

## РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ ДАННЫХ НА ОСНОВЕ МНОГОСЛОЙНОГО ПЕРСЕПТРОНА

© 2019 О. Ю. Лавлинская, В. О. Логвина

*Воронежский институт высоких технологий (г. Воронеж, Россия)*

*Большой объем информации, которая поступает на портал государственных услуг требует использования специальных методов ее обработки и последующего анализа. Классификация информации на ЕПГУ подразумевает анализ собранных данных различными средствами, к примеру на основе нейросетевых технологий, так нейронные сети используются для решения сложных задач, требующих аналитических вычислений подобных тем, что делает человеческий мозг.*

*Ключевые слова: интеллектуальный анализ, нейронная сеть, синапс, двухслойный персептрон.*

Интеллектуальный анализ данных – это некий процесс поиска различных сведений, которые подходят для использования на большом множестве информации. В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных. Обычно такие закономерности (последовательности) нельзя обнаружить при традиционном просмотре данных, поскольку связи слишком сложны, или из-за чрезмерного объема данных.

На основе анализа статистической информации можно составить прогнозы различных показателей на будущие периоды, выявить потребности пользователей в зависимости от классов (по личным данным, по месту проживания, по категории и прочим). При анализе статистических данных наиболее перспективным направлением является использование нейросетевых технологий.

Под искусственными нейронными сетями (ИНС) понимаются вычислительные структуры, которые моделируют биологические процессы, ассоциируемые с процессами человеческого мозга. Нейронная сеть, другими словами, представляет собой последовательность нейронов, соединенных между собой синапсами. Из-за свойств такой структуры, машина обретает способность анализировать и даже запоминать различную информацию.

Нейронная сеть используется для реализации сложных отношений, с помощью

которых можно решать широкий класс задач математики, управления, информационной безопасности и других.

Рассмотрим основные параметры нейрона:

- входные  $x_i$ ;
- вещественные весовые коэффициенты  $w_i$ ;
- уровень активации (потенциал) нейрона:

$$P = \sum_{i=1}^n w_i x_i \quad (1)$$

- функция активации  $f(P)$  – предназначена для вычисления выходного значения сигнала.

Выход нейрона есть функция его состояния  $y=f(S)$ .

Нейрон реализует функцию векторного аргумента:

$$P = \sum_{i=1}^n x_i w_i + b_i, y = f(S) \quad (2)$$

Пусть имеются  $n$  объектов  $I_1, \dots, I_n$ , представляющих собой пользователей портала государственных услуг. Каждый пользователь имеет свой набор из  $L$  параметров  $P_1, \dots, P_L$ .

Результат измерения  $i$ -го параметра объекта  $I_j$  обозначим  $x_{ij}$ . Вектор  $X_j = [x_{ij}]$  отвечает каждому ряду измерений (для каждого  $j$ -го пользователя). Таким образом, получаем, что для множества  $I$  объектов задано множество векторов  $X = \{x_1, \dots, x_n\}$  измерений, которые описывают множество  $I$ . Множество  $X$  может быть представлено как  $n$  точек в  $p$ -мерном евклидовом пространстве  $E_p$ . Совокупность значений  $L$  па-

Лавлинская Оксана Юрьевна – Воронежский институт высоких технологий, к. т. н., доцент ВИБТ, lavlin2010@yandex.ru.

Логвина Виктория Олеговна – Воронежский институт высоких технологий, студент, miss.logvina@gmail.com.

раметров могут быть представлены в виде матрицы  $X$ :

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{l1} & x_{l2} & \dots & x_{ln} \end{pmatrix} \quad (3)$$

Рассмотрим множество  $M$  классов  $M_1, \dots, M_M$ , характеризующих пользователя. Требуется определить принадлежность каждого из пользователей  $I_1, \dots, I_n$  одному из классов  $M_1, \dots, M_M$  на основе информации о значениях их параметров  $P_1, \dots, P_L$ , то есть построить отображение  $\alpha: I \rightarrow M$ , ставящее в соответствие состоянию  $I_j \in I$  определенный класс  $M_M \in M, j = 1, \dots, n$ .

