УДК 519.95

ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПОДБОРА ВЕСОВЫХ КОЭФФИЦИЕНТОВ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ С ПОМОЩЬЮ КВАЗИНЬЮТОНОВСКИХ МЕТОДОВ, ИСПОЛЬЗУЮЩИХ ФОРМУЛУ БРОЙДЕНА–ФЛЕТЧЕРА–ГОЛЬДФАРДА–ШЕННО

© О.В. Крючин, Е.В. Вязовова, А.А. Арзамасцев

Ключевые слова: искусственные нейронные сети; параллельные вычисления; квазиньютоновские методы; информационные ресурсы; повышение эффективности; аналитические выражения.

Описан параллельный квазиньютоновский метод, использующий формулу Бройдена–Флетчера–Гольдфарда–Шенно. Приводятся аналитические выражения, позволяющие оценить эффективность предложенных алгоритмов, и вычислительные эксперименты, проведенные для проверки этих выражений.

Введение. В настоящее время существуют различные методы подбора весовых коэффициентов искусственной нейронной сети (ИНС), наиболее распространенными из которых являются градиентные [1], базирующиеся на разложении целевой функции в ряд Тейлора. Суть этих методов заключается в вычислении градиента:

$$\nabla\varepsilon = \frac{\partial\varepsilon}{\partial w} = \left(\frac{\partial\varepsilon(w_0)}{\partial w_0}, \frac{\partial\varepsilon(w_1)}{\partial w_1}, \frac{\partial\varepsilon(w_2)}{\partial w_2}, \ldots, \frac{\partial\varepsilon(w_{l_w}-1)}{\partial w_{l_w}-1} \right) =$$

$$= \left(\frac{\varepsilon(w_0^{(I)} + \Delta w_0^{(I-1)}) - \varepsilon(w_0^{(I)})}{\Delta w_0^{(I-1)}}, \frac{\varepsilon(w_1^{(I)} + \Delta w_1^{(I-1)}) - \varepsilon(w_1^{(I)})}{\Delta w_1^{(I-1)}}, \ldots, \frac{\varepsilon(w_{l_w-1}^{(I)} + \Delta w_{l_w-1}^{(I-1)}) - \varepsilon(w_{l_w-1}^{(I)})}{\Delta w_{l_w-1}^{(I-1)}} \right) \quad (1)$$

и изменении весовых коэффициентов $\vec{w}$ в противоположном направлении. Здесь ε – невязка, вычисляемая целевой функцией:

$$\varepsilon = \varepsilon(\vec{w}), \quad (2)$$

$\vec{w}$ – весовые коэффициенты (веса); l_w – число весов; I – номер итерации [2, 3].

Следовательно, для вычисления одного элемента градиента $\nabla\varepsilon_i^{(I)}$ необходимо вычислить невязку при текущем значении весовых коэффициентов, а затем при измененном. После вычисления градиента происходит изменение весовых коэффициентов согласно формуле:

$$\vec{w}^{(I+1)} = \vec{w}^{(I)} - \vec{s}^{(I)}\nabla\varepsilon^{(I)}, \quad (3)$$

где $\vec{s}^{(I)}$ – коэффициент обучения на I -й итерации.

Изменение весовых коэффициентов можно вычислить как

$$\Delta\vec{w}^{(I)} = \vec{w}^{(I)} - \vec{w}^{(I-1)}. \quad (4)$$

Одним из частных случаев градиентных методов являются методы переменной метрики, известные также как квазиньютоновские, которые предполагают, что функция (2) аппроксимируется как квадратичная в области оптимума, и используют не только первую, но и вторую производную (гессиан) для поиска решения. В квазиньютоновских методах не требуется рассчитывать гессиан через вторую производную минимизируемой функции, поскольку эта матрица вычисляется на основе градиента (1) [2].

Согласно квазиньютоновским методам, значение вектора весовых коэффициентов можно вычислить по формуле:

$$\vec{w}^{(I)} = -\left[H\left(\vec{w}^{(I-1)}\right)\right]^{-1}\nabla\varepsilon^{(I-1)}, \quad (5)$$

где $H\left(\vec{w}^{(I-1)}\right)$ – гессиан, элементы которого вычисляются по формуле:

$$H\left(\vec{w}^{(I)}\right) = \begin{bmatrix} \frac{\partial^2\varepsilon}{\partial w_0^{(I)}\partial w_0^{(I)}} & \cdots & \frac{\partial^2\varepsilon}{\partial w_0^{(I)}\partial w_{l_w}^{(I)}-1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2\varepsilon}{\partial w_{l_w}^{(I)}-1\partial w_0^{(I)}} & \cdots & \frac{\partial^2\varepsilon}{\partial w_{l_w}^{(I)}-1\partial w_{l_w}^{(I)}-1} \end{bmatrix}. \quad (6)$$

Формула (5) представляет собой основу ньютоновского алгоритма оптимизации и является теоретическим выражением, т. к. требует положительной определенности гессиана на каждом шаге. Поскольку практически это неосуществимо, то гессиан $H(\vec{w})$ заменяют на его приближение $\tilde{H}$. На каждом шаге гессиан и обратная ему величина, которая получена на предыдущем шаге, модифицируются на величину некоторой поправки [4–5].

Целью данной работы является повышение эффективности квазиньютоновских методов подбора весовых коэффициентов при помощи параллельных вычислений, выведение аналитических выражений, показывающих значение эффективности разработанного алгоритма, и проведение вычислительных экспериментов для подтверждения этого значения.

Формула Бройдена–Флетчера–Гольдфарда–Шенно. В соответствии с формулой Бройдена–Флетчера–Гольдфарда–Шенно (Broyden–Fletcher–Goldfarb–Shanno) процесс уточнения значения матрицы $\tilde{H}$ можно описать рекуррентной зависимостью:

$$\tilde{H}^{(I+1)} = \tilde{H}^{(I)} + C(H1)^{(I)} - C(H2)^{(I)}; \qquad (7)$$

$$C(H1)^{(I)} = \frac{\left(\begin{array}{l}\left[\Delta\vec{w}^{(I)}\right]^T \Delta\nabla\varepsilon^{(I)} + \\ + \left[\Delta\nabla\varepsilon^{(I)}\right]^T \tilde{H}^{(I)} \Delta\nabla\varepsilon^{(I)}\end{array}\right)\left(\Delta\vec{w}^{(I)}\left[\Delta\vec{w}^{(I)}\right]^T\right)}{\left(\left[\Delta\vec{w}^{(I)}\right]^T \Delta\nabla\varepsilon^{(I)}\right)^2}; \qquad (8)$$

$$C(H2)^{(I)} = \frac{\tilde{H}^{(I)}\Delta\nabla\varepsilon^{(I)}\left[\Delta\vec{w}^{(I)}\right]^T + \Delta\vec{w}^{(I)}\left[\Delta\nabla\varepsilon^{(I)}\right]^T \tilde{H}^{(I)}}{\left[\Delta\vec{w}^{(I)}\right]^T \Delta\nabla\varepsilon^{(I)}}. \qquad (9)$$

