2.5 Многослойные искусственные нейронные сети

Возможности однослойных искусственных нейронных сетей ограничены. Для расширения возможностей ИНС они строятся как многослойные. Помимо входного и выходного слоёв вводятся промежуточные слои, которые называются ***скрытыми слоями***. Одним из примеров многослойных ИНС являются многослойные персептроны.

Все многослойные персептроны представляют собой сети, в которых передача информации происходит от входного слоя через K скрытых слоев к выходному слою (рис. 2.16). Количество нейронов в разных слоях может быть различным. Входной слой персептрона служит для приема входных сигналов X_1, X_2, …, X_V, и их передачи на нейроны скрытого слоя.

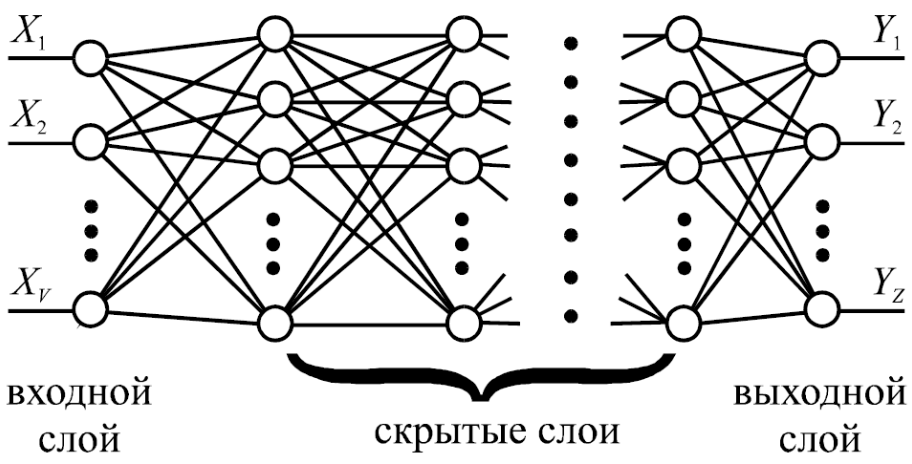

Рисунок 2.16 Структура многослойного персептрона

В общем случае каждый искусственный нейрон некоторого скрытого слоя соединён с выходами всех нейронов предыдущего слоя. Например, нейрон с номером i в k-ом слое соединён со всеми нейронами предыдущего k–1-го слоя. Связи имеют вес, и, в частности, $w_{ij}^{(k)}$ - коэффициент веса связи i –го нейрона в k-ом слое с j-м нейроном k–1-го слоя. Общее число слоёв равно K+1. Номера слоёв k изменяются от 0 до K+1, причём номером k = 0 обозначен входной слой, а номером k = K+1 – выходной слой сети.

В скрытых слоях происходит основное нелинейное преобразование информации. Функция активации, используемая в скрытых слоях, представляет собой дифференцируемую сигмоидную функцию.

Выходной слой служит для формирования выходных сигналов сети Y_1, Y_2, …, Y_Z путём суперпозиции взвешенных выходных сигналов последнего скрытого слоя.

2.6 Обучение многослойных искусственных нейронных сетей методом обратного распространения ошибки

Обучение многослойных искусственных нейронных сетей представляет собой более сложную задачу, чем обучение простейшего персептрона. Для его осуществления разработан метод обратного распространения ошибки (англ. *backpropagation*). Он представляет собой ***итеративный градиентный алгоритм обучения многослойного персептрона***.

Этот метод разработан для получения желаемых реакций многослойной искусственной нейронной сети в результате ***минимизации ошибок*** её работы. При этом обучение ИНС рассматривается как ***задача параметрической оптимизации*** связей между нейронами.

Впервые метод обратного распространения ошибки был описан в 1974 г. независимо и практически одновременно А.И.

Галушкиным и Полом Вербосом. Но в результате спада интереса к ИНС в семидесятых годах прошлого столетия этот метод долгое время оставался невостребованным. Дальнейшее развитие этого метода осуществили в 1986 г. исследователи Румельхарт, Хинтон и Вильямс, а независимо и практически одновременно с ними - Барцев и Охонин.

