

УДК 004.9:378:001.891

ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ DATA MINING В УЧЕБНЫХ КЛИНИЧЕСКИХ ИССЛЕДОВАНИЯХ

Пермякова А.В.

*Пермский государственный медицинский университет им. академика Е.А. Вагнера, Пермь,
e-mail: derucheva@mail.ru*

В статье рассматриваются возможности применения современных IT-технологий в учебной и исследовательской работе студентов в медицинском вузе на примере технологии искусственного интеллекта Data Mining. Представлен метод логистической регрессии в качестве инструмента по разработке математико-статистической модели прогноза вероятности наступления интересующего исследователя события на примере группы часто болеющих детей раннего возраста (всего 146 человек), продемонстрированы возможности применения алгоритма оптимального квантования при изучении клинических и лабораторных предикторов коклюшной инфекции у госпитализированных детей (56 человек). Применение в учебно-исследовательской работе студентов медицинского вуза метода математико-статистического моделирования на основе искусственного интеллекта позволяет проводить сложные виды анализа, формулировать их в доступной пониманию большинства форме и готовить на их основе клинические решения.

Ключевые слова: многомерные данные, data mining, медицинские исследования

THE POSSIBILITIES OF USING DATA MINING IN EDUCATIONAL CLINICAL STUDIES

Permyakova A.V.

Medical University PGMU them. Acad. E. Wagner MoH, Perm, e-mail: derucheva@mail.ru

The article examines the possibilities of applying modern IT technologies in educational and research work at a medical college. The method of logistic regression as a tool for developing a mathematical and statistical model for predicting the probability of an event of interest for an investigator is presented by researching of a group of frequently and long-term ill children (146 in all), and the possibility of using the optimal quantization algorithm in studying clinical predictors of pertussis infection in children (56 people). The method of mathematical-statistical modeling applied in the students' teaching and research work allows one to carry out the most complicated types of analysis, formulate them in an accessible understanding of the majority of forms, and prepare clinical solutions based on them.

Keywords: multidimensional data, data mining, medical research

Стремительное развитие современной науки напрямую связано с возрастающим ростом получаемой информации, переработать которую возможно только при помощи современных технологий. Одним из инструментов автоматического анализа данных является технология искусственного интеллекта Data Mining [1]. Обнаружение знаний в базах данных – таков наиболее близкий синоним этого термина при переводе на русский язык. В отличие от традиционной математической статистики, оперирующей средними данными по выборке, в технологию Data Mining заложена концепция *шаблонов*, позволяющая обнаруживать в данных скрытые и неуловимые для человеческого интеллекта взаимосвязи, так называемые *правила*. Возможности, заложенные в этой технологии, позволяют не только описывать данные, но и предсказывать их с определенной степенью точности. Технология Data Mining успешно реализуется при помощи разнообразных программных продуктов, предлагаемых производителями софта. Это такие пакеты, как SPSS (SPSS, Clementine), Statistica (StatSoft), Deductor Academic (Base Group Labs).

К сожалению, такие современные инструменты анализа данных пока еще не получили должного распространения в академической среде. Большинство высших медицинских учебных заведений имеют опыт обучения студентов в области классической статистики, хотя только статистикой анализ не ограничивается. Анализ данных – это гораздо более широкое понятие. Кроме того, руководств по анализу данных, особенно адаптированных к потребностям практиков, крайне мало, программы медицинских вузов подготовки специалистов по обработке баз данных не включают, так что решения в сфере здравоохранения принимаются в лучшем случае на основании результатов научных исследований и экспертных оценок, зачастую противоречивых.

По мнению отечественных авторов [2], совершенно очевидно назрела необходимость организации в каждом регионе аналитических центров, обобщающих информацию из разного рода источников на основе Data Mining (возможно, это будут Data-центры). Data mining – это именно то, что сейчас следует активно осваивать, адаптировать к задачам и возможностям заин-

тересованных в оценке здоровья населения ведомств и внедрять в их рутинную работу.