Нейронные сети являются универсальным средством приближения функций. Именно это свойство позволяет использовать НС в решении задач классификации. Как правило, ИНС оказываются наиболее эффективным способом классификации, так как они генерируют фактически большое число регрессионных моделей (которые используются в решении задач классификации различными статистическими методами).

Поэтому для решения поставленной задачи целесообразно использовать аппарат искусственных нейронных сетей. Решение задач нейросетевой классификации пользователей на группы (классы) включает следующие основные этапы:

- определение набора классификационных параметров (На данном этапе необходимо составить базу данных из примеров, характерных для данной задачи);
- определение типа и структуры нейронной сети (количества слоёв, количества нейронов в каждом слое, связей между нейронами);
- обучение нейронной сети (Выбрать алгоритм обучения сети);
- использование обученной нейронной сети для решения задач классификации.

Следует обратить внимание на структуру сети, проектирование которой будет проводиться. Для решения задачи классификации может быть использована сеть прямого распространения информации – многослойный персептрон.

Многослойный персептрон представляет собой сеть, в которой входной сигнал преобразуется в выходной, проходя последовательно через несколько слоев. Первый слой называют входным, последний - выходным. Эти слои содержат так называемые

вырожденные нейроны и иногда в количестве слоев не учитываются. Кроме входного и выходного слоев, в многослойном персептроне есть один или несколько промежуточных слоев, которые называют скрытыми. В этой модели персептрона должен быть хотя бы один скрытый слой. Присутствие нескольких таких слоев оправдано лишь в случае использования нелинейных функций активации. Такую сеть легко можно интерпретировать как модель вход-выход, в которой веса и пороговые значения (смещения) являются свободными параметрами модели.

Нейронные сети обладают рядом преимуществ, таких как устойчивость к шумам входных данных, адаптацией к изменениям, сверхвысоким быстродействием и отказоустойчивостью. Искусственные нейронные сети являются мощным и одновременно гибким методом имитации процессов и явлений. Современные искусственные нейронные сети представляют собой программно-аппаратные средства создания специализированных моделей и устройств, а также позволяют решать широкий круг задач диагностики на основе применения алгоритмов теории распознавания образов.

Для решения задач классификации пользователей портала госуслуг на группы выбран двухслойный персептрон. Обобщенная схема двухслойного персептрона представлена на рисунке 1.

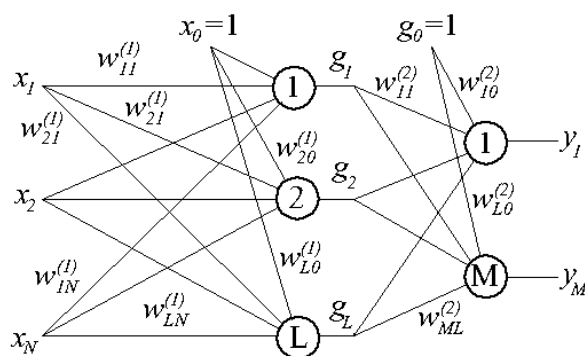


Рисунок 1. Двухслойный персептрон.

Обозначения на схеме:  $x_1, x_2, \dots, x_N$  – это входные сигналы нейронной сети, представляющие собой классификационные параметры, характеризующие каждого пользователя портала госуслуг.

Рассмотрим эти признаки: Пол пользователя, Возраст, Сфера деятельности, Категория пользователя, Место жительства, Семейное положение, Наличие детей, Количество подключенных услуг 1-й группы, Количество подключенных услуг 2-й группы,

Количество подключенных услуг 3-й группы, Количество подключенных услуг 4-й группы, Количество подключенных услуг 5-й группы, Количество подключенных услуг 6-й группы, Количество подключенных услуг 7-й группы, Количество подключенных услуг 8-й группы, Количество подключенных услуг 9-й группы.

