А. Л. Серебренников

ОБЗОР ВОЗМОЖНОСТЕЙ СРЕДЫ SIGNIFICO НА ПРИМЕРЕ РЕШЕНИЯ ПРИКЛАДНОЙ ЗАДАЧИ

ВВЕДЕНИЕ

Со временем мы сталкиваемся с увеличивающимся потоком информации в различных областях знаний. Порой полученные нами данные неудобны для восприятия или обработки. В основном, это данные, полученные с различных регистрирующих и сканирующих устройств. Другими словами, мы получаем некоторую проекцию реального мира, и тут-то начинаются сложности. Общеизвестно, что с помощью стандартных методов довольно трудно производить, например, распознавание образов, классификацию, тем более прогнозирование. Все эти задачи могут быть решены с помощью нейросетевых технологий, о которых и пойдет речь в этой статье.

Создано достаточно много нейросетевых пакетов, но, как показывает практика, их возможностей хватает для решения довольно ограниченного круга задач. Существующие нейросетевые пакеты обладают недостаточным инструментарием для предварительной обработки данных, не позволяют организовывать системы нейронных сетей и содержат недостаточное число алгоритмов обучения. В этих условиях специалистам приходится создавать специализированный комплекс программного обеспечения для решения задачи.

В статье описывается создаваемая интегрированная среда по работе с нейронными сетями. В среде Significo сделана попытка создать инструментарий, с помощью которого можно довольно удобно решать комплексные задачи. Используя мощную систему предварительной обработки данных, возможность построения систем нейронных сетей и обширный набор алгоритмов обучения, аналитик или специалист могут создать в результате специфичный конвейер обработки данных.

В статье представлен пример решения задачи с помощью среды. Предлагается полностью пройти цикл решения задачи. Это позволит показать

преимущества и недостатки нейросетевых технологий и нововведения, имеющие место в среде Significo.

В данном случае решение задачи позволит производить анализ крови человека, не производя забор крови, а основываясь только на отражённом спектре света от тела человека. Эта методика могла бы значительно увеличить скорость анализа (с нескольких часов до секунды), уменьшить требуемое оборудование (с лаборатории до портативного устройства). Также с помощью метода можно было бы открыть новые возможности диагностики, применяя динамический анализ крови.

Решение задачи будет проходить в пять этапов.

1. Описание физических процессов, проходящих при отражении света, и создание их физической модели.
2. Анализ спектров, в результате которого, нужно выделить необходимые для обучения нейросети участки спектра.
3. Предварительная обработка данных.
4. Выбор наиболее подходящей архитектуры сети и алгоритма её обучения.
5. Подведение итогов проведённой работы и оценка результатов.

1. ОПИСАНИЕ ЗАДАЧ И ЦЕЛЕЙ ЭКСПЕРИМЕНТА

Под воздействием различных электромагнитных полей промышленного и природного происхождения физиологические системы организма человека претерпевают сложные процессы перестройки. К естественным природным электромагнитным полям, в основном, относится солнечная радиация, воздействие на организм человека которой до сих пор не имеет четкой и ясной оценки.

Связь между внешней средой, поверхностными покровами и внутренними органами живого организма многообразна. Она осуществляется через сосудистую, лимфатическую и нервную системы. Кожа человека, являясь эктодермальной производной, связана теснейшим образом со всеми внутренними органами человека. Кожа сама является комплексным органом, покрывающим все тело человека, она соприкасается с внешней средой и,

прежде всего, на ней возникает реакция от внешнего воздействия, переходящая на внутренние органы [1].

Ранее влияние солнечной радиации на организм человека изучалось чаще всего либо с точки зрения адаптации к различным климатическим условиям [2], либо с точки зрения исследования биологических ритмов [3]. В биологии и медицине влияние солнечной радиации на организм человека изучается достаточно давно. Значительные успехи достигнуты в лечении людей ультрафиолетовым и красным светом. Хотя механизмы взаимодействия человеческого организма с электромагнитным полем оптического диапазона частот остаются до конца не изученными, следует отметить также отсутствие медицинской аппаратуры, регистрирующей это взаимодействие.

