УДК 004.032.26

Л.С. Ломакина, К.М. Носков

НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ ДИАГНОСТИРОВАНИЯ СОСТОЯНИЙ БИОЦЕНОЗА НА ОСНОВЕ АПРИОРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Нижегородский государственный технический университет им. Р. Е. Алексеева

Рассматривается проблема классификации состояний биоценозов как многомерных объектов. Предлагается алгоритм построения классификатора многомерных объектов, в частности, классификации биоценозов, с применением нейронных сетей. Рассмотрена структура нейронной сети на радиально-базисных функциях. Описан алгоритм обучения нейронной сети.

Ключевые слова: классификация, классификатор, нейронная сеть, биоценоз, радиально-базисные функции.

Введение

Медико-биологический объект (биоценоз) – это совокупность разных видов микроорганизмов, связанных между собой определенными отношениями и населяющими определенную биологическую нишу [1]. Частным случаем биоценоза является микрофлора желудочно-кишечного тракта (ЖКТ) человека. Путем сравнения состояния микрофлоры ЖКТ пациента с эталонными значениями можно выявить риск возникновения опасных заболеваний. Качественные и количественные изменения в составе микрофлоры кишечника отражаются на состоянии как самой микрофлоры кишечника, так и на состоянии всего организма в целом. Правильная трактовка результата может иметь решающий вклад в постановку диагноза и выбора соответствующего метода лечения.

Для обработки информации в многомерном признаковом пространстве необходимы новые, более эффективные методы классификации в сложной помеховой обстановке и в условиях априорной неопределенности. Создание системы автоматического диагностирования состояний биоценоза позволит повысить скорость и точность постановки диагноза врачом.

Задача классификации

Классификация – группировка множества объектов по некоторым классификационным признакам, отражающая степень сходства объектов между собой и принадлежность к заранее определенным классам. Необходимое условие для проведения классификации – существование заранее известных классов и характеризирующих их признаков [2].

Цель процесса классификации состоит в построении математической модели, которая принимает на вход прогнозирующие атрибуты (признаки), а в качестве результата на выходе класс, которому соответствуют такие атрибуты. Процесс классификации заключается в создании связей между множеством признаков объектов и множеством классов и состоит из двух основных этапов: конструирования модели и её использования.

Классификатор – реализация процесса классификации, которая определяет объект в один из предопределенных классов по входному вектору признаков.

Методы классификации

Все существующие методы классификации имеют достоинства и недостатки и применяются в зависимости от поставленной задачи. Их оценку следует проводить, принимая в расчет следующие характеристики:

- скорость – величина, характеризующая время, требуемое на создание классификатора и его использование;

- робастность – устойчивость к разного рода некорректным данным, зашумленным данным, а также пропусков некоторых переменных;
- интерпретируемость, т.е. прозрачность устройства для возможности анализа модели;
- надежность – классификация должна предусматривать работу с зашумленными данными и выбросами.

Нейронные сети можно вынести в отдельный большой класс методов классификации. Сети с прямой синоптической связью являются универсальным методом аппроксимации. Это свойство позволяет использовать их так же и в задачах классификации. Причем такие нейронные сети оказываются одним наиболее эффективных способов классификации.

Нейронные сети имеют следующие преимущества перед классическими методами классификации:

- гибкую структуру;
- позволяют строить нелинейные зависимости;
- параллельные вычисления;
- возможность модификации под конкретную задачу;
- работу с зашумленными данными;
- быстрые алгоритмы обучения;
- справляются с задачами, не имеющими явного решения.

Однако наряду с большим количеством положительных черт нейронных сетей, существует ряд проблем при их использовании.

Никогда заранее не известно, какая нужна сложность вычислительной сети.

М. Минский в работе «Персептроны» доказал, что только многослойные нейронные сети, имеющие скрытый слой, способны решать нелинейные задачи. Простейшая нейронная сеть, построенная из одного слоя персептронов, сможет решить только линейные задачи.

Базовая модель

Базовая модель, описывающая состояние ЖКТ, представляет собой *n*-мерное пространство признаков, которые априорно разделены на четыре класса, соответствующие степени дисбактериоза пациентов или его отсутствию. Результат отдельного пациента представлен в виде вектора $\vec{X} = (x_1, x_2, \dots, x_n)$в *n*-мерном евклидовом пространстве, координатами которого являются скалярные величины, каждая из которых равна количеству микроорганизмов данного вида. Модель построена на исходных данных по исследованию количественно-качественного состава микрофлоры желудочно-кишечного тракта, собранных Нижегородским научно-исследовательским институтом эпидемиологии и микробиологии имени И.Н. Блохиной. Каждая запись представлена совокупностью 29 признаков, характеризующих состояние микрофлоры желудочно-кишечного тракта по микроорганизмам 376 видов из 70 родов (табл. 1).