В качестве характерного примера рассмотрим трёхслойную искусственную нейронную сеть, имеющую один скрытый слой (рис. 2.17).

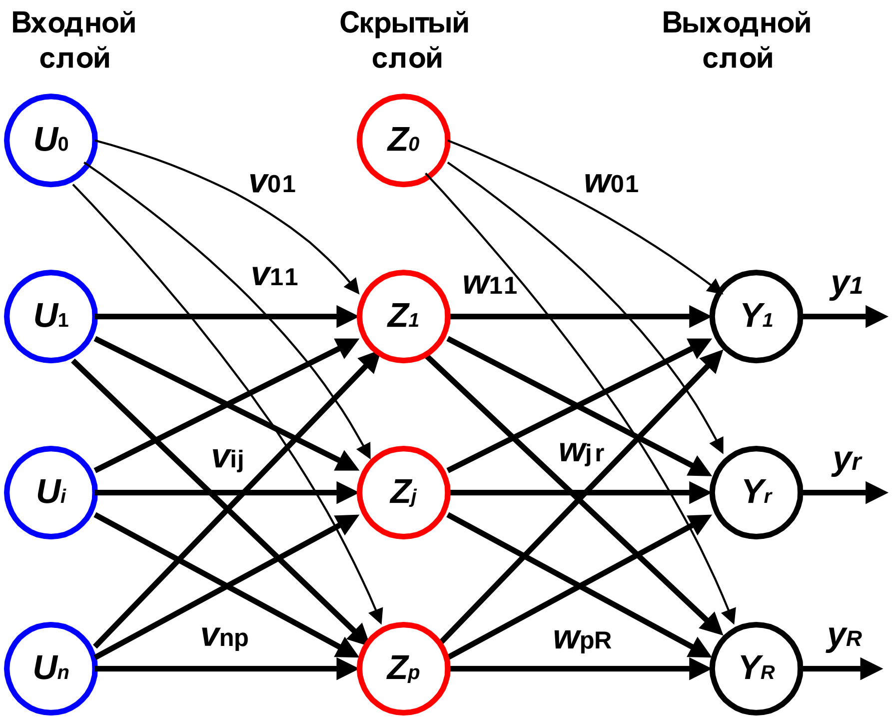

Рисунок 2.17. Схема трёхслойной искусственной нейронной сети

Во входном слое имеется n входных нейронов, каждый из которых связан с соответствующим ему источником входного сигнала. Входные нейроны обозначены $U_1, U_2, ..., U_i, ...U_n$ и транслируют на свои выходы входные сигналы $u_1, u_2, ..., u_i, ...u_n$ соответственно. Кроме того, имеется один дополнительный источник U_0 «опорного» сигнала u_0= -1. Таким образом, общее число входных сигналов равно $n+1$.

В скрытом слое находится p искусственных нейронов. Они обозначены $Z_1, Z_2,..., Z_j,...Z_p$ и формируют на своих выходах сигналы $z_1, z_2,..., z_j,...z_p$. Каждый нейрон скрытого слоя получает сигналы с выходов всех нейронов входного слоя и связан с источником U_0 «опорного» сигнала u_0= -1. Весовые коэффициенты связей выходов входных нейронов со входами нейронов скрытого слоя обозначены $v_{11}, v_{12},..., v_{np}$. Индексы ij в обозначении коэффициента v_{ij} означают, что выход i-го нейрона входного слоя связан со входом j-го нейрона скрытого слоя. $v_{01}, v_{02},..., v_{0p}$ – весовые коэффициенты связей нейронов скрытого слоя с источником «опорного» сигнала u_0. Количество связей нейронов входного и скрытого слоёв равно $(n+1)p$.