Для решения этой задачи был использован спектрофотометр «Спектротон», способный регистрировать отраженный сигнал в диапазоне от 380 до 720 нм. Спектральные измерения коэффициента отражения кожи проводились с использованием техники интегрирующей сферы. Такой метод позволяет собирать рассеянное кожей излучение с разных слоев кожи и количественно оценивать коэффициент отражения в широком спектральном диапазоне.

Световой сигнал прибора, посылаемый на кожу, действует кратковременно, что не приводит к изменению физических свойств кожи и не вызывает повышения температуры тканей в зоне облучения. При облучении, например, красным лазером повышение температуры сопровождается выделением свободной воды, а при облучении ультрафиолетовой лампой повышение температуры не сопровождается потоотделением, что говорит о различных механизмах взаимодействия отдельных составляющих спектра с кожей человека.

Для исследования был выбран контингент практически здоровых мужчин и женщин в возрасте от 40 до 80 лет, проживающих в одной географической зоне (г. Новосибирск). Исследования проводились несколько лет с ноября по февраль месяц в первой половине дня, с 10 до 14 часов местного времени. Всего было обследовано 111 человек, из которых 67 человек — женщины и только 44 — мужчины. Спектральная характеристика снималась с внутренней поверхности предплечья левой руки, при этом обследуемый человек находился в комфортных температурных условиях в положе-

нии сидя. Исследования проводились в затемненной комнате. Эта область руки была выбрана для исследования преднамеренно как зона, не содержащая активно функционирующих органов.

2. ОПИСАНИЕ ФИЗИКИ ПРОЦЕССА ОТРАЖЕНИЯ/ПОГЛОЩЕНИЯ СВЕТА ЧЕЛОВЕЧЕСКИМ ТЕЛОМ

Измеряемый параметр достаточно сложная величина. С оптической точки зрения кожа представляет собой поглощающую среду с ярко выраженными рассеивающими свойствами. Взаимодействие света с кожей носит сложный характер, который начинает проявляться уже при прохождении света сквозь границу раздела «воздух—кожа». Граница раздела воздух-кожа не является гладкой. Она представляет собой плотный слой кератиноцитов, на котором располагаются фрагменты эпидермиса, находящиеся в стадии десквамации. Падающее излучение частично отражается (френелевское отражение, составляющее величину порядка 5%), не меняя своего спектрального состава. Значительная часть света (95%) входит в кожу, где свет поглощается и рассеивается. Энергия поглощенного света тратится на протекание различных фотохимических реакций. Спектр поглощения любой биологической ткани, в том числе и кожи, определяется наличием у практически всех биологически активных молекул сопряженных двойных связей (хромофорных групп) [4].

В измеряемом спектральном диапазоне 380-720 нм. роль хромофоров выполняют окисленные формы флавиновых соединений (λ = 450нм.). При связывании флавина с белком максимум сдвигается в сторону больших длин волн и проявляется при λ = 480–490 нм. [5]. Кроме того, помимо флавиновых соединений доминирующим хромофором эпидермиса, определяющим его поглощающие свойства в видимой области спектра, является пигмент меланин [6–10]. Спектр меланина, содержащегося в коже, имеет максимум поглощения около 335 нм. проявляется во всем измеряемом диапазоне и плавно уменьшается с увеличением длины волны.

Основными компонентами дермы, определяющими ее поглощение в видимой области спектра, являются хромофоры дермальной крови (оксигемоглобин, деоксигемоглобин, билирубин), каротиноиды и порфирины. Оксигенированная и деоксигенированная формы гемоглобина поглощают

свет следующим образом. Оксигемоглобин наиболее сильно поглощает свет в области 405нм., с увеличением длины волны его поглощение уменьшается, при этом на длинах волн 535 и 575 нм. имеются характерные максимумы поглощения. Деоксигемоглобин наиболее сильно поглощает свет в области 430 нм. и менее сильно вблизи 550 нм. Поглощение обеих форм гемоглобина в области свыше 600 нм. мало [6, 11, 12]. Для билирубина характерна полоса поглощения с максимумом в диапазоне 450–460 нм.