Цель исследования

Показать возможность применения современных IT-технологий в учебной и исследовательской работе в медицинском вузе.

Кафедра детских инфекционных болезней ГБОУ ВО ПГМУ имени ак. Е.А. Вагнера успешно апробировала и применяет методы анализа данных Data mining в учебно-исследовательской работе студентов. За последние 3 года осуществлено 5 таких научных студенческих работ. Результаты всех УИРС были представлены на научных конференциях различного уровня (Пермь, Минск, Москва, Санкт-Петербург), и опубликованы в индексируемых ВАК и РИНЦ журналах.

Материалы и методы исследования

Для учебно-исследовательской работы студентов выбрана аналитическая платформа Deductor Academic (Base Group Labs), а именно Deductor Studio 5.3 – рабочее место аналитика. Программа является свободно распространяемым софтом, проста в применении и предназначена для визуального проектирования логики принятия решений. Все действия настраиваются при помощи всего 4-х мастеров: импорт, экспорт, обработка и визуализация. Deductor Studio позволяет автоматизировать рутинные операции по обработке данных, сосредоточиться на интеллектуальной работе и формализации правил принятия решений.

Одно из направлений исследовательской работы кафедры – построение прогностических моделей патологических состояний (инфекционных заболеваний) у детей. Модель всегда имитирует некоторый процесс, позволяя исследователю объяснить причины заболевания, оценить значимость патогенных факторов, а также спрогнозировать течение патологического процесса (заболевания). Для того чтобы построить модель, исходные данные подвергают обработке, классифицируют и далее анализируют математическими методами [3]. Используются следующие алгоритмы обработки данных: кластеризация, классификация, нейросеть, дерево решений, ассоциативные правила, карты Кохонена, линейная и логистическая регрессия и т.п. Авторами в ходе учебно-исследовательской работы построены математические модели тяжелой формы ротавирусной инфекции, синдрома цитолиза при инфекционном мононуклеозе, математически определены клинико-эпидемические предикторы коклюшной инфекции, предложен вариант диагностического алгоритма работы с часто болеющими детьми.

Результаты исследования и их обсуждение

В УИРС, посвященной часто болеющим детям, для построения математической модели использовали анамнестические и клинико-лабораторные данные, полученные в ходе исследования, в котором приняли участие 146 детей в возрасте 2–3 лет. Для

оценки частоты респираторных заболеваний использовали индекс резистентности (ИР), определяемый как отношение числа заболеваний к числу месяцев наблюдения, в процентном выражении. Использовался метод логистической регрессии, тип модели – одномерная, моделируемое событие (зависимая переменная) – частая заболеваемость острыми респираторными инфекциями (ОРИ), значения индекса резистентности более 50%, в качестве входных независимых переменных (предикторов) использовались анамнестические, клинические и лабораторные данные, полученные в исследовании. Логистическая регрессия применяется в случае, когда зависимая переменная является бинарной (т.е. может принимать только два значения) [4].

Интерпретация параметров логистической регрессии производилась на основе величины OR – отношение шансов, когда значения отношения шансов больше единицы, это означает, что переменная способствует увеличению шансов частых ОРИ, в то время как значения менее единицы указывают, что переменная отрицательно влияет на эти шансы. Составлены логистические регрессионные модели для прогнозирования частоты острых респираторных инфекций в группе часто болеющих детей. При помощи ROC-анализа были оценены чувствительность и специфичность предикторов.

