СИГНАЛЫ: ФОРМИРОВАНИЕ, ОБРАБОТКА И ПЕРЕДАЧА

УДК 004.032 26

АДАПТИВНОЕ НЕЙРОСЕТЕВОЕ СЖАТИЕ СИГНАЛОВ БОЛЬШОЙ РАЗМЕРНОСТИ НА ОСНОВЕ ВЗВЕШЕННОГО ИНФОРМАЦИОННОГО КРИТЕРИЯ

С. В. МАШТАЛИР, Е. С. САКАЛО

Предложен адаптивный оптимальный по быстродействию градиентный алгоритм самообучения двухслойной нейронной сети для выделения главных компонент в задачах понижения размерности пространства признаков, возникающих при обработке сигналов большой размерности. Рассматриваемая нейронная сеть и соответствующая процедура самообучения объединяют в себе достоинства последовательного и параллельного нейросетевого подходов и предназначены, прежде всего, для использования в системах видеообработки реального времени.

The optimal rate adaptive gradient algorithm for a two-layer self-learning neural network is offered for allocation of the principal components in problems of feature space dimension reduction which arise at large scale signals processing. The examined neural network and the corresponding self-learning procedure combine advantages of sequential and parallel neural network approaches and they are primarily intended for usage in the real time videoprocessing systems.

ВВЕДЕНИЕ

В задачах анализа и обработки сигналов высокой размерности, возникающих, прежде всего, при распознавании и обработке изображений, космических снимков, данных экологического мониторинга, биометрической информации и т.п., на предварительных этапах исследования достаточно часто появляется необходимость понижения размерности исходного пространства признаков, для чего традиционно используется стандартный анализ главных компонент. С теоретической точки зрения анализ главных компонент сводится к хорошо изученной алгебраической проблеме собственных значений, а ее решение в задачах, связанных с обработкой изображений, известно как преобразование Карунена-Лоэва. При этом предполагается, что исходная информация задана в виде фиксированного массива данных, образованного N n-мерными векторами $x(1), x(2), \ldots, x(k), \ldots, x(N)$, где $k = 1, 2, \ldots$ – номера наблюдения в исходном массиве, а результатом решения является набор доминантных собственных значений $\lambda_1 > \ldots > \lambda_j > \ldots > \lambda_m$ и соответствующих им собственных значений векторов $w_1, w_2, \ldots, w_j, \ldots, w_m$, $w_j = (w_{j_1}, w_{j_2}, \ldots, w_{j_n})^T$ корреляционной $(n \times n)$ – матрицы исходных данных

$$\begin{cases} R(N) = \dfrac{1}{N} \sum_{k=1}^{N} (x(k) - \bar{x}(N))(x(k) - \bar{x}(N))^T, \\ \bar{x}(N) = \dfrac{1}{N} \sum_{k=1}^{N} x(k). \end{cases} \quad (1)$$

Собственно компрессия (понижение размерности) исходного пространства производится с помощью отображения

$$y(k) = Wx(k),$$

где $y(k) = (y_1(k), y_2(k), \ldots, y_m(k))^T$, $W = (w_1, w_2, \ldots, w_{m-1}, w_m)^T$ – $(m \times n)$ – матрица проецирования, образованная доминантными собственными векторами корреляционной матрицы $R(N)$.

Существенные проблемы при решении этих задач возникают, когда исходная размерность n достаточно велика, k имеет смысл текущего дискретного времени, сами данные в виде векторов $x(k)$ поступают на обработку последовательно в реальном времени. В этой ситуации на первый план выходит подход, основанный на применении искусственных нейронных сетей [1–3].

К настоящему времени известно достаточно много нейросетевых архитектур и алгоритмов обучения, решающих задачи выделения главных компонент [4], которые можно условно разделить на две большие подгруппы: последовательные и параллельные системы.

