УДК 004.032.26

Б.А. СКОРОХОД

АЛГОРИТМЫ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ И НЕЙРО-ФАЗЗИ СИСТЕМ С СЕПАРАБЕЛЬНОЙ СТРУКТУРОЙ

Аннотация. Рассматриваются задачи обучения нейронных сетей и нейро-фаззи систем, приводящие к сепарабельным моделям — структурам, нелинейным относительно некоторых неизвестных параметров и линейным относительно других неизвестных. Предлагаются новые алгоритмы их обучения, в основе которых — нелинейная оптимизационная задача, включающая априорную информацию только о нелинейно входящих параметрах. Предполагается, что она может быть получена по обучающему множеству, распределению генерирующей выборки или лингвистической информации. Для решения задачи используются метод Гаусса–Ньютона с линеаризацией в окрестности последней оценки, асимптотические представления псевдоинверсий возмущенных матриц и сепарабельная структура моделей. Полученные алгоритмы обладают рядом важных свойств: не требуется подбора начальных значений для линейно входящих параметров, который может приводить к расходимости, но при этом нет необходимости находить частные производные от проекционной матрицы; могут быть использованы в режимах последовательной и пакетной обработки; как частный случай, из них следуют известные алгоритмы, а моделирование показывает, что разработанные алгоритмы могут превосходить известные по точности и скорости сходимости.

Ключевые слова: сепарабельная регрессия, нейронные сети, нейро-фаззи системы, алгоритмы обучения.

ВВЕДЕНИЕ

Рассматриваются задачи обучения нейронных сетей (НС) и нейро-фаззи систем (НФС), приводящие к сепарабельным моделям — структурам, нелинейным относительно некоторых неизвестных параметров и линейным относительно других, например таких, как многослойный персептрон (МП) с линейной функцией активации (ФА) на выходе, радиально-базисная НС (РБНС) или НФС Сугено.

Известно, что задача обучения представляет собой сложную, многоэкстремальную, часто плохо обусловленную нелинейную оптимизационную проблему [1]. Для ее решения были разработаны различные алгоритмы, превосходящие алгоритм обратного распространения ошибки и его многочисленные модификации по скорости сходимости, точности аппроксимации и способности к обобщению. В настоящей статье предлагаются алгоритмы, в их основе — использование специальной структуры (сепарабельной) моделей, разработке которых посвящены многочисленные публикации. Так, в [2] предложен VP-алгоритм (variable projection), в соответствии с которым исходная задача оптимизации преобразуется к новой задаче, но уже относительно только нелинейно входящих параметров. При выполнении определенных условий множества стационарных точек двух задач совпадают, однако при этом уменьшается размерность и, как следствие, исключается необходимость подбора начальных значений для линейно входящих параметров. Более того, новая оптимизационная задача лучше обусловлена [3–6], и если для оптимизации исходной и преобразованной задач используется один и тот же метод, то VP-алгоритм всегда сходится за меньшее количество итераций. Вместе с тем VP-алгоритм может быть реализован только в пакетном режиме; кроме того, существенно усложняется процедура определения частных производных преобразованного критерия по параметрам. В [7] предложена двухстадийная процедура обучения НФС, основанная на последовательном использо-

вании на каждой итерации рекуррентного метода наименьших квадратов (РМНК) для определения линейно входящих параметров и градиентного метода для нелинейных параметров. Двухстадийные (гибридные) алгоритмы обучения различных нейросетевых моделей с использованием метода Левенберга-Марквардта разработаны в [8–10]. В работе [11] представлены ELM-алгоритмы (extreme learning machine), с помощью которых обучаются только линейно входящие параметры, а нелинейно входящие параметры выбираются случайно, без учета обучающей выборки. Известно, что это может приводить к невысокой точности аппроксимации при относительно небольшом размере обучающего множества [12, 13]. Отметим также, что при инициализации гибридных и ELM-алгоритмов, использующих РМНК, необходим подбор начальных значений для матрицы, удовлетворяющей уравнению Риккати. Более того, поскольку априорная информация об оцениваемых параметрах отсутствует, ее элементы обычно полагаются пропорциональными большому параметру (например, 10^4 в [9] и $10^8 - 10^{16}$ в [14, с. 240]), что может приводить к расходимости даже в случае линейного РМНК [15]. В работе [13] исследована асимптотика расширенного фильтра Калмана с диффузной инициализацией и предложены робастные (по отношению к неизвестной априорной информации о линейно входящих параметрах) алгоритмы обучения МП. В отличие от VP-алгоритма они могут быть использованы в режимах пакетной и последовательной обработки и, кроме того, являются более простыми с точки зрения программной реализации.

В настоящей статье предполагается, что помимо обучающего множества задана априорная информация только о нелинейно входящих параметрах, которая может быть получена по распределению генерирующей выборки, обучающему множеству или лингвистической информации [11, 16]. Рассматривается задача минимизации квадратического критерия качества, включающего только эту информацию. Такая постановка задачи и использование метода Гаусса–Ньютона (ГН) приводят к новому алгоритму обучения, робастному по отношению к неизвестной априорной информации о линейно входящих параметрах. Показано, что ELM-алгоритмы и двухстадийные процедуры с робастной инициализацией следуют из предложенного алгоритма как частный случай и устанавливается его связь с диффузной постановкой задачи обучения.

1. СЕПАРАБЕЛЬНЫЕ МОДЕЛЬНЫЕ СТРУКТУРЫ

Рассмотрим нелинейную регрессию вида

$$y_t = \Phi(z_t, \beta)\alpha, \ t = 1, \dots, N, \tag{1}$$

где $z_t = (z_{1t}, \dots, z_{nt})^{\mathrm{T}} \in R^n$ — вектор входов, $y_t = (y_{1t}, \dots, y_{mt})^{\mathrm{T}} \in R^m$ — вектор выходов, $\alpha = (\alpha_1, \dots, \alpha_r)^{\mathrm{T}} \in R^r$, $\beta = (\beta_1, \dots, \beta_l)^{\mathrm{T}} \in R^l$ — векторы неизвестных параметров, $\Phi(z_t, \beta)$ — $(m \times r)$-матрица заданных нелинейных функций, R^l — пространство векторов размерности l, $(\cdot)^{\mathrm{T}}$ — операция транспонирования матрицы, N — размер обучающей выборки.

Покажем, что в зависимости от выбора матричной функции $\Phi(\cdot,\cdot)$ можно получить различные, хорошо известные типы НС и НФС. Многослойный параметр с линейной ФА на выходе описывается выражениями [1]

$$y_{it} = \sum_{k=1}^{p} w_{ik} \sigma\left(\sum_{j=1}^{q} a_{kj} z_{jt} + b_k \right), \ i = 1, \dots, m, \ t = 1, \dots, N, \tag{2}$$

где a_{kj}, b_k, $k=1,\dots,p$, $j=1,\dots,q$, — веса и смещения скрытого слоя, w_{ik}, $i=1,2,\dots,m$, $k=1,2,\dots,p$, — веса выходного слоя, $\sigma(x)$ — ФА скрытого слоя. Используя векторно-матричные обозначения, получаем

$$y_t = W(\sigma(a_1 z_t),\dots,\sigma(a_p z_t))^{\mathrm{T}} = \Phi(z_t,\beta)\alpha,\ t=1,\dots,N, \tag{3}$$

при этом $y_t = (y_{1t},\dots,z_{mt},1)^{\mathrm{T}} \in R^m$, $z_t = (z_{1t},\dots,z_{qt},1)^{\mathrm{T}} \in R^{q+1}$, $W = (w_1^{\mathrm{T}},\dots,w_m^{\mathrm{T}})^{\mathrm{T}} \in R^{m\times p}$, $w_i = (w_{i1},\dots,w_{ip})$, $i=1,2,\dots,m$, $a_k = (a_{k1},\dots,a_{kq},b_k)$, $k=1,\dots,p$, $\alpha = (w_1,\dots,w_m)^{\mathrm{T}} \in R^{mp}$, $\beta = (a_1^{\mathrm{T}},\dots,a_p^{\mathrm{T}})^{\mathrm{T}} \in R^{(q+1)p}$, $\Phi(z_t,\beta) = I_m \otimes \Sigma(z_t,\beta)$ — $(m\times mp)$-матрица, $\Sigma(z_t,\beta) = (\sigma(a_1 z_t),\dots,\sigma(a_p z_t))$, I_m — единичная $(m\times m)$-матрица, $\otimes$ — прямое произведение двух матриц.

