УДК 519.2
ББК 22.1

СИМУЛЯЦИЯ ВХОДНЫХ ДАННЫХ НЕЙРОННЫХ СЕТЕЙ НА ПРИМЕРЕ ЗАБОЛЕВАЕМОСТИ ВИРУСНЫМ ГЕПАТИТОМ

Муфтахов Т. И.[1], Гиниятуллин В. М.[2]
(*Уфимский государственный нефтяной технический университет, Уфа*)

В данной работе применена искусственная нейронная сеть для классификации вирусного гепатита на острый и хронический по результатам клинических лабораторных исследований: общего анализа крови и биохимического анализа крови.

Ключевые слова: нейронные сети, кластеризация, диагностика вирусного гепатита, персептрон.

1. Введение

Вирусный гепатит является одним из самых широко распространённых заболеваний в мире. По данным Всемирной организации здравоохранения (ВОЗ), ежегодно в мире погибают около 2 млн. человек от острых и хронических вирусных гепатитов [9].

Вирусный гепатит А отличает повсеместное распространение и неравномерная интенсивность на отдельных территориях. Данный вид гепатита протекает только в острой стадии, а также относительно редко заканчивается смертельным исходом [11].

Гепатиты В и С являются самыми опасными инфекционными заболеваниями в мире. По данным ВОЗ [11],

[1]*Тимур Ильдарович Муфтахов, студент (muftakhov.timur@mail.ru)*
[2]*Вахит Мансурович Гиниятуллин, канд. тех. наук, доцент (fentazer@mail.ru)*

предположительно 325 миллионов человек в мире живут с хронической инфекцией, вызванной вирусом гепатита В или вирусом гепатита С. Только от хронического гепатита В умирают 40% больных, инфицированных вирусом HBV [6]. При этом численность хронических больных вирусным гепатитом постоянно растёт.

В некоторых случаях причины болезни однозначно выявить не удается даже при использовании современных средств диагностики, тогда ставится следующий диагноз: хронический гепатит неуточненной этиологии. Для такого заболевания характерны воспалительно-деструктивные процессы.

При длительном течении хронический вирусный гепатит С может перейти в цирроз или начальные стадии рака печени, поэтому очень важна правильная и своевременная диагностика заболевания.

Наибольший интерес для практического здравоохранения представляют системы дифференциальной диагностики [10] заболеваний. При этом для принятия решений могут использоваться самые разнообразные данные: анамнез болезни, клинический осмотр, результаты лабораторных исследований, сложные функциональные методы и др. Для решения подобных задач могут использоваться [2] искусственные нейронные сети (ИНС), представляющие собой мощный и одновременно гибкий метод имитации процессов и явлений. Предметные области, в которых могут применяться ИНС, весьма обширны и разнообразны по постановкам задач [1, 7].

Отличительное свойство нейронных сетей состоит в их способности обучаться на основе экспериментальных данных предметной области. В медицине экспериментальные данные представляются в виде множества исходных признаков или параметров объекта и поставленного на основе них диагноза.

2. Описание выборки входных показателей

При выполнении работы была сформирована общая выборка из данных о 250 больных вирусными гепатитами А, В и С. Затем общая выборка была разделена на обучающую и тестовую выборку. В обучающей выборке содержатся данные о

225 пациентах, из которых 58 больны острыми вирусными гепатитами А, В и С и 167 – хроническими вирусными гепатитами В и С. В тестовую выборку попали данные о 3 больных острыми вирусными гепатитами А и В и данные о 22 больных хроническими гепатитами В и С.