Таблица 1

Признаки, характеризирующие состояние микрофлоры ЖКТ

1	**Количество Bifidobacterium spp.**
2	**Количество Lactobacillus spp.**
3	**Количество Lactococcus spp.**
4	**Количестводругиханаэробныхмикроорганизмов**
5	**Количество Bacteroides spp.**
6	**Количество E.coli (лак+)**
7	**Количество E.coli (л/д)**
8	**Количество E.coli (лак-)**
9	**Количество E.coli (гем+)**
10	**Количество E.coli(всего)**
11	**Количество Enterococcus spp.**
12	**Количество Enterococcus (гем+)**

13	**Количество Staphylococcus epidermidis**
14	**Количество Staphylococcus aureus**

Окончание табл. 1

1	2
15	**Количество Klebsiella spp.**
16	**Количество Enterobacter spp., Pantoea spp.**
17	**Количество Citrobacter spp.**
18	**Количество Proteus mirabilis**
19	**Количество Morganella morganii**
20	**Количество Proteus vulgaris**
21	**Количество Providencia spp. и др.**
22	**Количество Hafnia spp., Serratia spp.**
23	**Количество Pseudomonas aeruginosa**
24	**Количество НГОБ**
25	**Количество Candida spp.**
26	**Количестводрожжевыхклеток**
27	**Количество Shigella spp.**
28	**Количество Salmonella spp.**
29	**Количество ЭПКП**

Единицы измерения признаков не должны влиять на результаты нейросетевого моделирования. Поэтому следует провести предварительную нормировку исходных данных, т.е. привести все исходные данные к заданному диапазону.

Один из способов нелинейной нормировки – с использованием сигмоидной логистической функции или гиперболического тангенса [3]. Переход от традиционных единиц измерения к нормированным осуществляется следующим образом:

- при нормализации в пределах [0, 1]: $\overline{x_{ik}} = \frac{1}{e^{-\alpha(x_{ik}-x_{ci})}+1}$;
- при нормализации в пределах [–1, 1]: $\overline{x_{ik}} = \frac{e^{-\alpha(x_{ik}-x_{ci})}-1}{e^{-\alpha(x_{ik}-x_{ci})}+1}$,

где $x_{ci} = \frac{x_{min}+x_{max}}{2}$ – центр нормализуемого интервала измерения входных переменных;

α – параметр, влияющий на степень нелинейности изменения переменной в нормализуемом интервале.

Способ представления выходных значений

Одним из способов представления выходных значений является вектор, компоненты которого соответствуют классам. Результатом классификации будет класс под номером максимального компонента выходного вектора. Такой подход позволяет получить наряду с результатом вероятность принадлежности этому классу.

При классификации состояний микрофлоры ЖКТ, выходной вектор может принимать следующий вид:

$\vec{Y}$={здоров; болен} – при разделении на 2 класса;

$\vec{Y}$={здоров; 1 степень; 2 степень; 3 степень} – при разделении на 4 класса.

Структура сети

В качестве классификатора использована нейронная сеть на радиально-базисных функциях. Нейронные сети такого типа широко применяются при аппроксимации функций с множеством переменных и в качестве классификаторов[2]. В отличие от многослойных сетей, радиальные сети обладают свойствами, позволяющими производить более простое отображение характеристик. Сигмоидальные нейронные сети решают задачи глобальной ап-

проксимации, так как значение, отличное от нуля, применяемой активационной функции распространяется от некоторой точки в многомерном пространстве до бесконечности [4]. Радиальные функции имеют значения, отличные от 0, только в некоторой ограниченной области вокруг центра радиального элемента в виде сферы в многомерном признаковом пространстве.

Таким образом, радиальные элементы реализуют методы, связанные с локальными отображениями данных. Это является главной особенностью сетей на радиально-базисных функциях и позволяет существенно упростить структуру сети и соответственно ускорить её обучение [5].

Структура сети (рис. 1) содержит два слоя нейронов. Выходы первого (скрытого) слоя определяют степень близости входных значений к центрам радиально-базисных функций. Выходы нейронов второго слоя – это линейные комбинации выходов скрытого слоя.