Для РБНС соотношения, связывающие входы и выходы, задаются выражениями [1]

$$y_{it} = \sum_{k=1}^{p} w_{ik}\varphi(b_k \| z_t - a_k \|^2) + w_{i0},\ i=1,\dots,m,\ t=1,\dots,N, \tag{4}$$

где $z_t = (z_{1t},\dots,z_{nt})^{\mathrm{T}} \in R^n$; $a_k \in R^n$, $b_k \in R^+$, $k=1,\dots,p$, — центры и шкалирующие множители соответственно; R^+ — множество положительных действительных чисел; w_{ik}, $i=1,2,\dots,m$, $k=1,2,\dots,p$, — веса выходного слоя; w_{i0}, $i=1,2,\dots,m$, — смещения; $\varphi(.)$ — базисная функция. Используя векторно-матричные обозначения, получаем

$$y_t = W(1,\varphi(b_1\| z_t - a_1\|^2),\dots,\varphi(b_p\| z_t - a_p\|^2))^{\mathrm{T}} = \Phi(z_t,\beta)\alpha,\ t=1,\dots,N,$$

где $W = (w_1^{\mathrm{T}},\dots,w_m^{\mathrm{T}})^{\mathrm{T}} \in R^{m\times p}$; $w_i = (w_{i1},\dots,w_{ip})$, $i=1,2,\dots,m$; $\alpha = (w_1,\dots,w_m)^{\mathrm{T}} \in$ $\in R^{m(p+1)}$, $\beta = (b_1,\dots,b_p,a_1^{\mathrm{T}},\dots,a_p^{\mathrm{T}})^{\mathrm{T}} \in R^{(n+1)p}$; $\Phi(z_t,\beta) = I_m \otimes \Sigma(z_t,\beta)$ — $(m\times m(p+1))$-матрица, $\Sigma(z_t,\beta) = (1,\varphi(b_1\| z_t - a_1\|^2),\dots,\varphi(b_p\| z_t - a_p\|^2))$.

Пусть задана база знаний в форме Сугено

$$\text{if } (z_{1t} \text{ is } A_{1j}) \text{ and } \dots \text{ if } (z_{nt} \text{ is } A_{nj}) \text{ then } y_t = a_j^{\mathrm{T}} z_t + b_j,\ j=1,\dots,m, \tag{5}$$

где A_{ij}, $i=1,\dots,n$, $j=1,\dots,m$, — нечеткие множества (НМ) с параметризованными функциями принадлежности (ФП) $\mu_{ij}(z_{it},c_{ij})$, c_{ij}, $i=1,\dots,n$, $j=1,\dots,m$, — векторы неизвестных параметров, $a_j \in R^n$, $b_j \in R^1$, $j=1,\dots,m$. Тогда соотношение для выхода, определяемого по m правилам (5), представляется выражением [7]

$$y_t = \frac{\sum_{i=1}^{m}\prod_{j=1}^{n}\mu_{ij}(z_{jt},c_{ij})(a_i^{\mathrm{T}} z_t + b_i)}{\sum_{i=1}^{m}\prod_{j=1}^{n}\mu_{ij}(z_{jt},c_{ij})},\ t=1,\dots,N. \tag{6}$$

Отсюда следует представление (1) при $\alpha = (a_1^{\mathrm{T}},\dots,a_m^{\mathrm{T}},b_1,\dots,b_m)^{\mathrm{T}} \in R^{2m}$, $\beta = (c_{11}^{\mathrm{T}},\dots,c_{nm}^{\mathrm{T}})^{\mathrm{T}}$, $\Phi(z_t,\beta) = (q_1(\beta)z_t^{\mathrm{T}},\dots,q_m(\beta)z_t^{\mathrm{T}},q_1(\beta),\dots,q_m(\beta))^{\mathrm{T}}$,

$$q_i(\beta) = \frac{\prod_{j=1}^{n}\mu_{ij}(z_{jt},c_{ij})}{\sum_{i=1}^{m}\prod_{j=1}^{n}\mu_{ij}(z_{jt},c_{ij})},\ i=1,\dots,m.$$

Модельная структура (1) (сепарабельная регрессия [2]) линейна относительно вектора параметров α и нелинейна относительно β. Исходя из нейросетевой интерпретации это архитектура прямого распространения с одним скрытым слоем, n входами и m выходами [1, 16]. Нас интересует задача оценки параметров α и β по обучающему множеству $\{z_t, y_t\}$, $t = 1, \ldots, N$.

2. РЕШЕНИЕ СЕПАРАБЕЛЬНОЙ ОПТИМИЗАЦИОННОЙ ЗАДАЧИ

Будем полагать выполненными следующие условия относительно параметров α и β, входящих в описание (1).

1. Для вектора β известно вероятное значение $\overline{\beta}$, и возможные отклонения от $\overline{\beta}$ характеризуются функцией $\Gamma(\beta) = (\beta - \overline{\beta})^{\mathrm{T}} \overline{P}_{\beta}^{-1} (\beta - \overline{\beta})$, где $\overline{P}_{\beta} > 0$ — заданная положительно-определенная матрица.

2. Априорная информация относительно компонент вектора α отсутствует; они могут быть либо неизвестными постоянными, либо случайными величинами, статистические характеристики которых неизвестны.

Пусть задано обучающее множество $\{z_t, y_t\}$, $t = 1, \ldots, N$, и критерий качества в момент t определяется выражением

$$J_t(x) = \sum_{k=1}^{t} \lambda^{t-k} (y_k - \Phi(z_k, \beta)\alpha)^{\mathrm{T}} (y_k - \Phi(z_k, \beta)\alpha) + \lambda^t \Gamma(\beta), \ t = 1, \ldots, N, \quad (7)$$

где $\lambda \in (0,1]$ — параметр забывания, позволяющий уменьшить влияние предыдущих наблюдений, $x = (\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$. Вектор параметров $x \in R^{l+r}$ находится из условия минимума $J_t(x)$ и должен обновляться после поступления нового измерения.

Нам потребуется несколько вспомогательных утверждений, чтобы получить основной результат работы — теорему 1.

Лемма 1. Справедливо тождество

$$M_t^{+} \widetilde{C}_t^{\mathrm{T}} = [(I_{l+r} - K_t C_t) M_{t-1}^{+} \widetilde{C}_{t-1}, K_t], \ t = 1, \ldots, N, \quad (8)$$

где $C_t \in R^{m \times (l+r)}$, $\widetilde{C}_t = (\lambda^{(t-1)/2} C_1^{\mathrm{T}}, \ldots, C_t^{\mathrm{T}})^{\mathrm{T}} \in R^{mt \times (l+r)}$, $M_t^{+} = (\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M)^{+} \in$

$\in R^{(l+r)\times(l+r)}$, $K_t = M_t^{+} C_t^{\mathrm{T}} \in R^{(l+r)\times m}$, $M = \begin{pmatrix} \overline{P}_{\beta}^{-1} & 0_{l\times r} \\ 0_{r\times l} & 0_{r\times r} \end{pmatrix} \in R^{(l+r)\times(l+r)}$,

$\overline{P}_{\beta} \in R^{l\times l}$, $0_{l\times r}$ — $(l \times r)$-матрица с нулевыми элементами, $(.)^{+}$ — псевдообратная матрица соответствующей матрицы.

Доказательство. Так как $M_t^{+} \widetilde{C}_t^{\mathrm{T}} = M_t^{+} (\widetilde{C}_{t-1}^{\mathrm{T}}, C_t^{\mathrm{T}})^{\mathrm{T}}$, то последние $l+r$ столбцов в левой и правой частях (8) совпадают. Покажем, что

$$M_t^{+} \widetilde{C}_{t-1}^{\mathrm{T}} = (I_{l+r} - K_t C_t) M_{t-1}^{+} \widetilde{C}_{t-1}^{\mathrm{T}}.$$

Представляя это выражение в эквивалентной форме

$$[M_t^{+} (I_{l+r} - M_{t-1} M_{t-1}^{+}) - (I_{l+r} - M_t^{+} M_t) M_{t-1}^{+}] \widetilde{C}_{t-1}^{\mathrm{T}} = 0, \quad (9)$$

покажем, что

$$(I_{l+r} - M_t^{+} M_t) M_{t-1}^{+} = 0, \quad (10)$$

$$M_t^{+} (I_{l+r} - M_{t-1} M_{t-1}^{+}) \widetilde{C}_{t-1}^{\mathrm{T}} = 0. \quad (11)$$

Пусть матрица $Q_t = (\widetilde{C}_t^{\mathrm{T}}, (\lambda^t M)^{1/2})$ имеет ранг $q(t)$ и $l_{1,t}, \dots, l_{q(t),t}$ — ее произвольные, линейно независимые столбцы. Линейное пространство $C(Q_t)$, определяемое ими, совпадает с пространством, образованным столбцами M_t, что обеспечивает $C(M_{t-1}) \subseteq C(M_t)$. С использованием скелетного разложения матриц имеем $(\widetilde{C}_t^{T}, (\lambda^t M)^{1/2}) = L_t \Gamma_t$, где $L_t = (l_{1,t}, \dots, l_{q(t),t})$, $\operatorname{rank}(\Gamma_t) = q(t)$, $\operatorname{rank}(.)$ — ранг соответствующей матрицы.