Таблица 1. Входные параметры нейронной сети

№ п/п	Показатель	Диапазон изменения	Единица измерения
1	Пол	мужской, женский	-
2	Возраст	7… 81	лет/год
3	Гемоглобин	74… 193	г/л
4	Эритроциты	1,85… 5,91	$x10^{12}$/л
5	Лейкоциты	1,4… 16,8	$x10^{9}$/л
6	Тромбоциты	26… 417	$x10^{9}$/л
7	Скорость оседания эритроцитов (СОЭ)	1… 51	мм/ч
8	Глюкоза	3,3… 9,8	ммоль/л
9	Холестерин	1,8… 9,6	ммоль/л
10	Общий билирубин	5… 366,1	мкмоль/л
11	Тимоловая проба	0,25… 34	Ед.
12	Аспартатаминотрансфераза (АСТ)	15… 3630	Ед/л
13	Аланинаминотрансфераза (АЛТ)	9… 4014	Ед/л
14	Гамма-глютамилтранс-пептидаза (ГГТ)	11… 1436	Ед/л
15	Щелочная фосфатаза (ЩФ)	31… 1581	Ед/л
16	Общий белок	60… 100	г/л
17	Вид вирусного гепатита	A, B, C	-
18	Форма тяжести	Легкая, среднетяжелая, тяжелая	-
19	Альбумин	32,48… 57,2	г/л
20	Концентрация α_1	1,99… 18,48	%
21	Концентрация α_2	4,92… 17,07	%

22	Концентрация β	7,11… 23,0	%
23	Концентрация γ	15,45… 42,0	%

Для входных параметров нейронной сети были включены 23 показателя, которые представлены в таблице 1.

Из таблицы 1 видно, что такие показатели, как пол, вид гепатита и форма тяжести являются нечисловыми данными. Они были представлены в дискретном виде: пол – бинарный (мужской = 1, женский = -1), вид гепатита – четырехзначный, но гепатит неуточненной этиологии в данной работе не рассматривается, поэтому виды гепатита кодируются так: A = 0, B = -1, C = 1. Форма тяжести – исчисляемый параметр и кодируется следующим образом: легкая = -1, среднетяжелая = 0, тяжелая = 1. Остальные показатели масштабированы в диапазон [-1; 1]. Выходной слой кодируется так: острое течение = -1, хроническое течение = 1.

3. Структура персептрона

Для обучения нейронной сети использовалось свободное программное обеспечение Multiple Back-Propagation [8]. Исходя из многократных проведенных тестов, была выбрана следующая структура персептрона: два нейрона в скрытом слое, выходной слой состоит из одного нейрона. В качестве функции активации в скрытом и выходном слоях используется функция гиперболического тангенса. Обучение искусственной нейронной сети осуществлялось с помощью метода обратного распространения ошибки (back propagation).

В результате обучения получаются матрицы весов скрытого и выходного слоя персептрона (рис. 1).

to the 1th hidden layer	1th neuron	2th neuron
bias	-0,928338	-0,337032
1th weight	0,307831	0,294374
2th weight	0,797554	1,02618
3th weight	1,2414	-0,357926
4th weight	-0,83936	-1,66756
5th weight	-1,19906	-0,714341
6th weight	0,31524	-1,72561
7th weight	-1,14925	0,545223
8th weight	0,0759451	1,49098
9th weight	1,50956	-0,0692936
10th weight	0,45035	-4,61832
11th weight	-0,846072	-2,08323
12th weight	-0,771772	0,227809
13th weight	-4,73377	0,0928176
14th weight	1,39388	-0,914482
15th weight	-2,57924	1,16982
16th weight	0,248778	-0,0510175
17th weight	0,534629	0,501062
18th weight	4,48537	-0,618889
19th weight	0,00148284	-0,0956685
20th weight	0,00759095	-0,220889
21th weight	-0,649486	1,56455
22th weight	-0,261958	-0,683136
23th weight	0,301353	0,411919

to the output layer	1th neuron
bias	-3,2349
1th weight	3,4011
2th weight	3,13924

Рис. 1. Веса скрытого и выходного слоя персептрона

4. Подмена функции активации на пороговую

Распишем подробнее рабочий ход персептрона, выполнив следующие действия в электронных таблицах:

1. Результат выполнения скалярного умножения вектора входов на матрицу весов скрытого слоя приведен на рис. 2 (столбцы ps1 и ps2).

2. Результат применения функции активации (гиперболический тангенс) к результату скалярного умножения представлен в столбцах p1 и p2.

3. Результат скалярного умножения матрицы из столбцов p0…p2 и вектора весов выходного слоя представлен в столбце ys1.

