DOI: 10.18720/SPBPU/2/id20-59

РУССКОЕ ПРЕДИСЛОВИЕ

***Ксения Александровна Найденова**, к.т.н., ст. научный сотрудник, Военно-медицинская академия имени С.М.Кирова, Россия, Санкт-Петербург, ул. Академика Лебедева, д. 6, индекс 194044, ksennaidd@gmail.com.*

***Константин Владимирович Швецов**, профессор, Санкт-Петербургский политехнический университет Петра Великого, Россия, Санкт-Петербург, ул. Политехническая, дом 29, индекс 195251, Konstantin.Shvetsov@spbstu.ru.*

***Александр Викторович Яковлев**, к.т.н., ст. научный сотрудник, Военно-медицинская академия имени С.М.Кирова, Россия, Санкт-Петербург, ул. Академика Лебедева, д. 6, индекс 194044; доцент, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Россия, Санкт-Петербург, ул. Большая Морская, д. 67, индекс 190000.*

***Владимир Андреевич Пархоменко**, программист, Санкт-Петербургский политехнический университет Петра Великого, Россия, Санкт-Петербург, ул. Политехническая, дом 29, индекс 195251, Vladimir.Parkhomenko@spbstu.ru.*

Данная книга посвящена машинному обучению. Мы рассматриваем *машинное обучение* как класс методов в области интеллектуальной обработки данных, характерной чертой которых является обучение по примерам применения решений множества сходных задач, а не априорное задание способа решения задачи. Машинное обучение исследует построение и изучение алгоритмов, которые могут обучаться и делать прогнозы на данных. Методы машинного обучения могут быть отнесены к прогнозной аналитике или прогнозному моделированию.

Для построения алгоритмов машинного обучения используются следующие методы: математическая статистика, численные методы, методы оптимизации, теория вероятностей, теория графов, извлечение из данных логических правил для распознавания заданных классов объектов, построение онтологий, баз знаний и различного типа структур знаний и другие подходы.

Различают два типа обучения: обучение по прецедентам и выявление эмпирических закономерностей в данных. Во втором случае имеется множество объектов (ситуаций) и множество возможных ответов (откликов, реакций). Существует некоторая зависимость между ответами и объектами, но она неизвестна. Известна только конечная совокупность пар «объект, ответ», называемая обучающей выборкой. На основе этих данных требуется построить алгоритм, способный для любого возможного входного объекта выдать достаточно точный классифицирующий ответ.

Эта зависимость не обязательно выражается аналитически, например, нейросети реализуют аппроксимацию практически любой сложной функции, которую нельзя выразить аналитически. Важной особенностью при этом является способность обучаемой системы к обобщению, то есть к адекватному отклику на данные, выходящие за пределы имеющейся обучающей выборки.

Данная постановка является обобщением классических задач аппроксимации функций. В классических задачах аппроксимации объектами являются действительные числа, векторы, временные последовательности и классификации объектов как, например, в подходе к машинному обучению на основе конструирования хороших классификационных (диагностических) тестов (см. главу 9 в книге). В реальных прикладных задачах входные данные об объектах могут быть неполными, неточными, нечисловыми, разнородными. Эти особенности приводят к большому разнообразию методов машинного обучения.

Методы машинного обучения достаточно хорошо развиты. Однако современные технические средства измерений, экспериментальных исследований, наблюдений привели к накоплению огромных объёмов данных в науке, производстве, бизнесе, транспорте, здравоохранении. Возникающие при этом задачи прогнозирования, управления и принятия решений часто затруднены, так как необходима предварительная обработка данных, их структурирование, декомпозиция, выделение или формирование классификационных признаков, дискретизация признаков, построение оценочных шкал и многое другое. Таким образом, центр работ в области компьютерных наук

сместился в сторону обработки и анализа больших данных, в которых машинное обучение зачастую представляет собой «вершину громадного айсберга».

Сегодня термин «большие данные» приобрел большую популярность и активно используется в различных сферах. Однозначного понимания содержания этого термина до сих пор не существует. Однако все определения согласны с тем, что «большие данные» — это технология анализа данных, направленная на извлечение полезных новых знаний из таких объемов данных, с которыми не справляется человек. Анализ литературы показывает, что «большим данным», кроме их большого объема, присущи два основополагающих признака:

- объединение данных из разнообразных источников; это может быть совокупность различных баз данных, наблюдений или измерений, цифровые архивы медицинских изображений, скрининг населения на основе телемедицинской системы, Интернет и т. п.;
- необходимость использования принципиально новых и сложных методов анализа таких данных в большей степени, чем ранее основанных на принципах машинного обучения.