Имеем $M_t = L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}}$, где $\widetilde{\Gamma}_t = \Gamma_t \Gamma_t^{\mathrm{T}}$. Так как L_t — матрица полного ранга по столбцам, то $L_t^{+} = (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}}$ и

$$M_t^{+} = (L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}})^{+} = (L_t^{\mathrm{T}})^{+} \widetilde{\Gamma}_t^{-1} (L_t)^{+}, \; M_t^{+} M_t = L_t (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}},$$

$$(I_{l+r} - L_t (L_t^{\mathrm{T}} L_t)^{-1} L_t^{\mathrm{T}}) L_{t-1} = 0,$$

что обусловливает (10). Докажем равенство (11). Поскольку существует матрица $\overline{\Gamma}_t$ такая, что $\widetilde{C}_{t-1}^{\mathrm{T}} = L_{t-1} \overline{\Gamma}_t$, то

$$M_t^{+} (I_{l+r} - M_{t-1} M_{t-1}^{+}) \widetilde{C}_{t-1}^{\mathrm{T}} = M_t^{+} (I_{l+r} - L_{t-1} (L_{t-1}^{\mathrm{T}} L_{t-1})^{-1} L_{t-1}^{\mathrm{T}}) L_{t-1} \overline{\Gamma}_t = 0.$$

Лемма 2. Существует асимптотическое представление

$$(\varepsilon U_t + M_t)^{-1} = 1/\varepsilon A_{-1} + A_0 + O(\varepsilon), \; \varepsilon \to 0, \; t = 1, \dots, N, \tag{12}$$

где $\varepsilon > 0$ — малый параметр, в котором

$$A_0 = M_t^{+} = (\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M)^{+}, \; U_t = \begin{pmatrix} 0_{l \times l} & 0_{l \times r} \\ 0_{r \times l} & \lambda^t I_r \end{pmatrix} \in R^{(l+r) \times (l+r)}.$$

Доказательство. Представление (12) следует из разложения [17]

$$(H^{\mathrm{T}} H + 1/\varepsilon G^{\mathrm{T}} G)^{+} = (\overline{H} H)^{+} + \varepsilon (I - \overline{H}^{+} H)(G^{\mathrm{T}} G)^{+} (I - \overline{H}^{+} H)^{\mathrm{T}} + O(\varepsilon^2), \tag{13}$$

где H и G — произвольные матрицы соответствующих размерностей, I — единичная матрица, $\overline{H} = H(I - G^{+} G) = H(I - (G^{\mathrm{T}} G)^{+} G^{\mathrm{T}} G)$ при $H = U_t$, $G = M_t^{1/2}$.

Данное утверждение выполняется при условии, что

$$M_t^{+} = (I_{l+r} - \overline{H}_t^{+} U_t) M_t^{+} (I_{l+r} - \overline{H}_t^{+} U_t)^{\mathrm{T}}, \tag{14}$$

где $\overline{H}_t = U_t (I_{l+r} - M_t^{+} M_t)$, или, что то же самое,

$$\overline{H}_t^{+} U_t M_t^{+} = 0. \tag{15}$$

Найдем вначале представление для $\overline{H}_t^{+}$. Пусть $T_t = (t_{1t}, \dots, t_{l+rt})$ — ортогональная матрица такая, что $M_t = T_t \Lambda_t T_t^{\mathrm{T}}$, где $\Lambda_t = \operatorname{diag}(\lambda_{1t}, \dots, \lambda_{l+rt})$, $\operatorname{diag}(.)$ — диагональная матрица, λ_{it}, $i = 1, \dots, l+r$, — собственные числа матрицы M_t. Так как $M_t^{+} = T_t \Lambda_t^{+} T_t^{\mathrm{T}}$, то $\overline{H}_t = U_t T_t (I_{l+r} - \Lambda_t^{+} \Lambda_t) T_t^{\mathrm{T}}$. Определим структуру матрицы T_t. Имеем

$$\operatorname{rank}(M_t) = \operatorname{rank}(\lambda^t M + (\widetilde{C}_t^{\beta})^{\mathrm{T}} \widetilde{C}_t^{\beta}) + \operatorname{rank}(S_t),$$

где $S_t = (\widetilde{C}_t^{\alpha})^{\mathrm{T}} \widetilde{C}_t^{\alpha} - (\widetilde{C}_t^{\beta})^{\mathrm{T}} \widetilde{C}_t^{\alpha} (I_l + (\widetilde{C}_t^{\beta})^{\mathrm{T}} \widetilde{C}_t^{\beta})^{-1} (\widetilde{C}_t^{\alpha})^{\mathrm{T}} \widetilde{C}_t^{\beta}$, $\widetilde{C}_t = (\widetilde{C}_t^{\beta}, \widetilde{C}_t^{\alpha})$,

$\widetilde{C}_t^{\beta} \in R^{m\times l}$, $\widetilde{C}_t^{\alpha} \in R^{m\times r}$. Отсюда следует, что $q(t) = \operatorname{rank}(M_t) \geq l$. Если $\operatorname{rank}(M_t) = l+r$, то утверждение леммы очевидно. Пусть $l \leq q(t) < l+r$. Собственные векторы матрицы M_t (столбцы матрицы T_t), соответствующие нулевым собственным числам, определяются из системы $M_t x_t = 0$. Используя блочное представление $x_t = (x_{1t}^{\mathrm{T}}, x_{2t}^{\mathrm{T}})^{\mathrm{T}}$, получаем

$$\lambda^t \overline{P}_{\beta}^{-1} x_{1t} + (\widetilde{C}_t^{\beta})^{\mathrm{T}} \widetilde{C}_t x_t = 0,\ (\widetilde{C}_t^{\alpha})^{\mathrm{T}} \widetilde{C}_t x_t = 0.$$

В результате имеем $x_{1t} = 0$, $x_{2t} = (I_r - (\widetilde{C}_t^{\alpha})^{+} \widetilde{C}_t^{\alpha}) f$, где $f \in R^r$ — произвольный вектор. Отсюда без ограничения общности, полагая, что $\Lambda_t = \operatorname{diag}(\lambda_{1t}, \ldots, \lambda_{q(t)t}, 0, \ldots, 0)$, $\lambda_{it} > 0$, $i = 1, \ldots, q(t)$, получим

$$T_t = \begin{pmatrix} T_{1t} & 0_{l\times(l+r-q(t))} \\ T_{2t} & T_{3t} \end{pmatrix},\ T_{1t} \in R^{l\times q(t)},\ T_{2t} \in R^{r\times q(t)},\ T_{3t} \in R^{r\times(l+r-q(t))}.$$

Так как матрица $M_t^{+} M_t$ идемпотентна, то

$$T_t (I_{l+r} - \Lambda_t^{+} \Lambda_t) T_t^{\mathrm{T}} = \begin{pmatrix} 0_{l\times l} & 0_{l\times r} \\ 0_{r\times l} & T_{3t} T_{3t}^{\mathrm{T}} \end{pmatrix},$$

отсюда следует $\overline{H}_t^{+} U_t M_t^{+} = U_t T_t (I_{l+r} - \Lambda_t^{+} \Lambda_t) T_t^{\mathrm{T}} U_t T_t \Lambda_t^{+} T_t^{\mathrm{T}} = 0$.

Лемма 3. Справедливо представление

$$M_t^{+} = (\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M)^{+} = \widetilde{S}_t + \widetilde{V}_t W_t^{+} \widetilde{V}_t^{\mathrm{T}},\ t = 1, \ldots, N, \tag{16}$$

$$\widetilde{S}_t = \begin{pmatrix} S_t & 0_{l\times r} \\ 0_{r\times l} & 0_{r\times r} \end{pmatrix},\ \widetilde{V}_t = \begin{pmatrix} V_t \\ I_r \end{pmatrix},\ S_t \in R^{l\times l},\ V_t \in R^{l\times r}, \tag{17}$$

$$S_t = S_{t-1}/\lambda - S_{t-1}(C_t^{\beta})^{\mathrm{T}} (\lambda I_m + C_t^{\beta} S_{t-1} (C_t^{\beta})^{\mathrm{T}})^{-1} C_t^{\beta} S_{t-1}/\lambda,\ S_0 = \overline{P}_{\beta}, \tag{18}$$

$$V_t = (I_l - S_{t-1}(C_t^{\beta})^{\mathrm{T}} (\lambda I_m + C_t^{\beta} S_{t-1} (C_t^{\beta})^{\mathrm{T}})^{-1} C_t^{\beta}) V_{t-1} -$$
$$- S_{t-1}(C_t^{\beta})^{\mathrm{T}} (\lambda I_m + C_t^{\beta} S_{t-1} (C_t^{\beta})^{\mathrm{T}})^{-1} C_t^{\alpha},\ V_0 = 0_{l\times r}, \tag{19}$$

$$W_t = \lambda W_{t-1} + \lambda (C_t^{\beta} V_{t-1} + C_t^{\alpha})^{\mathrm{T}} (\lambda I_m + C_t^{\beta} S_{t-1} (C_t^{\beta})^{\mathrm{T}})^{-1} \times$$
$$\times (C_t^{\beta} V_{t-1} + C_t^{\alpha}),\ W_0 = 0_{r\times r}. \tag{20}$$