В выходном слое находится R нейронов. Они обозначены $Y_1, Y_2,..., Y_r,...Y_R$ и образуют на своих выходах сигналы $y_1, y_2,..., y_r,...y_R$ соответственно. Каждый нейрон выходного слоя связан с выходами всех нейронов скрытого слоя и с элементом Z_0. Весовые коэффициенты связей выходов нейронов скрытого слоя со входами нейронов выходного слоя обозначены $w_{11}, w_{12},..., w_{pR}$. Обозначение коэффициента v_{jr} означает, что выход j-го нейрона скрытого слоя связан со входом r-го нейрона выходного слоя. $w_{01}, w_{02},..., w_{0R}$ – весовые коэффициенты связей нейронов выходного слоя с источником «опорного» сигнала z_0. Общее количество связей нейронов скрытого и выходного слоёв равно $(p+1)R$.

Для осуществления настройки сети строится ***функция оценки*** качества её работы. Она позволяет количественно определить успешность решения сетью поставленных ей задач. В этом случае обучение нейронных сетей можно представить как задачу их ***параметрической оптимизации***. Нужно подобрать такие значения параметров v_{ij} и w_{jr} для всех i от 0 до n, j от 0 до p, r от 1 до R, при которых функция качества имеет минимальное значение.

Чаще всего функция S представляет собой сумму квадратов разностей между фактическими и желаемыми выходными сигналами сети.

$$S = \frac{1}{2}\sum_{r=1}^{R}(y_r - y_{Жr})^2,$$

где R – количество выходов ИНС; y_r – фактический выходной сигнал ИНС на её r-ом выходе: $y_{Жr}$ – желаемый выходной сигнал ИНС на её r-ом выходе, соответствующий правильному (желаемому) поведению ИНС; $r = 1,2,...,R$. Отметим, что такая функция оценки явно зависит от выходных сигналов ИНС $y_1, y_2,..., y_R$ и неявно зависит от значений всех весовых коэффициентов $w_{01},..., w_{11}, w_{12},..., w_{pR}$ и $v_{01},..., v_{11}, v_{12},..., v_{np}$.

Для достижения минимального значения функции оценки S организуется итерационный процесс по методу ***градиентного спуска***, в котором шаг за шагом корректируются значения всех весовых коэффициентов ИНС.

Чаще всего применяется метод ***стохастического градиентного спуска***, который является модификацией классического метода градиентного спуска. Его особенность состоит в том, что корректировка весовых коэффициентов производится ***на каждом шаге обучения*** после определения реакции ИНС на предъявленный сети обучающий набор входных сигналов. Изменение значений весовых коэффициентов связей между нейронами производится в сторону, противоположную ***градиенту*** функции S.

Градиент функции S – это вектор, направление которого указывает направление наибольшего возрастания функции S. Модуль этого вектора равен быстроте роста функции оценки S в указанном направлении. Как указано выше, функция S зависит от $d = p(R+n+1)+R$ переменных $w_{01},...,w_{11},w_{12},...,w_{pR}$ и $v_{01},...,v_{11},v_{12},...,v_{np}$. Поэтому её градиентом называется d - мерный вектор, компонентами которого являются частные производные функции S по всем её аргументам.

$$grad \quad S = \left[\frac{\partial S}{\partial w_{01}},...,\frac{\partial S}{\partial w_{11}},\frac{\partial S}{\partial w_{12}},...,\frac{\partial S}{\partial w_{pR}},\frac{\partial S}{\partial v_{01}},...,\frac{\partial S}{\partial v_{11}},\frac{\partial S}{\partial v_{12}},...,\frac{\partial S}{\partial v_{np}}\right]^T$$

Например, если n = 5, p = 4 и R = 2, общее число весовых коэффициентов связей ИНС равно 34. Такое же значение частных производных функции оценки содержит её градиент. Это значит, что в результате параметрической оптимизации ИНС предстоит настроить 34 её параметра.