4. Применение гиперболического тангенса к столбцу ys1 и округление до целого числа - столбец y1.

ps1	ps2	p0	p1	p2	ys1	y1
6,399	1,522	1	0,999994	0,909031	3,019849	1
6,769	-4,061	1	0,999997	-0,99941	-2,97118	-1
0,524	2,508	1	0,48091	0,986813	1,498568	1
1,398	3,530	1	0,884815	0,998285	2,908303	1
1,187	2,688	1	0,82975	0,99079	2,697492	1
0,321	-2,234	1	0,309995	-0,97732	-5,24861	-1
4,944	-3,939	1	0,999898	-0,99924	-2,97101	-1
1,887	4,476	1	0,955145	0,999741	3,152069	1
0,659	2,497	1	0,5777	0,986524	1,82685	1
1,437	4,667	1	0,893123	0,999823	2,941386	1
-2,032	-0,707	1	-0,96622	-0,60859	-8,43162	-1
0,948	5,698	1	0,738701	0,999978	2,416664	1
4,028	-1,303	1	0,999365	-0,86261	-2,5439	-1
1,309	3,501	1	0,864014	0,998181	2,837231	1
-6,678	-0,547	1	-1	-0,49844	-8,2007	-1
-0,606	-3,590	1	-0,54113	-0,99848	-8,20978	-1
-2,544	0,379	1	-0,98775	0,362229	-5,4572	-1
4,336	1,760	1	0,999657	0,942449	3,123608	1

Рис. 2. Фрагмент таблицы результатов рабочего хода

Сравнение ожидаемого выхода обучающей выборки с результатом рабочего хода персептрона демонстрирует их совпадение. Это означает, что данный персептрон обучен правильно. Тестовая выборка также полностью распознается без ошибок.

Подмена гиперболической функции активации на троичную пороговую функцию вида [4]:

$$y=f(a)=\begin{cases} -1, \text{прих} < -0,5; \\ 0, \text{при} -0,5 \le x \le 0,5; \\ 1, \text{прих} > 0,5, \end{cases}$$

выполняется следующим образом: значения в столбцахр1 и р2 округляются до целого числа, затем копируются в текстовый редактор, благодаря этому, данные теряют форматирование и преобразуются в текст в виде целых чисел, далее они копируются из текстового редактора в ячейки р1 и р2, это заменяет исходные числа на их округленные значения (рис. 3).

ps1	ps2	p0	p1	p2	ys1	y1
6,399	1,522	1	1	1	3,30544	1
6,769	-4,061	1	1	-1	-2,97304	-1
0,524	2,508	1	0	1	-0,09566	1
1,398	3,530	1	1	1	3,30544	1
1,187	2,688	1	1	1	3,30544	1
0,321	-2,234	1	0	-1	-6,37414	-1
4,944	-3,939	1	1	-1	-2,97304	-1
1,887	4,476	1	1	1	3,30544	1
0,659	2,497	1	1	1	3,30544	1
1,437	4,667	1	1	1	3,30544	1
-2,032	-0,707	1	-1	-1	-9,77524	-1
0,948	5,698	1	1	1	3,30544	1
4,028	-1,303	1	1	-1	-2,97304	-1
1,309	3,501	1	1	1	3,30544	1
-6,678	-0,547	1	-1	0	-6,636	-1
-0,606	-3,590	1	-1	-1	-9,77524	-1
-2,544	0,379	1	-1	0	-6,636	-1
4,336	1,760	1	1	1	3,30544	1

Рис. 3. Фрагмент результатов подмены на функцию активации

После подмены функции активации на пороговую значения в столбце ys1 изменились, при этом столбец y1 остался прежним. В некоторых строках значения ys1 совпадают, что соответствуют равным значениям p1 и p2.

Таким образом, нейроны скрытого слоя в неявном виде реализуют логическую функцию, значность результата которой равна трём. При этом нейрон выходного слоя (столбец y1) реализует смешанную 3-2 логику, аргументы которой троичны, а результат бинарный [3].