Обозначим произведение изменения вектора весовых коэффициентов $\Delta\vec{w}$ на транспонированное изменение вектора весовых коэффициентов $\left[\Delta\vec{w}\right]^T$ как $\tilde{w}$:

$$\tilde{w}^{(I)} = \Delta\vec{w}^{(I)}\left[\Delta\vec{w}^{(I)}\right]^T = \begin{pmatrix} \Delta w_0^{(I)} \\ \vdots \\ \Delta w_{l_w-1}^{(I)} \end{pmatrix}\left(\Delta w_0^{(I)},\ldots,\Delta w_{l_w-1}^{(I)}\right) = \qquad (10)$$

$$= \begin{pmatrix} \left(\Delta w_0^{(I)}\right)^2 & \cdots & \Delta w_0^{(I)}\Delta w_{l_w-1}^{(I)} \\ \vdots & \ddots & \vdots \\ \Delta w_{l_w-1}^{(I)}\Delta w_0^{(I)} & \cdots & \left(\Delta w_{l_w-1}^{(I)}\right)^2 \end{pmatrix}$$

произведение транспонированного вектора приращения весовых коэффициентов $\left[\Delta\vec{w}\right]^T$ на вектор приращения градиента $\Delta\nabla\varepsilon$ как $C(H3)$:

$$C(H3)^{(I)} = \left[\Delta\vec{w}^{(I)}\right]^T \Delta\nabla\varepsilon^{(I)} = \left(\Delta w_0^{(I)},\ldots,\Delta w_{l_w-1}^{(I)}\right)\begin{pmatrix} \Delta\nabla\varepsilon_0^{(I)} \\ \vdots \\ \Delta\nabla\varepsilon_{l_w-1}^{(I)} \end{pmatrix} = \qquad (11)$$

$$= \sum_{i=0}^{l_w-1}\left(\Delta w_i^{(I)}\Delta\nabla\varepsilon_i^{(I)}\right);$$

а произведение транспонированного вектора приращения градиента $\left[\Delta\nabla\varepsilon\right]^T$ на матрицу $\tilde{H}$ как $\left[\vec{C}(H4)\right]^T$:

$$\left[\vec{C}(H4)^{(I)}\right]^T = \left[\Delta\nabla\varepsilon^{(I)}\right]^T \tilde{H}^{(I)} = \left(\Delta\nabla\varepsilon_0^{(I)},\ldots,\Delta\nabla_{l_w-1}^{(I)}\right)\times$$
$$\times\begin{pmatrix} \tilde{H}_{0,0}^{(I)} & \cdots & \tilde{H}_{0,l_w-1}^{(I)} \\ \vdots & \ddots & \vdots \\ \tilde{H}_{l_w-1,0}^{(I)} & \cdots & \tilde{H}_{l_w-1,l_w-1}^{(I)} \end{pmatrix} = \qquad (12)$$
$$= \left(\sum_{i=0}^{l_w-1}\Delta\nabla\varepsilon_i^{(I)}\tilde{H}_{i,0}^{(I)},\ldots,\sum_{i=0}^{l_w-1}\Delta\nabla\varepsilon_i^{(I)}\tilde{H}_{i,l_w-1}^{(I)}\right).$$

В этом случае формула (7) может быть записана как

$$\tilde{H}^{(I+1)} = \tilde{H}^{(I)} + \frac{\left(C(H3)^{(I)} + \left[\vec{C}(H4)^{(I)}\right]^T \Delta\nabla\varepsilon^{(I)}\right)\tilde{w}^{(I)}}{\left(C(H3)^{(I)}\right)^2} - \frac{\tilde{H}^{(I)}\Delta\nabla\varepsilon^{(I)}\left[\Delta\vec{w}^{(I)}\right]^T + \Delta\vec{w}^{(I)}\left[\vec{C}(H4)^{(I)}\right]^T}{C(H3)^{(I)}}. \qquad (13)$$

Подбор весовых коэффициентов при использовании большого числа информационных ресурсов классическим градиентным алгоритмом. Пусть имеется *n* элементов информационных ресурсов (ИР-элементов), имеющих распределенную память, пронумерованных от 0 до *n* – 1. Первый ИР-элемент называется управляющим, поскольку отвечает за синхронизацию вычислений. В качестве такого информационного ресурса может быть использована кластерная система или вычислительная сеть.

Как можно заметить, для вычисления нового значения одного из весовых коэффициентов $\vec{w}^{(I+1)}$ градиентные алгоритмы используют значения остальных

весовых коэффициентов, которые были на предыдущей итерации $\vec{w}^{(I)}$. Исходя из этого, можно сделать вывод, что элементы градиента $\nabla\varepsilon^{(I)}$ могут быть вычислены одновременно, следовательно, этот вектор может быть разделен на n частей, каждая из которых вычисляется отдельным ИР-элементом, которому для этого необходимо лишь передать текущие значения весовых коэффициентов $\vec{w}^{(I)}$. После окончания вычисления ИР-элементы не возвращают полученные результаты на управляющий, а изменяют значения приписанных к ним весовых коэффициентов, а уже после этого возвращают результат (новые весовые коэффициенты). Следовательно, каждый ИР-элемент, за исключением управляющего, вычисляет

$$\hat{l}_w = \begin{cases} \dfrac{l_w}{n}, & l_w \mod n = 0; \\ \dfrac{l_w}{n-1}, & l_w \mod n \neq 0 \end{cases} \qquad (14)$$

весовых коэффициентов (и соответствующих им элементов градиента), а ведущий вычисляет

$$\hat{N}_w = \begin{cases} \hat{l}_w, & l_w \mod n = 0; \\ l_w - \hat{l}_w(n-1), & l_w \mod n \neq 0 \end{cases} \qquad (15)$$

элементов, где n – число используемых ИР-элементов; l_w – количество весовых коэффициентов [1] (рис. 1).

Таким образом, на каждой итерации происходит только две передачи данных – всех весовых коэффициентов на все ИР-элементы и части из них со всех ИР-элементов на управляющий [1, 6].

Подбор весовых коэффициентов модифицированным квазиньютовским методом. Модифицированный для использования большого числа ИР-элементов квазиньютоновский алгоритм может быть реализован следующим способом (рис. 2).