В соответствии с методом ***стохастического градиентного спуска*** алгоритм корректировки значений коэффициентов связей на одном шаге процесса настройки ИНС выглядит следующим образом:

$$w_{jr} := w_{jr} - \Delta w_{jr},$$

$$v_{ij} := v_{ij} - \Delta v_{ij},$$

где

$$\Delta w_{jr} = \eta \frac{\partial S}{\partial w_{jr}},$$

$$\Delta v_{ij} = \eta \frac{\partial S}{\partial v_{ij}},$$

η – коэффициент, удовлетворяющий условию $0 < \eta < 1$ и влияющий на устойчивость и быстроту процесса настройки параметров ИНС; $i = 0,...,n$; $j = 0,..., p$; $r = 1,..., R$.

Сначала рассмотрим особенности оптимизации коэффициентов связи нейронов скрытого слоя с нейронами выходного слоя. Для этого определим частные производные функции S по переменным $w_{01},..., w_{pR}$, представив функцию S в развёрнутом виде:

$$S = \frac{1}{2}\sum_{r=1}^{R}[f(\sum_{j=0}^{p} w_{jr} z_j) - y_{Жr}]^2,$$

где $f(q)$ – активационная функция искусственного нейрона выходного слоя; q – аргумент этой функции. Считаем, что все нейроны выходного и скрытого слоёв имеют одинаковые активационные функции. Тогда вычисление частных производных производится по формуле

$$\frac{\partial S}{\partial w_{jr}} = (\sum_{m=1}^{R} \delta_m)\frac{\partial f(q_r)}{\partial w_{jr}},$$

где q_r– величина, характеризующая состояние r-го нейрона выходного слоя; δ_m – ошибка сети по m-му выходу, причём $j = 0,..., p$; $r = 1,..., R$, $m = 1,..., R$

$$q_r = \sum_{j=0}^{p} w_{jr} z_j,$$

$$\delta_m = f(q_m) - y_{Жm}.$$

Из формулы для определения частной производной следует, что для применения метода обратного распространения ошибки активационная функция нейронов должна быть ***дифференцируемой***.

Отметим, что функция одной переменной является дифференцируемой в некоторой точке тогда и только тогда, когда в этой точке она имеет производную.

В качестве активационной функции часто используют сигмоидную функцию (рис. 2.18)

$$f(x) = \frac{1}{1+e^{-x}}.$$

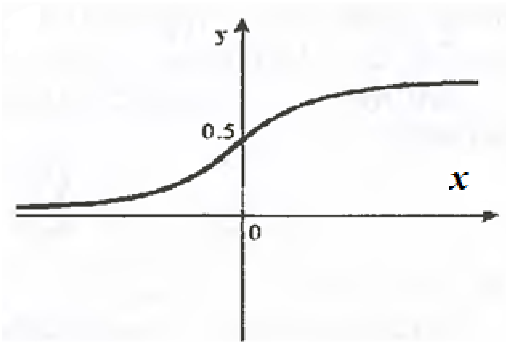

Рисунок 2.18 Сигмоидная функция

При изменении аргумента от $-\infty$ до $+\infty$ её значения лежат в диапазоне от 0 до 1. Такая функция дифференцируема при всех значениях её аргумента. Производная этой функции имеет вид

$$\frac{d}{dx} f(x) = f(x)[1 - f(x)] = \frac{e^{-x}}{(1+e^{-x})^2}.$$

Отсюда следует, что искомая частная производная $\partial S / \partial w_{jr}$ определяется по формуле

$$\frac{\partial S}{\partial w_{jr}} = (\sum_{m=1}^{R} \delta_m) \frac{e^{-q_r}}{(1+e^{-q_r})^2} z_j.$$

Таким образом, изменение коэффициента w_{jr} на одном шаге процесса обучения должно производиться на величину

$$\Delta w_{jr} = \eta(\sum_{m=1}^{R} \delta_m) \frac{e^{-q_r}}{(1+e^{-q_r})^2} z_j.$$

В частности,

$$\Delta w_{11} = \eta(\sum_{m=1}^{R} \delta_m) \frac{e^{-q_1}}{(1+e^{-q_1})^2} z_1,$$

$$\Delta w_{pR} = \eta(\sum_{m=1}^{R} \delta_m) \frac{e^{-q_R}}{(1+e^{-q_R})^2} z_p.$$

Для оптимизации коэффициентов $v_{01},...,v_{np}$ связи нейронов скрытого слоя с нейронами входного слоя используется тот же подход, который применён при определении коэффициентов связи $w_{01},...,w_{pR}$.