Окончательную математическую модель составили 3 предиктора, со значимостью $p \leq 0,05$, а именно: дисбиоз слизистых рото- и носоглотки 2 степени, аллергические заболевания респираторного тракта и гипоиmuноглобулинемия. Таким образом, для каждого из полученных предикторов ($x_{1,2,3}$) составлено уравнение регрессии, и вычислена вероятность зависимой переменной – индекса резистентности (ИР):

Так, например, для x_1 (дисбиоз 2 степени):

$$p = e^{-7,6+5,34/1 + e^{-7,6+5,34}} = 9,5/10,5 = 0,9 (90,0\%),$$

где p – теоретическая вероятность значений $ИР \geq 50\%$, то есть высокой частоты острых респираторных инфекционных заболеваний. Подобным образом рассчитываются теоретические значения вероятностей для всех предикторов, а также для возможных вариантов их взаимосочетаний. Экспоненты коэффициентов уравнения регрессии для предикторов интерпретированы как отношения шансов $OR = e^b$: дисбиоз 2 степени – OR 208 (ДИ 8,7; 4983), респираторные аллергические заболевания – OR 38 (ДИ 1,3; 1104), гипоиmuноглобулинемия – OR 14 (ДИ 1,3; 147). В проведенном исследовании показано, что ведущими предикторами высокой частоты ОРИ у часто болеющих

детей в возрасте до 3 лет являются: нарушение микробиоценоза верхних дыхательных путей, персистирующее аллергическое воспаление, дефицит гуморального звена иммунитета [5].

Таким образом, математически обоснована программа комплексного обследования часто болеющих детей раннего возраста, включающая оценку гуморального иммунитета и состава микрофлоры верхних дыхательных путей, с целью назначения иммунотропной и местной антибактериальной терапии.

В предсказательной аналитике существует ряд подходов, позволяющих провести предварительную обработку данных с целью улучшения работы классификаторов, а также решить ряд сопутствующих задач: исследовать значимость входных переменных и в той или иной форме проверить гипотезы о причинных связях между ними [6]. Одним из таких подходов является *оптимальное квантование* (Fine & Coarse Classing), или метод «начальных и конечных классов». Процедура формирования конечных классов представляет собой уменьшение числа значений исходного набора данных за счет их объединения в пределах некоторого интервала с использованием информации о целевой переменной. В результате такого преобразования число значений переменной должно уменьшиться без существенного ущерба для информативности данных [7].

Алгоритм квантования был использован в исследовательской работе студентов, посвященной анализу клинико-лабораторных предикторов коклюшной инфекции у госпитализированных детей. В ходе исследования эмпирические предикторы диагноза коклюшной инфекции были подвергнуты анализу, где каждому из них соответствовала бинарная выходная переменная (диагноз да/нет – *событие/не-событие*). В качестве клинико-эпидемических критериев диагноза было выбрано 9 признаков: – возраст (в годах), прививочный анамнез (привит/не привит), сведения о контакте с длительно кашляющим человеком (да/нет), длительность заболевания до поступления в стационар (в днях), особенности анамнеза до заболевания (частые ОРВИ, перинатальная отягощенность), лейкоцитоз более 15×10^9 лимфоцитарного характера при нормальной СОЭ. Учитывалась также тяжесть заболевания, наличие осложнений, число койко-дней госпитализации. Критерии были выбраны эмпирическим путем без предварительных расчетов.

Затем произвели разбиение всего диапазона изменения признака на несколько

начальных классов, для каждого из которых вычислили коэффициент WoE:

$$WoE_i = \ln(F^-/F^+),$$

где i – индекс начального класса;

F^- – относительная частота появления не событий в классе;

F^+ – относительная частота появления событий в классе.

На основе коэффициентов WoE вычисляется величина, определяющая значимость признака в модели бинарной классификации, называемая информационным индексом (вес доказательства, *informationvalue*, $I_{inf}V$) по формуле

$$I_{inf}V = \sum_{i=1}^k \times \{(N_i/N - P_i/P) \times WoE_i\}.$$

Информационный индекс всегда является положительной величиной, на его основе определяется значимость признака по следующей методике:

$I_{inf}V < 0,02$ – значимость отсутствует;

$0,02 \leq I_{inf}V < 0,1$ – значимость низкая;

$0,1 \leq I_{inf}V < 0,3$ – значимость средняя;

$I_{inf}V \geq 0,3$ – значимость высокая.

Методом WoE-анализа определена значимость эмпирически выбранных диагностических предикторов для обеих групп исследования: вычислены значения информационного индекса $I_{inf}V$ и определен их ранг.