В последовательных системах [5–7], как следует из названия, главные компоненты выделяются последовательно одна за одной, при этом каждая следующая компонента вычисляется с помощью уже вычисленной предыдущей. Преимуществом таких систем является возможность адаптивно изменять количество нейронов в сети (а следовательно, и число необходимых главных компонент) по ходу процесса обучения, а основным недостатком – низкая скорость настройки, ограничивающая использование этих нейросетей в системах видеообработки реального времени.

В параллельных системах [8–11] все нейроны сети работают одновременно, благодаря чему достигается увеличение быстродействия, однако результатом вычисления являются не собственно главные компоненты, а, так называемые, главные подпространства, натянутые на доминантные собственные векторы. Такие системы хотя и обеспечивают эффективное сжатие данных, физичес-

кая интерпретация результатов при этом становится невозможной.

В основе настройки большинства из указанных сетей лежит минимизация критерия, основанного на квадрате ошибки обучения, в силу чего эти процедуры так или иначе являются ближайшими родственниками алгоритмов обучения аппроксимирующих нейронных сетей. Несколько в стороне стоит сеть Мяо-Хуа [11], обучение которой основано на использовании не квадратичного (в смысле ошибки) информационного критерия (NIC), благодаря чему достигается некоторое увеличение быстродействия по сравнению с другими указанными выше нейросетями. К сожалению, нейросеть Мяо-Хуа обеспечивает лишь анализ главных подпространств вместо желаемого анализа главных компонент.

В связи с этим представляется достаточно актуальной задача построения быстродействующей параллельной искусственной нейронной сети реального времени, обеспечивающей нахождение главных компонент корреляционной матрицы данных последовательно поступающих на обработку.

1. ОПТИМАЛЬНЫЙ АЛГОРИТМ САМООБУЧЕНИЯ НА ОСНОВЕ ВЗВЕШЕННОГО ИНФОРМАЦИОННОГО КРИТЕРИЯ

В основе предлагаемого алгоритма лежат подходы, изложенные в [8, 11, 12]. Так, Оя с соавторами в [8] для однослойной нейронной сети, образованной адаптивными линейными ассоциаторами и реализующей отображение

$$y(k) = W(k-1)x(k),$$

предложили в качестве критерия самообучения (энергетической функции) использовать конструкцию

$$\min_W \{E^{oow}(W) = -\frac{1}{2} Sp(WR(k)W^T)\} \qquad (2)$$

при ограничениях

$$WW^T = A, \qquad (3)$$

где $A = diag\{a_i\}$ – произвольная $(m \times m)$ – диагональная матрица с элементами $a_1 > a_2 > \ldots > a_j > \ldots > a_m$. Авторами [8] показано, что решение задачи оптимизации (2), (3) приводит к получению набора из m доминантных собственных векторов w_j корреляционной матрицы $R(N)$, однако эффективного алгоритма решения предложено не было.

Мяо и Хуа в [11] ввели так называемый новый информационный критерий (NIC), имеющий вид

$$\min_W \{E^{NIC}(W) = -\frac{1}{2} Sp(\log(WR(N)W^T) - Sp(WW^T))\} \qquad (4)$$

и градиентную процедуру его минимизации, в результате чего могут быть найдены главные подпространства, натянутые на доминантные собственные векторы матрицы $R(N)$. К сожалению, минимизация критерия (4) не приводит к нахождению главных компонент.

На основе (2), (3) и (4) был введен взвешенный информационный критерий (WINC)

$$\min_W \{E^{WINC}(W) = -\frac{1}{2} Sp(\log(WRW^T A) - Sp(WW^T))\} \qquad (5)$$

и алгоритм обучения нейронной сети на его основе вида

$$W(k) = W(k-1) - \eta G(k) = W(k-1) + \eta(A^{-1}W(k-1) \times R(k)W^T(k-1)A)^{-1}W(k-1)R(k) - W(k-1)), \qquad (6)$$