Миссией редакторов был поиск таких работ авторов, в которых акцентировано внимание на следующих проблемах для решения прогностических и диагностических задач в медико-биологических и социально-экономических исследованиях:

- практическое применение алгоритмов машинного обучения;
- подходы и практические методы формирования шкал признаков для решения задач машинного обучения;
- сбор и предварительная обработка мульти-модальных данных большого объема.

Публикационная политика. Приём работ осуществлялся через систему Easychair. Все главы, включенные в монографию, прошли две итерации слепого рецензирования. Каждая глава была рецензирована одним из редакторов книги (К. А. Найдёновой, К. В. Швецовым, А. В. Яковлевым).

Отдельные главы были дополнительно рецензированы авторами других глав. Часть поданных работ, к сожалению, не вошли в сборник. Данные работы, как правило затрагивали вопросы смежных, но не приоритетных на-

правлений книги. Редакторы благодарны всем, кто подавал свою работу для включения в данное издание.

Содержание рукописи находится в *свободном доступе* через Интернет, используя DOI. DOI книги указан в её библиографическом описании. Каждая глава имеет собственный DOI, работающий как url-ссылка, в эпиграфе перед названием каждой главы. Эта ссылка приводит к соответствующему pdf-файлу.

Оглавление книги представлено на английском и русском языке (также как и предметный указатель, предисловие и некоторые другие элементы книги). В оглавлении не отображаются такие специальные разделы каждой работы как «введение», «заключение», «благодарности» и «библиографический список».

Далее мы приводим краткое содержание глав, условно разделённых по двум частям. Авторы постарались представить большое число новых и ценных материалов, относящихся к области обработки данных и машинного обучения.

0.1. Машинное обучение в анализе биомедицинских данных

Первую часть книги начинает глава 1 «Использование методов машинного обучения в медицине» авторы Олег Валентинович Сенько и Анна Викторовна Кузнецова. Они дают обзор технологий машинного обучения для решения задач диагностики и прогнозирования в медицине. Данные технологии позволяют генерировать алгоритмы, вычисляющие диагностические или прогнозные решения в автоматическом режиме по накопленным массивам клинических данных. В главе приводятся примеры использования различных методов машинного обучения для решения конкретных медицинских задач. Наряду с распространёнными технологиями рассматриваются также оригинальные методы, основанные на принятии коллективных решений на основе выделенных закономерностей. Обсуждаются вопросы корректной оценки эффективности алгоритмов диагностики и прогноза.

Глава 2 «Импульсные рекуррентные нейронные сети для классификации электрокардиограмм по типу аритмии» Кирилла Вячеславовича Никитина

открывает читателю одну из активно развивающихся областей в теории нейронных вычислений. Большинство всех существующих моделей нейронных сетей (далее — НС) можно условно разделить на два класса в зависимости от модели нейрона — искусственные НС и биологически ориентированные НС. В первом случае в качестве нейронов используются математические модели, описывающие некоторые особенности своих прототипов — как правило, входные сигналы умножаются на веса, складываются и к ним применяется передаточная функция, чаще всего сигмоидальная. Во втором случае нейроны моделируются как можно более правдоподобно, с учетом динамики, физических и химических свойств и процессов, происходящих в реальных нейронах. Одно из основных таких свойств — генерация нейронами коротких вспышек активности — импульсов. Поэтому часто такие модели НС называют импульсными.

При моделировании импульсных НС для каждого нейрона составляется система дифференциальных уравнений и далее эта система приближенно решается во временной области. Во многих работах показано, что по своим возможностям импульсные НС обладают большим потенциалом по сравнению с искусственными, особенно при работе с изменяющимися во времени сигналами. Поэтому в настоящее время импульсным НС уделяется большое внимание в самых различных областях. В главе описывается разработанная автором программная среда моделирования импульсных НС, приводятся результаты её исследования и анализ различных моделей импульсных нейронов и химических синапсов. Полученные результаты используются для построения методики соответствующих НС в составе машины неустойчивых состояний для решения задач распознавания динамических образов. Рассматривается пример применения методики моделирования НС для классификации электрокардиограмм по наличию и типу аритмии.