1. Поскольку первые шаги, вычисляющие значения $\nabla\varepsilon^{(I)}$, $\Delta\nabla\varepsilon^{(I)}$ и $\tilde{w}^{(I)}$, нуждаются только в значениях, полученных на предыдущих шагах, то каждый ИР-элемент может вычислять часть значений этих векторов и строк матриц. После этого неуправляющие ИР-элементы передают на управляющий вычисленные ими части изменения градиента $\Delta\nabla\varepsilon^{(I)}$ и матрицы $\tilde{w}^{(I)}$ (значения градиента $\nabla\varepsilon^{(I)}$ не требуются управляющему ИР-элементу).

2. Управляющий ИР-элемент рассылает новые значения вектора $\Delta\nabla\varepsilon^{(I)}$ на все прочие (матрица $\tilde{w}^{(I)}$ необходима только управляющему ИР-элементу).

Каждый ИР-элемент, за исключением управляющего, вычисляет соответствующие строки матрицы

$$\begin{aligned} C\left(\tilde{H}_{B0}\right)^{(I)} &= \Delta\nabla\varepsilon^{(I)}\left[\Delta\vec{w}^{(I)}\right]^T = \\ &= \begin{pmatrix} \Delta\nabla\varepsilon_0^{(I)} \\ \vdots \\ \Delta\nabla\varepsilon_{l_w-1}^{(I)} \end{pmatrix}\left(\Delta\vec{w}_0^{(I)},\ldots,\Delta\vec{w}_{l_w-1}^{(I)}\right) = \\ &= \begin{pmatrix} \Delta\nabla\varepsilon_0^{(I)}\Delta w_0^{(I)} & \cdots & \Delta\nabla\varepsilon_0^{(I)}\Delta w_{l_w-1}^{(I)} \\ \vdots & \ddots & \vdots \\ \Delta\nabla\varepsilon_{l_w-1}^{(I)}\Delta w_0^{(I)} & \cdots & \Delta\nabla\varepsilon_{l_w-1}^{(I)}\Delta w_{l_w-1}^{(I)} \end{pmatrix} \end{aligned} \qquad (16)$$

и вектора $\vec{C}(H4)^{(I)}$. Строки матрицы $C\left(\tilde{H}_{B0}\right)^{(I)}$ и элементы вектора $\vec{C}(H4)^{(I)}$ распределяются таким образом, что все ИР-элементы с второго по предпоследний вычисляют

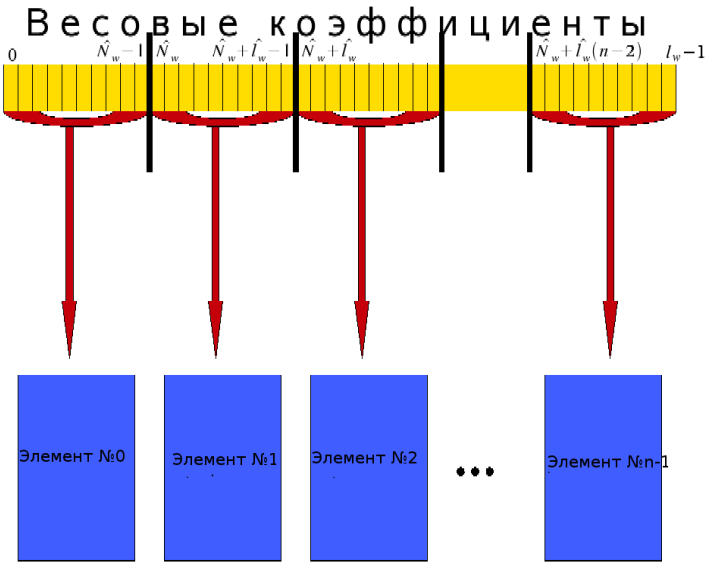

Рис. 1. Схема размещения вектора весовых коэффициентов по ИР-элементам

$$\tilde{l}_w = \begin{cases} \left[\frac{l_w}{n-1}\right], & l_w \mod (n-1) = 0; \\ \left[\frac{l_w}{n-2}\right], & l_w \mod (n-1) \neq 0, \end{cases} \quad (17)$$

строк/элементов, а последний –

$$\hat{\tilde{l}}_w = \begin{cases} \hat{l}_w, & l_w \mod (n-1) = 0; \\ l_w - \hat{l}_w(n-2), & l_w \mod (n-1) \neq 0. \end{cases} \quad (18)$$

Управляющий ИР-элемент вычисляет значение $C(H3)^{(I)}$.

4. Неуправляющие ИР-элементы передают на управляющий вычисленные ими значения матрицы $C(H_{B0})^{(I)}$ и вектора $\vec{C}(H4)^{(I)}$.

5. Управляющий ИР-элемент рассылает значения матрицы $C(\tilde{H}_{B0})^{(I)}$ и вектора $\vec{C}(H4)^{(I)}$ на прочие ИР-элементы.

6. Управляющий ИР-элемент вычисляет

$$C(\tilde{H}_{B1})^{(I)} = \left[\vec{C}(H4)^{(I)}\right]^T \Delta\nabla\varepsilon^{(I)} = \left(\vec{C}(H4)_0^{(I)},\ldots,\vec{C}(H4)_{l_w-1}^{(I)}\right)\begin{pmatrix} \Delta\nabla\varepsilon_0^{(I)} \\ \vdots \\ \Delta\nabla\varepsilon_{l_w-1}^{(I)} \end{pmatrix} = \sum_{i=0}^{l_w-1}\left(C(H4)_i^{(I)}\Delta\nabla\varepsilon_i^{(I)}\right) \quad (19)$$