Помимо поглощения кожная ткань характеризуется значительным светорассеянием. Причиной рассеяния является неоднородность показателя преломления в объеме биоткани, отражающей ее физическую неоднородность [8, 16]. В кожной ткани есть рассеиватели, размеры которых меньше длины волны света, порядка длины волны, и рассеиватели, размеры которых значительно превышают длину волны света. В результате, в коже имеют место различные виды рассеивания: от рэлеевского рассеяния до рассеивания Ми.

3. ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

Предварительная обработка данных имеет очень большое значение при работе с нейросетевыми технологиями, по сути, задачи решаются именно на этом этапе. Искусственные нейронные сети практически помогают в компенсации различного рода помех, распознавании образов, классификации и прогнозировании. Таким образом, на этапе предварительной обработки требуется изучить сущность решаемой проблемы, описать физический смысл имеющихся данных, построить физико-математическую модель обследуемого объекта и относительно полученной модели извлечь из данных принципиально важные параметры, которые впоследствии будут применяться при обучении нейронных сетей.

Среда Significo представляет достаточно обширный инструментарий для осуществления предварительной обработки данных. В Significo представляется возможным составлять из различных элементов среды систему, которой под силу осуществлять предварительную обработку данных в автоматическом режиме, не используя дополнительных инструментариев. В данном аспекте среда Significo имеет ряд принципиальных отличий по

сравнению с имеющимися нейросетевыми пакетами, они заключаются в следующем.

- Возможность построения системы из различных элементов, представляющих собой отдельные специальные алгоритмы, предназначенные для предварительной обработки. В эти алгоритмы, помимо детерминированно направленных методов, входят и некоторые нейронные сети, которые способны компенсировать помехи, находящиеся в данных.
- Наличие значительно расширенного инструментария предварительной обработки, который включает в себя методы линейного статистического анализа, элементы спектрального разложения и др.
- Реализация прозрачности системы предобработки при работе пользователя на этапе обучения сети.

Модель облучаемых биологических тканей целесообразно представлять в упрощенном виде, так как реальная система очень сложна, а множество деталей не известны. Таким образом, представим модель тканей как четырёхслойную структуру.

Первый роговой слой обладает рассеивающими свойствами, и при его прохождении теряется порядка 5% излучения. Этот слой не изменяет состава спектра.

Второй слой представляет собой однородную светопроводящую среду с наличием клеток, которые поглощают отдельные участки спектров, энергия которых уходит на прохождение многочисленных химических реакций.

Третий слой — это совокупность светонепроницаемых образований, расстояние между которыми на порядок ниже длин волн видимого спектра. Таким образом, этот слой представляет собой интерференционную решётку, которая разлагает попавшее на неё излучение на спектр.

Четвёртый слой подобен второму, за исключением того, что в силу другого характера химических реакций поглощает другие участки спектра.

В результате становится ясно, как в рамках принятой нами модели могут изменяться полученные спектральные данные.

4. ОБЩИЙ АНАЛИЗ СПЕКТРОВ ОТНОСИТЕЛЬНО ГЕМОГЛОБИНА

Представленные в литературе относительно гемоглобина данные являются недостаточными, так как в этом случае мы имеем отраженный спектр с поверхности кожи, а известная информация была получена непосредственно из спектра крови. Поэтому необходимо провести общий анализ имеющихся спектров.

С помощью средств предобработки среды Significo строим следующий набор данных. В нём имеют место закономерности:

- все строки упорядочены по возрастанию гемоглобина,
- первый столбец — это изменения гемоглобина между двумя смежными столбцами, все остальные столбцы построены по аналогии, только для различных длин волн спектра,
- для более удобного графического представления первый столбец умножен на 10.

В таблице 3.1 включены участки спектров, которые отмечаются в литературе при анализе спектров крови. Применяя включённый в систему предобработки алгоритм получаем данные, находящиеся в таблице 3.2, которая была получена при анализе данных. Выбранные участки спектров соответствуют длинам волн, которые наибольшим образом коррелируют с параметром динамики гемоглобина.

В результате, из представленных таблиц можно сделать вывод, что большая часть представленных в источниках участков спектров или сильно зашумлены процессами, имеющими место в тканях человека, или сдвинулись по тем же причинам. В итоге совпало менее 50% участков спектра.