Доказательство. Матрица $P_t^{-1} = \varepsilon U_t + M_t$ удовлетворяет разностному уравнению

$$P_t^{-1} = \lambda P_{t-1}^{-1} + C_t^{\mathrm{T}} C_t,\ P_0^{-1} = \operatorname{block\,diag}(\overline{P}_{\beta}^{-1}, \varepsilon I_r),\ t = 1, \ldots, N.$$

Воспользуемся матричным тождеством

$$A^{-1} = (B^{-1} + CD^{-1}C^{\mathrm{T}})^{-1} = B - BC(D + C^{\mathrm{T}} BC)^{-1} C^{\mathrm{T}} B \tag{21}$$

для определения P_t. Полагая $B = P_{t-1}$, $D = \lambda I_m$, $C = C_t^{\mathrm{T}}$, получаем

$$P_t = (\lambda P_{t-1}^{-1} + C_t^{\mathrm{T}} C_t)^{-1} = (P_{t-1} - P_{t-1} C_t^{\mathrm{T}} (\lambda I_m + C_t P_{t-1} C_t^{\mathrm{T}})^{-1} C_t P_{t-1})/\lambda =$$

$$= P_{t-1}/\lambda - 1/\lambda^2 P_{t-1} C_t^{\mathrm{T}} (I_m + 1/\lambda C_t P_{t-1} C_t^{\mathrm{T}})^{-1} C_t P_{t-1}, \tag{22}$$

$$P_0 = \operatorname{block\,diag}(\overline{P}_\beta, I_r/\varepsilon),\ t = 1, \ldots, N.$$

Обозначим $A = I_{r+l}/\lambda^{1/2}$, $\widetilde{C}_t = C_t/\lambda^{1/2}$. Тогда (22) примет вид

$$P_t = AP_{t-1}A - AP_{t-1}\widetilde{C}_t^{\mathrm{T}} (I_m + \widetilde{C}_t P_{t-1} \widetilde{C}_t^{\mathrm{T}})^{-1} \widetilde{C}_t P_{t-1} A. \tag{23}$$

Пусть $\widetilde{P}_t$ и $\overline{P}_t$ — два произвольных решения этого уравнения. Тогда разность $Q_t = \widetilde{P}_t - \overline{P}_t$ будет удовлетворять уравнению [18]

$$Q_t = (A_t Q_{t-1} A_t^{\mathrm{T}} - A_t Q_{t-1} \widetilde{C}_t^{\mathrm{T}} (I_m + \widetilde{C}_t \widetilde{P}_{t-1} \widetilde{C}_t^{\mathrm{T}})^{-1} \widetilde{C}_t Q_{t-1} A_t^{\mathrm{T}}) =$$

$$= (A_t Q_{t-1} A_t^{\mathrm{T}} - A_t Q_{t-1} C_t^{\mathrm{T}} (\lambda I_m + C_t \widetilde{P}_{t-1} C_t^{\mathrm{T}})^{-1} C_t Q_{t-1} A_t^{\mathrm{T}})/\lambda, \tag{24}$$

$$Q_0 = \widetilde{P}_0 - \overline{P}_0,\ t = 1, \ldots, N,$$

где

$$\widetilde{A}_t = A - A\widetilde{P}_{t-1} \widetilde{C}_t^{\mathrm{T}} (I_m + \widetilde{C}_t \overline{P}_{t-1} \widetilde{C}_t^{\mathrm{T}})^{-1} C_t,$$

$$A_t = I_{r+l} - \overline{P}_{t-1} C_t^{\mathrm{T}} (\lambda I_m + C_t \overline{P}_{t-1} C_t^{\mathrm{T}})^{-1} C_t. \tag{25}$$

Пусть $\widetilde{P}_0 = P_0$, $\overline{P}_0 = \operatorname{block\,diag}(\overline{P}_\beta, 0_{r\times r})$. Тогда $\widetilde{P}_t = P_t$, $\overline{P}_t = \widetilde{S}_t$,

где

$$\widetilde{S}_t = \widetilde{S}_{t-1}/\lambda - \widetilde{S}_{t-1} C_t^{\mathrm{T}} (I_m + 1/\lambda C_t \widetilde{S}_{t-1} C_t^{\mathrm{T}})^{-1} C_t \widetilde{S}_{t-1}/\lambda, \tag{26}$$

$$\widetilde{S}_0 = \operatorname{block\,diag}(\overline{P}_\beta, 0_{r\times r}).$$

Отсюда следует блочное представление для $\widetilde{S}_t$ и уравнение (18).

Покажем, что

$$Q_t = \widetilde{V}_t L_t^{-1} \widetilde{V}_t^{\mathrm{T}}, \tag{27}$$

где

$$L_t = L_{t-1}/\lambda + (C_t^\beta V_{t-1} + C_t^\alpha)^{\mathrm{T}} (\lambda I_m + C_t^\beta S_{t-1} (C_t^\beta)^{\mathrm{T}})^{-1} \times$$

$$\times (C_t^\beta V_{t-1} + C_t^\alpha)/\lambda,\ L_0^{-1} = I_r/\varepsilon. \tag{28}$$

Подставляя (27) в (24), получаем

$$\widetilde{V}_t L_t^{-1} \widetilde{V}_t^{\mathrm{T}} = (A_t \widetilde{V}_{t-1} L_{t-1}^{-1} \widetilde{V}_{t-1}^{\mathrm{T}} A_t^{\mathrm{T}} - A_t \widetilde{V}_{t-1} L_{t-1}^{-1} \widetilde{V}_{t-1}^{\mathrm{T}} C_t^{\mathrm{T}} \times$$

$$\times (\lambda I_m + C_t \widetilde{P}_{t-1} C_t^{\mathrm{T}})^{-1} C_t \widetilde{V}_{t-1} L_{t-1}^{-1} \widetilde{V}_{t-1}^{\mathrm{T}} A_t^{\mathrm{T}})/\lambda\,. \tag{29}$$

Это равенство будет выполняться, если L_t и $\widetilde{V}_t$ удовлетворяют уравнениям

$$\widetilde{V}_t = A_t \widetilde{V}_{t-1},\ \widetilde{V}_0 = (0_{r\times l}, I_r), \tag{30}$$

$$L_t^{-1} = (L_{t-1}^{-1} - L_{t-1}^{-1} \widetilde{V}_{t-1}^{\mathrm{T}} C_t^{\mathrm{T}} (\lambda I_m + C_t \widetilde{P}_{t-1} C_t^{\mathrm{T}})^{-1} C_t \widetilde{V}_{t-1} L_{t-1}^{-1})/\lambda,\ L_0^{-1} = I_r \varepsilon, \tag{31}$$

начальные условия для которых следуют из выражений (16), (18), (24). Так как

$$\overline{P}_{t-1}C_t^{\mathrm{T}} = \widetilde{S}_{t-1}C_t^{\mathrm{T}} = \begin{pmatrix} \widetilde{S}_{t-1}(C_t^{\beta})^{\mathrm{T}} \\ 0_{r\times l} \end{pmatrix},\ C_t\widetilde{P}_{t-1}C_t^{\mathrm{T}} = C_t^{\beta}S_{t-1}(C_t^{\beta})^{\mathrm{T}},$$

$$A_t = I_{r+l} - \begin{pmatrix} S_{t-1}(C_t^{\beta})^{\mathrm{T}}N_t^{-1}C_t^{\beta} & S_{t-1}(C_t^{\beta})^{\mathrm{T}}N_t^{-1}C_t^{\alpha} \\ 0_{r\times l} & 0_{r\times r} \end{pmatrix},$$

где $N_t = \lambda I_m + C_t^{\beta}S_{t-1}(C_t^{\beta})^{\mathrm{T}}$, то отсюда следует (19).

Преобразуем (31), используя матричное тождество (21). Пусть

$$B = L_{t-1},\ C = \widetilde{V}_{t-1}^{\mathrm{T}}C_t^{\mathrm{T}},\ D = \lambda I_m + C_t^{\beta}S_{t-1}(C_t^{\beta})^{\mathrm{T}}.$$

Тогда, учитывая, что $\widetilde{P}_t = Q_t - \widetilde{S}_t$, получаем

$$L_t^{-1} = (L_{t-1} - \widetilde{V}_{t-1}^{\mathrm{T}}C_t^{\mathrm{T}}(\lambda I_m + C_t\widetilde{S}_{t-1}C_t^{\mathrm{T}})^{-1}C_t\widetilde{V}_{t-1})^{-1}/\lambda.$$

Но так как $C_t\widetilde{V}_{t-1} = C_t^{\beta}V_{t-1} + C_t^{\alpha}$, то отсюда следует (28).

Используя лемму 3 (см. [13]), из (28) находим

$$L_t^{-1} = 1/\varepsilon(I_r - W_tW_t^T) + W_t^+O(\varepsilon),\ \varepsilon\to 0.$$

Поскольку

$$P_t = (U_t + M_t)^{-1} = Q_t + \widetilde{S}_t = \widetilde{V}_tL_t^{-1}\widetilde{V}_t^{\mathrm{T}} + \widetilde{S}_t =$$

$$= 1/\varepsilon\widetilde{V}_t(I_r - W_tW_t^{\mathrm{T}})\widetilde{V}_t^{\mathrm{T}} + \widetilde{V}_tW_t^+\widetilde{V}_t^{\mathrm{T}} + \widetilde{S}_t + O(\varepsilon),\ \varepsilon\to 0, \tag{32}$$

(16) следует из леммы 2.