где η – постоянный параметр шага поиска, $G = -((A^{-1}WRW^T A)^{-1}WR - W) = \{\frac{\partial E^{WINC}}{\partial w_{ji}}\}$ – $(m \times n)$ матрица, образованная частными производными (5) по настраиваемым параметрам w_{ji}, $R(k)$ – корреляционная матрица, вычисляемая по k наблюдениям. В случае, если обработка данных проводится в реальном времени, для вычисления этой матрицы вместо (1) могут быть использованы рекуррентные соотношения

$$\begin{cases} R(k) = \frac{k-1}{k} R(k-1) + \frac{1}{k}(x(k) - \bar{x}(k))(x(k) - \bar{x}(k))^T, \\ \bar{x}(k) = \frac{k-1}{k}\bar{x}(k-1) + \frac{1}{k}x(k) \end{cases}$$

для стационарного случая и

$$\begin{cases} R(k) = \alpha R(k-1) + (1-\alpha)(x(k) - \bar{x}(k))(x(k) - \bar{x}(k))^T, \\ \bar{x}(k) = \alpha\bar{x}(k-1) + (1-\alpha)x(k) \end{cases}$$

для нестационарного. Здесь $0 < \alpha < 1$ – эмпирически выбираемый параметр забывания устаревшей информации. Авторами алгоритма (6) доказана его сходимость к доминантным собственным векторам. При этом интересно заметить, что при единичной матрице A $(A = \mathrm{I}_m)$ алгоритм (6) минимизирует критерий NIC (4), т. е. с помощью одного алгоритма можно находить и главные компоненты и главные подпространства.

Скорость сходимости алгоритма (6), как и всех градиентных алгоритмов оптимизации, существенным образом зависит от скалярного параметра шага поиска η, который в данном случае полагается постоянным. Ясно, что оптимальный выбор этого параметра позволит повысить эффективность решения задачи в целом.

Введя в рассмотрение апостериорную ошибку восстановления

$$\tilde{e}(k) = e(k) + G(k)y(k), \qquad (7)$$

где $G(k) = -((A^{-1}W(k-1)R(k)W^T(k-1)A)^{-1}W(k-1) \times R(k) - W(k-1))$, можно ввести вспомогательный

критерий оптимизации

$$\min_{\eta}\{E(\eta)=\|\tilde{e}(k)\|^2\}, \qquad (8)$$

минимизация которого обеспечивает максимально возможную скорость сходимости алгоритма (6).

Записав квадрат нормы вектора ошибок (7)

$$\|\tilde{e}(k)\|^2=\|e(k)\|^2-2\eta e^T(k)G(k)y(k)+\eta^2\|(G(k)y(k)\|^2$$

и решив уравнение

$$\frac{\partial\|\tilde{e}(k)\|^2}{\partial\eta}=-e^T(k)G(k)y(k)+\eta\|(G(k)y(k)\|^2=0,$$

получаем оптимальное значение параметра шага поиска в виде

$$\eta(k)=\frac{e^T(k)G(k)y(k)}{\|(G(k)y(k)\|^2}.$$

Тогда окончательно оптимальный алгоритм самообучения на основе взвешенного информационного критерия может быть записан в виде

$$W(k)=W(k-1)+$$
$$+\frac{((x(k)-W^T(k-1)y(k))^TG(k)y(k)}{\|G(k)y(k)\|^2}G(k). \qquad (9)$$

2. АРХИТЕКТУРА ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ ДЛЯ НАХОЖДЕНИЯ ГЛАВНЫХ КОМПОНЕНТ

Архитектура нейронной сети, решающая рассматриваемую здесь проблему, представлена на рис. 1. Сеть имеет два слоя, образованных m (в первом скрытом слое) и n (в выходном слое) адаптивными линейными ассоциаторами. В первом скрытом слое, синаптические веса которого образуют $(m\times n)$ – матрицу $W=\{w_{ji}\}$, производится сжатие информации, при этом на его выходе вычисляются значения главных компонент $y_1,y_2,...,y_m$. Выходной слой служит для восстановления входного сигнала, с помощью $(n\times m)$ матрицы синаптических весов $W^T=\{w_{ij}\}$, в связи с чем на его выходе появляются значения $\hat{x}_1(k),\hat{x}_2(k),...,\hat{x}_n(k)$, являющиеся оценками компонент входного сигнала:

$$\hat{x}(k)=W^T(k-1)y(k)=W^T(k-1)W(k-1)x(k).$$

Понятно, что такое восстановление возможно при $m=n$, однако при $m<n$ обеспечивается восстановление с максимально возможной точностью в смысле критерия (8)

$$E(\eta)=\|\tilde{e}(k)\|^2=\|x(k)-W^T(k)y(k)\|^2=$$
$$=\|x(k)-W^T(k)W(k)x(k)\|^2,$$

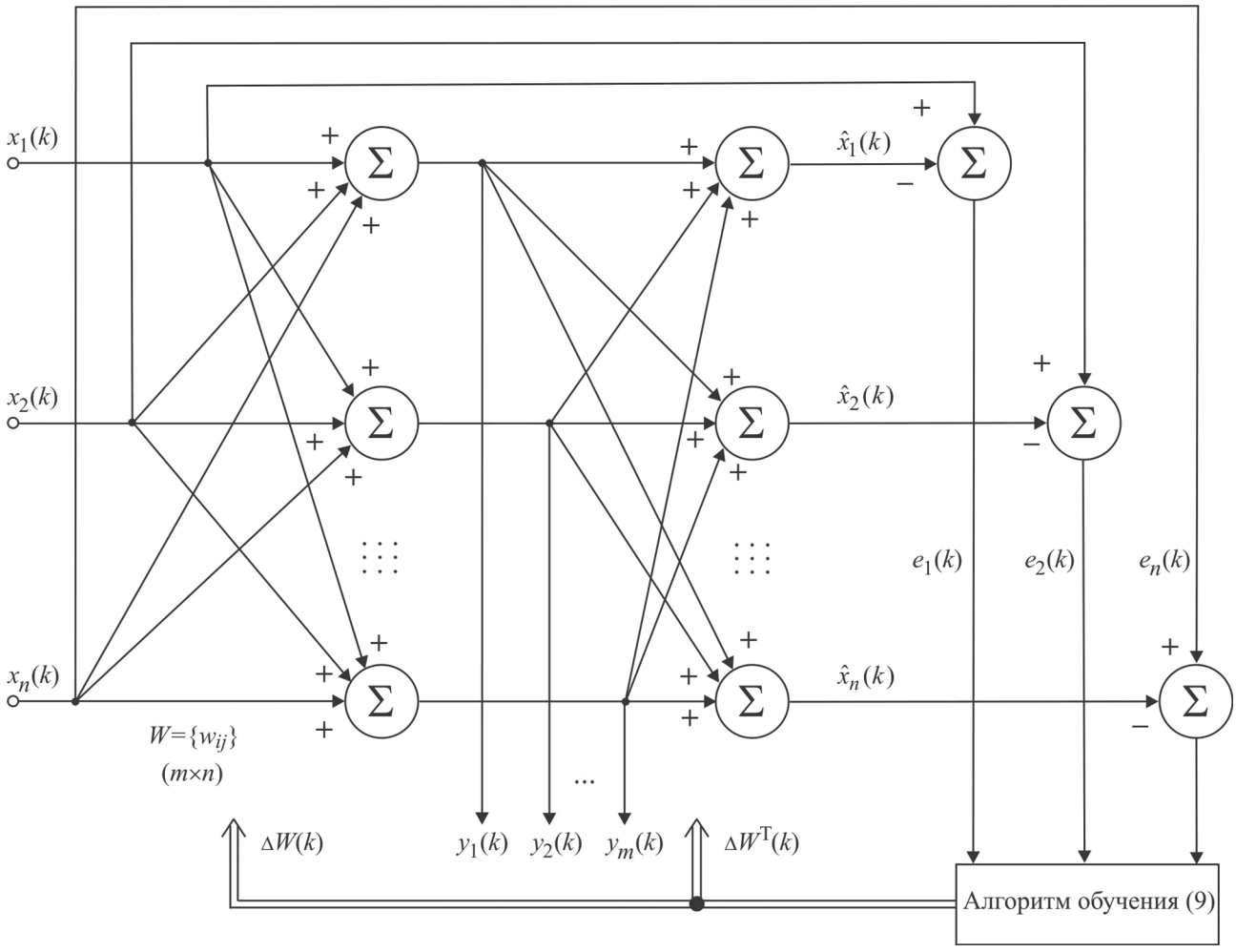

Рис. 1. Параллельная искусственная нейронная сеть для нахождения главных компонент

что выгодно отличает рассматриваемые здесь конструкции от нейросети, предложенной в [12].

ВЫВОДЫ

Предложен оптимальный по быстродействию адаптивный алгоритм самообучения двухслойной параллельной нейронной сети для нахождения главных компонент и доминантных собственных векторов корреляционных матриц данных высокой размерности. Алгоритм прост в численной реализации и позволяет обрабатывать данные в реальном времени по мере их поступления.

Литература.

[1] *Mao J. Jain A.K.* Artificial neural networks for feature extraction and multivariate data projection // IEEE Trans. on Neural Networks. – 1995. – 6. – P.296-317.

[2] *Kohonen T.* Self-organizing Maps. – Berlin: Springer-Verlag. – 1995. – 362 p.

[3] *Zhang D., Zuo W.* Computational intelligence-based biometric technologies // IEEE Computational Intelligence Magazine. – 2002. – 2. – №2. – P. 26-36.

[4] *Ham F.M., Kostanic I.* Principles of Neurocomputing for Science & Engineering. – N.Y.: McGraw-Hill, Inc., – 2001. – 642 p.

[5] *Sanger T.D.* Optimal unsupervised learning in a single-layer linear feedforward neural network // Neural Network. – 1989. – 2. – P.459-473.

