

## ГИБРИДНЫЕ АЛГОРИТМЫ ОБУЧЕНИЯ RBF-СЕТЕЙ

### Аннотация

Разработан комбинированный подход к обучению нейронных RBF-сетей. Весовые коэффициенты нейронов скрытого слоя определяются на основе известных процедур обучения без учителя. Для определения параметров выходного слоя и структуры RBF-сети предложены алгоритмы, основанные на развитой байесовской методологии.

Известно, что результат обучения нейронных сетей существенно зависит от их структуры [1 – 5]. Единую и последовательную концептуальную основу для обучения нейронных сетей представляет байесовская методология [2 – 11]. На основе этой методологии в [3] получены алгоритмы структурно-параметрического синтеза многослойных нейронных сетей с последовательными связями. Способность к быстрому обучению RBF-сетей, по сравнению, например, с многослойными перцептронами, позволила успешно применить их для решения ряда практических задач [1].

С учетом специфики RBF-сетей в рамках байесовской методологии в [4] синтезированы алгоритмы определения параметров выходного и скрытого слоев нейронной сети, а также разработаны процедуры модификации структуры сети, основанные на добавлении и удалении нейронов скрытого слоя. Предложенные процедуры позволили получить семейство алгоритмов обучения RBF-сетей. В то же время синтез эффективных алгоритмов обучения, обладающих невысокой вычислительной сложностью для отдельных видов нейронных сетей остается актуальной задачей.

**Нейронные RBF-сети.** Для решения задачи обучения нейронной сети воспользуемся моделью наблюдения

$$z(n) = f_s(\mathbf{x}(n), \mathbf{w}) + \xi(n), \quad n = \overline{1, N}, \quad (1)$$

Здесь  $\mathbf{x} - d_x$ -вектор факторных переменных;  $z$  – скалярная результирующая переменная;  $\xi$  – шум наблюдения, с нулевым средним значением и дисперсией  $D_\xi$ ;  $N$  – объем обучающей выборки;  $f_s(\mathbf{x}, \mathbf{w})$  – характеристика вход-выход RBF-сети, которую запишем в виде

$$f_s(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^d w_i \varphi_i(\mathbf{x}) = \boldsymbol{\varphi}^T(\mathbf{x}) \mathbf{w}, \quad (2)$$

где  $\mathbf{w} - d$ -вектор весовых коэффициентов выходного слоя;  $\boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x}), \dots, \varphi_d(\mathbf{x})]^T$  – вектор-функция,  $\varphi_i(\mathbf{x})$  – характеристика вход-выход  $i$ -го нейрона скрытого слоя сети;  $s$  – структура RBF-сети. В рассматриваемом случае исключение связи между нейроном выходного слоя и нейроном скрытого слоя приводит к исключению этого нейрона. Поэтому структура  $s$  RBF-сети (2) определяется количеством базисных функций  $d$ .

При использовании гауссовой функции активации характеристика вход-выход нейронов скрытого слоя RBF-сети (базисная функция), принимает вид

$$\varphi_i(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{c}_i)^T \mathbf{G}_i (\mathbf{x} - \mathbf{c}_i)\right), \quad i = \overline{1, d},$$

где  $\mathbf{c}_i - d_x$ -вектор, определяющий положение центра базисной функции,  $\mathbf{G}_i$  – симметричная неотрицательно определенная квадратная матрица порядка  $d_x$ , элементы которой представляют собой весовые коэффициенты.

**Параметрическая оптимизация.** Основу для синтеза гибридных алгоритмов обучения нейронных RBF-сетей составляет комбинированное применение как процедур обучения с учителем, так и процедур обучения без учителя. При этом параметры линейного выходного слоя определяются на основе обучения с учителем, а для определения параметров нейронов скрытого слоя  $\mathbf{c}_i$  и  $\mathbf{G}_i$  могут быть использованы известные процедуры обучения без учителя [1]. В частном случае центры могут совпадать с выборочными значениями факторной переменной. Так полная RBF-сеть состоит из  $d = N$  нейронов скрытого слоя с центрами  $\mathbf{c}_i = \mathbf{x}(i)$ ,  $i = \overline{1, d}$ . В наиболее простом случае матрицы  $\mathbf{G}_i = g^2 \mathbf{I}$ . При этом единственный весовой коэффициент может быть определен согласно [1] в виде  $g = \sqrt{2d} / \Delta \mathbf{c}_{\max}$ , где  $\Delta \mathbf{c}_{\max}$  – максимальное расстояние между центрами, которое определяется выражением  $\Delta \mathbf{c}_{\max} = \max_{\substack{i, j = \overline{1, d} \\ i \neq j}} \|\mathbf{c}_i - \mathbf{c}_j\|$ .