Таблица 1

160	10	10	14	20	22	32	35
30	-45	-37	-33	-52	-59	-56	-63
30	63	48	39	53	57	37	32
0	-21	-8	-6	-7	-8	-2	-6
30	4	3	8	13	16	18	27
20	-79	-76	-74	-79	-78	-50	-29
0	130	123	120	123	116	91	59
10	2	-4	-9	4	13	12	42
10	-39	-32	-31	-40	-41	-56	-65
0	-73	-64	-54	-87	-101	-51	-79
40	79	70	61	89	98	57	70
0	-1	-2	2	1	2	6	22
40	-57	-49	-46	-43	-38	-56	-82
20	51	44	43	38	30	55	64
40	-33	-36	-41	-40	-37	-47	-30
20	2	10	19	9	4	21	14
70	67	63	51	71	77	66	65
	400	410	430	450	460	550	600

Таблица 2

160	12	10	10	14	25	26
30	-65	-45	-37	-33	-59	-46
30	96	63	48	39	54	9
0	-47	-21	-8	-6	-10	5
30	8	4	3	8	21	1
20	-96	-79	-76	-74	-75	8
0	152	130	123	120	113	6
10	11	2	-4	-9	18	58
10	-43	-39	-32	-31	-22	-58
0	-116	-73	-64	-54	-134	-55
40	120	79	70	61	111	49
0	-4	-1	-2	2	-2	45
40	-84	-57	-49	-46	-43	-71
20	77	51	44	43	27	17
40	-24	-33	-36	-41	-32	-10
20	-21	2	10	19	1	16
70	83	67	63	51	95	29
	380	400	410	430	480	720

Полученные участки спектров являются лишь первоначальной предобработкой данных, заключающейся в выделении необходимых и достаточных данных для решения задачи.

Допустим, что в результате последующей нормировки все входные и выходные переменные отображаются в единичном кубе.

Задача нейросетевого моделирования, найти статистически достоверные зависимости между входными и выходными переменными. Единственным источником информации для статистического моделирования являются примеры из обучающей выборки. Чем больше бит информации принесет каждый пример, тем лучше используются имеющиеся в распоряжении данные.

Рассмотрим произвольную компоненту нормированных (предобработанных) данных — $\tilde{x}_i$. Среднее количество информации, приносимой каждым примером $\tilde{x}_i^\alpha$, равно энтропии распределения значений этой компоненты $H(\tilde{x}_i)$. Если эти значения сосредоточены в относительно небольшой области единичного интервала, информационное содержание такой компоненты мало. В пределе нулевой энтропии, когда все значения переменной совпадают, эта переменная не несет никакой информации. Напротив, если значения переменной $\tilde{x}_i^\alpha$ равномерно распределены в единичном интервале, информация такой переменной максимальна. Общий принцип предобработки данных для обучения, таким образом, состоит в максимизации энтропии входов и выходов. Этим принципом следует руководствоваться и на этапе кодирования нечисловых переменных.

5. ВЫБОР ОПТИМАЛЬНОЙ АРХИТЕКТУРЫ, ДОПОЛНИТЕЛЬНЫЕ СРЕДСТВА ДЛЯ УСПЕШНОГО РЕШЕНИЯ ЗАДАЧИ

Среда Significo к своему завершению должна иметь обширный набор нейросетевых парадигм, который мог бы позволить решать разнообразные задачи, связанные с компенсацией шумов, аппроксимацией, распознаванием образов, кластеризацией, классификацией. Для решения всех этих задач существуют наиболее удачные нейросетевые парадигмы. Предполагается, что в среде будет реализовано 6 архитектур, которые можно будет применять для решения вышеприведённых задач. Рассмотрим их кратко.

