УДК 004.8:616-006

Построение классификационной модели онкологических заболеваний с применением методов искусственного интеллекта

А.А. Шайдуров, А.С. Бубликов

Алтайский государственный университет (Барнаул, Россия)

Development of a Cancer Diseases Classification Model Using Artificial Intelligence

A.A. Shaidurov, A.S. Bublikov

Altai State University (Barnaul, Russia)

Рассматривается анализ возможности построения модели классификации онкологических заболеваний с помощью слоистых искусственных нейронных сетей. Исследовались данные, полученные при помощи технологии IMMUNOSIGNATURE. Для получения иммуносигнатуры капля крови обследуемого, содержащая в себе клетки иммунной системы, наносится на биочип, разделенный на сектора, в каждом из которых находится уникальная аминокислотная последовательность. В зависимости от того, с какой интенсивностью в конкретных ячейках проявится иммунная реакция, формируется профиль (сигнатура) для конкретного человека. Сложность анализа данных заключается в стохастическом влиянии внешних факторов на получаемые результаты при помощи технологии IMMUNOSIGNATURE. Рассматриваются вопросы коррекции получаемых данных при помощи корреляционного анализа для избавления от случайных артефактов. Изучен вопрос нормализации данных для минимизации влияния системных искажений, возникающих вследствие влияния таких внешних факторов, как температура, влажность, концентрация вещества и т. д. В качестве классификатора используется искусственная нейронная сеть обратного распространения ошибки.

Ключевые слова: искусственная нейронная сеть, метод доверительных интервалов, корреляционный анализ, технология IMMUNOSIGNATURE, онкология, рак молочной железы.

The paper deals with the possibility analysis of developing a cancer diseases classification model based on layered artificial neural networks. We studied the data obtained by the IMMUNOSIGNATURE technology. Immunosignaturing requires a drop of test subject blood that contains cells of immune system to be applied to a biochip. The biochip is divided into sectors with unique amino acid sequences. Depending on the intensity of the immune response manifestation in specific cells, a specific profile (signature) is formed for a certain individual. The complexity of immunosignature data analysis lies in external stochastic influence on the obtained results. In the paper, aspects of correlation analysis data correction to get rid of random artifacts are considered. Data normalization procedure to minimize the impact of systemic distortions due to the influence of external factors, such as temperature, humidity, substance concentration, etc., is discussed. Classification is performed by the backpropagation artificial neural network.

Keywords: artificial neural network, a method of confidence intervals, correlation analysis, technology IMMUNOSIGNATURE, oncology, breast cancer.

DOI 10.14258/izvasu(2015)1.2-32

Введение. С развитием компьютерной техники и информационных технологий у человечества появились мощные инструменты обработки. К этим инструментам относятся:

- статистические методы;
- искусственные нейронные сети (ИНС);
- Data Mining.

Существует ряд направлений деятельности человечества, в которых востребованы информационные методы скоростной обработки данных и алгоритмы решения неалгоритмических задач. Одним из таких направлений является медицина.

В настоящий же момент компьютеры в медицинских учреждениях используются в большинстве случаев либо как терминал для вывода информации с диагностических устройств, либо как «печатная машинка». Тем не менее активно развиваются и внедряются в медицинскую практику скрининговые и диагностические информационные системы. Благодаря новейшим информационным и технологическим инструментам можно использовать современные методы для анамнестического обследования пациентов и соответствующей обработки больших массивов данных, что позволит формировать точные прогнозы о протекании различных заболеваний.

Одним из направлений медицины является онкология. Это то направление, в котором применять подобные методы просто необходимо, ведь часто жизнь пациента зависит от того, как оперативно был поставлен диагноз. А в силу того, что медицинские показатели обладают большой лабильностью клинических проявлений, бывает трудно провести топическую диагностику онкологического заболевания. Поэтому одним из путей решения данной проблемы является внедрение математического аппарата и применение проблемно-ориентированных систем обработки информации.