5. Распределение кластеров в непрерывном пространстве

После сортировки строк по возрастанию столбца ys1 выделяются кластера – это строки с одинаковыми значениями в

столбцах p1, p2 и, как следствие, в столбце ys1 (рис. 4). Возможно существование 9 кластеров, в данном случае обнаружено 7, кластера (1; 0) и (0; 0) остались незаселенными.

ps1	ps2	p1	p2	ys1	
-1,154	-0,595	-1	-1	-9,775	1 кластер
-2,032	-0,707	-1	-1	-9,775	
-0,606	-3,590	-1	-1	-9,775	
-6,678	-0,547	-1	0	-6,636	2 кластер
-2,544	0,379	-1	0	-6,636	
-2,367	0,290	-1	0	-6,636	
0,321	-2,234	0	-1	-6,374	3 кластер
-2,903	3,954	-1	1	-3,497	4 кластер
-2,642	3,604	-1	1	-3,497	
-2,382	0,701	-1	1	-3,497	
-8,712	4,291	-1	1	-3,497	
2,778	-5,260	1	-1	-2,973	5 кластер
9,330	-6,811	1	-1	-2,973	
6,015	-4,737	1	-1	-2,973	
4,542	-3,920	1	-1	-2,973	
9,095	-5,139	1	-1	-2,973	
0,524	2,508	0	1	-0,096	6 кластер
-0,368	5,821	0	1	-0,096	
0,512	5,334	0	1	-0,096	
-0,393	7,250	0	1	-0,096	
0,946	3,462	1	1	3,305	7 кластер
0,936	5,304	1	1	3,305	
0,811	2,852	1	1	3,305	
0,939	4,238	1	1	3,305	
1,061	2,702	1	1	3,305	
7,213	1,139	1	1	3,305	

Рис. 4. Фрагмент таблицы с кластерами

Каждому кластеру присваивается свой символ: представители первого кластера обозначены синими ромбами, второго – фиолетовыми прямоугольниками, третьего – черными треугольниками, четвертого – коричневыми крестиками, пятого

– голубыми снежинками, шестого – красными кругами и седьмого – зелеными квадратами. Распределение представителей кластеров в непрерывном пространстве ps (столбцы ps1 и ps2) – результат скалярного умножения входного вектора на матрицу весов скрытого слоя представлено на рис. 5.

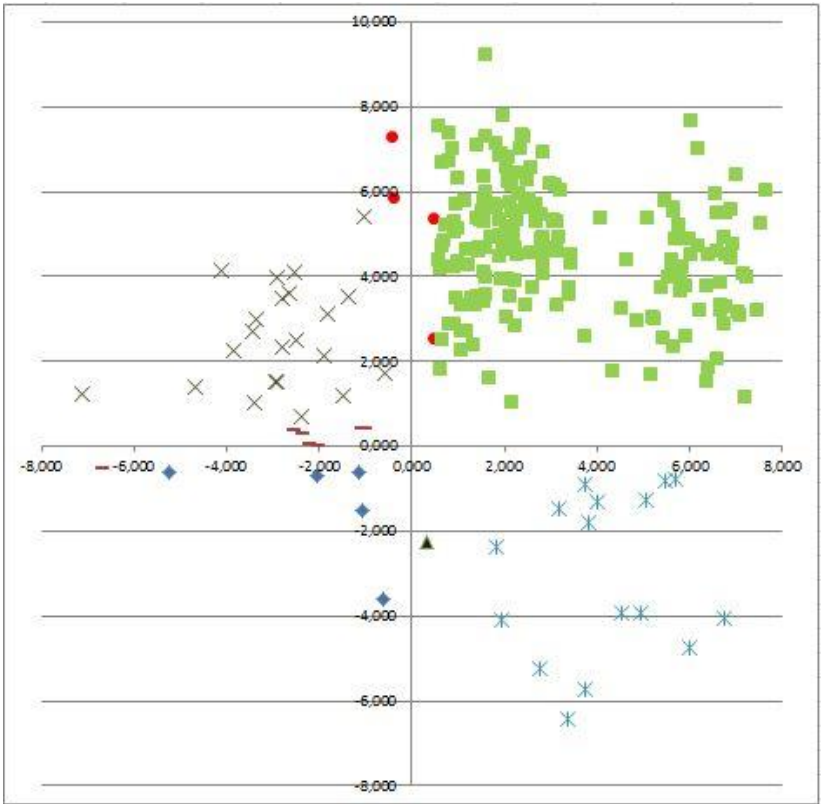

Рис. 5. Распределение кластеров в непрерывном пространстве

Из рис. 5 видно, что в непрывном пространстве ps кластер (-1; -1) заселен неравномерно и отсутствуют строки с «большими минусами» в строках ps1, ps2 одновременно. Естественно, возникает вопрос о существованиии таких строк, но в терминах прикладной области.

