ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 04.032.26

КЛАССИФИКАЦИЯ СОСТОЯНИЙ БИОЦЕНОЗА НА ОСНОВЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ

© 2014 г. ***Л.С. Ломакина, А.С. Пожидаева, Я.П. Губернаторов***

Нижегородский государственный технический университет им. Р.Е. Алексеева

yaroslav.gubernatorov@gmail.com

Поступила в редакцию 06.03.2014

Рассматривается задача классификации состояния микрофлоры желудочно-кишечного тракта человека (ЖКТ) для диагностики дисбактериоза. В качестве классификатора используется искусственная нейронная сеть, построенная по архитектуре трехслойного персептрона.

Ключевые слова: классификация, дисбиоз, диагностирование, персептрон, нейронные сети.

Введение

Изучение состояния микробиоты ЖКТ человека является одним из актуальных направлений в современной медицине и биологии. Используя качественную и количественную оценку состава микрофлоры, можно установить, что пациент здоров, либо имеются отклонения от нормы, т.е. наблюдается дисбиоз. Существующие методы трактовки результатов медико-биологического исследования могут носить противоречивый характер из-за сильного влияния субъективных факторов, поэтому актуальной задачей является разработка систем автоматизированного анализа состава микрофлоры ЖКТ человека.

Постановка задачи

Имеются данные по медико-биологическому исследованию людей различных возрастных групп, представленные в виде совокупности значений 29 признаков, характеризующих состояние микрофлоры желудочно-кишечного тракта. Необходимо разработать нейронную сеть, позволяющую различать нормальный состав микрофлоры ЖКТ и дисбактериоз I–III степеней по указанным признакам.

Выбор метода решения задачи

Поставленная задача является задачей классификации и предполагает выделение признаков для отнесения рассматриваемых объектов к заданным группам.

Система основана на нейросетевой модели. Искусственная нейронная сеть в отличие от других известных моделей и методов классификации (регрессионного анализа, классификатора Байеса, метода максимальной энтропии, деревьев принятия решений):

- позволяет моделировать сложные закономерности благодаря большому числу связей;
- реализует нормирование и усреднение, поэтому устойчива к шумам во входных данных;
- основана на вычислении взвешенной суммы, что не требует трудоёмких вычислений и может быть выполнено параллельно на нескольких процессорах ЭВМ.

Анализ исходных данных. Модель внешней среды

Имеются данные по исследованию микрофлоры желудочно-кишечного тракта 2576 человек. Сбор данных произведён Нижегородским научно-исследовательским институтом эпидемиологии и микробиологии им. И.Н. Блохиной. Данные по медико-биологическому исследованию для каждого человека представлены в виде совокупности значений 29 признаков, характеризующих состояние микрофлоры желудочно-кишечного тракта по микроорганизмам 376 видов из 70 родов.

Приведённые признаки можно разделить на четыре группы. В первую группу входят те микроорганизмы, присутствие которых оказывает безусловно положительное влияние на состояние микрофлоры. Причем это влияние усиливается с увеличением количества и

разнообразия этих микроорганизмов. Во вторую группу входят те микроорганизмы, присутствие которых в микробиоте имеет ограниченно отрицательное влияние. Это означает, что наличие их в количествах, меньших 1×10^5, не несет никакой угрозы здоровью человека и даже способствует улучшению воздействия первой группы микроорганизмов на ЖКТ. Присутствие же в больших количествах может осложнить течение основного заболевания. К третьей группе относятся микроорганизмы, которые оказывают безусловно отрицательное влияние на состояние микрофлоры, которое усиливается с увеличением количества и разнообразия представителей этой группы. Четвертая группа состоит из двух видов бактерий, но их влияние на состояние микрофлоры желудочно-кишечного тракта человека таково, что при обнаружении хотя бы одного микроорганизма можно без проведения дальнейшего анализа диагностировать дисбиоз III степени, так как они являются возбудителями таких заболеваний, как дизентерия и сальмонеллез.