Рассмотрим вспомогательную линейную задачу оптимизации

$$(\alpha^*, \beta^*) = \arg\min J_t(\alpha,\beta)\ ,\ \alpha\in R^r,\ \beta\in R^l, \tag{33}$$

где

$$J_t(x) = \sum_{k=1}^{t}\lambda^{t-k}(y_k - C_k^{\beta}\beta - C_k^{\alpha}\alpha)^{\mathrm{T}}(y_k - C_k^{\beta}\beta - C_k^{\alpha}\alpha) + \lambda^t\Gamma(\beta),\ t=1,\ldots,N. \tag{34}$$

Лемма 4. Решение задачи (33), (34) может быть представлено в следующей рекуррентной форме:

$$\beta_t = \beta_{t-1} + K_t^{\beta}(y_t - C_t^{\beta}\beta_{t-1} - C_t^{\alpha}\alpha_{t-1}),\ \beta_0 = \overline{\beta}, \tag{35}$$

$$\alpha_t = \alpha_{t-1} + K_t^{\alpha}(y_t - C_t^{\beta}\beta_{t-1} - C_t^{\alpha}\alpha_{t-1}),\ \alpha_0 = 0_{r\times 1},\ t=1,\ldots,N, \tag{36}$$

где

$$K_t^{\beta} = (S_t + V_tW_t^+V_t^{\mathrm{T}})(C_t^{\beta})^{\mathrm{T}} + V_tW_t^+(C_t^{\alpha})^{\mathrm{T}}, \tag{37}$$

$$K_t^{\alpha} = W_t^+V_t^{\mathrm{T}}(C_t^{\beta})^{\mathrm{T}} + W_t^+(C_t^{\alpha})^{\mathrm{T}}. \tag{38}$$

Доказательство. Полагая $x = (\alpha^T,\ \beta^T)^T$, перейдем к более компактной форме представления критерия

$$J_t(x) = (Y_t - \widetilde{C}_tx)^{\mathrm{T}}(Y_t - \widetilde{C}_tx) + \lambda^t(x-\bar{x})^{\mathrm{T}}M(x-\bar{x}),\ t=1,\ldots,N, \tag{39}$$

где $\widetilde{C}_t = (\lambda^{(t-1)/2} C_1^{\mathrm{T}}, \dots, C_t^{\mathrm{T}})^{\mathrm{T}} \in R^{mt\times(l+r)}$, $Y_t = (\lambda^{(t-1)/2} y_1^{\mathrm{T}}, \dots, y_t^{\mathrm{T}})^{\mathrm{T}} \in R^{mt\times 1}$, $\bar{x} = (\bar{\beta}^{\mathrm{T}}, 0_{1\times r})^{\mathrm{T}}$, матрица M определена в лемме 1. Вводя замену $z = x - \bar{x}$, получаем

$$J_t(z+\bar{x}) = ||(\widetilde{Y}_t - \widetilde{C}_t z)||^2 + \lambda^t z^{\mathrm{T}} M z = \left\| \begin{pmatrix} \widetilde{Y}_t \\ 0_{(l+r)\times 1} \end{pmatrix} - \begin{pmatrix} \widetilde{C}_t \\ \lambda^{t/2} M^{1/2} \end{pmatrix} z \right\|^2,$$

где $\widetilde{Y}_t = Y_t - \widetilde{C}_t \bar{x}$, $||.||$ — евклидова норма вектора. Стационарные точки задачи определяются из системы нормальных уравнений

$$\begin{pmatrix} \widetilde{C}_t \\ \lambda^{t/2} M^{1/2} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \widetilde{C}_t \\ \lambda^{t/2} M^{1/2} \end{pmatrix} z = \begin{pmatrix} \widetilde{C}_t \\ \lambda^{t/2} M^{1/2} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \widetilde{Y}_t \\ 0_{(l+r)\times 1} \end{pmatrix}$$

или с использованием эквивалентной формы записи

$$(\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M) z = \widetilde{C}_t^{\mathrm{T}} \widetilde{Y}_t. \tag{40}$$

Будем рассматривать только решение с минимальной нормой

$$z^* = (\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M)^+ \widetilde{C}_t^{\mathrm{T}} \widetilde{Y}_t. \tag{41}$$

Покажем, что оно может быть найдено рекуррентно. Обозначим $z_t = z^*$. С использованием леммы 1 имеем

$$z_t = (\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M)^+ \widetilde{C}_t^{\mathrm{T}} (\widetilde{Y}_{t-1}^{\mathrm{T}}, \widetilde{y}_{t-1}^{\mathrm{T}})^{\mathrm{T}} =$$

$$= [(I_{l+r} - K_t C_t) M_{t-1}^+ \widetilde{C}_{t-1}, K_t] (\widetilde{Y}_{t-1}^{\mathrm{T}}, \widetilde{y}_t^{\mathrm{T}})^{\mathrm{T}} =$$

$$= (I_{l+r} - K_t C_t) z_{t-1} + K_t \widetilde{y}_t, \ z_0 = 0, \ t = 1, \dots, N,$$

где $K_t = M_t^+ C_t^{\mathrm{T}} = (\widetilde{C}_t^{\mathrm{T}} \widetilde{C}_t + \lambda^t M)^+ C_t^{\mathrm{T}}$.

Возвращаясь к исходной переменной $x_t = z_t + \bar{x}$, получим

$$x_t = x_{t-1} + K_t (y_t - C_t x_{t-1}), \ x_0 = (\bar{\beta}^{\mathrm{T}}, 0_{1\times r})^{\mathrm{T}}, \ t = 1, \dots, N. \tag{42}$$

Выражения (37) и (38) для K_t следуют из леммы 3.

Рассмотрим теперь решение нелинейной оптимизационной задачи с критерием (7).

Теорема 1. Решение задачи минимизации критерия (7) методом ГН может быть получено рекуррентно

$$x_t = x_{t-1} + K_t (y_t - h_t(x_{t-1})), \ x_0 = (\bar{\beta}^{\mathrm{T}}, 0_{1\times r})^{\mathrm{T}}, \ t = 1, \dots, N, \tag{43}$$

где $h_t(x_{t-1}) = \Phi(z_t, \beta_{t-1})\alpha_{t-1}$, $K_t = ((K_t^\beta)^{\mathrm{T}}, (K_t^\alpha)^{\mathrm{T}})^{\mathrm{T}}$ определяется в лемме 4 при

$$C_t^\beta = C_t^\beta(x_{t-1}) = \partial[\Phi(z_t, \beta_{t-1})\alpha_{t-1}]/\partial\beta_{t-1}, \ C_t^\alpha = C_t^\alpha(x_{t-1}) = \Phi(z_t, \beta_{t-1}). \tag{44}$$

Доказательство. При линеаризации невязок $e_t = y_t - \Phi(z_t, \beta)\alpha$ в окрестности точки x_{t-1} имеем

$$e_t = y_t - \Phi(z_t, \beta_{t-1})\alpha_{t-1} - \Phi(z_t, \beta_{t-1})(\alpha - \alpha_{t-1}) -$$

$$-\partial[\Phi(z_t, \beta_{t-1})\alpha_{t-1}]/\partial\beta_{t-1}(\alpha - \alpha_{t-1}) = \widetilde{y}_t - C_t x,$$

где $\widetilde{y}_t = y_t + C_t x_{t-1} - \Phi(z_t, \beta_{t-1})\alpha_{t-1}$, $x = (\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$.

Подстановка этого выражения в критерий (7) приводит к линейной оптимизационной задаче (33), (34), решение которой имеет вид (см. лемму 4)

$$x_t = x_{t-1} + K_t(\tilde{y}_t - C_t x_{t-1}),\ x_0 = (\bar{\beta}^{\mathrm{T}}, 0_{1\times r})^{\mathrm{T}},\ t = 1, \dots, N.$$

Отсюда следует (43).

Приведем итерационную модификацию этого алгоритма обучения, ориентированную на пакетную обработку,

$$x_t^{i-1} = x_{t-1}^{i-1} + K_t^i(y_t - h_t(x_{t-1}^{i-1})),\ t = 1, \dots, N,\ i = 1, 2, \dots, \tag{45}$$

с инициализацией

$$x_0^i = x_N^{i-1},\ S_0^i = S_N^{i-1},\ V_0^i = V_N^{i-1},\ W_0^i = W_N^{i-1},$$

$$x_N^0 = (\bar{\beta}^{\mathrm{T}}, 0_{1\times r})^{\mathrm{T}},\ S_N^0 = \bar{P}_\beta,\ V_N^0 = 0_{l\times r},\ W_N^0 = 0_{r\times r}.$$

Здесь i — номер итерации, $h_t(x_{t-1}^{i-1}) = \Phi(z_t, \beta_{t-1}^{i-1})\alpha_{t-1}^{i-1}$, а $K_t^i = ((K_t^\beta)^{\mathrm{T}}, (K_t^\alpha)^{\mathrm{T}})^{\mathrm{T}}$ определяется в лемме 4 при $C_t^\beta = C_t^\beta(x_{t-1}^{i-1}) = \partial[\Phi(z_t, \beta_{t-1}^{i-1})\alpha_{t-1}^{i-1}]/\partial\beta_{t-1}^{i-1}$, $C_t^\alpha = C_t^\alpha(x_{t-1}^{i-1}) = \Phi(z_t, \beta_{t-1}^{i-1})$.