и

$$C(\tilde{H}_{\tilde{w}})^{(I)} = \left(C(H3)^{(I)} + C(\tilde{H}_{B1})^{(I)}\right)\tilde{w}^{(I)} = \begin{pmatrix} \left(C(H3)^{(I)} + C(\tilde{H}_{B1})^{(I)}\right)\tilde{w}_{0,0}, & \cdots & \left(C(H3)^{(I)} + C(\tilde{H}_{B1})^{(I)}\right)\tilde{w}_{0,l_w-1} \\ \vdots & \ddots & \vdots \\ \left(C(H3)^{(I)} + C(\tilde{H}_{B1})^{(I)}\right)\tilde{w}_{l_w-1,0}, & \cdots & \left(C(H3)^{(I)} + C(\tilde{H}_{B1})^{(I)}\right)\tilde{w}_{l_w-1,l_w-1} \end{pmatrix}; \quad (20)$$

$$C(H1)^{(I)} = \frac{C(\tilde{H}_{\tilde{w}})^{(I)}}{C(H3)^2}, \quad (21)$$

а все прочие – соответствующие строки матриц

$$C(\tilde{H}_{B2})^{(I)} = \left(\tilde{H}^{(I)}\right)\left(C(\tilde{H}_{B0})^{(I)}\right) = \begin{pmatrix} \tilde{H}_{0,0}^{(I)} & \cdots & \tilde{H}_{0,l_w-1}^{(I)} \\ \vdots & \ddots & \vdots \\ \tilde{H}_{l_w-1,0}^{(I)} & \cdots & \tilde{H}_{l_w-1,l_w-1}^{(I)} \end{pmatrix} \begin{pmatrix} C(\tilde{H}_{B0})_{0,0}^{(I)} & \cdots & C(\tilde{H}_{B0})_{0,l_w-1}^{(I)} \\ \vdots & \ddots & \vdots \\ C(\tilde{H}_{B0})_{l_w-1,0}^{(I)} & \cdots & C(\tilde{H}_{B0})_{l_w-1,l_w-1}^{(I)} \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{l_w-1}\left(H_{0,i}C(\tilde{H}_{B0})_{i,0}^{(I)}\right) & \cdots & \sum_{i=0}^{l_w-1}\left(H_{l_w-1,i}C(\tilde{H}_{B0})_{i,0}^{(I)}\right) \\ \vdots & \ddots & \vdots \\ \sum_{i=0}^{l_w-1}\left(H_{l_w-1,i}C(\tilde{H}_{B0})_{i,0}^{(I)}\right) & \cdots & \sum_{i=0}^{l_w-1}\left(H_{l_w-1,i}C(\tilde{H}_{B0})_{i,l_w-1}^{(I)}\right) \end{pmatrix}; \quad (22)$$

$$\tilde{N}(\tilde{H}_{B3})^{(I)} = \Delta\vec{w}^{(I)}\left[\vec{C}(H4)^{(I)}\right]^T = \begin{pmatrix} \Delta\vec{w}_0^{(I)} \\ \vdots \\ \Delta\vec{w}_{l_w-1}^{(I)} \end{pmatrix}\left(\vec{C}(H4)_0^{(I)},\ldots,\vec{C}(H4)_{l_w-1}^{(I)}\right) = \begin{pmatrix} \Delta w_0^{(I)}C(H4)_0^{(I)} & \cdots & \Delta w_0^{(I)}C(H4)_{l_w-1}^{(I)} \\ \vdots & \ddots & \vdots \\ \Delta w_{l_w-1}^{(I)}C(H4)_0^{(I)} & \cdots & \Delta w_{l_w-1}^{(I)}C(H4)_{l_w-1}^{(I)} \end{pmatrix} \quad (23)$$

и

$$C(H2)^{(I)} = \frac{C(\tilde{H}_{B2})^{(I)} + C(\tilde{H}_{B3})^{(I)}}{C(H3)}. \quad (24)$$

При этом все ИР-элементы кроме последнего вычисляют $\tilde{l}_w$, а последний – $\hat{\tilde{l}}_w$ строк.

7. Неуправляющие ИР-элементы передают на управляющий $C(H2)^{(I)}$ (значения матриц $C(\tilde{H}_{B2})^{(I)}$ и $C(\tilde{H}_{B3})^{(I)}$ не требуются на управляющем ИР-элементе).

8. Управляющий ИР-элемент вычисляет гессиан и новые значения весовых коэффициентов $\vec{w}^{(I+1)}$ и рассылает их на все прочие ИР-элементы.

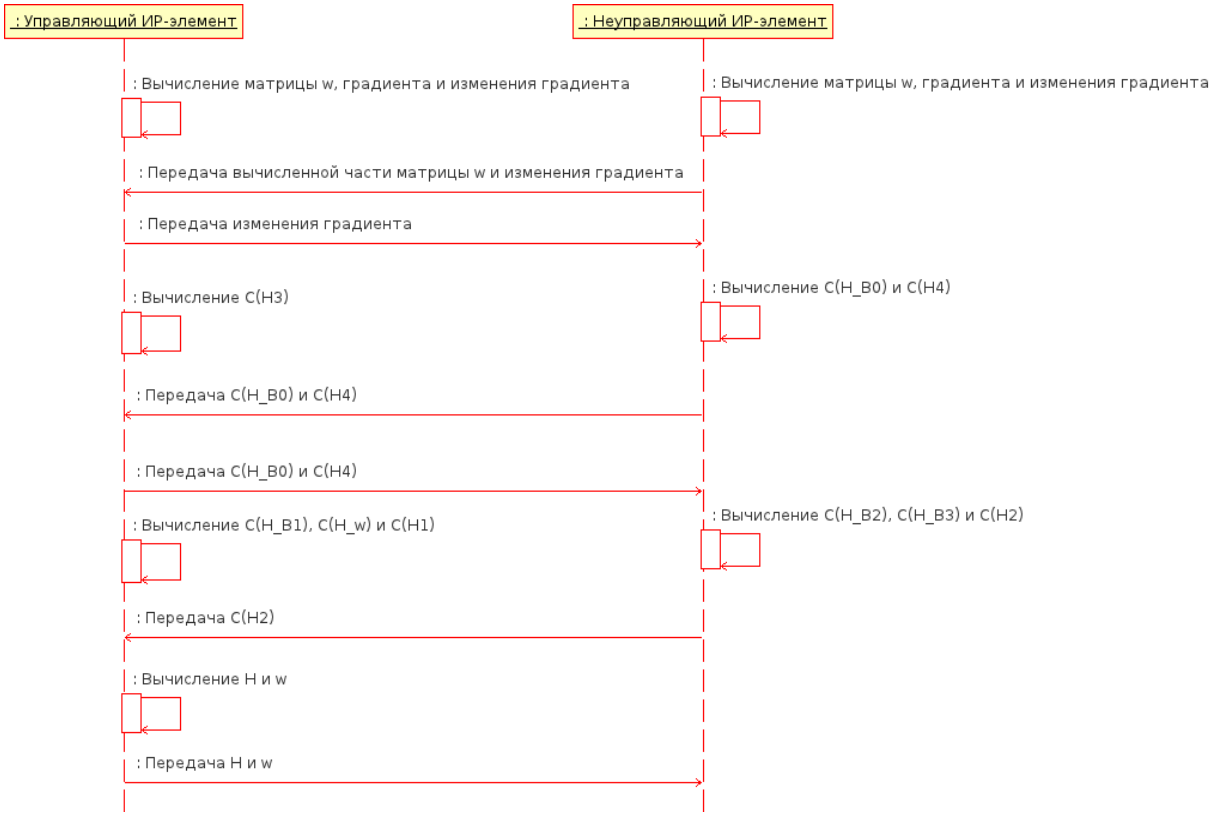

Рис. 2. Диаграмма последовательности выполнения параллельного вычисления весовых коэффициентов