Применим развитую [4, 5] байесовскую методологию для определения параметров выходного слоя и структуры RBF-сети. Для нейронной сети со структурой  $s$  воспользуемся критерием максимума апостериорной плотности вероятности. При этом параметры выходного слоя находятся из решения непрерывной экстремальной задачи

$$\hat{\mathbf{w}}_s = \arg \max_{\mathbf{w}} p(\mathbf{w}/\mathbf{Z}, s) = \arg \min_{\mathbf{w}} J(\mathbf{w}, s), \quad (3)$$

где  $J(\mathbf{w}, s) = -\ln(p(\mathbf{Z}/\mathbf{w}, s)p(\mathbf{w}/s))$  – целевая функция параметрической оптимизации,  $p(\mathbf{w}/s)$  – априорная плотность вероятности параметров  $\mathbf{w}$ ,  $p(\mathbf{Z}/\mathbf{w}, s)$  – функция правдоподобия,  $\mathbf{Z}$  – совокупность значений  $\mathbf{z}(n)$ ,  $n = \overline{1, N}$ .

Для случая гауссова шума наблюдения  $\xi$  функция правдоподобия, с учетом (1) и (2), принимает вид

$$p(\mathbf{Z}/\mathbf{w}, s) = (2\pi D_\xi)^{-N/2} \exp\left(-\frac{\|\mathbf{e}_s(\mathbf{w})\|^2}{2D_\xi}\right), \quad (4)$$

Здесь  $\mathbf{e}_s(\mathbf{w}) = \mathbf{Z} - \Phi_s \mathbf{w}$  – вектор невязок,  $\mathbf{Z} = [z(1), \dots, z(N)]^T$ ,  $\Phi_s$  –  $(N \times d)$ -матрица, строками которой являются базисные векторы  $\Phi^T(\mathbf{x}(n))$ ,  $n = \overline{1, N}$ .

Следуя [2, 8], при решении задачи обучения нейронной сети воспользуемся гауссовой априорной плотностью вероятности

$$p(\mathbf{w}/s) = (2\pi)^{-d/2} (\det \mathbf{K}_s)^{-1/2} \exp(-\mathbf{w}^T \mathbf{K}_s^{-1} \mathbf{w}/2), \quad (5)$$

с нулевым вектором средних значений и диагональной ковариационной матрицей  $\mathbf{K}_s = \text{diag}[D_{w1}, \dots, D_{wd}]$ . При этом и апостериорная плотность вероятности  $p(\mathbf{w}/\mathbf{Z}, s) = p(\mathbf{Z}/\mathbf{w}, s)p(\mathbf{w}/s)/p(\mathbf{Z}/s)$  является гауссовой с ковариационной матрицей

$$\mathbf{G}_s = (D_\xi^{-1} \Phi_s^T \Phi_s + \mathbf{K}_s^{-1})^{-1} \quad (6)$$

и вектором средних значений

$$\hat{\mathbf{w}}_s = D_\xi^{-1} \mathbf{G}_s \Phi_s^T \mathbf{Z} = (\Phi_s^T \Phi_s + \mathbf{M})^{-1} \Phi_s^T \mathbf{Z}. \quad (7)$$

Здесь  $\mathbf{M} = D_\xi \mathbf{K}_s^{-1}$  – диагональная матрица параметров регуляризации. Выражение (7) определяет оценку параметров выходного слоя, соответствующую решению задачи (3).

**Структурная оптимизация.** Для определения структуры нейронной сети применим критерий максимума апостериорной вероятности

$$\hat{s} = \arg \max_s P(s/\mathbf{Z}) = \arg \min_s L(s). \quad (8)$$

Здесь  $L(s) = -\ln(p(\mathbf{Z}/s)P(s))$  – целевая функция структурной оптимизации,  $P(s)$  – априорная вероятность,  $p(\mathbf{Z}/s)$  – функция правдоподобия структуры модели, которая определяется выражением

$$p(\mathbf{Z}/s) = \int_{\mathbf{w}} p(\mathbf{Z}/\mathbf{w}, s) p(\mathbf{w}/s) d\mathbf{w}. \quad (9)$$