2.7 Искусственные нейронные сети на основе радиально-базисных функций (RBF сети)

Рассматривается двухслойная искусственная нейронная сеть с прямым распространением сигналов (без обратных связей), которая содержит скрытый слой нейронов, функционирующих на основе радиально-базисных (радиально-симметричных) функций. Эти нейроны называются радиально симметричными нейронами, а вся сеть – радиальной базисной нейронной сетью (Radial Basis Function Network, сеть RBF).

Искусственные нейронные сети на основе радиально-симметричных (радиально-базисных) функций могут использоваться для решения широкого круга задач, например, задач аппроксимации зависимостей и классификации объектов.

Радиально-симметричные функции относятся к специальному классу функций. Их особенность состоит в том, что значение такой функции монотонно убывает при отклонении аргумента в ту или иную сторону от некоторого центрального значения. К числу таких функций относится, например, функция Гаусса. Если аргументом такой функции является скалярная величина x, то функция Гаусса $f(x)$ (рис. 2.19) имеет вид

$$f(x) = ae^{-\frac{(x-c)^2}{2\sigma^2}},$$

где c – коэффициент сдвига; σ – коэффициент масштаба, всегда имеющий положительное значение; a – коэффициент, определяемый по формуле $a = \dfrac{1}{\sigma\sqrt{2\pi}}$.

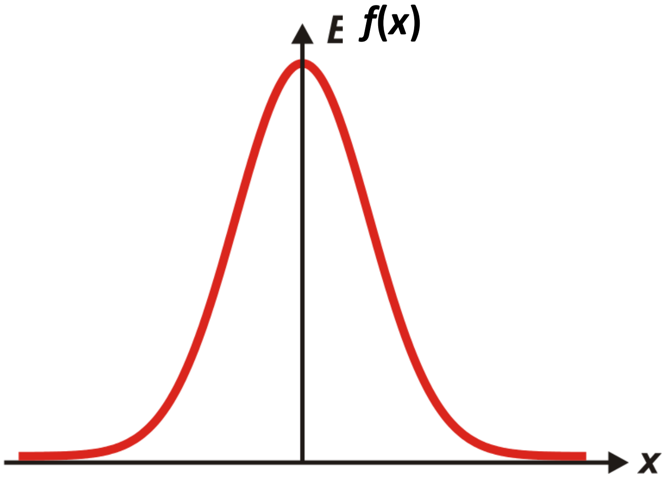

Рисунок 2.19 График функции Гаусса при c = 0.

При построении искусственной нейронной сети чаще используют следующую форму описания радиально-симметричной функции векторного аргумента x для i - го нейрона ($i = 1,...,m$):

$$h_i(x) = \exp(-\frac{\|x - c_i\|^2}{2r_i^2}), \qquad (2.1)$$

где m – количество нейронов сети, функционирующих на основе радиально-базисных функций;

$x = (x_1, x_2,..., x_M)^T$ – M x1 - ***вектор входных значений***, подаваемых на искусственную нейронную сеть;

$h(x) = [h_1(x), h_2(x),..., h_m(x)]^T$ – m x1 - ***вектор выходных сигналов нейронной сети***;

$c_i = (c_{1i}, c_{2i},..., c_{Mi})^T$ – M x1 - ***вектор координат центра*** радиально-симметричной функции;

$\|x - c_i\|$ - ***норма вектора*** отклонений входной векторной переменной x от центра радиально-симметричной функции i - го нейрона. Параметр r_i называется радиусом рассеяния входных переменных для i - го нейрона.

Норма вектора $\|x - c_i\|$ рассчитывается как евклидово расстояние

$$\|x - c_i\| = \sqrt{(x_1 - c_{1i})^2 + (x_2 - c_{2i})^2 + ... + (x_M - c_{Mi})^2}\,.$$

На рис. 2.20 приведена типовая структура искусственной нейронной сети на основе радиально-симметричных функций.