3. ДИФФУЗНАЯ ИНИЦИАЛИЗАЦИЯ АЛГОРИТМОВ

Пусть при выводе алгоритма используется диффузная инициализация, т.е. α предполагается случайной величиной с $E\alpha = 0$, $E(\alpha^{\mathrm{T}}\alpha) = I_r/\varepsilon$, где $\varepsilon > 0$ — малый параметр, подбираемый в процессе моделирования, и α и β некоррелированы. Тогда в (43) $K_t = P_t C_t^{\mathrm{T}}$, где $C_t = (C_t^\beta, C_t^\alpha)$, а P_t определяется из (22).

Теорема 2. Для любого конечного N справедливы представления

$$P_t = \tilde{V}_t(I_{rm} - W_t W_t^+)\tilde{V}_t^{\mathrm{T}}/\varepsilon + \tilde{S}_t + \tilde{V}_t W_t^+ \tilde{V}_t^{\mathrm{T}} + O(\varepsilon), \tag{46}$$

$$K_t = (\tilde{S}_t + \tilde{V}_t W_t^+ \tilde{V}_t^{\mathrm{T}})C_t^{\mathrm{T}} + O(\varepsilon),\ C_t \neq 0,\ \varepsilon \to 0,\ t = 1, \dots, N. \tag{47}$$

Доказательство следует из выражения (16), установленного при доказательстве леммы 3, а также лемм 2, 3 из [13].

Из (46), (47) следует, что ошибки численной реализации алгоритма при $t < tr$, $tr = \min_t\{t: W_t > 0,\ t = 1, \dots, N\}$ могут приводить к расходимости алгоритма при малых значениях ε. Действительно, пусть δW_t^+ — ошибка, связанная с вычислением псевдообратной матрицы W_t^+. Тогда использование (46) дает

$$K_t = P_t C_t^{\mathrm{T}} = [\tilde{V}_t(I_{rm} - W_t(W_t^+ + \delta W_t^+))]\tilde{V}_t^{\mathrm{T}} C_t^{\mathrm{T}}/\varepsilon + O(1),\ \varepsilon \to 0,\ t = 1, \dots, N.$$

Указанные особенности численной реализации отсутствуют в приведенных в разд. 2 алгоритмах, поскольку в их конструкции отсутствуют диффузные компоненты — величины, пропорциональные большому параметру; более того, даже при $t \geq tr$ неоправданным может оказаться переход к использованию представления

$$P_t = P_{t-1}/\lambda - 1/\lambda^2 P_{t-1} C_t^{\mathrm{T}}(I_m + 1/\lambda C_t P_{t-1} C_t^{\mathrm{T}})^{-1} C_t P_{t-1},$$

так как матрица P_{tr} может быть по-прежнему плохо обусловленной и к тому же отличной от диагональной.

Отметим также еще одно важное преимущество предложенных в статье алгоритмов — нет необходимости подбирать величину ε.

4. СВЯЗЬ С ELM-ПОДХОДОМ

При удачном выборе априорной информации для β и параметров модели можно ожидать быструю сходимость алгоритма к одной из приемлемых точек минимума критерия. Рассмотрим сначала предельный случай: $\overline{P}_\beta = 0_{l\times l}$. Из теоремы 1 следует, что

$$\alpha_t = \alpha_{t-1} + K_t(y_t - C_t^\alpha \alpha_{t-1}),\ x_0 = 0_{r\times 1},\ t = 1,\dots,N, \tag{48}$$

где $K_t = W_t^+ (C_t^\alpha)^{\mathrm{T}}$, $W_t = \lambda W_{t-1} + (C_t^\alpha)^T C_t^\alpha$, $W_0 = 0_{r\times r}$.

Теорема 3. Пусть существует $tr = \min_t \{t: W_t > 0,\ t = 1,\dots,N\}$. Тогда ошибка оценивания вектора $e_t = \alpha_t - \alpha$ обращается в ноль при $t \ge tr$.

Доказательство. Вектор e_t удовлетворяет уравнению

$$e_t = (I_r - K_t C_t^\alpha) e_{t-1} = A_t e_t,\ e_0 = -\alpha. \tag{49}$$

Покажем, что его переходная матрица определяется выражением

$$\Phi_{t,s} = I_r - W_t^{-1} \sum_{i=s+1}^{t} \lambda^{t-i} (C_i^\alpha)^{\mathrm{T}} C_{is}^\alpha,\ t \ge tr,\ t > s. \tag{50}$$

Продемонстрируем вначале, что решения матричных уравнений

$$G_{t-1,s} = A_t^{\mathrm{T}} G_{t,s},\ G_{s,s} = I_r,\ s \ge tr,\ t < s, \tag{51}$$

$$Z_{t-1,s} = Z_{t,s} - (C_t^\alpha)^{\mathrm{T}} C_t^\alpha W_t^{-1} \lambda^{s-t},\ Z_{s,s} = I_r,\ s \ge tr,\ t < s, \tag{52}$$

совпадают. Имеем

$$Z_{t,s} = I_r - \sum_{i=t+1}^{s} \lambda^{s-i} (C_i^\alpha)^{\mathrm{T}} C_i^\alpha W_s^{-1}. \tag{53}$$

При сравнении правых частей (53) и (51) устанавливаем выполнимость условия $K_t^{\mathrm{T}} Z_{t,s} = C_t^\alpha W_s^{-1}$; подставляя в него (53), получаем

$$C_t^\alpha W_t^+ \left(I_r - \sum_{i=t+1}^{s} \lambda^{s-i} (C_i^\alpha)^{\mathrm{T}} C_i^\alpha W_s^{-1} \right) =$$

$$= C_t^\alpha W_t^+ \left(W_s - \sum_{i=t+1}^{s} \lambda^{s-i} (C_i^\alpha)^{\mathrm{T}} C_i^\alpha W_s^{-1} \right) W_s^{-1} = C_t^\alpha W_t^+ W_t W_s^{-1} = C_t^\alpha W_s^{-1}. \tag{54}$$

При выводе (54) использовались скелетное разложение матрицы $C_t = L_t \Gamma_t$ и выражение $W_t = L_t \widetilde{\Gamma}_t L_t^{\mathrm{T}}$, где $\widetilde{\Gamma}_t = \Gamma_t \Gamma_t^{\mathrm{T}}$. Так как $\Phi_{t,s} = A_t \dots A_{s+1}$ и $G_{t,s} = A_{t+1}^{\mathrm{T}} \dots A_s^{\mathrm{T}}$, то $\Phi_{t,s} = G_{s,t}^{\mathrm{T}}$. Отсюда следует (50), а также, что $\Phi_{t,0} = 0$ при $t \ge tr$.

Рассмотренный предельный случай иллюстрирует еще одну важную особенность алгоритма — начальный этап обучения ($t < tr$) может оказывать существенное влияние на количество итераций, необходимое для сходимости нелинейной задачи обучения.

Более интересной представляется ситуация, при которой значения нормы $\overline{P}_\beta$ отличны от нуля, но относительно невелики. В этом случае возможны различные варианты построении упрощенных алгоритмов обучения. Пусть, например, $\overline{P}_\beta = O(\varepsilon), \varepsilon \to 0$. Тогда $S_t = O(\varepsilon), V_t = O(\varepsilon), \varepsilon \to 0$ равномерно по t при конечных N. Пренебрегая членами второго порядка малости в (19), (20), получаем двухэтапный алгоритм обучения

$$x_t = x_{t-1} + K_t(y_t - h_t(x_{t-1})),\ x_0 = (\overline{\beta}^{\mathrm{T}}, 0_{1\times r})^{\mathrm{T}},\ t = 1, \ldots, N, \tag{55}$$

где $K_t = ((K_t^\beta)^{\mathrm{T}}, (K_t^\alpha)^{\mathrm{T}})^{\mathrm{T}},\ K_t^\beta = S_t(C_t^\beta)^{\mathrm{T}},\ K_t^\alpha = W_t^+(C_t^\alpha)^{\mathrm{T}},$

$$S_t = S_{t-1}/\lambda - S_{t-1}(C_t^\beta)^{\mathrm{T}}(\lambda I_m + C_t^\beta S_{t-1}(C_t^\beta)^{\mathrm{T}})^{-1} C_t^\beta S_{t-1}/\lambda,\ S_0 = \overline{P}_\beta, \tag{56}$$

$$W_t = \lambda W_{t-1} + (C_t^\alpha)^{\mathrm{T}} C_t^\alpha,\ W_0 = 0_{r\times r}. \tag{57}$$

Отсюда, полагая $S_t = \varepsilon I_l$, получаем аналог гибридного алгоритма, предложенного в [7] для НФС.