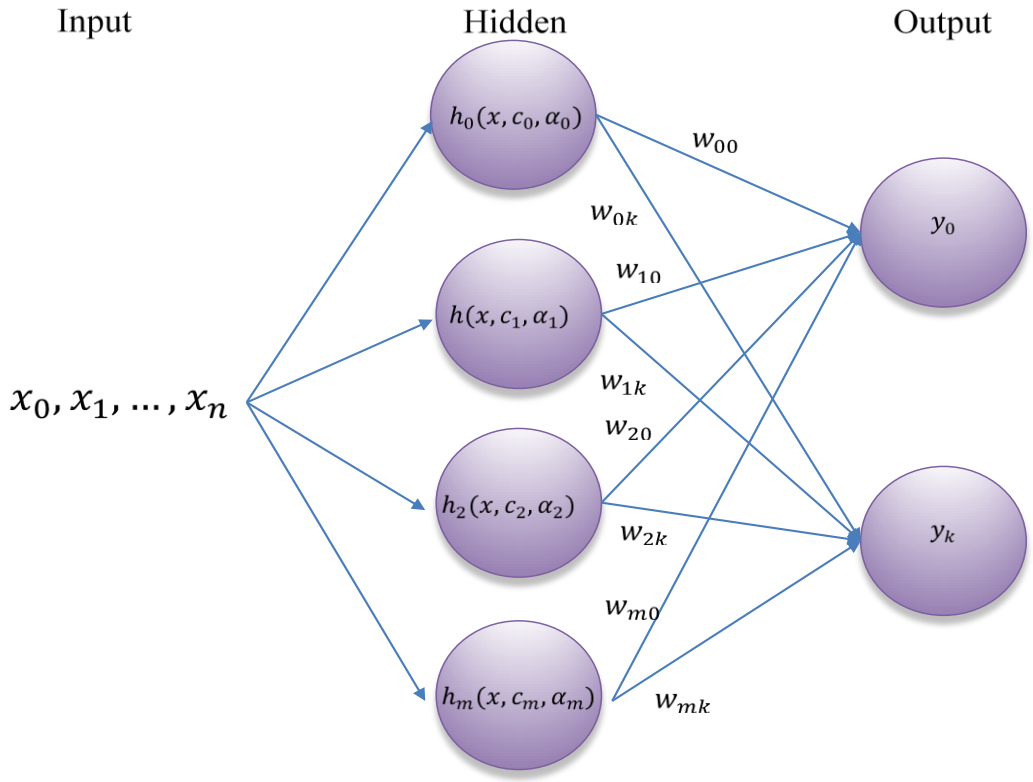

Рис. 1. Архитектура используемой нейронной сети на радиально-базисных функциях

Наиболее широко применяющиеся на практике Гауссовские функции (1) имеют локальный характер и позволяют установить зависимость между областью данных в многомерном признаковом пространстве и конкретным радиальным элементом [5]:

$$\vec{h}(x) = \exp(-\alpha \cdot \|x - \vec{c}\|^2), \quad (1)$$

где $\vec{c}$ - вектор центров множества радиально симметричных функций;

$\|x - \vec{c}\|$ - норма вектора отклонений входной переменной от центров радиально-симметричных функций. Параметр α связан с радиусом рассеяния входных переменных, ri может быть заменен в выражении (1) на соответствующее отношение: $\alpha = \frac{1}{2r^2}$.

Норма разности векторов рассчитывается как евклидово расстояние:

$$\|x - \vec{c}\| = \sqrt{(x - c_1)^2 + (x - c_2)^2 + \ldots + (x - c_m)^2}.$$

Алгоритм обучения нейронной сети

На стадии обучения можно выделить три этапа: подбор центров и радиусов радиально-симметричных функций и оптимизация синоптических коэффициентов линейного выходного слоя.

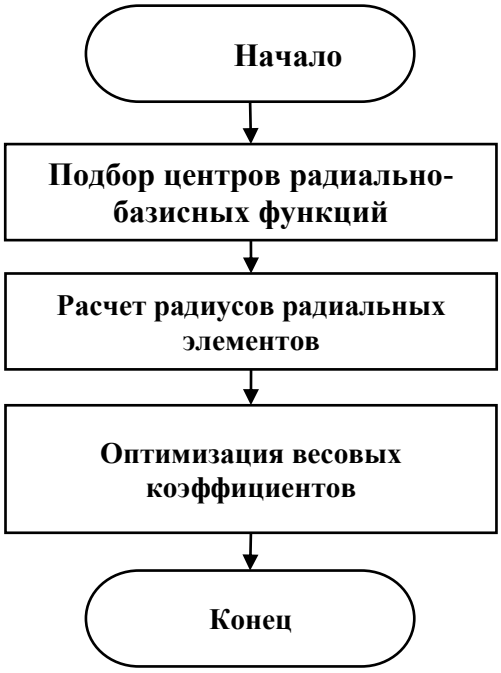

Рис. 2. Блок-схема алгоритма обучения нейронной сети

1. Подбор центров радиально-симметричных функций

При большом объеме обучающих примеров можно воспользоваться несколькими способами подбора центров радиальных элементов:

- в качестве центров можно использовать отдельные случайные входные вектора из обучающей выборки;
- подбор наилучших центров из множества обучающих примеров, основанный на выборе максимально далеко расположенных друг от друга векторов из обучающей выборки;
- использование различных алгоритмов кластеризации в том числе другие нейронные сети, например, сеть Кохонена.