Высокая значимость (ранг) информационного индекса ($I_{inf}V \geq 0,3$) получена для следующих диагностических предикторов:

отсутствие вакцинации против коклюша – $I_{inf}V = 3,44$ (1 ранговое место), в основной группе (ОГ) ни один ребенок не был привит против коклюша при 50,0% охвате вакцинопрофилактикой в группе сравнения (ГС), $p = 0,07$;

возраст до 1 года – $I_{inf}V = 2,93$ (2 ранговое место), средний возраст детей ОГ составил 1,14 (ДИ 0,96; 1,32), в ГС – 3 года (ДИ 1,87; 4,13), $p = 0,004$;

лейкоцитоз $Le > 15 \times 10^9$ – $I_{inf}V = 0,93$ (3 ранговое место), в общем анализе крови у детей ОГ лейкоцитоз выше 15×10^9 лимфоцитарного характера при нормальной СОЭ был у 64,2% детей против 18,7% случаев в ГС, $p = 0,02$;

сведения о контакте в анамнезе – $I_{inf}V = 0,4$ (4 ранговое место), контакт с длительно кашляющим лицом имел место у всех детей ОГ против 50,0% в ГС, $p = 0,07$.

Таким образом, в результате применения исследовательского алгоритма представлены следующие важные практические рекомендации: при отрицательном результате бактериологического/серологического

исследования и наличии типичной клинической картины коклюша можно использовать в качестве диагностических следующие признаки – отсутствие вакцинации против коклюша, возраст до 1 года, выраженный лейкоцитоз ($Le > 15 \times 10^9$) лимфоцитарного характера при нормальной СОЭ и контакт с длительно кашляющим человеком [8].

В заключение следует заметить, что примененный в учебно-исследовательской работе студентов метод математико-статистического моделирования является лишь малой долей того потенциала, который представляют современные инновационные технологии. Кроме того, очевидна необходимость подготовки специалистов умеющих не только проводить сложнейшие виды анализа, но, главное, интерпретировать результаты, формулировать их в доступной пониманию большинства форме и готовить на их основе клинические решения. Очевидно, что готовить таких специалистов должны профильные вузы, в данном случае медицинские.

Список литературы

1. Deductor. Руководство аналитика Версия 5.2 © 1995–2009. Компания BaseGroup™ Lab [Электронный ресурс]. – URL: <http://www.basegroup.ru/deductor/manual/guide-analyst-520>.
2. Гудинова Ж.В. О применении элементов Data Mining (обнаружения полезных знаний в базах данных) в гигиенических исследованиях и социально-гигиеническом мониторинге // Гигиена и санитария. – 2012. – № 5. – С. 78–81.
3. Карякина О.Е., Добродеева Л.К., Мартынова Н.А., Красильников С.В., Карякина Т.И. Применение математических моделей в клинической практике // Экология человека. – 2012. – С. 55–64.
4. Григорьев С.Г., Лобзин Ю.В., Скрипченко Н.В. Роль и место логистической регрессии и ROC-анализа в решении медицинских диагностических задач // Журнал инфектологии. – 2016. – № 8(4). – С. 36–45.
5. Дерюшева А.В., Львова И.И., Леготина Н.С. Обоснование программы лечебно-профилактических мероприятий у часто болеющих детей раннего возраста // Детские инфекции. – 2017. – № 1 (16). – С. 15–20.
6. Паклин Н.Б. Оптимальное квантование для повышения качества бинарных классификаторов // Искусственный интеллект. – 2013. – № 4. – С. 392–400.
7. Weed D.L. Weight of Evidence: A Review of Concept and Methods // Risk Analysis. – 2005. – Vol. 25(6). – P. 1545–1557.
8. Дерюшева А.В., Львова И.И., Леготина Н.С., Банько Н.В., Уварова Е.Ю. Предикторы клинико-эпидемического диагноза коклюша у детей // Пермский медицинский журнал. – 2017. – № 1(34). – С. 58–62.