В качестве групп будем использовать основные категории, на которые можно разделить услуги.

Входные параметры составляют входной слой;  $g_1, g_2, \dots, g_L$  – выходные сигналы первого слоя, которые образуют первый скрытый слой;  $y_1, \dots, y_M$  – выходные сигналы, определяющие принадлежность пользователя к одному из классов и образующие выходной слой;  $w_{ij}^{(m)}$  – значения весов. В схеме верхние индексы в скобках ( $m$ ) означают номер слоя нейрона.

Выходные сигналы нейронных слоев определяются по формулам:

$$g_l = f\left(\sum_{j=0}^N w_{lj}^{(1)} x_j\right), l = \overline{1, L} \quad (4)$$

$$y_i = f\left(\sum_{l=0}^L w_{il}^{(2)} g_l\right) = f\left(\sum_{l=0}^L w_{il}^{(2)} f\left(\sum_{j=0}^N w_{lj}^{(1)} x_j\right)\right) \quad (5)$$

В формулах (4) и (5)  $f$  обозначает функцию активации, которая является одинаковой для всех нейронов сети.

Задача обучения нейронной сети состоит в определении таких значений весовых коэффициентов, которые обеспечивали бы максимальное совпадение выходных сигналов сети с эталонными значениями. Обучение с учителем осуществляется на основе обучающей выборки.

$$\langle X^k, D^k \rangle, k = \overline{1, p}, \quad (6)$$

где  $p$  – количество обучающих пар,

$X^k$  – вектор значений входных сигналов сети;

$D^k$  – вектор ожидаемых значений выходных сигналов.

Задача обучения перцептрона сводится к задаче минимизации целевой функции, зависящей от вектора весовых коэффициентов  $W$ . Пусть  $Y^k$  – вектор выходных сигналов нейронной сети, полученный для входного вектора  $X^k$ , тогда:

$$F(W) = \frac{1}{2} \sum_{k=0}^p \sum_{i=1}^M (y_i - d_i)^2 \rightarrow \min \quad (7)$$

где  $M$  – количество выходных сигналов.

Если мы рассматриваем единичную обучающую выборку  $\langle X, D \rangle$ , то крите-

рий (7) будет выглядеть следующим образом:

$$F(W) = \frac{1}{2} \sum_{i=1}^M (y_i - d_i)^2 \rightarrow \min \quad (8)$$

Используя формулы (4) и (5) составим полную формулу критерия оптимальности:

$$F(W) = \frac{1}{2} \sum_{k=1}^p \sum_{i=1}^M \left( f\left(\sum_{l=0}^L w_{il}^{(2)} f\left(\sum_{j=0}^N w_{lj}^{(1)} x_j\right)\right) - d_i \right)^2 \rightarrow \min \quad (9)$$

Критерии, которые формируются на основе этих выражений являются сложной структурой, а также характеризуются нелинейностью, высокой размерностью. Из этого следует повышение требований к используемым алгоритмам обучения.

При решении задачи классификации количество нейронов скрытого слоя сделаем равным  $2/3N$  или  $1/2N$  для избегания переобучения. Компоненты выходного вектора нейронной сети можно представить в виде:

$$y_i(X) = \begin{cases} 1, & I \in K_i \\ 0, & \text{иначе} \end{cases}, \quad (10)$$

Во всех слоях сети в качестве функции активации возьмем сигмоидальную функцию, которая является монотонно возрастающей и всюду дифференцируемой S-образная нелинейной функцией с насыщением. Сигмоид позволяет усиливать слабые сигналы и не насыщаться от сильных сигналов. Формула представлена ниже:

$$f(y) = \frac{1}{1 + e^{-\alpha y}} \quad (11)$$

где  $\alpha$  – параметр наклона сигмоидальной функции активации. Изменяя этот параметр, можно построить функции с различной крутизной.

