УСКОРЕННЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Аксак Н.Г., Соколец Е.В.

Харьковский национальный университет радиоэлектроники

Колоссальные потоки информации в разных областях научной и производственной деятельности потекли благодаря совершенствованию технологий записи и хранения данных. Функционирование любой организации (научно-исследовательской, производственной, врачебной, коммерческой и т.д.) включает, кроме основных действий, регистрацию и запись всех тонкостей ее деятельности. Почти все компании имеют собственные страницы в социальных сетях, а анализ поступающих комментариев и сообщений является важной составляющей для их развития.

В качестве примера можно выделить анализ полученных о компании сообщений в Twitter, которые можно классифицировать как отзывы положительные, отрицательные и нейтральные, а затем с течением времени анализировать ситуацию.

Важно оптимизировать процесс получения прибыли при размещении информации на сайте и провести анализ - когда, где и какую рекламу следует располагать. Для этого определяются популярные клики, время посещения, глубина просмотра сайта и т.п. Подобные задачи возникают в сферах предложения услуг.

В связи с этим появляется множество задач анализа данных, одна из которых - кластеризация пользователей (например, категории пользователей медицинского Web-ресурса – врачи или пациенты). В таком случае для каждого кластера вырабатывается оптимальная стратегия, позволяющая учитывать предпочтения пользователей, автоматически настраивать контент, ускорять процесс поиска и повышать заинтересованность пользователя в посещаемости данного сайта.

В задаче классификации, например, требуется определить, воспользуется ли клиент услугами, предлагаемыми в рассылке. Часть признаков описывает клиентов: пол, идентификационный номер, регион. Часть – специфику: услуги\товары: стоимость, скидка, категория и т.д. Оставшиеся – поведение клиента: сколько рассылки ему делалось, сколькими услугами он воспользовался и т.д.

Поскольку «ручная» обработка и анализ информации при огромных объемах не представляется возможным, особое значение приобретает как разработка быстрых и точных методов их обработки и анализа, так и выбор эффективного инструментария, позволяющего в режиме реального времени производить автоматический сбор и анализ данных.

Карты Кохонена позволяют решать большой спектр задач, связанных с визуализацией, классификацией, кластеризацией данных, а также с задачами распознавания образов.

Для кластеризации пользователей Web-ресурсов предлагается адаптировать самоорганизующиеся карты Кохонена под SMP системы.

Пусть $X=\{X^1,...,X^N\}$ - множество рассматриваемых образцов $X^c=\left(x_1^c,...,x_n^c\right)^T$, $c=\overline{1,N}$, N - количество образцов. Сеть состоит из одного слоя, имеет n входных нейронов, соответствующих координатам рассматриваемых образцов, и s^2 выходных нейронов, представляющих собой квадратную решетку размером $s\times s$.

Параллельная реализация на системах с общей памятью, включающих p вычислителей, основана на одновременной работе максимально возможного количества нейронов в одной группе $G_{max}(p,s)$.

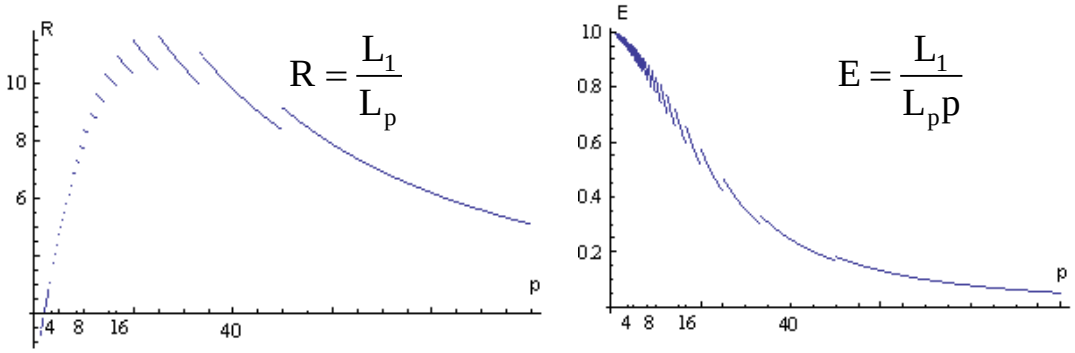
Количество операций последовательного алгоритма обучения нейронной сети выражается следующим соотношением

$$L_1=6+T\left(8+N\left(2+21s^2n+3\sum_{i=2}^{s^2}\frac{1}{i}\right)\right),$$

а параллельного, соответственно, выражением

$$L_p=6+T\left(8+N\left(2+21G_{max}(p,s)n+3\sum_{i=2}^{G_{max}(p,s)}\frac{1}{i}+3\sum_{i=2}^{p}\frac{1}{i}+(3n+2)p\right)\right).$$

На рисунке 1 приведены графики ускорения и эффективности параллельного алгоритма обучения нейронной сети.



(а) (б)

Рисунок 1. Графики ускорения (а) и эффективности (б)

Экспериментальный анализ показал, что увеличение числа вычислителей обеспечивает соответствующее снижение времени выполнения только до определенного значения. Целесообразно выбирать значения p кратные s^2, так как в этих точках наблюдается разрыв функций ускорения и эффективности

Таким образом, правильная конфигурация вычислительной системы зависит не только от операции, которая будет выполнена, но и от количества входных данных. Необходимо учитывать баланс между объемом обрабатываемых данных и количеством вычислителей для достижения наилучших результатов.