Следует отметить, что каждому кластеру двумерного дискретного пространства p соответствует n–мерный выпуклый многогранник в пространстве входов – симплекс [5]. В данном случае имется 9 симплексов в 23–мерном пространстве входов. В таблице 1 имеется 3 дискретных параметра и 20 числовых, изменяющихся в диапазоне [-1; 1]. При их полном переборе с шагом в 0,1 число возможных комбинаций превышает цифру в 20^{20}. Симуляция всевозможных комбинаций входных параметров невозможна, т.к. проверку на корректность симулированных данных приходиться делать вручную. Поэтому симуляция входных данных производилась случайным образом.

Вопреки ожиданиям, 1 строка из 50 симулированных попала в кластер (0; 0) – это строка 6 (рис. 6). Кроме того, найдены 15 строк, которые попадают в кластер (-1; -1), в том числе и в область «больших минусов». В симулированной выборке незаселенными оказались кластера (0; -1) и (1; 0).

№ п/п	пол	возраст	общий билирубин	АСТ	АЛТ	гепатит	форма	F	p1	p2
1	мужской	36	296,3	347	3332	A	среднетяж.	хронический	-1	-1
2	женский	12	264,8	2458	3963	A	легкая	хронический	-1	-1
3	мужской	17	283,5	2556	3737	A	среднетяж.	хронический	-1	-1
4	женский	17	109,8	391	3004	A	среднетяж.	хронический	-1	-1
5	мужской	29	353,7	2778	789	A	легкая	хронический	-1	-1
6	женский	34	277,6	2424	2367	A	среднетяж.	хронический	0	0
7	женский	36	264,6	152	2024	A	среднетяж.	острый	-1	-1
8	мужской	71	271,0	231	3036	A	легкая	острый	-1	-1
9	женский	57	125,6	3401	4000	B	среднетяж.	острый	-1	-1
10	женский	32	267,0	1777	2873	B	среднетяж.	острый	-1	-1
11	женский	74	288,7	1523	3806	B	среднетяж.	хронический	-1	-1
12	женский	74	321,4	2205	2994	B	среднетяж.	острый	-1	-1
13	женский	12	354,2	298	1930	B	среднетяж.	острый	-1	-1
14	женский	64	351,5	1509	718	B	легкая	острый	-1	-1
15	женский	9	298,5	705	2530	C	среднетяж.	хронический	-1	-1
16	мужской	22	356,9	3440	1311	C	среднетяж.	хронический	-1	-1

Рис. 6. Симуляция входных данных

Представители кластера (1; 0) были сгенерированы целенапраленно. Десятый вес второго нейрона матрицы весов скрытого слоя (рис. 1) имеет наибольшее значение, следовательно, при изменении десятого входного параметра значение столбца ps2 будет изменяться сильнее, чем при изменении других входов. Перебором десятого входного параметра с шагом в 0,1 были обнаружены две строки, принадлежащие к искомому кластеру.

Таким образом было показано, что все 9 симплексов полностью или частично расположены в границах единичного 23–мерного гиперкуба.

Из 50 симулированных строк 15 относятся к кластеру (-1; -1). Среди них лишь одна строка (номер 9) соответствует и показателям, и диагнозу: у данного больного острый вирусный гепатит В (повышены показатели АСТ и АЛТ) среднетяжелой формы (значение общего билирубина лежит в диапазоне 100-200 мкмоль/л, что характерно для данной формы тяжести). Остальные строки оказались некорректными, например, у первых шести пациентов диагностирован хронический гепатит А, тогда как гепатит А принимает только острое течение. Это означает, что даже правильно обученный персептрон может выдавать неверные результаты.