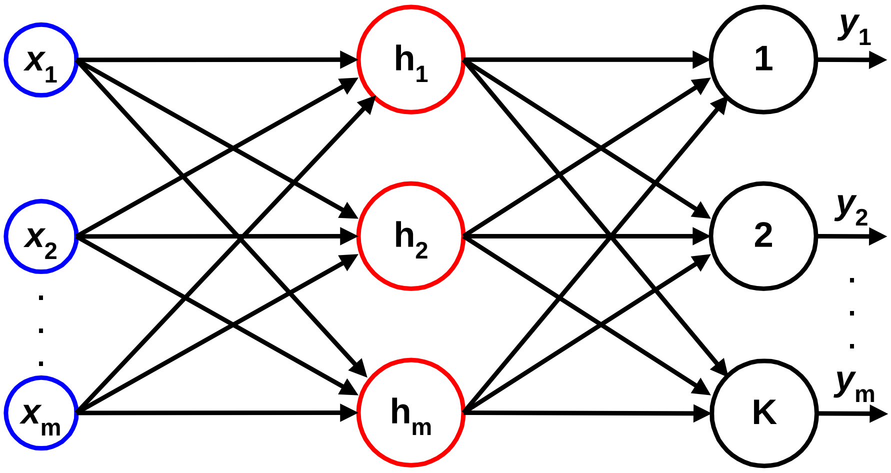

Рисунок 2.20 Структура нейронной сети на основе радиально-симметричных функций

Данная структура содержит два слоя нейронов. Сеть имеет M входов и m нейронов в первом (скрытом) слое. Поэтому входные воздействия можно представить в виде M – мерного вектора $x=(x_1, x_2,..., x_M)^T$. Все компоненты входного вектора влияют на выходные сигналы всех нейронов скрытого слоя. Выходы нейронов первого слоя $h_1(x), h_2(x),..., h_m(x)$ активируются радиально-симметричными функциями (2.1) и зависят от степени близости компонентов вектора входных значений x к центрам радиально-симметричных функций.

Во втором слое нейронной сети используется K нейронов, каждый из которых соединён с выходами всех нейронов первого слоя. Синаптические веса связей выходов нейронов первого слоя со входами нейронов второго слоя представлены $m \times K$ - матрицей $W=\{w_{ij}\}$.

Компоненты вектора выходов нейронов второго слоя, представляющего собой вектор выходов всей нейронной сети

$$Y = (y_1, y_2, ..., y_M)^T,$$

являются линейными комбинациями выходов нейронов первого слоя и для всех j от 1 до M вычисляются по формуле

$$Y = Wh(x)$$

или в развёрнутом виде – по формуле

$$y_j = \sum_{i=1}^{m} w_{ij} h_i(x).$$

Состав и количество входов и выходов определяются классом решаемой задачи. В целом можно констатировать, что нейронные сети с радиально-базисными функциями, как правило, имеют большее количество нейронов, чем другие типы сетей с прямой передачей сигналов, обучаемых методом обратного распространения ошибки. Но они оказываются более эффективными, если для их обучения может использоваться достаточно большое число обучающих векторов. Кроме того, как показывает практика, для обучения радиально-базисных сетей требуется меньше времени, чем для сетей, обучаемых по методу обратного распространения ошибки.

При решении задачи аппроксимации зависимостей, определяемых входными данными, входы сети рассматриваются в качестве аргументов аппроксимирующей зависимости. Выходы сети представляют собой значения функции, полученные в результате аппроксимации.

При решении задачи классификации данных на входы сети поступают характеристические признаки исследуемых объектов. По ним различаются объекты, относимые к тем или иным классам. Значения выходных сигналов сети указывают на соответствующий входам класс, к которому отнесён исследуемый объект.

Выбор количества m нейронов в скрытом слое зависит от решаемой задачи. Если речь идёт о задаче аппроксимации данных, то это количество может быть любым. Но в случае классификации данных m должно соответствовать количеству образов классов.