Следует также отметить, что видовой и количественный состав нормальной микрофлоры у детей и взрослых имеет существенные различия. Было выделено 8 возрастных групп обследуемых. Такое деление позволило увеличить точность при отделении людей с нормальной микрофлорой от тех, у кого имеются какие-либо отклонения от нормы. По ОСТ 91500.11.0004-2003 при определении степени дисбиоза выделяют 3 возрастные группы людей:

1. 0 – 1 год;
2. 1 год – 60 лет;
3. Более 60 лет.

Исследуя анализы ЖКТ для различных возрастных групп пациентов, мы выяснили, что качественный и количественный состав микрофлоры желудочно-кишечного тракта внутри этих групп различен.

Очевидно, что видовой и количественный состав микроорганизмов всех возрастных групп имеет существенные различия, причем критерии разделения «здоровья» и «болезни» тоже различны, следовательно, не вполне корректно объединять их в одну группу для определения степени дисбиоза. Кроме того, следует отметить, что все дети в возрасте до 23 часов являются здоровыми и у них можно различить только норму и дисбиоз I степени. Таким образом, кроме рассмотренных выше 29 признаков необходимо в качестве входного параметра учитывать возраст пациента.

Так как значения признаков могут варьироваться в пределах от 0 до 10^{12}, а работа сети не должна зависеть от порядков входных параметров, перед построением нейросетевого классификатора необходимо выполнить нормирование входных параметров. Предлагается следующий алгоритм:

1. Возбудители заболеваний объединяются в один параметр

$$V=\begin{cases}1, & (shigella>0)\vee(salmonella>0),\\ 0, & (shigella=0)\wedge(salmonella=0).\end{cases}$$

Возрастной признак принимает значения согласно номеру возрастной группы (табл. 1)

2. От значений остальных признаков $\tilde{V}$ берётся десятичный логарифм:

$$V=\begin{cases}\log(\tilde{V}), & \tilde{V}>0,\\ 0, & \tilde{V}=0.\end{cases}$$

Предложенный алгоритм приводит входные величины к одинаковой размерности, что позволит повысить скорость и качество процесса обучения проектируемой нейросети. Объединение группы возбудителей заболеваний (четвёртая группа) в один параметр позволило добавить в качестве входного параметра возраст пациента, не увеличив количество входных параметров (29 входных параметров).

Вектор выходных параметров согласно условиям поставленной задачи имеет вид

Y={норма, дисбиоз I степени, дисбиоз II степени, дисбиоз III степени},

что позволяет получить однозначный результат диагностирования.

Полагается, что выходной вектор проектируемой нейронной сети будет иметь вид, представленный в табл. 2.

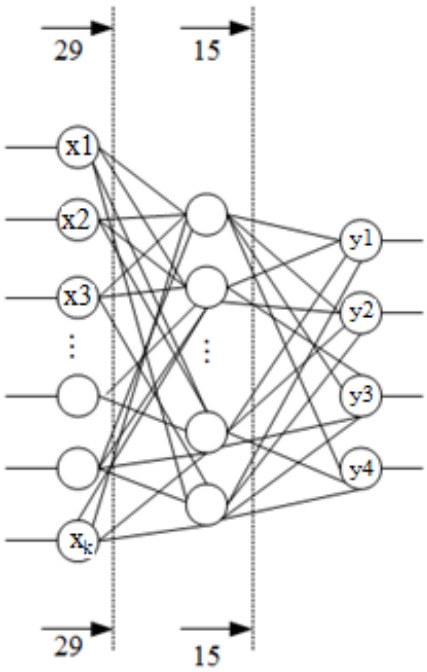

Рис. 1. Разработанная архитектура нейронной сети

Построение искусственной нейросети

Сеть построена по архитектуре трёхслойного персептрона. Она имеет 29 входов по количеству признаков, значения которых находятся в диапазоне [0, 12], и четыре выхода, значения которых 0 или 1.