Для обучения нейронной сети целесообразно использовать алгоритм обратного распространения ошибки. Алгоритм, представленный далее, применим к нейронной сети с одним скрытым слоем, что является допустимой и адекватной ситуацией для большинства приложений. Основная идея этого метода состоит в распространении сигналов ошибки от выходов сети к её входам, в направлении, обратном прямому распространению сигналов в обычном режиме работы.

Все нейроны в сети получают векторы входных значений, производят их взвешенное суммирование и преобразуют полученную сумму  $s$  с помощью активационной функции. Общая ошибка сети формируется на выходном слое.

Функция ошибки (5) минимизируется методом градиентного спуска. Суть метода заключается в том, что двигаясь в направлении, противоположном градиенту функции  $\Delta w_{ji}$ , мы приближаемся к минимуму функции  $E(w)$ :

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}}, \quad (12)$$

Здесь  $0 < \eta < 1$  задает скорость обучения;  $w_{ji}$  – коэффициент связи  $i$  нейрона слоя  $n - 1$  с  $j$  нейроном слоя  $n$ .

Покажем способ вычисления градиента  $\partial E / \partial w$ . Запишем обобщенно схему работы персептрона, в которой каждый нейрон рассчитывает взвешенную сумму своих входов:

$$a_i = \sum_n w_{ji} z_i, \quad (13)$$

Здесь  $z_i$  – это входы нейрона и соответственно выходы предыдущего слоя нейрон, как показано на рисунке 2.

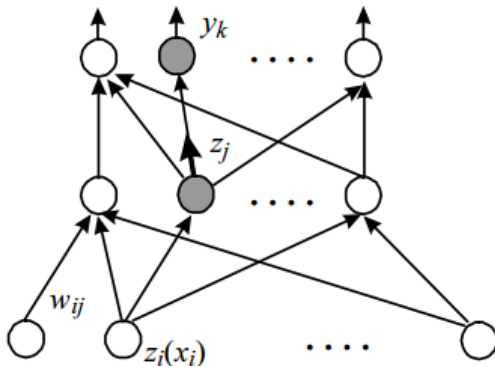


Рисунок 2. Часть многослойной нейронной сети.

Выход нейрона  $j$  – это преобразование суммы  $a_j$  пороговой функцией  $g$ :

$$z_j = g(a_j) \quad (14)$$

Вернемся к  $\partial E / \partial w$  и замети, что  $E(w)$  является сложной функцией, для которой  $w = w(a_j)$  является функцией от  $a_j$ . Отсюда по правилу вычисления сложной функции имеем:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}}, \quad (15)$$

где  $\frac{\partial E}{\partial w_{ij}}$  – это значение производной функции ошибки по весу  $w_{ij}$ , между  $i$  и  $j$  нейронами.

Преобразуем формулу, используя выражение (13):

$$\frac{\partial a_j}{\partial w_{ij}} = \frac{\partial \sum w_{ij} z_i}{\partial w_{ij}} = z_i. \quad (16)$$

Выражение  $\delta_j = \partial E / \partial a_j$  назовем ошибкой на слое  $j$ . Проведя анализ можно увидеть, что для того, чтобы вычислить градиент на каждом слое  $i$ , необходимо вычислить произведение величины ошибки предыдущего  $i$ , находящегося сверху слоя  $j$ , на входное значение слоя:

$$\frac{\partial E}{\partial w_{ji}} = \delta_j z_i. \quad (17)$$

Теперь перейдем к вычислению ошибки  $\delta_j$  для каждого слоя нейронной сети. Для выходного слоя  $\delta_k$  формула будет следующей:

$$\delta_k = \frac{\partial E}{\partial a_k} = \left| \begin{array}{l} a_k = \sum_j w_{jk} z_j \\ y_k = g(a_k) \end{array} \right| = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial a_k} \quad (18)$$