В главе 3 «Метод композиции алгоритмов машинного обучения на основе Oracle Data Mining для прогнозирования сердечно-сосудистых заболеваний» авторы Александра Дмитриевна Соболева и Олег Юрьевич Сабинин рассматривают существующие методы машинного обучения, решающие задачу прогнозирования и анализируют их недостатки. В главе предложен и обоснован метод, решающий задачу прогнозирования посредством агрегирования

результатов двух алгоритмов машинного обучения (Обобщенная линейная модель и Машина опорных векторов), противоположных по природе. Рассмотрена проблема раннего диагностирования сердечно-сосудистых заболеваний. Проведено исследование метода на примере трех различных наборов данных анализов пациентов и их анамнеза для оценки рисков кардиологических заболеваний, подтвердившее эффективность разработанного подхода для решения задачи ранней диагностики сердечно-сосудистых заболеваний.

Глава 4 «Возможности применения методов вычислительной гидродинамики в изучении гемодинамических особенностей интракраниальных аневризм» является примером комплексного подхода к сложнейшим медицинским проблемам, требующим интеграции знаний в области аналитической медицины, диагностике, математике, физике, обработке данных и программировании. Авторы главы Андрей Васильевич Гаврилов, Дарья Дмитриевна Долотова, Евгения Романовна Благосклонова, Иван Владимирович Архипов, Елена Владимировна Григорьева, Наталья Алексеевна Полунина и Владимир Викторович Крылов применяют для оценки гемодинамических особенностей интракраниальных аневризм современные методы обработки 2D/3D изображений, а также методы математического моделирования на основе вычислительной гидродинамики.

Выбор схемы лечения пациентов с подобной патологией является сложной задачей, в которой необходимо учитывать соотношение рисков разрыва аневризмы при ее консервативном лечении с интраоперационными рисками при различных способах хирургического лечения. До сих пор оценка риска разрыва аневризмы является неразрешенной проблемой. Авторы предлагают систему анализа рисков и диагностики аневризмы, базирующуюся на пациенто-специфических трехмерных реконструкциях участков сосудистого русла с аневризмой. При этом используются данные КТ-ангиографии с последующим выполнением гидродинамических расчетов с учетом скоростных характеристик потока крови конкретного пациента. При вычислениях риска учитываются, как морфометрические, так и гемодинамические показатели, оценить которые существующими методами диагностики in vivo не представляется возможным. Глава представляет богатейший материал сбора, анализа и обработки большого количества данных в конкретной лечебной

задаче. Применение методов машинного обучения является перспективным развитием предлагаемой аналитической диагностической системы.

Глава 5 «Автоматизированная система сбора и обработки данных для машинного обучения при оценке функционального состояния специалистов» Александра Викторовича Яковлева, Вячеслава Олеговича Матыцина, Ксении Александровны Найденовой и Владимира Андреевича Пархоменко посвящена проблеме синхронного измерения большого количества первичных показателей различной природы, их сохранения в базе данных для последующей обработки с целью выделения вторичных информативных показателей для оценки функционального состояния специалистов. Диагностики и прогнозирования функционального состояния и профессиональной надежности специалиста является актуальной задачей во многих областях человеческой деятельности: управление сложными человеко-машинными системами, вождение автотранспорта, пилотирование самолетов, работа в космосе и многое другое.

Авторами разработан комплекс мультимодальной регистрации данных о функциональном состоянии специалиста. В предлагаемом решении применяются технические и программные средства для регистрации больших массивов данных трех видов: аудиоданных, видеоданных и физиологических данных, получаемых с датчиков полиграфа. Были выбраны соответствующие алгоритмы, а преобразования первичных данных протестировано, чтобы получить структурированные мультимодальные индикаторы, выраженные в числовой форме. Произведён пилотный эксперимент, чтобы оценить характеристики разработанной базы данных и возможностей стенда. Стенд может быть использован не только для диагностических задач, но и для фундаментальных исследований в физиологии.

Глава 6 «Формирование внешнего критерия для машинного обучения на основе медико-биологических данных» Нэллы Алексеевны Щукиной касается важнейшей проблемы машинного обучения, а именно задания внешнего критерия, то есть внешней эталонной классификации объектов, при формировании обучающей выборки. В медико-биологических исследованиях серьёзным препятствием для применения методов машинного обучения является неопределённость внешнего критерия как основы «хорошо понимаемо-

го образа» в задачах распознавания. В работе предлагается метод создания содержательной шкалы, устанавливающей соответствие между количественным признаком и его качественными оценками при отсутствии заданного чёткого внешнего критерия.