Число мультипликативных операций, необходимых для вычисления нового значения весовых коэффициентов. Традиционным способом определения эффективности алгоритма является подсчет числа сделанных им мультипликативных операций. Поскольку на современных компьютерах мультипликативная операция обычно занимает немногим больше времени, чем аддитивная, то для построения аналитических выражений будем производить подсчет и мультипликативных, и аддитивных операций. Для приведения аддитивных операций к мультипликативным введем коэффициент σ, значение которого прямо пропорционально времени, затрачиваемому на одну аддитивную операцию, и обратно пропорционально времени, затрачиваемому на одну мультипликативную:

$$\sigma = \frac{t_a}{t_m}, \qquad (25)$$

где t_a – время, затрачиваемое на одну аддитивную операцию; t_m – время, затрачиваемое на одну мультипликативную операцию.

Таким образом, одна аддитивная операция занимает в σ раз больше времени, чем мультипликативная, и, следовательно, одну аддитивную операцию можно заменить σ мультипликативными и наоборот [8]. Обозначим количество мультипликативных операций, необходимых для вычисления значения невязки, как Z_ε. Тогда для определения эффективности необходимо вычислить количество операций, которые требуются для вычисления нового значения вектора весовых коэффициентов $\vec{w}^{(I+1)}$ по формуле Бройдена–Флетчера–Гольдфарда–Шенно – $z\left(\tilde{H}\right)$.

Обозначим количество операций, необходимых для вычисления произведения вектора, содержащего l_w элементов, на транспонированный вектор, содержащий l_w элементов, как

$$z_{v0} = l_w^2 + \sigma l_w^2, \qquad (26)$$

количество операций, необходимых для вычисления произведения транспонированного вектора на вектор, как

$$z_{v1} = 2\sigma l_w + \sigma + l_w, \qquad (27)$$

количество операций, необходимых для вычисления произведения матриц размерностью l_w, как

$$z_{v2} = z_{v1} l_w^2 + \sigma l_w^2, \qquad (28)$$

а количество операций, необходимых для произведения транспонированного вектора на матрицу, как

$$z_{v3} = z_{v1} l_w + \sigma l_w \qquad (29)$$

(размерность матрицы и вектора равна l_w).

Для получения количества мультипликативных операций, требуемых для вычисления нового значения вектора весовых коэффициентов $\vec{w}^{(I+1)}$, необходимо вычислить количества операций для каждого из шагов:

1) вычисление по формуле (1) значения градиента $\nabla\varepsilon^{(I)}$. Для этого необходимо $l_w(z_\varepsilon + 3\sigma + 1) + l_w$ мультипликативных операций;

2) вычисление приращения градиента $\Delta\nabla\varepsilon^{(I)}$, требующее $2\sigma l_w$ операций;

3) вычисление по формуле (10) матрицы $\tilde{\tilde{w}}^{(I)}$. Для вычисления $\tilde{\tilde{w}}^{(I)}$ необходимо z_{v0} операций;

4) вычисление по формуле (11) $C(H3)^{(I)}$. Для этого необходимо z_{v1} мультипликативных операций;

5) вычисление по формуле (12) $\left[\vec{C}(H4)^{(I)}\right]^T$, требующее z_{v3} операций;

6) произведение приращения градиента $\Delta\nabla\varepsilon^{(I)}$ на транспонированный вектор приращения весовых коэффициентов $\left[\vec{w}^{(I)}\right]^T$ – вычисление $C\left(\tilde{H}_{B0}\right)^{(I)}$ по формуле (16). Для этого необходимо z_{v0} мультипликативных операций;

7) умножение транспонированного вектора $\left[\vec{C}(H4)^{(I)}\right]^T$ на вектор приращения градиента $\Delta\nabla\varepsilon^{(I)}$ ($C\left(\tilde{H}_{B1}\right)^{(I)}$ – формула (19)), требующее z_{v1} мультипликативных операций;

8) вычисление по формуле (22) значения $C\left(\tilde{H}_{B2}\right)^{(I)}$, представляющего собой произведение матриц $\tilde{H}^{(I)}$ и $C\left(\tilde{H}_{B0}\right)^{(I)}$. Для вычисления $C\left(\tilde{H}_{B2}\right)^{(I)}$ необходимо z_{v2} операций;

9) произведение вектора приращения весовых коэффициентов на транспонированный вектор $\left[\vec{C}(H4)^{(I)}\right]^T$, т. е. вычисление по формуле (23) $C\left(\tilde{H}_{B3}\right)^{(I)}$. Для этого необходимо z_{v1} мультипликативных операций;

10) умножение матрицы $\tilde{\tilde{w}}^{(I)}$ на число $C(H3)^{(I)} + C\left(\tilde{H}_{B0}\right)^{(I)}$ (в результате чего по формуле (20) вычисляется значение $C\left(\tilde{H}_{\tilde{\tilde{w}}}\right)^{(I)}$) требует $l_w^2\sigma + l_w^2 + 1$ операций;

11) вычисление по формуле (21) $C(H1)^{(I)}$, требующее $\sigma l_w^2 + l_w^2$ операций;

12) вычисление по формуле (24) $C(H2)^{(I)}$, требующее $\sigma l_w^2 + l_w^2 + 2\sigma l_w^2 + 1$ операций;

13) вычисление по формуле (7) значения матрицы $\tilde{H}^{(I+1)}$. Для этого необходимо $3\sigma l_w^2$ операций;

14) вычисление новых значений весовых коэффициентов требует $z_{v0}l_w + \sigma l_w$ операций.