- Самой популярной архитектурой на данный момент является многослойный персептрон. Это наиболее универсальная архитектура, позволяющая применять к ней большой спектр обучающих алгоритмов. Эта сеть в основном применяется для распознавания различных образов.
- Сети с обратными связями. Это несколько преобразованный многослойный персептрон, который имеет обратные связи. Применяются данные сети для задач прогнозирования.
- Рекуррентные сети. Эти сети наиболее популярны в задачах прогнозирования, так как обладают свойством долговременной памяти, т.е. запоминают последовательности.
- Сети Кохонена. Данные сети наиболее популярны в задачах классификации. Это самоорганизующаяся сеть, и обучается она без учителя, т.е. позволяет проводить автоматическую классификацию.
- Вероятностные нейронные сети. Этот тип сетей знаменит своей способностью обучаться на ограниченном наборе данных. Может быть использован для классификации и кластеризации.
- Нейронные сети с общей регрессией. Данный тип нейронных сетей отличается тем, что может обучаться на ограниченном наборе данных и используется обычно для аппроксимации непрерывных функций.

Принципиальными отличиями среды Significo по сравнению с существующими на данное время нейросетевыми пакетами являются следующие.

- Способность объединять отдельные нейросети в системы, что позволяет снизить время обучения, по сравнению с применением одной нейросети. Каждую локальную задачу можно решать с помощью специфичной для этой задачи нейросетевой архитектуры.

- Возможность контрастирования нейронных сетей. Это позволит представлять сети не как чёрный ящик, а как логически понятную структуру. Операция контрастирования также минимизирует количество связей и нейронов в сети, что при моделировании снижает время обучения сети и количество потребляемых ресурсов. Сеть становится наиболее предсказуемой в зависимости от степени контрастирования.
- Реализовано большее количество алгоритмов обучения, что особо полезно при практическом использовании среды.
- Сделаны оригинальные доработки алгоритмов обучения, которые позволяют более гибко управлять процессом обучения, анализировать его и ускорять.

Вернёмся к решаемой задаче. В данном случае от нейросети требуется работа с зашумленными данными по распознаванию образов. Существует достаточно много нейросетевых парадигм для решения данной задачи, но целесообразнее выбрать именно многослойный персептрон по следующим причинам:

- успешно работает с зашумлёнными данными,
- применяется в задачах распознавания образов,
- для обучения многослойного персептрона имеется очень большое количество обучающих алгоритмов,
- возможно применять операцию контрастирования,
- легко анализировать процесс обучения.

При работе с многослойным персептроном целесообразно использовать сначала алгоритм обратного распространения ошибки, так как он менее ресурсоёмкий и для обучения требуется меньше времени. Даже в случае неудачного обучения можно извлечь пользу, например, оценить насколько сложные закономерности существуют в обучающих векторах, на это указывает уровень локальных минимумов ошибки. Если сеть не выходит из локального минимума, в процессе обучения можно применять различные

средства для исправления ситуации, например метод шока сети, или перейти на один из стохастических алгоритмов обучения, например Больцмановское обучение или обучение Коши.

6. ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

На завершающем этапе работы производится обучение сети и подведение итогов.

Начнём с краткого анализа процесса обучения и указания особенностей работы среды в нём. Особенностями Significo являются следующие возможности.

- Выбор различных алгоритмов формирования списка калибровочных векторов: последовательный выбор, псевдослучайный выбор или ручной.
- Процесс обучения имеет несколько видов управления: ручной, по истечению определённого числа итераций, при достижении определённого значения ошибки на тренировочных или на калибровочных данных.
- Возможность отката процесса обучения на произвольную позицию с учётом выбираемого пользователем шага.

На рис. 1 мы видим картину процесса обучения, график серого цвета отображает ошибку обучения, черным цветом, отображается ошибка на калибровочных данных.

График на рис. 1 говорит о том, что на начальном этапе обучения сеть зашла в локальный минимум, через 40000 итераций вышла и далее продолжала обучаться в штатном режиме. По завершению обучения ошибка на тренировочных векторах составляла порядка 3%, а на калибровочных — 9%.