Учитывая (4), (5), после интегрирования находим

$$p(\mathbf{Z}/s) = p(\mathbf{Z}/\hat{\mathbf{w}}_s, s) p(\hat{\mathbf{w}}_s/s) (2\pi)^{d/2} \sqrt{\det \mathbf{G}_s}, \quad (10)$$

Полагая, что различные структуры моделей априори равновероятны, приходим к целевой функции структурной оптимизации

$$L(s) = -\ln p(\mathbf{Z}/s) = J(\hat{\mathbf{w}}_s, s) - H(s). \quad (11)$$

Здесь  $H(s) = \ln\left((2\pi)^{d/2} \sqrt{\det \mathbf{G}_s}\right) = \frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln \det \mathbf{G}_s$  – дифференциальная энтропия апостериорного распределения. Второе слагаемое в выражении (11) можно рассматривать, как штраф, препятствующий переобучению в процессе структурной оптимизации.

Подставляя (4), (5) в (11), получаем

$$L(s) = \frac{1}{2} \left( N \ln(2\pi D_\xi) + D_\xi^{-1} \|\mathbf{Z} - \Phi_s \hat{\mathbf{w}}_s\|^2 + \hat{\mathbf{w}}_s^T \mathbf{K}_s^{-1} \hat{\mathbf{w}}_s + \ln\left(\frac{\det \mathbf{K}_s}{\det \mathbf{G}_s}\right) \right). \quad (12)$$

Выражение (12), аналогично [9, 10] можно представить в другом виде

$$L(s) = \frac{1}{2} \left( \mathbf{Z}^T \mathbf{C}_s^{-1} \mathbf{Z} + \ln \det \mathbf{C}_s + N \ln(2\pi) \right). \quad (13)$$

Здесь  $\mathbf{C}_s = D_\xi \mathbf{I} + \Phi_s \mathbf{K}_s \Phi_s^T$ . Лемма об обращении матриц (формула Вудбери) позволяет определить  $\mathbf{C}_s^{-1} = D_\xi^{-1} (\mathbf{I} - D_\xi^{-1} \Phi_s \mathbf{G}_s \Phi_s^T)$ . Таким образом, при вычислении (13) не требуется выполнять обращение матрицы  $\mathbf{C}_s$  порядка  $N$ .

Поиск структуры нейронной сети, доставляющей наименьшее значение целевой функции  $L(s)$ , может проводиться с использованием различных алгоритмов, основанных на процедурах усложнения и (или) упрощения структуры модели [4, 5].

Обозначим  $I_s$  и  $J_s$  множества индексов базисных функций включенных и не включенных в состав нейронной сети со структурой  $s$  соответственно;  $s_{-i}$  и  $s_{+j}$  – структуры RBF-сетей, которые отличаются от сети со структурой  $s$  отсутствием  $i$ -ой и наличием дополнительной  $j$ -ой базисных функций соответственно. При этом  $i \in I_s$  и  $j \in J_s$ .

Выделим две базовые операции модификации структуры RBF-сети. Операция удаления ( $R_i$ -операция) состоит в исключении из нейронной сети  $i$ -й базисной функции  $\phi_i(\mathbf{x})$  с последующим определением значений параметров  $\hat{\mathbf{w}}_{s_{-i}}$ . Операция добавления ( $A_j$ -операция) заключается в добавлении  $j$ -й базисной функции  $\phi_j(\mathbf{x})$  и определении новых значений параметров модели  $\hat{\mathbf{w}}_{s_{+j}}$ . Процедура структурной модификации основана на многократном применении базовых операций и выполняется до тех пор, пока происходит уменьшение целевой функции  $L(s)$ .

Для обеспечения направленного (беспоискового) характера структурной оптимизации определим приращение целевой функции структурной оптимизации. При применении  $R_i$ -операции значение целевой функции уменьшается на величину  $\Delta L_i(s) = L(s) - L(s_{-i})$ . Используя (12), аналогично [3 – 5] находим

$$\Delta L_i(s) = -\frac{1}{2} \left( \frac{\hat{w}_{si}^2}{G_{sii}} + \ln(D_{wi}^{-1} G_{sii}) \right), \quad i = \overline{1, d}. \quad (14)$$

Выбор удаляемой базисной функции осуществляется из условия  $\hat{i} = \arg \max_{i \in I_s} \Delta L_i(s)$ . При этом операция удаления выполняется, если  $\Delta L_i(s) > 0$ . В противном случае происходит переход к  $A_j$ -операции, при применении которой значение целевой функции уменьшается на величину  $\Delta \bar{L}_j(s) = L(s) - L(s_{+j}) = -\Delta L_j(s_{+j})$ . Если для некоторой структуры  $s$  применение операций модификации структуры RBF-сети не приводит к уменьшению  $L(s)$ , то процедура структурно-параметрического синтеза завершается.