2. Подбор радиусов радиальных элементов

Выбор радиусов определяется требуемым видом радиально-симметричной функции. При больших значениях параметра α график функции слишком острый, а это значит, что сеть не сможет корректно интерполировать данные между известными точками на достаточно большом удалении от них, так как теряет способность к обобщению обучающих данных. Наоборот, при чрезмерно малых значениях параметра α сеть становится невосприимчивой к отдельным деталям.

Радиусы радиальных элементов могут задаваться как вручную при проектировании нейронной сети или автоматически рассчитываться по среднему расстоянию до нескольких (в зависимости от общего объема обучающей выборки и количества скрытых нейронов) ближайших примеров.

3. Оптимизация весовых коэффициентов

Использование псевдообратной матрицы является наиболее простым и быстрым алгоритмом оптимизации весовых коэффициентов:

1) рассчитывается характеристическая матрица $\overline{\overline{H}}$ значений радиально-симметричных элементов всех обучающих примеров:

$$\overline{\overline{H}} = \begin{bmatrix} h_{1(x_1)} & h_{2(x_1)} & \dots & h_m(x_1) \\ h_1(x_2) & h_2(x_2) & \dots & h_m(x_2) \\ \dots & \dots & \dots & \dots \\ h_1(x_n) & h_2(x_n) & \dots & h_m(x_n) \end{bmatrix},$$

где n – количество обучающих примеров; m – количество радиальных элементов;

2) методами линейной алгебры рассчитывается матрица весовых коэффициентов выходного слоя нейронов $\overline{\overline{W}}$:

$$\overline{\overline{W}} = \left(\overline{\overline{H}}^T \cdot \overline{\overline{H}}\right)^{-1} \cdot \overline{\overline{H}}^T \cdot \overline{\overline{Y}},$$

где $\overline{\overline{Y}}$ - матрица выходов обучающих примеров;

$$\overline{\overline{Y}} = \begin{bmatrix} y_{11} & y_{21} & \dots & y_{k1} \\ y_{12} & y_{22} & \dots & y_{k2} \\ \dots & \dots & \dots & \dots \\ y_{1n} & y_{2n} & \dots & y_{kn} \end{bmatrix},$$

n – количество обучающих примеров; k – количество выходов нейронной сети.

Результаты

Создана программная реализация классификатора, построенного на основе нейронной сети. В качестве обучающих и тестовых примеров использованы данные по бактериологическому исследованию микрофлоры кишечника, предоставленные Нижегородским научно-исследовательским институтом эпидемиологии и микробиологии им. Академика И. Н. Блохиной.

Для оценки точности и полноты классификации применен метод кросс-валидации, при котором оценка точности основывается на поочередном разбиении всего множества исходных данных на тестовое и тренировочные множества, пока тестовое множество не охватит весь набор данных. Делении множества на обучающие и тестовые примеры осуществляется путем разбиения всего множества данных на заданное k число частей. Метод состоит из k итераций, в ходе выполнения каждой из которых тестовым множеством называется одна из k частей, которая до этого не являлась тестовым множеством. Такой метод позволяет наиболее полно провести оценку точности разбиения данных по классам.