5. ВЫБОР ПАРАМЕТРА ЗАБЫВАНИЯ

Стандартные рекомендации по выбору λ состоят в следующем. На начальном этапе обучения величина λ может быть достаточно малой, обеспечивая этим высокую скорость сходимости. Далее предлагается постепенно ее увеличивать до единицы для достижения необходимой точности полученного решения. Один из вариантов выбора λ, обеспечивающих сходимость, предложен в [19]:

$$0 \le 1 - (\lambda_i)^N \le c/i,\ c > 0,\ i = 1, 2, \ldots \tag{58}$$

Опираясь на этот результат, приведем условия сходимости итерационного алгоритма, предложенного в работе. Пусть:

1) для некоторого $L > 0$ выполняется условие Липшица

$$||\nabla g_t(x) g_t(x) - \nabla g_t(y) g_t(y)|| \le L||x - y||\ \ \forall x, y \in R^{l+r},\ t = 1, \ldots, N,$$

где $g_t(x) = y_t - h_t(x)$, $\nabla g_t(x)$ есть $((l+r)\times m)$-матрица градиентов $g_t(x)$;

2) последовательность векторов $x_t^{i-1},\ t = 1, \ldots, N,\ i = 1, 2, \ldots,$ ограничена;

3) существует вектор x_t^{i-1} такой, что $W_t = W_t(x_t^{i-1}) > 0$.

Теорема 4. При выполнении условий (58), а также пп. 1–3 последовательность

$$J_t(x_N^i) = \sum_{k=1}^{N} (y_k - h_t(x_N^i)^{\mathrm{T}} (y_k - h_t(x_N^i))$$

сходится и каждая предельная точка последовательности $\{x_N^i\}$ является стационарной точкой $J_t(x)$.

Утверждение будет следовать из [19, proposition 2], если

$$P_t = P_t(x_t^{i-1}) = \begin{pmatrix} S_t + V_t W_t^{-1} V_t^{\mathrm{T}} & V_t W_t^{-1} \\ W_t^{-1} V_t^{\mathrm{T}} & W_t^{-1} \end{pmatrix} > 0.$$

Так как $S_t + V_t W_t^{-1} V_t^{\mathrm{T}} > 0$, то $|P_t| = |S_t + V_t W_t^{-1} V_t^{\mathrm{T}}| \times |W_t^{-1} - W_t^{-1} V_t^{\mathrm{T}} (S_t + V_t W_t^{-1} V_t^{\mathrm{T}})^{-1} V_t W_t^{-1}|$.

Преобразуя второй сомножитель в этом выражении с помощью тождества (21) и полагая, что $B = W_t^{-1}$, $C = V_t$, $D = S_t$, получим $|P_t| = |S_t + V_t W_t^{-1} V_t^{\mathrm{T}}| \times$ $\times |W_t + V_t S_t^{-1} V_t^{\mathrm{T}}| > 0$.

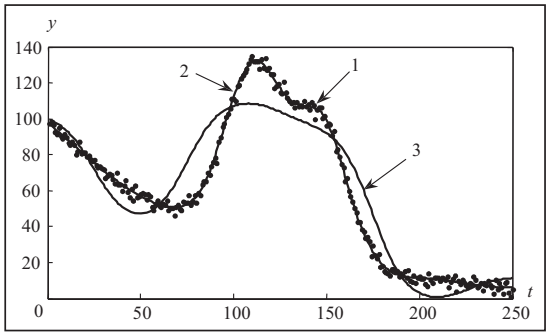

Рис. 1. Графики зависимостей гибридного алгоритма системы ANFIS (*1*), алгоритма из теоремы 1 (*2*), двухэтапного алгоритма, описываемого выражениями (55)–(57) (*3*), от числа наблюдений

Пример 1. Рассмотрим задачу идентификации статического объекта по вход–выходным данным [20], показанным на рис. 1 точками, с помощью системы нечеткого логического вывода Сугено нулевого порядка. Имеются два пика, наблюдаемые на фоне затухающего тренда и широкополосного шума. Использовались шесть функций принадлежности. Всего неизвестных параметров $n = 18$, из них $q = 12$ входят в описание ФП. В предположении, что ФП гауссовские, нетрудно определить матрицу C_t, исходя из (6). Центры и ширина ФП задавались, используя стандартную методику начального (до оптимизации) расположения параметров ФП [21] (рис. 2), а отклонения от них полагались равными 5 и 0.01 соответственно. На рис. 3 показаны функции принадлежности после оптимизации. Несмотря на значительные изменения ФП после оптимизации, первоначальный лингвистический порядок на этих рисунках сохранился. На рис. 4 приведены зависимости среднеквадратической ошибки оценивания от количества итераций $RMSE(M)$ (эпох) с параметром забывания $\lambda_i = \{\max(1 - 0.05/i, 0.99)\}$. Графики кривых *1*, *2*, *3* построены с помощью гибридного алгоритма системы ANFIS, алгоритма из теоремы 1 и двухэтапного алгоритма, описываемого выражениями (55)–(57) соответственно. Видно, что предложенные алгоритмы значительно превосходят по быстродействию гибридный алгоритм системы ANFIS.

Следует отметить, что влияние параметра забывания на скорость сходимости алгоритма существенна. Действительно, при $\lambda_i = 1$ и 40 итерациях имеем $RMSE(40) = 6.8$.

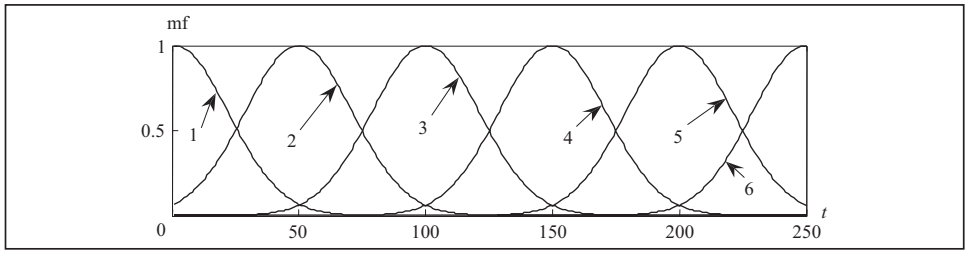

Рис. 2. Шесть функций принадлежности до оптимизации в зависимости от числа наблюдений

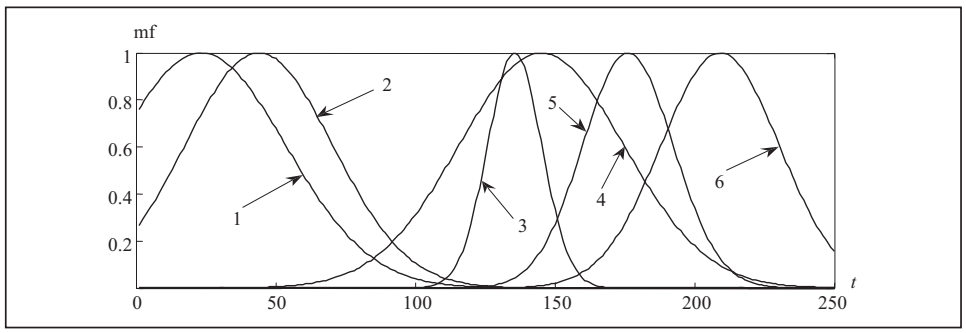

Рис. 3. Шесть функций принадлежности после оптимизации в зависимости от числа наблюдений

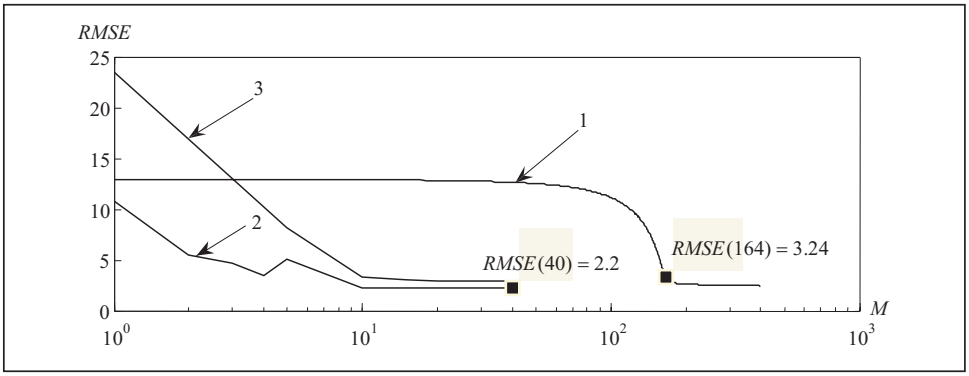

Рис. 4. Графики зависимостей среднеквадратической ошибки оценивания *RMSE* от количества итераций *M*