Технология IMMUNISIGNATURE. На основании современных исследований о биочипах появилась технология IMMUNOSIGNATURE, простой и недорогой способ, используемый для ранней диагностики онкологических заболеваний.

Суть метода иммуносигнатур заключается в отслеживании иммунных реакций. Каждое заболевание вызывает ответ иммунной системы, по которому можно узнать, болен ли человек и чем именно.

Для получения иммуносигнатуры капля крови обследуемого, содержащая в себе клетки иммунной системы, наносится на биочип, разделенный на сектора, в каждом из которых находится уникальная аминокислотная последовательность. Всего на пластинке — от 10 000 до 330 000 секторов, и в зависимости от того, с какой интенсивностью в конкретных ячейках проявится иммунная реакция, формируется профиль (сигнатура) для конкретного человека. Естественно, у каждого человека он индивидуален. Сама пластинка имеет небольшой размер, аналогичный размеру предметного стекла для светового микроскопа. Ее удобно перевозить и для нее не требуется большого объема исследуемой крови. Иммуносигнатура показывает, какие последовательности аминокислот активируют иммунную систему пациента, к чему у него выработано больше антител, т. е. с чем организм борется в данный момент. Располагая такими сведениями, можно понять, чем человек болен.

Революционный прорыв в данной технологии заключается в том, что стало возможно совершить продвижение с постсимптоматической медицины (когда опухоль уже развилась в организме) к предсимптоматической медицине.

Суть технологии — в том, чтобы выявить иммуносигнатуры, характерные для какого-то одного заболевания. Технология интенсивно изучается, исследования обширны и дают хорошие результаты в плане точности диагностики. Особенно метод перспективен для ранней диагностики рака, возможной еще на тех стадиях, когда речи не идет о постановке страшного диагноза [1].

Предварительный анализ экспериментальных данных. В ходе экспериментов были получены результаты исследования иммуносигнатур для двух групп людей. Одна — контрольная группа людей с невыявленной онкологией (25 человек), а другая группа — пациенты с диагнозом С50 «Злокачественное новообразование молочной железы» (10 человек). Чтобы получить достоверные данные, для каждого пациента проводилось три эксперимента в различные моменты времени на разных чипах. Такой подход должен позволить нам исключить случайные экстремальные значения и оценить влияние окружающей среды на подложку микрочипа.

Полученный результат представляет собой набор различных показателей более чем по десяти тысячам пептидам. К основным показателям, используемым нами, можно отнести: среднее и медианное значения светимости пептида, логарифм отношения для каждого пептида, среднее и медианное значения светимости фона в окрестности пептида.

Среднее значение светимости может дать искаженное представление, так как оно является слишком чувствительным параметром к так называемым «артефактам» — нехарактерным для изучаемой выборки. Медианное значение светимости более устойчиво к «артефактам». В случае же нормального распределения медиана совпадает со средним значением [2].

Логарифм отношения нужен нам для того, чтобы увидеть, есть ли существенное различие в показателях средней и медианной светимостей. И если таковое имеется, можно сказать о том, что результаты содержат в себе «артефакты», не характерные для данной выборки.

При анализе результатов исследования иммуносигнатур было обнаружено, что одни и те же показатели в разных экспериментах могут довольно сильно отличаться. Это поставило под вопрос адекватность полученных данных. Для того чтобы разобраться в причинах появления столь значимых различий, было выдвинуто предположение, что на результат эксперимента оказывают влияние микрочастицы пыли, находящиеся в окружающем воздухе. Пылинки, попадая на сектора с аминокислотными последовательностями, вызывают сильное свечение соответствующих ячеек. Этот эффект может существенно влиять