Кроме того, очевидно, что значения выходных параметров, относящиеся к различным пациентам, не коррелируют друг с другом, поэтому нет необходимости строить рекуррентную сеть с «памятью». Чтобы сохранить способность сети к обобщению, был использован только один скрытый слой нейронов (иначе сеть просто «запомнит» всю обучающую выборку).

Первый слой нейронов используется в качестве входа нейронной сети, формируя входной вектор $X=\{x_1,x_2,\ldots,x_k\}$, где $x_1,x_2,\ldots,x_k$ – представленные в нормированном виде значения признаков, характеризующих состояние микрофлоры, а также возраст пациента, k – число нейронов во входном слое. Количество нейронов входного слоя равно 29 – по числу признаков.

Второй слой нейронной сети (так называемый скрытый слой) осуществляет преобразование вида

$$s_j = f\left(\sum_{k=1}^{K} \overline{w}_{jk} x_k\right) + \overline{\theta}_j, \quad j = 1,2,3\ldots,J,$$

где s_j – выходные состояния нейронов второго слоя, w_{jk} – коэффициенты матрицы межнейронных связей первого и второго слоев, определяющие связь между k-м нейроном первого слоя и j-м нейроном второго слоя, θ_j – пороговые уровни скрытого слоя, J – эмпирически подбираемая функция активации, в качестве которой использовалась сигмоидальная зависимость. Количество нейронов в промежуточном слое равно 15.

Выходной слой состоит из 4 нейронов, формируя выходной вектор $Y=\{y_1,y_2,y_3,y_4\}$, и выполняет функцию линейного преобразования вида

$$y_i = \sum_{j=1}^{J} w_{ij} s_j + \theta_j, \quad j = 1,2,3,\ldots,L,$$

где y_j – активации нейронов выходного слоя, w_{ij} – коэффициенты матрицы межнейронных связей второго и третьего слоев, θ_i – пороговые уровни для выходного слоя, L – число нейронов в каждом из слоёв.

Алгоритм обучения

Так как плотности распределений входных параметров не были известны, то в качестве алгоритма обучения был использован метод обратного распространения ошибки. Данный алгоритм относится к алгоритмам обучения с учителем. В качестве обучающего множества будем использовать совокупность пар $\{x^s, d^s\}$, где x – набор входных параметров, d – известное решение для данного набора входных параметров. Количество элементов S в обучающем множестве должно быть достаточным для обучения сети, чтобы под управлением алгоритма сформировать набор параметров сети, дающий нужное отображение $x \rightarrow y$, y – код, соответствующий поставленному диагнозу. Количество пар в обучающем множестве не регламентируется. Выберем один из векторов x и подадим его на вход сети. На выходе получится некоторый вектор y. Тогда ошибкой сети можно считать $E_s = \|d^s - y^s\|$ для каждой пары (x^s, d^s). С помощью алгоритма обучения минимизируется суммарная квадратичная ошибка, которая имеет вид

$$E = \frac{1}{2}\sum_j \sum_s (y^s - d^s)^2,$$

где j – число нейронов в выходном слое сети.

Задача обучения ставится следующим образом: подобрать такие значения параметров сети, чтобы ошибка E была минимальной для данного обучающего множества. Рассматриваемый метод обучения является итерационным. Параметрам сети (весовым коэффициентам и пороговым уровням) присваиваются малые начальные значения. Затем параметры изменяются так, чтобы ошибка E убывала. Изменения продолжаются до тех пор, пока ошибка не достигнет заданного значения. Ошибка E нейронной сети является функцией параметров сети, то есть

имеем некоторую функцию $E(P)$, где P – параметр сети. Параметр P является вектором: $P = (W)$, где W – вектор, компоненты которого – все весовые коэффициенты сети. Таким образом, на каждой итерации будем корректировать параметры в направлении антиградиента

$$\Delta P = \varepsilon(-\text{grad}(E(P)).$$

Коррекцию параметров сети необходимо рассчитывать на каждой итерации. Алгоритм обратного распространения ошибки позволяет ускорить расчёт градиентов. Общий принцип работы алгоритма заключается в следующем:

1. Представить $E(P)$ в виде сложной функции.