[6] *Bannour S., Azimi-Sadjadi M.R.* Principal component extraction using recursive least squares learning // IEEE Trans. on Neural Network. – 1995. – 6. – P. 457-468.

[7] *Yang B.,* Projection approximation subspace tracking // IEEE Trans. on Signal Processing. – 1995. – 43. – P. 95-107.

[8] *Oja E., Ogawa H., Wangviwattana J.* Principal component analysis by homogeneous neural networks – Part 1: Weghted subspace criterion // IEICE Trans. Inform. Syst. – 1992. – E75-D. – №3. – P. 366-375.

[9] *Xu L.* Least mean square error reconstruction principal components for self-organizing neural nets // Neural Networks. – 1993. – 6. – P. 627–648.

[10] *Kung S.Y., Diamantras K.I., Taur J.S.* Adaptive principal component extraction (APEX) and applications // IEEE Trans. on Signal Processing. – 1994. – 42. – P.1202-1217.

[11] *Miao Y.F., Hua Y.B.* Fast subspace tracking and neural networks learning by a novel information criterion // IEEE Trans. on Signal Processing. – 1998. – 46. – P. 1962-1979.

[12] *Ouyang S., Bao Z.* Fast principal component extraction by a weighted information criterion // IEEE Trans. on Signal Processing. – 2002. – 50. – P. 1994-2002.

Поступила в редколлегию 29.05.2008

Машталир Сергей Владимирович, канд. техн. наук, доцент кафедры информатики Харьковского национального университета радиоэлектроники. Область научных интересов: методы обработки и распознавания изображений.

Сакало Евгений Сергеевич, аспирант, ассистент кафедры системотехники Харьковского национального университета радиоэлектроники. Область научных интересов: методы обработки и распознавания изображений.