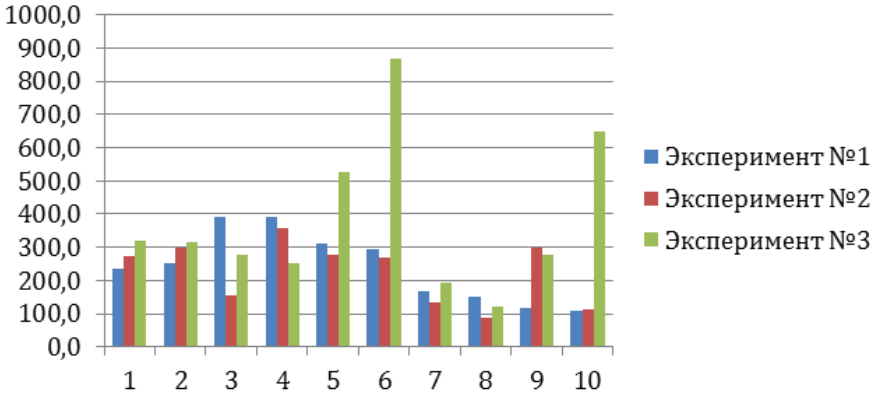

Рис. 1. Светимости пептидов EMPTY по трем экспериментам для группы пациентов

на конечный результат. На рисунке 1 приведена гистограмма распределения средней светимости пептидов EMPTY по трем экспериментам для десяти пациентов с раком молочной железы. Теоретически пептиды EMPTY должны давать единый минимальный уровень светимости (светимость EMPTY должна быть на уровне светимости фона). Однако, как видно на гистограмме светимости EMPTY, для одного пациента в трех экспериментах показатели могут отличаться в несколько раз.

Таким образом, на начальном этапе исследований необходимо было выявить, как сильно внешнее воздействие пыли может влиять на конечный результат исследований. Влияние такого внешнего фактора было решено оценить методом корреляционного анализа. Чтобы проверить, верно ли предположение о попадании пыли на подложку и влиянии ее на конечный результат, были взяты анализы группы из 10 пациентов людей по трем экспериментам и рассчитан коэффициент корреляции для трех экспериментов по каждому пациенту.

На первом этапе велся расчет по всем пептидным последовательностям. В результате у пяти пациентов коэффициент корреляции довольно сильно отклонялся от нормы и находился в пределах от 0,623 до 0,835. В то же время у оставшихся пациентов он был в диапазоне от 0,915 до 0,97. Во избежание нежелательных отклонений мы поставили ограничение на логарифм отношения средней светимости ячейки к медианной светимости

$$\log(\text{Mean}/\text{Median}) \leq 0{,}1.$$

В результате те пептиды, в ячейках которых наблюдалась чрезмерная светимость, отсеялись. Корреляция несколько улучшилась. Но чтобы окончательно отсеять всю пыль, ограничение пришлось увеличить:

$$\log(\text{Mean}/\text{Median}) \leq 0{,}01.$$

В результате были исключены все пептиды, светимость которых превышала норму. На результаты пациентов, у которых корреляция изначально соответствовала норме, это практически не повлияло. Но для пациентов группы с показателями, отклоненными от нормы, она значительно улучшилась. Это подтвердило предположение, что пыль, попавшая на подложку, оказывает значительное воздействие на результат анализа. Результаты эксперимента приведены в таблице.

Значения коэффициента корреляции при различных порогах учета пыли

Номер пациента	Пыль не учитывается	Log (Mean/Median) <0,1	Log (Mean/Median) <0,001
1	0,92	0,92	0,92
2	0,97	0,97	0,97
3	0,93	0,93	0,93
4	0,96	0,96	0,96
5	0,91	0,92	0,92
6	0,83	0,84	0,86
7	0,82	0,84	0,91

Окончание таблицы

Номер пациента	Пыль не учитывается	Log (Mean/Median) <0,1	Log (Mean/Median) <0,001
8	0,71	0,72	0,89
9	0,62	0,65	0,69
10	0,72	0,73	0,81

На следующем этапе необходимо было осуществить нормализацию данных для минимизации влияния внешних факторов (температура и влажность окружающей среды, концентрация компонентов и т. д.) при протекании реакции связывания пептидов в процессе осуществления технологии IMMUNOSIGNATURE. На рисунке 2 изображено типичное распределение средней светимости пептидов EMPTY и остальных пептидов по трем экспериментам для одного пациента.