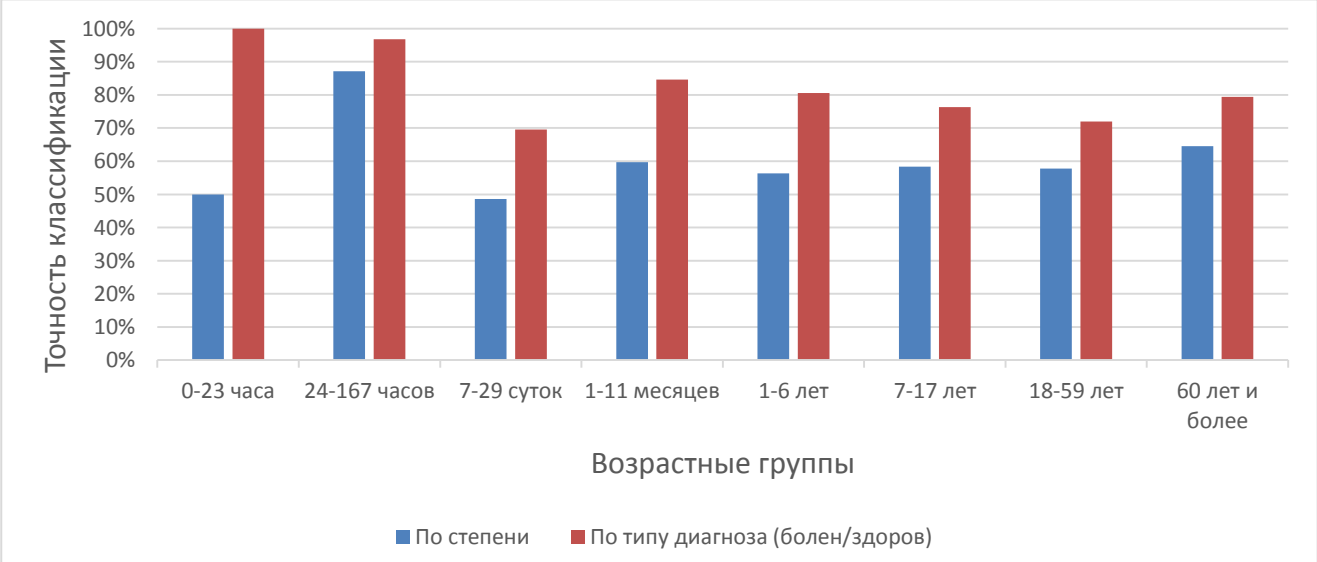

Рис. 3. Точность классификации по возрастным группам

Точность построенного классификатора (рис. 3) различается в зависимости от возрастных групп пациентов, а также от количества обучающих примеров. Общая точность классификации по всем возрастным группам: 62.55% при классификации по степени дисбак-

териоза (4-компонентный выходной вектор) и 83.69% при классификации по типу диагноза (2-компонентный выходной вектор).

Заключение

Рассмотрена возможность применения нейронных сетей на радиально-базисных функциях для построения классификатора многомерных объектов, в частности, биоценозов.

Описана структура и алгоритм обучения используемой нейронной сети на радиально-базисных функциях.

Выполнена оценка точности классификатора, построенного на основе нейронной сети на радиально-базисных функциях.

Библиографический список

1. **Lomakina, L.S.** Expert system forbiocoenosis diagnosis basedon / L.S. Lomakina, I.V. Solovieva, S.A. Zelentsov // Bayesian data analys isandfuzzy productionrule system (Scopus); The 5thBioMedPub 26-27 aug. 2017, Bandung, Indonesia. – P. 6657–6664.
2. **Хливенко, Л.В.** Практика нейросетевого моделирования: монография / Л.В. Хливенко. – Воронеж: Воронежский государственный технический университет, 2015. – 214 с.
3. **Ломакина, Л.С.** Модели и алгоритмы диагностирования состояний биоценоза на основе априорных статистических данных / Л.С. Ломакин [и др.] // Научно-технический вестник Поволжья. Казань: НИКГУ. – 2013. – № 5. – С. 251–256.
4. **Рутковская, Д.** Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский. – М.: Горячая линия - Телеком, 2006. – 452 с.
5. **Осовский, С.** Нейронные сети для обработки информации: [пер. с польского И.Д. Рудинского] / С. Осовский. – М.: Финансы и статистика, 2002. – 344 с.

Дата поступления
в редакцию 31.01.2018

L.S. Lomakina, K.M. Noskov

NEURAL NETWORK TECHNOLOGIES DIAGNOSING BIOCENOSIS STATES BASED ON A PRIORI STATISTICS DATA

Nizhny Novgorod state technical university n.a. R.E. Alekseev

Purpose: Development of models and algorithms for data processing of the gastrointestinal tract microflora bacteriological studies. This allows increasing the speed and quality of diagnosis and choice of treatment

Design/methodology/approach: An algorithm with the use of neural networks for constructing a classifier for multidimensional objects, in particular, classification of biocenoses, is proposed.

Findings: The structure of a neural network based on radial-basis functions is considered. Described neural network training algorithm.

Research limitations/implications: The methods described in the work can find practical application in automated systems for diagnosing biocenosis. The results of work are in demand by scientific institutions and organizations of medical and biological profile

Originality/value: Developed models and algorithms for creating decision support systems for diagnosing biocenosis.

Key words: classification, classifier, neural network, biocenosis, radial basis functions.