Метод иллюстрируется примером отбора респондентов на специальность по уровню их физической подготовленности. Моделируется внешний критерий, основанный на качественной оценке их анаэробной способности, которая физиологически связана с искомой степенью физической подготовленности. В статье описаны этапы эмпирического моделирования внешнего критерия, на основе опосредованного шкалирования физиологического свойства респондентов. Эта работа также может найти широкое применение не только в медико-биологических, но и в социологических исследованиях.

0.2. Машинное обучение в анализе социально-экономических данных

Вторую часть книги начинает глава 7 «Машинное обучение для выявления подгрупп индивидов, сильно реагирующих на воздействие» авторы Алексей Владимирович Бузмаков, посвящена проблеме оценки эффекта от воздействия на индивидуальном уровне, актуальной во многих областях знаний от маркетинга до медицины. Например, в медицине эта проблема связана с определением, на кого как действует какое-либо лекарство. В маркетинге отправление предложений только тем людям, которых интересует конкретный товар, возможно снизит издержки рекламной кампании. В данной главе рассматриваются существующие математические методы оценки эффекта от воздействия на индивидуальном уровне, дается их критический анализ и показывается необходимость создания новых эффективных методов в этой области знаний.

В глава 8 «Функция желательности и шкала предпочтений Харрингтона в психологических исследованиях» Нэллы Алексеевны Щукиной предлагается оригинальное решение проблемы интегральной оценки психологического профиля респондентов в области профессионального отбора. Сложность состоит в формировании комплексного показателя из тестовых показателей разной смысловой и психофизической природы.

В данной работе предлагается для решения проблемы ввести единую искусственную метрику, основанную на функции желательности Харрингтона, и поставить ей в соответствие стандарт в виде безразмерной шкалы (шкалы предпочтений Харрингтона). На этой основе был разработан способ формирования искомого интегрального показателя с использованием машинного обучения для кодирования тестовых оценок респондентов. В статье изложены принципы формирования интегрального показателя и правила его применения. Предлагаемый способ формирования интегрального показателя может быть успешно применен для широкого круга проблем в психофизиологических, социологических и социально-экономических исследованиях.

Глава 9 «Применение логико-комбинаторной нейроподобной сети в задачах символьного машинного обучения» Ксении Александровны Найденовой, Владимира Андреевича Пархоменко и Константина Владимировича Швецова знакомит читателя с одним из направлений символьного машинного обучения на основе конструирования хороших классификационных (диагностических) тестов. В работе формулируются основные задачи символьного машинного обучения, цель которых выделение из данных логических правил и закономерностей, включающих функциональные, импликативные зависимости, ассоциативные правила, паттерны, удовлетворяющие различным ограничениям и многие другие. Приводится Apriori-алгоритм, решающий перечисленные задачи символьного машинного обучения. Предлагается логико-комбинаторная нейроподобная обучаемая сеть, позволяющая эффективно реализовать Apriori-подобный универсальный индуктивный метод выделения зависимостей из данных.

Функционирование нейроподобной логико-комбинаторной сети описывается на примерах выделения из данных хороших максимально избыточных и безизбыточных классификационных тестов, в том числе и в практической задаче выявления взаимосвязей между личностными характеристиками и динамикой интеллектуального развития женщин кадетов. Предлагаются новые методы снижения вычислительной сложности алгоритмов, легко реализуемые на логико-комбинаторных нейроподобных сетях.

Благодарности

Редакторы благодарны авторам за подачу работ, рецензентам за все усилия по улучшению содержания книги, руководству и коллективу Санкт-Петербургского политехнического университета Петра Великого (далее — СПбПУ) за поддержку издательского процесса.

Редакторы выражают глубокое уважение и благодарят члена-корреспондента Российской академии наук, проректора по научной работе СПбПУ В. В. Сергеева за предоставленную возможность попробовать новый русско-английский формат монографии для её большего доступа для англоязычных читателей.

Редакторы выражают отдельную благодарность доценту СПбПУ А. В. Щукину, без ценных замечаний и поддержки которого данная книга вряд ли бы появилась на свет.

Редакторы очень признательны следующим сотрудникам СПбПУ за помощь в разработке требований к оформлению издания: А. В. Ваньковичу, В. М. Якубсон, Т. В. Бабошиной, Н. В. Соколовой, В. А. Лысенко, С. В. Шутовой, Т. П. Наумовой.

Процесс издания книги был частично поддержан Проектом академического превосходства 5-100, предложенным Санкт-Петербургским политехническим университетом Петра Великого.

Искренне Ваши,

редакторы