Таким образом, на эксперимент влияют внешние параметры. При этом суммарная светимость для одного пациента может изменяться в несколько раз при проведении нескольких экспериментов.

На основе эмпирического анализа был предложен метод нормализации данных. Цель нормализации заключается в том, чтобы в результате нормализации все светимости EMPTY имели одинаковые и минимальные значения.

Соответственно, для достижения данного результата необходимо:

1. Рассчитать нормировочный коэффициент по формуле 1/(средняя светимость EMPTY).

2. Все светимости пептидов преобразовать следующим образом:

- если светимость больше удвоенной средней светимости для EMPTY, то она умножается на нормировочный коэффициент;
- если светимость меньше или равна удвоенной средней светимости для EMPTY, то она приравнивается к нулю;
- если пептид является артефактным, то он исключается из дальнейших расчетов.

На следующем этапе нормализованные данные исследовались методом доверительных интервалов. То есть целью численного эксперимента стало выявление наиболее значимых пептидных последовательностей, которые на протяжении всех проведенных экспериментов имеют стабильные значения светимости с небольшим отклонением. Выявление таких последовательностей должно позволить осуществить классифицирование между разными группами пациентов.

В результате проведенных исследований были выявлены пептиды, которые для всех пациентов в пределах одной группы имеют стабильное значение светимости с малым отклонением. Таким образом, для группы пациентов с диагнозом С50 «Злокачественное новообразование молочной железы» было выявлено 20 характерных пептидов, однозначно относящих пациента к данной нозологической группе.

Нейросетевой анализ данных. На данном этапе анализ данных был осуществлен при помощи слоистых нейронных сетей на основе нейропарадигмы Backpropagation.