**Оценки параметров распределений.** В подавляющем большинстве случаев обучение нейронных сетей выполняется в условиях априорной неопределенности относительно априорного распределения параметров и распределения шума наблюдения. Преодолеть априорную неопределенность можно с помощью одного из вариантов иерархического байесовского подхода. В зарубежной литературе такой подход обычно называют maximum marginal likelihood (MML) [9 – 11]. При фиксированной структуре RBF-сети оценки параметров распределений могут быть найдены на основе максимизации функции правдоподобия  $p(\mathbf{Z}/s; \mathbf{K}_s, D_\xi)$  или минимизации целевой функции  $-\ln p(\mathbf{Z}/s; \mathbf{K}_s, D_\xi)$ , которая определяется аналогично (12) или (13).

Используя метод простых итераций, приходим к выражению для нового приближения дисперсии шума наблюдения

$$\hat{D}_\xi = \frac{\|\mathbf{Z} - \Phi_s \tilde{\mathbf{w}}_s\|^2}{N - \tilde{\gamma}}, \quad (15)$$

где

$$\tilde{\mathbf{w}}_s = \tilde{D}_\xi^{-1} \tilde{\mathbf{G}}_s \Phi_s^T \mathbf{Z}, \quad \tilde{\mathbf{G}}_s = (\tilde{D}_\xi^{-1} \Phi_s^T \Phi_s + \tilde{\mathbf{K}}_s^{-1})^{-1}, \quad \tilde{\gamma} = d - \text{tr}(\tilde{\mathbf{K}}_s^{-1} \tilde{\mathbf{G}}_s) -$$

эффективное количество параметров,  $\tilde{\mathbf{K}}_s$  и  $\tilde{D}_\xi$  – оценки параметров распределений, полученные на предыдущей итерации.

Процедура оценки элементов ковариационной матрицы априорного распределения  $\mathbf{K}_s$  зависит от ее вида. Так алгоритм RVM (Relevance vector machine) [9], полученный для диагональной матрицы  $\mathbf{K}_s$ , включает процедуру определения гиперпараметров  $\alpha_i = D_{wi}^{-1} = \mathbf{K}_{sii}^{-1}$ ,  $i = \overline{1, d}$ . Соответствующее выражение для оценки априорной дисперсии параметров сети принимает вид  $\hat{D}_{wi} = \tilde{w}_{si}^2 + \tilde{G}_{sii}$ ,  $i = \overline{1, d}$ . К аналогичным результатам приводит и другой итерационный алгоритм  $\hat{D}_{wi} = \tilde{w}_{si}^2 / \tilde{\gamma}_i$ , где  $\tilde{\gamma}_i = 1 - \tilde{D}_{wi}^{-1} \tilde{G}_{sii}$ . При этом  $\tilde{\gamma} = \sum_{i=1}^d \tilde{\gamma}_i$ .

В [10, 11] предложен подход к оценке параметров априорного распределения, основанный на покоординатной оптимизации целевой функции. На каждом шаге для очередного гиперпараметра находится его оптимальное значение и, затем выполняется коррекция оценок параметров нейросетевой модели  $\tilde{\mathbf{w}}_s$ . Так при фиксированных значениях  $\tilde{D}_{wj}$ ,  $j = \overline{1, d}$ ,  $j \neq i$  функции правдоподобия  $p(\mathbf{Z}/s; \mathbf{K}_s, D_\xi)$  достигает максимального значения для  $\hat{D}_{wi} = (\tilde{w}_{si}^2 - \tilde{\gamma}_i \tilde{G}_{sii}) / \tilde{\gamma}_i^2$ , если выполняется условие  $\tilde{w}_{si}^2 / \tilde{G}_{sii} > \tilde{\gamma}_i$ . В противном случае (если условие не

выполняется) функция правдоподобия достигает максимума при  $\hat{D}_{wi} = 0$  и  $i$ -й нейрон скрытого слоя исключается из RBF-сети.