Перед началом использования искусственной нейронной сети она должна быть настроена на решение определённых задач. Настройка сети производится в ***два этапа***. Сначала определяется структура сети и выбираются значения параметров радиально-базисных функций нейронов скрытого слоя. Затем проводится обучение сети в результате которого определяются оптимальные значения синаптических коэффициентов связей выходов нейронов скрытого слоя со входами нейронов второго (выходного) слоя.

На первом этапе настройки рассматриваемой нейронной сети необходимо определить центры c и радиусы r радиально-базисных функций, определяющих работу нейронов скрытого слоя.

Выбор центров целесообразно осуществлять с учётом следующих соображений. В случае, если количество эталонных образцов для обучения относительно невелико, рекомендуется использовать входные векторы – образцы в качестве центров радиально-симметричных функций нейронов. Если объём обучающей выборки достаточно велик, то в качестве центров могут быть использованы отдельные векторы из состава обучающей выборки, выбираемые случайным образом. Этот подход наиболее эффективен при большом количестве нейронов в скрытом слое.

Выбор значений радиусов радиально-базисных функций нейронов чаще всего осуществляется разработчиком нейронной сети на основе эвристического метода. Значения параметра r влияет на чувствительность нейронной сети. Поэтому для достижения наибольшей точности работы нейронной сети может потребоваться рассмотрение качества работы сети при нескольких вариантах значений параметров c и r нейронов скрытого слоя.

На этапе оптимизации весовых коэффициентов линейного выходного слоя последовательно выполняются следующие действия.

Рассчитывается характеристическая (интерполяционная) $N \times m$ - матрица значений радиально-симметричных элементов всех обучающих примеров:

$$H = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \ldots & h_m(x_1) \\ h_1(x_2) & h_2(x_2) & \ldots & h_m(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ h_1(x_N) & h_2(x_N) & \ldots & h_m(x_N) \end{bmatrix}.$$

Количество строк данной матрицы равно количеству N примеров в обучающей выборке. Количество столбцов – количеству m нейронов в скрытом слое.

Рассчитывается матрица W весовых коэффициентов выходного слоя нейронов. С целью минимизации среднего квадрата ошибки применяют метод определения псевдообратных матриц, в соответствии с которым матрица W рассчитывается по формуле

$$W = \left(H^T H\right)^{-1} H^T Y_0, \qquad (2.2)$$

где $N \times K$ - матрица Y_0 выходов обучающих примеров содержит строки в количестве, равном числу N обучающих примеров, и столбцы в количестве, соответствующем числу K выходов нейронной сети:

$$Y_0 = \begin{bmatrix} y_{11} & y_{21} & \cdots & y_{K1} \\ y_{21} & y_{22} & \cdots & y_{K2} \\ \cdots & \cdots & \cdots & \cdots \\ y_{1N} & y_{2N} & \cdots & y_{KN} \end{bmatrix}.$$

Искусственные нейронные сети с радиально-базисными функциями имеют ряд преимуществ, к числу которых относятся следующие. Во-первых, для аппроксимации сложных нелинейных зависимостей достаточно иметь единственный скрытый слой. Во-вторых, следует отметить простоту оптимизации значений синаптических коэффициентов. Используемый для этого метод обучения свободен от проблем, связанных с локальными минимумами. Третье, и существенное преимущество заключается в высокой скорости обучения. Она существенно выше скорости обучения многослойных сетей по методу обратного распространения ошибки.

Но у нейронных сетей на основе радиально-базисных функций имеются и недостатки. Прежде всего, следует упомянуть отсутствие методов однозначного определения значений параметров радиально-базисных функций и необходимость применения эвристических подходов и рассмотрения множества вариантов настройки сети. Кроме того, такие сети, хорошо справляясь с задачей интерполяции, обладают относительно плохими экстраполирующими свойствами и оказываются весьма громоздкими, требующими значительного числа нейронов при большой размерности вектора входных воздействий.