Таким образом, количество операций, необходимых для вычисления значений $\nabla\varepsilon$, $\Delta\nabla\varepsilon$, $\tilde{\tilde{w}}$, $C(H3)$ и $[C(H4)]^T$, может быть вычислено по формуле:

$$\begin{aligned} z(C(H0)) &= l_w(z_\varepsilon + 3\sigma + 1) + l_w + 2\sigma l_w + \\ &+ z_{v0} + z_{v1} + z_{v3} = \\ &= l_w z_\varepsilon + 5\sigma l_w + 2l_w + z_{v0} + z_{v1} + z_{v3}, \end{aligned} \quad (30)$$

количество операций, необходимых для вычисления $C(H1)$ ($C\left(\tilde{H}_{B0}\right)$ и $C\left(\tilde{H}_{\tilde{\tilde{w}}}\right)$), – по формуле:

$$\begin{aligned} z(C(H1)) &= z_{v0} + \sigma l_w^2 + l_w^2 + 1 + \sigma l_w^2 + l_w^2 = \\ &= z_{v0} + 2l_w^2 + 2l_w^2 + 1, \end{aligned} \quad (31)$$

а количество операций, необходимых для вычисления $C(H2)$ ($C\left(\tilde{H}_{B1}\right), C\left(\tilde{H}_{B2}\right)$ и $C\left(\tilde{H}_{B3}\right)$), – по формуле:

$$\begin{aligned} z(C(H2)) &= z_{v1} + z_{v2} + z_{v1} + 2\sigma l_w^2 + \sigma l_w^2 + l_w^2 \\ &+ 1 = 2z_{v1} + z_{v2} + 3\sigma l_w^2 + l_w^2 + 1. \end{aligned} \quad (32)$$

Следовательно, общее число операций z может быть вычислено по формуле:

$$\begin{aligned} z &= z(C(H0)) + z(C(H1)) + z(C(H2)) + \\ &+ 3\sigma l_w^2 + z_{v0}l_w + \sigma l_w = l_w z_\varepsilon + 5\sigma l_w + 2l_w + \\ &+ 2z_{v0} + z_{v1} + z_{v3} + z_{v0} + 2l_w^2 + 2l_w^2 + 1 + \\ &+ 2z_{v1} + z_{v2} + 3\sigma l_w^2 + l_w^2 + 1z_{v0}l_w + \sigma l_{=w} = \\ &= l_w z_\varepsilon + 6\sigma l_w + 2l_w + 3z_{v0} + 3z_{v1} + z_{v2} + \\ &+ z_{v3} + 5l_w^2 + 2 + 3\sigma l_w^2 + z_{v0}l_w. \end{aligned} \quad (33)$$

Число мультипликативных операций, необходимых для параллельного вычисления гессиана. Обозначим количество мультипликативных операций, необходимых для передачи l элементов с одного ИР-элемента на другой при скорости интерконекта ν как $\gamma(l, \nu)$ (значение $\gamma(l, \nu)$ зависит от используемой платформы). Тогда, учитывая, что для подготовки одного элемента к передаче между ИР-элементами и обработке его после приема необходимо $1 + 2\sigma$ операций, для каждого из шагов алгоритма необходимо следующее количество операций.

1. Для вычисления элементов градиента и его приращения необходимо $\hat{N}_w z_\varepsilon + 5\sigma\hat{N}_w + 2\hat{N}_w$ мультипликативных операций на управляющем ИР-элементе и $\hat{l}_w z_\varepsilon + 5\sigma\hat{l}_w + 2\hat{l}_w$ на прочих. Для вычисления части строк матрицы $\tilde{\tilde{w}}^{(I)}$ все неуправляющие ИР-элементы выполняют по $l_w\hat{l}_w + 2\sigma l_w\hat{l}_w$ операций (управляющий выполняет $l_w\hat{N}_w + 2\sigma l_w\hat{N}_w$ операций). Для передачи $\hat{l}_w$ элементов вектора $\Delta\nabla\varepsilon^{(I)}$ необходимо $\hat{l}_w + 2\sigma\hat{l}_w$ операций, а для передачи $\hat{l}_w$

строк матрицы $\tilde{w}^{(I)}$ – $l_w(\hat{l}_w + 2\sigma\hat{l}_w)$. Для приема вычисленных значений управляющему ИР-элементу необходимо выполнить $(n-1)$ $\left(\hat{l}_w + 2\sigma\hat{l}_w + l_w\hat{l}_w + 2\sigma l_w\hat{l}_w\right)$ операций, но прием может начаться лишь после отправки, следовательно, количество операций, выполняемых на управляющем ИР-элементе, вычисляется по формуле:

$$C_B^{(1,0)} = max\begin{pmatrix} \hat{N}_w z_\varepsilon + 5\sigma\hat{N}_w + 2\hat{N}_w + \\ l_w\hat{N}_w + 2\sigma l_w\hat{N}_w, \\ \max\limits_{k=1..n-1}\left(C_{wGD}^{(1,k)} + \gamma\left(\hat{l}_w, v\right)\right) \end{pmatrix} + \\ + n\hat{l}_w + 2\sigma n\hat{l}_w - \hat{l}_w - 2\sigma\hat{l}_w, \quad (34)$$

где $\gamma\left(\hat{l}_w, v\right)$ – число мультипликативных операций, выполняемых для передачи $\hat{l}_w$ элементов при скорости интерконекта v, а на прочих – по формуле:

$$C_B^{(2,0)} = \hat{l}_w z_\varepsilon + 7\sigma\hat{l}_w - 3\hat{l}_w + l_w\hat{l}_w + 2l_w\hat{l}_w + \\ l_w\hat{l}_w + 2\sigma l_w\hat{l}_w = \hat{l}_w z_w + 7\sigma\hat{l}_w + 3\hat{l}_w + \\ + 4l_w\hat{l}_w + 2\sigma l_w\hat{l}_w. \quad (35)$$

2. Для рассылки значений $\Delta\nabla\varepsilon^{(I)}$ ведущий ИР-элемент выполняет

$$C_B^{(2,0)} = nl_w + 2\sigma nl_w - l_w - 2\sigma l_w, \quad (36)$$

операций, а прочие – по

$$C_B^{(2,k)} = l_w + 2\sigma l_w + \gamma(l_w, v) + kl_w + 2\sigma kl_w. \quad (37)$$

3. Для вычисления части строк матрицы $C\left(\tilde{H}_{B0}\right)$ все неуправляющие ИР-элементы, за исключением последнего, выполняют по $l_w\tilde{l}_w + 2\sigma l_w\tilde{l}_w$ операций (последний выполняет $l_w\hat{\tilde{l}}_w + 2\sigma l_w\tilde{l}_w$ операций). Для вычисления части элементов вектора $\vec{C}(H4)$ неуправляющие ИР-элементы выполняют $z_{v1}(l_w)\tilde{l}_w + \sigma\tilde{l}_w$ и $z_{v1}(l_w)\hat{\tilde{l}}_w + \sigma\hat{\tilde{l}}_w$ операций, соответственно. Таким образом, общее число мультипликативных операций для неуправляющих ИР-элементов –

$$C_B^{(3,k)} = l_w\tilde{l}_w + 2\sigma l_w\tilde{l}_w + z_{v1}(l_w)\tilde{l}_w + \sigma\tilde{l}_w \quad (38)$$

(для непоследних) и

$$C_B^{(3,n\text{-}1)} = l_w\hat{\tilde{l}}_w + 2\sigma l_w\hat{\tilde{l}}_w + z_{v1}(l_w)\hat{\tilde{l}}_w + \sigma\hat{\tilde{l}}_w \quad (39)$$

(для последнего), а управляющий ИР-элемент выполняет (для вычисления $C(H3)^{(I)}$)

$$C_B^{(3,0)} = z_{v1} \quad (40)$$

операций.