Процедура обучения нейросетевой модели упрощается при использовании ковариационной матрицы априорного распределения в виде  $\mathbf{K}_s = D_w \mathbf{I}$ . В этом случае, максимизация функции правдоподобия с помощью метода простых итераций приводит к выражению  $\hat{D}_w = \|\tilde{\mathbf{w}}_s\|^2 / \tilde{\gamma}$ , где  $\tilde{\mathbf{w}}_s = (\Phi_s^T \Phi_s + \tilde{\mu} \mathbf{I})^{-1} \Phi_s^T \mathbf{Z}$ ,  $\tilde{\mu} = \tilde{D}_\xi / \tilde{D}_w^{-1}$ ,  $\tilde{\gamma} = d - \tilde{\mu} \text{tr}(\Phi_s^T \Phi_s + \tilde{\mu} \mathbf{I})^{-1}$ . Таким образом, оценка параметров выходного слоя RBF-сети сопровождается определением единственного параметра регуляризации  $\mu$ . Так в [5] получен итерационный алгоритм

$$\hat{\mu} = \frac{\tilde{\gamma}}{N - \tilde{\gamma}} \frac{\|\mathbf{Z} - \Phi_s \tilde{\mathbf{w}}_s\|^2}{\|\tilde{\mathbf{w}}_s\|^2}. \quad (16)$$

По завершении итерационного процесса находятся оценки параметров распределений, формируется оценка параметров выходного слоя RBF-сети  $\hat{\mathbf{w}}_s$  и выполняется переход к процедуре структурной оптимизации. При этом приращения (14) целевой функции структурной оптимизации определяются с использованием полученных оценок распределений  $\hat{D}_\xi$  и  $\hat{D}_w$ . Модификация структуры сети сопровождается изменением параметров скрытого слоя RBF-сети.

Таким образом, с использованием байесовской методологии получено семейство гибридных алгоритмов структурно-параметрического синтеза нейронных RBF-сетей. Предложенные процедуры могут быть использованы при построении разнообразных регрессионных моделей. В частности, алгоритм определения параметра регуляризации позволяет находить решение задач гребневой регрессии.

Результаты статистического моделирования подтверждают работоспособность и позволяют судить о сравнительной эффективности разработанных алгоритмов.

#### *Список литературы*

1. Осовский С. Нейронные сети для обработки информации. – М.: Финансы и статистика, 2002. – 344 с.
2. Bishop C.M. Neural Networks for pattern recognition. – Oxford: Oxford University Press, 1995. – 504 p.
3. Милов В.Р. Структурно-параметрический синтез нейронных сетей с последовательными связями на основе байесовской методологии // Нейроинформатика-2005. Сборник научных трудов. Ч. 1. – М.: МИФИ, 2005. – С. 18 – 25.
4. Милов В.Р., Махмудов Я.Я. Обучение нейронных RBF-сетей на основе байесовской методологии и решение задачи восстановления зависимостей // Нейрокомпьютеры: разработка, применение. – 2005. – № 4. – С. 23 – 31.
5. Баранов В.Г., Костюков В.Е., Милов В.Р. Байесовская методология синтеза регрессионных моделей // Труды НГТУ: Информационные технологии/ НГТУ. – Н.Новгород. – 2004. – Т. 48, Вып. 1. – С. 135 – 141.
6. Шумский С.А. Байесова регуляризация обучения // Лекции по нейроинформатике. Часть II. – М.: МИФИ, 2002. – С. 30 – 93.
7. Матвейкин В.Г., Фролов С.В. Использование байесовского подхода в обучении нейронных сетей // Информационные технологии. – 1998. – № 10. – С. 27 – 35.
8. MacKay D.J.C. Bayesian framework for backpropagation networks // Neural Computation. – 1992. – V. 4, N. 3. – P. 448 – 472.
9. Tipping M.E. Sparse Bayesian learning and the relevance vector machine // Journal of Machine Learning Research. – 2001. – V. 1. – P. 211 – 244.
10. Bishop C.M., Tipping M.E. Bayesian regression and classification. In «Advances in learning theory: methods, models and applications». NATO-ASI Series in Computer and Systems Sciences. – IOS Press, 2003. – P. 267 – 288.
11. Tipping M.E., Faul A.C. Fast Marginal likelihood maximization for sparse Bayesian models // 9 International workshop on Artificial Intelligence and Statistics, Kew West, 2003.