Таблица 1

Размеры обучающего (*N*) и тестирующего (*M*) множеств	Количество нейронов скрытого слоя	Оценки 90-й процентили ошибок аппроксимации выхода объекта					
		Точность алгоритмов при обучении			Точность алгоритмов при тестировании		
		№1	№2	№3	№1	№2	№3
$N = 2000$, $M = 500$	5	0.13	0.15	0.57	0.06	0.1	0.59
	10	0.12	0.13	0.23	0.05	0.05	0.18
	15	0.13	0.12	0.14	0.05	0.04	0.08
	20	0.13	0.12	0.12	0.05	0.04	0.04
$N = 500$, $M = 500$	4	0.16	0.21	0.83	0.11	0.2	0.83
	5	0.15	0.19	0.58	0.1	0.15	0.58
	6	0.14	0.18	0.48	0.1	0.12	0.46
	10	0.14	0.12	0.23	0.09	0.08	0.19

Пример 2. Рассмотрим задачу идентификации нелинейного динамического объекта, описываемого разностным уравнением [22]

$$y_t = y_{t-1} y_{t-2} (y_{t-1} + 2.5) / (1 + y_{t-1}^2 + y_{t-2}^2) + u_{t-1}.$$

Значения обучающей выборки входа u_t равномерно распределены на интервале $[-2,2]$, а значения тестирующей выборки определяются выражением $u_t = \sin(2\pi t / 250)$. Модель объекта находится в виде нелинейной модели авторегрессии и скользящего среднего $y_t = f(y_{t-1}, y_{t-2}, u_{t-1})$, где $f(.)$ — многослойный персептрон с сигмоидной ФА. Веса скрытого слоя и смещений выбираются из равномерных распределений на интервалах $[-1,1]$ и $[0,1]$ соответственно.

В табл. 1 приведены оценки по 500 реализациям 90-й процентили среднеквадратических ошибок оценивания выхода объекта, полученные с помощью трех сравниваемых алгоритмов обучения. В столбцах под номером 1 приведены результаты моделирования алгоритма из теоремы 1, под номером 2 — двухэтапного алгоритма, описываемого выражениями (55)–(57), под номером 3 — пакетной версии ELM-алгоритма [11]. Предложенные в работе алгоритмы использовались в режиме последовательной обработки (количество эпох равно единице) при $\lambda_i = 1$.

Видно, что при большом размере выборки и 20 нейронах скрытого слоя все три алгоритма дают приблизительно одинаковые результаты. При уменьшении размера выборки или количества нейронов скрытого слоя предложенные алгоритмы существенно превосходят ELM-алгоритм по быстродействию, а алгоритм из теоремы 1 превосходит двухэтапный алгоритм. На рис. 5 показаны графики

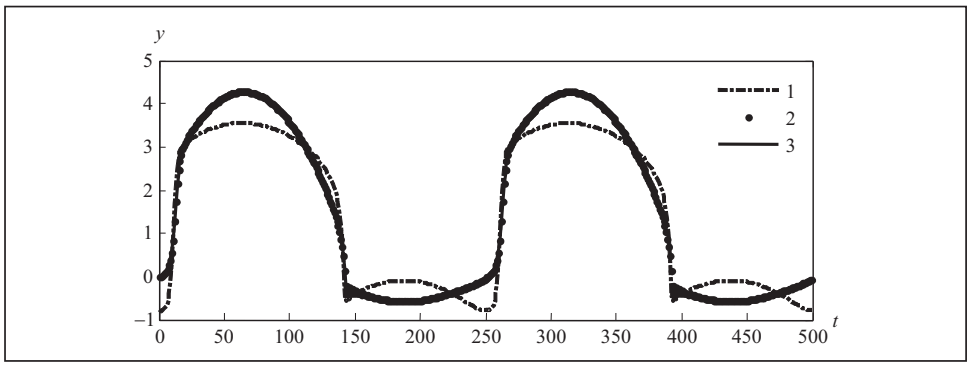

Рис. 5. Графики зависимоси выхода EML-алгоритма (*1*), выхода объекта (*2*), выхода модели (*3*) от числа реализаций

зависимости выходов объекта и моделей с пятью нейронами скрытого слоя для некоторой реализации: выхода ELM-алгоритма, выхода объекта, выхода модели, построенного с помощью алгоритма из теоремы 1.

ЗАКЛЮЧЕНИЕ

Рассмотрены задачи обучения НС и НФС, приводящие к сепарабельной регрессии, и предложены новые алгоритмы их обучения. Обучение формулируется, как нелинейная оптимизационная задача, в постановку которой вводится априорная информация только о нелинейно входящих параметрах. Для ее решения используется метод ГН с линеаризацией в окрестности последней оценки и асимптотические представления псевдоинверсий возмущенных матриц. Реализация алгоритмов не требует подбора начальных значений для линейно входящих параметров, который может приводить к расходимости. Предложенные алгоритмы могут быть использованы в режимах последовательной и пакетной обработки. Как частный случай, из них следуют алгоритмы, предложенные в [7], а моделирование показывает, что разработанные алгоритмы могут превосходить их по точности и скорости сходимости. В дальнейшем предполагается обобщить полученные результаты на динамические нейронные сети.

СПИСОК ЛИТЕРАТУРЫ

1. Haykin S. Neural networks and learning machines. — Upper Sadle River (N.J.): Prentice Hall, 2009. — 916 p.
2. Golub G.H., Pereyra V. The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate // SIAM J. Numer. Anal. — 1973. — **10**. — P. 413–432.
3. Pereyra V, Scherer G., Wong F. Variable projections neural network training // Math. and Comput. Simulation. — 2006. — **73**. — P. 231–243.
4. Sjoberg J., Viberg M. Separable non-linear least-squares minimization and possible improvements for neural net fitting // IEEE Workshop in Neural Networks for Signal Processing. — Florida (USA), 1997. — P. 345–354.
5. Ngia L. Separable Nonlinear least-squares methods for on-line estimation of neural nets hammerstein models // Proc. of the IEEE Signal Processing Society Workshop. — Sydney, 2000. — **1**. — P. 65–74.
6. Parisi P., Claudio D. Di., Orlandi G., Rao B.D. A generalized learning paradigm exploiting the structure of feedforward neural networks // IEEE Trans. Neural Networks. — 1996. — **7**. — P. 1450–1460.
7. Jang J.S.R. ANFIS: adaptive-network-based fuzzy inference system // IEEE Trans. on Systems, Man and Cybernetics. — 1993. — N 3. — P. 665–685.

8. Bodyanskiy Ye., Kolodyazhniy V.V. and Stephan A. An adaptive learning algorithm for a neuro-fuzzy network. Computational Intelligence. Theory and Applications. — Berlin; Heidelberg; New York: Springer, 2001. — P. 68–75.
9. Bodyanskiy Ye., Kolodyazhniy V.V. and Otto P. Neuro-fuzzy Kolmogorov's network for time series prediction and pattern classification // Lect. Notes Comput. Sci. — 2005. — P. 191–202.
10. Bodyanskiy Ye.V., Pliss I., Vynokurova O. Adaptive wavelet-neuro-fuzzy network in the forecasting and emulation tasks // Intern. J. Inform. Theories & Appl. — 2008. — **15**. — P. 47–55.
11. Huang G.B., Wang D.H., Lan Y. Extreme learning machines: A survey // Intern. J. of Machine Leaning and Cybernetics. — 2011. — **2**, N 2. — P. 107–122.
12. Cheol-Taek Kim and Ju-Jang Lee. Training two-layered feedforward networks with variable projection method // IEEE Trans. Neural Networks. — 2008. — **19**, N 2. — P. 371–375.
13. Скороход Б.А. Диффузные алгоритмы обучения нейронных сетей прямого распространения // Кибернетика и системный анализ. — 2013. — № 3. — С. 14–26.
14. Kecman V. Learning and soft computing, support vector machines, neural networks, and fuzzy logic models. — MIT Press, Cambridge (Mass.): MIT Press, 2001. — 361 p.
15. Скороход Б.А. Асимптотика линейной рекуррентной регрессии при диффузной инициализации // Международный научно-технический журнал «Проблемы управления и информатики». — 2009. — N 3. — С. 98–107.
16. Jang R., Sun C., Mizutani E. Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. — Prentice Hall, 1997. — 601 p.
17. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. — М.: Наука, 1977. — 224 с.
18. Wimmer H.R. Stabilizing and unmixed solutions of the discrete time algebraic Riccati equa-tion // Proc. Workshop on the Riccati Equation in Control, Systems, and Sygnals. — 1989. — Italy. — P. 95–98.
19. Bertsekas D.P. Incremental least squares methods and the extended Kalman filter / SIAM J. Optimization. — 1996. — **3**. — P. 807–822.
20. Curve Fitting Toolbox 3. MathWorks, Inc.
21. Fuzzy Logic Toolbox. The MathWorks, Inc.
22. Narendra K.S. and Parthasarathy K. Identification and control of dynamical systems using neural networks // IEEE Trans. Neural Networks. — 1990. — **1**, N 1. — P. 4–27.

Поступила 20.09.2013