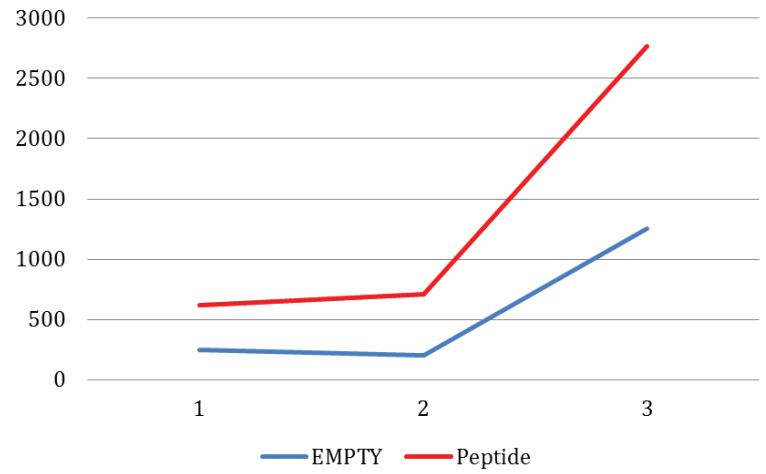

Рис. 2. Средние значения светимостей по трем экспериментам для одного пациента

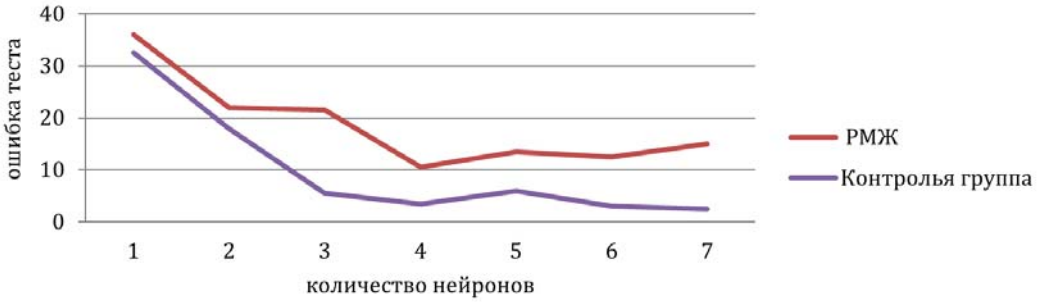

Рис. 3. Динамика ошибки обобщения искусственной нейронной сети при изменении количества скрытых нейронов

Backpropagation на данный момент является очень распространенной моделью построения нейронных сетей, и скорость ее работы достаточна высока для обработки больших массивов информации.

В ходе проектирования искусственной нейронной сети (ИНС) Backpropagation была выбрана оптимальная архитектура, позволяющая распознать рак молочной железы. На рисунке 3 представлен сравнительный анализ архитектур ИНС [3].

Эксперимент проходил следующим образом. В исходной выборке присутствовало 10 пациентов с диагнозом и 25 человек контрольной группы. Поскольку по каждому пациенту было осуществлено три эксперимента, то общее число записей выборки составило 105. Исходная выборка была разделена на две части: обучающая (35 записей) и тестовая (70 записей). В обучающую выборку были включены записи первых экспериментов, а в тестовую — записи вторых и третьих экспериментов.

В ходе нейросетевого исследования число скрытых нейронов ИНС менялось от 2 до 7, в результате чего был получен график зависимости ошибки обобщения от числа нейронов в скрытом слое. Эксперимент показал, что сети с четырьмя нейронами в скрытом слое оптимально классифицирует пациентов из выборки. В настоящее время также ведутся исследования по анализу данных при помощи ИНС Хопфилда, которая позволяет обрабатывать данные в виде изображений. В силу того что данные, получаемые в результате обработки при помощи технологии IMMUNOSIGNATURE, могут быть представлены как в числовом варианте, так и в виде изображения высокой четкости, то использование сети Хопфилда дает возможность для решения поставленной задачи без использования предварительного анализа данных [4, 5].

Назначение применения ИНС Хопфилда состоит в том, то необходимо сформировать образ «идеального» больного. Нейросетевые алгоритмы обработки изображения (в том числе сеть Хопфилда) позволят без предварительной подготовки данных осуществлять классификацию пациентов.

Выводы. В ходе эксперимента были предложены методы использования ИНС при решении задачи классификации данных. Разработанная модель анализа данных технологии IMMUNOSIGNATURE на основе применения традиционных и нейросетевых алгоритмов позволила с достаточной точностью (90 %) классифицировать пациентов с онкологическим диагнозом С50 «Злокачественное новообразование молочной железы».

Проводимые в настоящее время исследования по выявлению оптимальных значений светимостей пептидных последовательностей при помощи ИНС Хопфилда позволят сформировать банк характерных сигнатур для рассматриваемых онкологических диагнозов.

Библиографический список

1. Alexa K Hughes, Zbigniew Cichacz, Adrienne Scheck, Stephen W. Coons, Stephen Albert Johnston, Phillip Stafford Immunosignaturing Can Detect Products from Molecular Markers in Brain Cancer, 2012, DOI:10.1371/journal.pone.0040201

2. Реброва О.Ю. Среднее или все же медиана? // Троицкий вариант. — 2011. — № 90.

3. Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. — Новосибирск, 1996.

4. Золин А.Г., Силаева А.Ю. Применение нейронных сетей в медицине // Актуальные проблемы науки, экономики и образования XXI века : материалы II Междунар. науч.-практ. конф., 5 марта — 26 сентября 2012 года : в 2 ч. Ч. 2 / отв. ред. Е.Н. Шереметьева. — Самара, 2012.

5. Осовский С. Нейронные сети для обработки информации / пер. с польск. — М., 2004.