2. Последовательно рассчитать все частные производные.

Алгоритм обратного распространения ошибки делится на два этапа. На первом этапе на вход сети подаётся некоторый вектор из обучающего множества и производится расчёт выходных параметров сети. На втором этапе для каждого выходного параметра вычисляется ошибка сети δ, и начинается её обратное распространение от выходного слоя к входному.

Алгоритм обратного распространения основан на обобщённом дельта-правиле. Запишем частную производную суммарной квадратичной ошибки по весовым коэффициентам:

$$\Delta w_{ijl} = -\varepsilon \frac{\partial E}{\partial w_{ijl}},\ w_{ijl}' = w_{ijl} + \Delta w_{ijl},$$

w_{ijl} – значение веса на текущей итерации, w_{ijl}' – значение веса на следующей итерации.

Обозначим NET значение взвешенной суммы нейрона (сумма произведений значений входов на веса). $OUT = f(NET)$ обозначим значение функции активации. Тогда производная ошибки может быть представлена в виде

$$\frac{\partial E}{\partial w_{ijl}} = \frac{\partial E}{\partial OUT_{jl}} \frac{\partial OUT_{jl}}{\partial NET_{jl}} \frac{\partial NET_{jl}}{\partial w_{ijl}}.$$

Определим величину δ_{jl} с помощью формулы

$$\delta_{il} = -\frac{\partial E}{\partial NET_{jl}},$$

которую можно переписать в виде

$$\delta_{il} = -\frac{\partial E}{\partial OUT_{jl}} \frac{\partial OUT_{jl}}{\partial NET_{jl}}.$$

Для одной пары из обучающей выборки $E = \frac{1}{2}\sum_j \left(d_j - OUT_{jl}\right)^2$. Для функции активации выходом является $OUT = f(NET)$, поэтому для производной f получаем

$$\frac{\partial OUT_{jl}}{\partial NET_{jl}} = f'\left(NET_{jl}\right).$$

Таким образом, $\delta_{il} = (d_j - OUT_{ij}) f'\left(NET_{jl}\right)$.

Для нахождения комбинированного ввода используется обычное взвешенное суммирование: $NET_{jl} = \sum_i x_i w_{ijl}$, поэтому $\frac{\partial NET_{jl}}{\partial w_{ijl}} = x_i$.

Рассмотрев произведение полученных производных, получим

$$\frac{\partial E}{\partial w_{ijl}} = -(d_j - OUT_{ij}) f'\left(NET_{jl}\right) x_i.$$

Мы получили частную производную суммарной квадратичной ошибки от веса нейрона выходного слоя.

С учётом того, что вес должен изменяться в направлении, противоположном тому, которое указывает производная поверхности ошибок, и с учётом скорости обучения ε, изменение веса для элемента должно вычисляться по формуле $\Delta w_{ijl} = \varepsilon \delta_{jl} x_i$. Для удобства в качестве функции активации будем использовать сигмоиду, тогда $f'(NET_{jl}) = f(NET_{jl})(1 - f(NET_{jl}))$. При этом для выходного слоя ошибку δ_{jl} можно записать в виде

$$\delta_{il} = (d_j - OUT_{ij}) f(NET_{jl})(1 - f(NET_{jl})).$$

Указанная выше ошибка δ соответствует ошибке выходного элемента, но ошибка скрытого элемента не связана с целевым выходным значением непосредственно. Поэтому весовые значения скрытого элемента следует скорректировать пропорционально его вкладу в величину ошибки следующего слоя (т.е. выходного слоя в случае сети с одним скрытым слоем). В сети с одним скрытым слоем при распространении сигналов ошибок в обратном направлении ошибка каждого выходного элемента вносит свой вклад в ошибку каждого элемента скрытого слоя. Этот вклад для элемента скрытого слоя зависит от величины ошибки выходного элемента и весового коэффициента связи, соединяющей элементы. Другими словами, выходной элемент с большей ошибкой делает больший вклад в ошибку того элемента скрытого слоя, который связан с данным выходным элементом большим по величине весом. Для скрытого элемента ошибка вычисляется по формуле