4. Для оправки вычисленных значений ($C\left(\tilde{H}_{B0}\right)$ и $\vec{C}(H4)$) неуправляющие ИР-элементы выполняют

$$C_B^{(4,k)} = l_w\tilde{l}_w + \tilde{l}_w + 2\sigma l_w\tilde{l}_w + 2\sigma\tilde{l}_w \quad (41)$$

(для непоследних) и

$$C_B^{(4,n\text{-}1)} = l_w\hat{\tilde{l}}_w + \hat{\tilde{l}}_w + 2\sigma l_w\hat{\tilde{l}}_w + 2\sigma\hat{\tilde{l}}_w \quad (42)$$

(для последнего) операций, а управляющий ИР-элемент – для получения

$$C_B^{(4,0)} = (n-2)(l_w\tilde{l}_w + \tilde{l}_w + 2\sigma l_w\tilde{l}_w + 2\sigma\tilde{l}_w) + \\ + l_w\hat{\tilde{l}}_w + \hat{\tilde{l}}_w + 2\sigma l_w\hat{\tilde{l}}_w + 2\sigma\hat{\tilde{l}}_w \quad (43)$$

операций.

5. Для рассылки элементов матрицы $C\left(\tilde{H}_{B0}\right)$ вектора $\vec{C}(H4)$ управляющий ИР-элемент выполняет

$$C_B^{(5,0)} = (n-1)(l_w^2 + 2\sigma l_w^2 + l_w + 2\sigma l_w) \quad (44)$$

мультипликативных операций, а прочие ИР-элементы – по

$$C_B^{(5,k)} = l_w^2 + 2\sigma l_w^2 + l_w + 2\sigma l_w + \gamma(l_w^2 + l_w, v) + \\ + k(l_w^2 + 2\sigma l_w^2 + l_w + 2\sigma l_w) \quad (45)$$

мультипликативных операций.

6. Для вычисления строк матриц $C\left(\tilde{H}_{B2}\right)^{(I)}$, $C\left(\tilde{H}_{B3}\right)^{(I)}$ и $C(H2)^{(I)}$ последний неуправляющий ИР-элемент выполняет

$$C_{B,}^{(6,n\text{-}1)} = z_{v1}l_w\hat{\tilde{l}} + 2\sigma\hat{\tilde{l}}_w + \sigma + \hat{\tilde{l}} + \\ + l_w\hat{\tilde{l}}_w + 6\sigma l_w\hat{\tilde{l}}_w + 1 \quad (46)$$

операций (из них $z_{v1}l_w\hat{\tilde{l}} + \sigma l_w\hat{\tilde{l}}_w$ операций для вычисления $C\left(\tilde{H}_{B2}\right)^{(I)}$, $2\sigma\hat{\tilde{l}}_w + \sigma + \hat{\tilde{l}}_w$ операций для вычисления $C\left(\tilde{H}_{B3}\right)^{(I)}$ и $5\sigma l_w\hat{\tilde{l}}_w + l_w\hat{\tilde{l}} + 1$ для вычисления $C(H2)^{(I)}$), остальные неуправляющие – по

$$C_{B,}^{(6,n-1)} = z_{v1} l_w \tilde{l}_w + 2\sigma \tilde{l}_w + \sigma + \tilde{l}_w + l_w \tilde{l} + 6\sigma l_w \tilde{l}_w + 1 \quad (47)$$

(из которых $z_{v1} l_w \tilde{l}_w + \sigma l_w \tilde{l}_w$ операций для вычисления $C\left(\tilde{H}_{B2}\right)^{(I)}$, $2\sigma \tilde{l}_w + \sigma + \tilde{l}_w$ операций для вычисления $C\left(\tilde{H}_{B3}\right)^{(I)}$ и $5\sigma l_w \tilde{l}_w + l_w \tilde{l}_w + 1$ для вычисления $C(H2)^{(I)}$), а управляющий (для вычисления $C\left(H_{B1}\right)^{(I)}$, $C\left(H_{\tilde{w}}\right)^{(I)}$ и $C(H1)^{(I)}$) –

$$C_{B,}^{(6,0)} = z_{v1} + 2\sigma l_w^2 + 2l_w^2 + 1 \quad (48)$$

(z_{v1} операций для вычисления $C\left(H_{B1}\right)^{(I)}$, $\sigma l_w^2 + l_w^2 + 1$ операций для вычисления $C\left(H_{\tilde{w}}\right)^{(I)}$ и $\sigma l_w^2 + l_w^2$ для вычисления $C(H1)^{(I)}$).

7. Для передачи $C(H2)^{(I)}$ неуправляющие ИР-элементы выполняют по

$$C_B^{(7,n-1)} = l_w \hat{\tilde{l}}_w + 2\sigma l_w \hat{\tilde{l}}_w \quad (49)$$

(последний ИР-элемент) и

$$C_B^{(7,k)} = l_w \tilde{l}_w + 2\sigma l_w \tilde{l}_w \quad (50)$$

(непоследние неуправляющие ИР-элементы) мультипликативных операций, а управляющий –

$$C_B^{(7,0)} = (n-2)(l_w \tilde{l}_w + 2\sigma l_w \tilde{l}_w) + l_w \hat{\tilde{l}}_w + 2\sigma l_w \hat{\tilde{l}}_w \quad (51)$$

операций.