Получается, что  $\partial E / \partial a_k$  – это производная пороговой функции  $\partial E / \partial a_k = g'(a)$ . Для логического сигмоида такая производная будет иметь следующую формулу:

$$g'(a) = g(a)(1 - g(a)), \quad (19)$$

Второй множитель  $\partial E / \partial y_k$  вычисляется по формуле (7) и приводится к следующему виду:

$$\frac{\partial E}{\partial y_k} = (y_k - t_k) \quad (20)$$

Таким образом, общий вид ошибки для выходного слоя будет определяться формулой:

$$\delta_k = g'(y_k - t_k) \quad (21)$$

Формула (22) приведена по аналогичному принципу, но для промежуточного слоя:

$$\delta_j = \frac{\partial E}{\partial a_j} = \left| \begin{array}{l} a_k = \sum_j w_{jk} z_j \\ z_i = g(a_j) \end{array} \right| = \sum \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_j} \frac{\partial z_j}{\partial a_j}, \quad (22)$$

Суммирование происходит по всем  $k$ , к которым нейрон  $j$  посылает сигнал, как показано на рисунке 3.

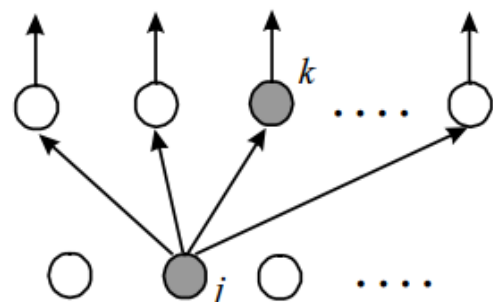


Рисунок 3. Распространение ошибки.

В формуле  $\partial E / \partial u_k$  представляет собой тоже самое, что и  $\delta_k$ , а  $\partial z_j / \partial a_j$  – это производная пороговой функции  $g'(a_j)$ . От множителя  $\partial a_k / \partial z_j$  остается просто вес  $w_{jk}$ . Таким образом, совместив все эти понятия получим довольно простую формулу (23):

$$\delta_j = g'(a_j) \sum_k \delta_k w_{jk} \quad (23)$$

Таким образом, ошибка на каждом слое вычисляется через значения ошибки на предыдущих слоях: коррекция ошибки распространяется как будто обратно по нейронной сети.

Запишем кратко приведенный алгоритм:

1. Перед применением алгоритма весам  $w_{ji}$  присваиваются случайные значения.
2. На вход нейронной сети (персептрона) поступает вектор  $x^n$  и на выходе получаем значение  $y^n$  по выражениям (6) и (7).
3. Производится вычисление ошибки  $\delta_k$  для выходного слоя сети по формуле (21).
4. Производится вычисление ошибки  $\delta_j$  для скрытых слоев по выражению (23).
5. По формуле (12) находим значения градиента.

6. Корректируем значения синаптических весов.

7. Производится вычисление значения ошибки (5). Если найденная величина ошибки не устраивает, то осуществляем переход на шаг 2.

В итоге, делаем вывод о том, что применение нейросетевых технологий, в частности использование сети прямого распространения информации непосредственно подходит для решения задачи классификации данных на портале госуслуг, так как многослойный персептрон имеет достаточно гибкую структуру и для его обучения имеется широкий диапазон алгоритмов.

#### ЛИТЕРАТУРА

1. Галушкин, А. И. Нейронные сети / А. И. Галушкин. – Горячая линия-Телеком, 2010. – 496 с.
2. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский. – М: Горячая линия Телеком, 2013. – 384 с.

### SOLVING THE TASK OF DATA CLASSIFICATION BASED ON A MULTILAYERED PERCEPTRON

© 2019 O. Yu. Lavlinskaya, V. O. Logvina

Voronezh Institute of High Technologies (Voronezh, Russia)

*A large amount of information that enters the portal of public services requires the use of special methods for its processing and subsequent analysis. Classification of information of the EPGU involves analyzing the collected data by various means, for example, based on neural network technologies, so neural networks are used to solve complex problems that require analytical calculations similar to what the human brain does.*

*Keywords: intellectual analysis, neural network, synapse, double-layer perceptron.*