$$\delta_{j(l-1)} = f(NET_{j(l+1)})(1 - f(NET_{j(l+1)}))\left(\sum_i x_i w_{ijl}\right).$$

Изменение веса для нейрона скрытого слоя вычисляется по той же формуле, что и для нейрона выходного слоя: $\Delta w_{ij(l-1)} = \varepsilon \delta_{j(l-1)} x_i$, где x_i – вход нейрона данного скрытого слоя. На первой

стадии происходит инициализация весов малыми случайными значениями – например, значениями из диапазона между –0.3 и +0.3. Обучение продолжается до тех пор, пока изменение средней квадратичной ошибки не окажется меньше некоторого допустимого значения при переходе от одной итерации к следующей. Например, допустимое значение 0.01 означает, что средняя квадратичная ошибка соседних итераций не должна отличаться более чем на ±0.01. Если в процессе обучения наступает момент, когда ошибка в сети попадает в рамки допустимого изменения, говорят, что наблюдается сходимость. Другим критерием окончания обучения можно считать наступление момента, когда выход для каждого учебного образца оказывается в рамках допустимого отклонения от соответствующего целевого выходного образца, либо суммарная квадратичная ошибка E стала меньше какой-то заранее известной величины.

Эксперимент

Нейросетевая модель реализована программно на языке C++. Обучение сети выполнено на выборке из 100 медико-биологических исследований ЖКТ. В качестве критерия останова при обучении сети использовалось условие $\Delta \leq 20\%$, где Δ – средняя ошибка по слоям. Работа обученной нейронной сети проверялась на контрольной выборке из 1000 медико-биологических исследований ЖКТ. Построенная нейронная сеть позволила классифицировать состояние микрофлоры ЖКТ с ошибкой 18%.

Результаты

- Разработана автоматизированная система классификации состояния ЖКТ.
- Обоснован выбор нейросетевой модели.

Разработана архитектура нейронной сети, позволяющая покрыть все выделенные нами возрастные группы и различать нормальный состав микрофлоры ЖКТ и дисбактериоз I–III степеней с ошибкой 18%.

Список литературы

1. Ломакина Л.С., Соловьева И.В., Зеленцов С.А., Пожидаева А.С. Модели и алгоритмы диагностирования состояний биоценоза на основе априорных статистических данных // Научно-технический вестник Поволжья. 2013. № 5. Казань: НИКГУ. С. 251–256.

2. Уоссерман Ф. Нейрокомпьютерная техника: Теория и практика. М.: Мир, 1992.

CLASSIFICATION OF BIOCENOSIS STATES BASED ON NEURAL NETWORK TECHNOLOGIES

L.S. Lomakina, A.S. Pozhidaeva, Ya.P. Gubernatorov

Classification of microflora states of the human gastrointestinal tract (GI tract) for dysbiosis diagnostics is considered. An artificial neural network built by a three-layer perceptron architecture is used as a classifier.

Keywords: classification, dysbiosis, diagnostics, perceptron, neural network.

References

1. Lomakina L.S., Solov'eva I.V., Zelencov S.A., Pozhidaeva A.S. Modeli i algoritmy diagnostirovaniya sostoyanij biocenoza na osnove apriornyh statisticheskih dannyh // Nauchno-tekhnicheskij vestnik Povolzh'ya. 2013. № 5. Kazan': NIKGU. S. 251–256.

2. Uosserman F. Nejrokomp'yuternaya tekhnika: Teoriya i praktika. M.: Mir, 1992.