8. Для вычисления гессиана управляющий ИР-элемент выполняет $3\sigma l_w^2$ мультипликативных операций, а для вычисления новых значений весовых коэффициентов – $z_{v0} l_w + \sigma l_w$ операций. Для рассылки вычисленных значений управляющий ИР-элемент выполняет $(n-1)(l_w^2 + \sigma l_w^2 + l_w + \sigma l_w)$, а прочие для приема – по $l_w^2 + \sigma l_w^2 + l_w + \sigma l_w$ операций. Таким образом, общее число мультипликативных операций, выполняемых на управляющем ИР-элементе на данном шаге, –

$$C_B^{(8,0)} = \hat{\tilde{C}} + (n-1)(l_w^2 + \sigma l_w^2 + l_w + \sigma l_w), \quad (52)$$

где $\hat{\tilde{C}} = 3\sigma l_w^2 + z_{v0} l_w + \sigma l_w$, а на прочих – по

$$C_B^{(8,0)} = l_w^2 + \sigma l_w^2 + l_w + \sigma l_w + \gamma(l_w^2 + l_w, v) + k(l_w^2 + \sigma l_w^2 + l_w + \sigma l_w) + \hat{\tilde{C}}. \quad (53)$$

Как можно видеть, для выполнения первых четырех шагов максимально-выполняемое число мультипликативных операций составляет

$$\hat{C}_B^{(1)} = C_B^{(1,0)} + \max(C_B^{(2,0)} + C_B^{(3,0)}, \max_{k=1..n-2}(C_B^{(2,k)} + C_B^{(3,k)} + C_B^{(4,k)} + \gamma(\tilde{l}_w l_w + \tilde{l}_w, v)), C_B^{(2,n-1)} + C_B^{(3,n-1)} + C_B^{(4,n-1)} + \gamma(\hat{\tilde{l}}_w l_w + \hat{\tilde{l}}_w, v)) + C_B^{(4,0)}, \quad (54)$$

для пятого, шестого и седьмого –

$$\hat{C}_B^{(2)} = \max(C_B^{(5,0)} + C_B^{(6,0)}, \max_{k=1..n-2}(C_B^{(5,k)} + C_B^{(6,k)} + C_B^{(7,k)} + \gamma(\tilde{l}_w l_w, v)), C_B^{(5,n-1)} + C_B^{(6,n-1)} + C_B^{(7,k)} + \gamma(\hat{\tilde{l}}_w l_w, v)) + C_B^{(7,0)}, \quad (55)$$

а для восьмого –

$$\hat{C}_B^{(3)} = \max_{k=0..n-1}(C_B^{(8,k)}). \quad (56)$$

Исходя из вышесказанного, общее число мультипликативных операций, выполняемых для параллельного вычисления весовых коэффициентов на одной итерации, составляет

$$\hat{C}_B = \sum_{\iota=1}^{3} \hat{C}_B^{(\iota)}. \quad (57)$$

Аналитические и эмпирические значения коэффициента эффективности. Для проверки эффективности разработанных параллельных алгоритмов введем коэффициент эффективности, который вычисляется по формуле:

$$\alpha(Z) = \frac{z_{w0} I_G z}{n z_{w0} + n I_G Z} 100\,\%, \quad (58)$$

где I_G – число итераций алгоритма обучения; z_{w0} – количество операций, необходимых для инициализации; Z – число операций, необходимых для параллельного вычисления весовых коэффициентов, вычисляемое по формуле:

$$Z = \hat{C}_B + 2\sigma l_w. \quad (59)$$

Для определения эмпирического значения коэффициента эффективности необходимо вычислить временные затраты на последовательный t и параллельный $\tau(n)$ алгоритм. Таким образом, эмпирическое значение коэффициента эффективности может быть вычислено по формуле:

Таблица 1

Аналитические и эмпирические значения коэффициента эффективности

Число ИР-элементов (процессоров)	Аналитический коэффициент эффективности, %	Эмпирический коэффициент эффективности, %
2	90,02	89,99
4	90,01	89,99
6	90,00	89,97
8	90,00	89,96
10	89,98	89,96
12	89,96	89,96
14	89,95	89,95

$$\alpha(t) = \frac{t}{n\tau(n)} 100\,\%. \qquad (60)$$

Вычислительные эксперименты. Вычислительные эксперименты были проведены на кластерной системе Тамбовского государственного университета им. Г.Р. Державина. Для проведения эксперимента была использована сеть каскадной корреляции Фальмана [7]. Аналитические и эмпирические значения коэффициента эффективности приведены в табл. 1.

Выводы. Приведенные выше результаты говорят о том, что данная технология распараллеливания способна значительно снизить временные затраты на обучение ИНС, что обусловливает более эффективное использование. Таким образом, поставленную цель повышения эффективности квазиньютоновских методов при помощи параллельных вычислений можно считать достигнутой.

ЛИТЕРАТУРА

1. *Крючин О.В.* Разработка параллельных градиентных алгоритмов обучения искусственной нейронной сети // Исследовано в России. 2009. С. 1208-1221. URL: // http://zhurnal.ape.relarn.ru /articles/2009/096.pdf. Загл. с экрана.
2. *Gill P., Murray W., Wrights M.H., Wrights M.* Practical Optimisation. N. Y.: Academic Press, 1981. 420 p.
3. *Bonnans J.F., Gilbert J.Ch., Lemarechal C., Sagastizbal C.A.* Numerical optimization, theoretical and numerical aspects. Springer, 2006. 437 p.
4. *Byrd R.H., Lu P., Nocedal J.A.* Limited Memory Algorithm for Bound Constrained Optimization // SIAM J. on Scientific and Statistical Computing. 1995. P. 1190-1208. URL: // http://www.ece.northwestern.edu/~nocedal/PSfiles/limited.ps.gz. Загл. с экрана.
5. *Крючин О.В., Арзамасцев А.А., Королев А.Н., Горбачев С.И., Семенов Н.О.* Универсальный симулятор, базирующийся на технологии искусственных нейронных сетей, способный работать на параллельных машинах // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2008. Т. 13. Вып. 5. С. 372-375.
6. *Крючин О.В., Арзамасцев А.А.* Сравнение эффективности последовательных и параллельных алгоритмов обучения искусственных нейронных сетей на кластерных вычислительных системах // Вестник Тамбовского университета. Серия Естественные и технические науки. Тамбов, 2010. Т. 15. Вып. 6. С. 372-375.
7. *Fahlman S.E., Lebiere C.* The cascade-correlation learning architecture: tech. rep. CMU-CS-90-100. School of Computer Science. Carnegie Mellon University. August 1991. 18 p.

Поступила в редакцию 31 августа 2012 г.

Kryuchin O.V., Vyazovova E.V., Arzamastsev A.A. INFORMATION PROCESSES OF EFFICIENCY INCREASING OF SELECTION OF WEIGHT COEFFICIENTS OF ARTIFICIAL NEURAL NETWORK BY MEANS OF QUASI-NEWTON METHODS USING BROYDEN-FLETCHER-GOLDFARB-SHANNO FORMULA

In this paper the parallel quasi-Newton method which uses the Broyden-Fletcher-Goldfarb-Shanno formula is presented. It describes analytic equations allowing estimating the presented algorithm efficiency and experiments which were done for these equations checking.

Key words: artificial neural networks; parallel calculations; quasi-Newton methods; information resources; efficiency increasing; analytic equations.