. Для этого необходимо подобрать оптимальные значения параметров, которых предоставляется несколько сотен [1].

Большинство таких параметров настраиваются по предоставленному руководству один раз администратором, но рекомендованные значения могут негативно влиять на конкретную выполняемую задачу. Также обычно не учитывается топология архитектуры кластеров, текущая нагрузка, обязательные параллельные задачи, которые запускаются самой системой обработки больших данных (например, удаление мусора, перераспределение данных между узлами, очистка буферов и др.). Важно понимать, что большинство параметров непредсказуемо взаимосвязаны, например, при увеличении значения одного параметра можно получить и положительное, и отрицательное влияние на систему другого параметра. Чтобы не задумываться обо всех смежных процессах, система рассмотрена в виде “черного ящика”, который моделировался нейронной сетью.

В качестве архитектуры нейронной сети использовали три полносвязных слоя с оптимизатором Adam [2] на объеме данных в 1млн записей, которые отражают статистики выполнения одной map-reduce задачи с варьированием параметров конфигурации кластера, параметров системы Hadoop и присущих конкретной задачи характеристик. Это позволило получить модель для определения длительности выполнения задачи. Такую обученную нейросеть предполагается встроить в процессы запуска с целью предварительной оптимизации желаемого времени выполнения, что позволит целесообразно использовать доступные ресурсы в условиях высокой нагрузки на кластер.

Список использованных источников:

1. Muhammad Bilal and Marco Canini. 2017. Towards automatic parameter tuning of stream processing systems. In Proceedings of the 8th ACM Symposium on Cloud Computing (SoCC’17). ACM, 189–200.

2. Diederik P. Kingma, Jimmy Ba Adam: A Method for Stochastic Optimizatiot [Электронный ресурс] // arXiv.org. 2015. URL: https://arxiv.org/abs/1412.6980