6. Выводы и направления дальнейших исследований

Подмена непрерывной функции активации на пороговую позволяет производить кластеризацию входных данных не только по значению выходного нейрона, но и по значениям нейронов скрытого слоя.

При необходимости можно заселять кластера симулированными входными данными. В данном случае оказалось, что часть симулированных данных некорректна, притом, что персептрон распознаёт тестовую выборку без ошибок. Следовательно, исходная выборка неполная.

Возникает предположение, что «полной» обучающей выборкой следует считать такую выборку, у которой плотность распределения представителей кластеров в каждом симплексе равномерна. Более того, формализация термина «полная выборка» может привести к созданию критерия адекватности математической модели процесса, реализуемого ИНС.

Необходимо разработать способ доказательства принадлежности каждого симплекса входного пространства единичному гиперкубу.

Объемы симплексов предполагается оценивать методом Монте-Карло.

Литература

1. АРСЛАНОВ И.Г., ДМИТРИЕВ Г.Ю., ГИНИЯТУЛЛИН В.М., ЗАЙЦЕВА А.А., КИРЛАН С.А. *Прогнозирование химических соединений с комплексом необходимых свойств* / Башкирский химический журнал. 2015. Т. 22, № 2. С. 80 – 85.
2. ВЕДЕНЯПИН Д.А., ЛОСЕВ А.Г. *Об одной нейросетевой модели диагностики венозных заболеваний* / Управление большими системами. Выпуск 39. М.: ИПУ РАН, 2012. С. 219 – 229.
3. ГИНИЯТУЛЛИН В.М. *Троичная логика в нейросетевом базисе* / В сборнике: Теоретико-методологические проблемы естественнонаучных методов в гуманитарных науках. Сборник статей. Международной научно-практической конференции, 2014. С. 318 – 325.
4. ГИНИЯТУЛЛИН В.М., СКРЫПИН А.Р., ТАЙСИН Р.Р. *Линейная разделимость функций троичной логики* / В сборнике: Актуальные проблемы науки и техники – 2015. Материалы VIII международной научно-практической конференции молодых ученых, 2015. С. 116 – 119.
5. ПАЛИЙ И.А. *Линейное программирование*: Учебное пособие / И.А. Палий. – М.: Эксмо, 2008. – 256 с. – (Техническое образование).
6. ШУВАЛОВА Е.П., БЕЛОЗЕРОВ Е.С., БЕЛЯЕВА Т.В., ЗМУШКО Е.И. *Инфекционные болезни*: Учеб. пособие для вузов. Под ред. Е.П. Шуваловой / Серия «Учебники и учебные пособия» – Ростов н/Д: Изд-во «Феникс», 2001. – 960 с.
7. AKHMETSHIN R.M., GINIYATULLIN V.M., KIRLAN S.A. *Identification of Structures of Organic Substances by Means of Complex-valued Perceptron* / Optical Memory & Neural Networks (Information Optics). 2012. Vol. 21, № 1. P. 11 – 25.
8. http://mbp.sourceforge.net (дата обращения: 26.11.2017).
9. http://www.tiensmed.ru/illness/gepatit3.html (дата обращения: 03.12.2017).
10. https://vocabulary.ru/termin/diagnostika-differencialnaja.html (дата обращения: 04.12.2017).

11. http://www.who.int/mediacentre/news/releases/2017/global-hepatitis-report/ru/ (дата обращения: 03.12.2017).

SIMULATION INPUT DATA OF NEURAL NETWORKS ON THE EXAMPLE OF THE INCIDENCE OF VIRAL HEPATITIS

T. Muftakhov, Ufa State Petroleum Technological University, Ufa, student (muftakhov.timur@mail.ru).
V. Giniyatullin, Ufa State Petroleum Technological University, Ufa, Candidate of Science, Associate Professor (fentazer@mail.ru).

Abstract: In this work applied artificial neural network for the classification of viral hepatitis into acute and chronic according to the results of clinical laboratory tests: general blood analysis and biochemical blood analysis.

Keywords: Neural networks, clustering, diagnostics of viral hepatitis, a perceptron.