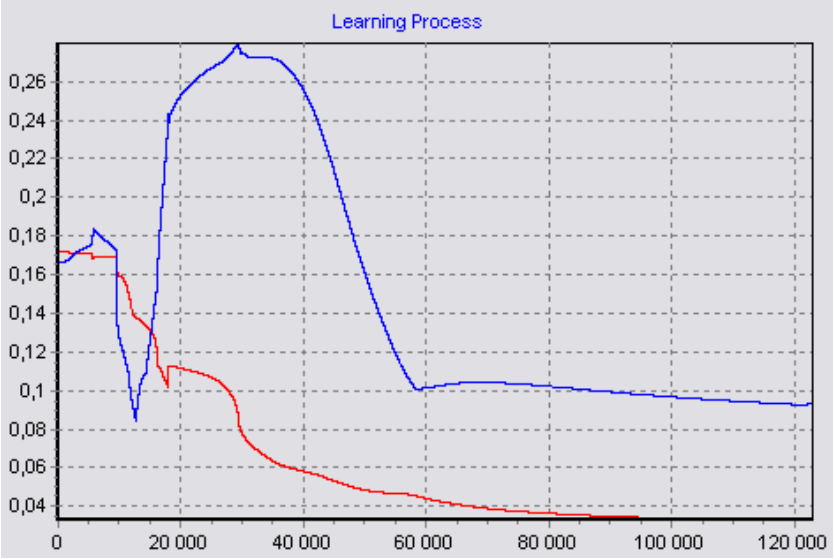

Рис. 1. График процесса обучения нейронной сети

Рассмотрим полученные результаты, они приведены в таблице 3. Вся таблица разделена на отдельные элементы, в которых указаны: прогнозируемое значение, значение, полученное классическими методами, абсолютная ошибка прогноза. Учитывая, что допустимая ошибка в стандартных анализах крови может достигать 15мг, все прогнозируемые значения гемоглобина удовлетворяют поставленным условиям. При обучении для калибровки было выбрано 20% векторов, 2 из которых не имеют отклонений, а 2 были спрогнозированы с достаточно большой ошибкой в 10%. Это объяснятся тем, что в наборе тренировочных векторов не было сходных с ними векторов. В тренировочном наборе имеются два прогноза с умеренной ошибкой, скорее всего их статистическая составляющая была довольно низкой. Основная же часть прогнозов была предельно точной.

Таким образом, это небольшое исследование можно считать успешным. Для достижения лучшего результата необходимо большее количество обучающих векторов.

Таблица 3

Результаты обучения нейронной сети

136,9993 137 -0,0007		135,0345 135 0,034522		125,0992 136 -10,9008	8,80%	142,055 142 0,054997	
131,0209 131 0,020897		144,4476 148 -3,55237	2,40%	118,9158 119 -0,08418		163,6293 164 -0,37074	
124,8767 125 -0,12329		138,9955 139 -0,00455		134,9771 135 -0,02287		127,0232 127 0,023163	
145,0121 145 0,012106		142,0218 142 0,021806		131,0731 131 0,073131			
124,5533 112 12,55333	10%	124,5512 137 -12,4488	10%	121,0053 121 0,005333			

общее число векторов	18
предельно корректный результат	12
хороший результат	2
Удовлетворительный результат	1
результат укладывающийся в норму 15мг	3

СПИСОК ЛИТЕРАТУРЫ

1. Владимиров Ю.А., Потапенко А.Я. Физико-химические основы фотобиологических процессов. — М.: Высш.шк., 1989. — 189 с.
2. Кочубей В.И., Конюхова Ю.Г. Методы спектрального исследования крови и костного мозга. — Саратов: Изд-во Сарат. ун-та., 2000. — 72 с.

3. А.Н.Горбань, В.Л.Дунин-Барковский, А.Н.Кирдин и др. Нейроинформатика. — Новосибирск: Наука. Сибирское предприятие РАН, 1998. — 296 с.
4. Ф. Уоссермен Нейрокомпьютерная техника, теория и практикаю — Norwood, 1992. — 118 с.
5. Масалович А.И. От нейрона к нейрокомпьютеру. — Журнал доктора Добба. — 1992. — 48 с.
6. Poli R., Cagnoni S., Livi R. et al. A Neural Network Expert System for Diagnosing and Treating Hypertension. — Computer, 1991. — 71 с.
7. Baxt W.G. A neural network trained to identify the presence of myocardial infarction bases some decisions on clinical associations that differ from accepted clinical teaching. — Med. Decis. Making, 1994. — 222 с.