УДК 517.977 DOI: 10.14529/ctcr200210

РАЗРАБОТКА МОДЕЛИ УПРАВЛЕНИЯ ПОТОКОМ ПАЦИЕНТОВ С СЕРДЕЧНО-СОСУДИСТЫМИ ЗАБОЛЕВАНИЯМИ МЕТОДАМИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

И.П. Болодурина[1], А.М. Назаров[2], Д.И. Кича[3], Л.С. Забродина[1], А.Ю. Жигалов[1]

[1] Оренбургский государственный университет, г. Оренбург, Россия,
[2] Оренбургская областная клиническая больница, г. Оренбург, Россия,
[3] Российский университет дружбы народов, г. Москва, Россия

Введение. В настоящее время развитие технологий Big Data и методов интеллектуального анализа больших данных открыло возможность исследования своевременности, доступности и эффективности проводимой терапии при обработке всей доступной информации о практике лечения. Методы персонифицированной и профилактической медицины, основанные на удаленном мониторинге пациентов и интеллектуальном анализе аналогичной практики лечения, приведут к существенному сокращению затрат и повышению качества жизни. Одними из эффективных методов исследования данных о пациентах и их электронных медицинских карт являются методы машинного обучения.

Цель исследования. Данное исследование направлено на построение модели управления потоком пациентов с сердечно-сосудистыми заболеваниями на основе анализа персонифицированных карт данных больных.

Материалы и методы. Определение прогноза на предмет обращения в поликлинику с заболеваниями сердца проведено методом логистической регрессии, алгоритмом построения деревьев решений ID3 и методом обучения ансамбля – случайные леса. В рамках экспериментального исследования проведена оценка эффективности применения рассмотренных методов для прогнозирования на основе анализа ROC-кривой и метрики AUC.

Результаты. Эксперименты на массиве электронных персонифицированных данных о медицинских услугах в территориальном фонде обязательного медицинского страхования (ТФОМС) и медицинском информационно-аналитическом центре г. Оренбурга показали, что для краткосрочного прогнозирования на 1 месяц более высокие результаты показал алгоритм ID3 построения решающих деревьев, а при увеличении рассматриваемого периода до 3 месяцев более эффективным оказался метод логистической регрессии.

Заключение. Предложенный подход к прогнозированию обращений пациентов позволяет повысить качество управления клинико-организационной системой здравоохранения при оказании медицинской помощи, а также спланировать объем и количество отдельных медицинских услуг.

Ключевые слова: логистическая регрессия, деревья решений, случайный лес, сердечно-сосудистые заболевания, алгоритмы обучения.

Введение

В настоящее время технологии Big Data стали одним из доминирующих направлений в развитии информационных технологий. Предполагается, что работа с колоссальными объемами неструктурированных данных окажет наибольшее влияние на производство [1], госуправление [2], торговлю [3] и медицину [4].

Благодаря методам интеллектуального анализа больших данных появилась возможность исследования эффективности проводимой терапии при обработке всей доступной информации о практике лечения [5]. На основе анализа известных историй болезни и диагностики в практику врачей входит широкое использование систем поддержки принятия решений, позволяющих предоставить доступ к опыту тысяч коллег по всей стране.

Методы персонифицированной и профилактической медицины, основанные на удаленном мониторинге пациентов, приведут к существенному сокращению затрат и повышению качества жизни [6]. Распространение различных сенсоров активностей человеческого организма, подключаемых к гаджетам, позволяет сократить необходимость проведения лабораторных исследова-

ний, предотвратить неожиданные осложнения, а автоматическое напоминание о необходимости проведения самостоятельных лечебно-профилактических манипуляций повысит качество назначенного лечения [7].

Одним из эффективных методов исследования данных о пациентах и их электронных медицинских карт являются методы машинного обучения. В последнее десятилетие из-за недоступности персональных данных о пациентах методы, основанные на интуиции и эвристике, были использованы для решения задач прогнозирования – постановки диагноза. В настоящее время каждый пациент занесен в базу данных поликлиники и имеет свою историю посещений, которая позволяет подсчитывать статистические характеристики, такие как среднее количество обращений с отдельными типами болезней, среднюю продолжительность пребывания в поликлинике, диагнозы и другие. Благодаря накопленному массиву данных врачи могут отойти от эвристического определения болезни и, используя опыт коллег, быстрее, своевременнее, а главное – более точно поставить правильный диагноз [8].

Подходы к анализу данных, основанные на машинном обучении, включают в себя несколько хорошо известных методов решения проблем, которые позволяют анализировать несбалансированные наборы данных для классификации и прогнозирования. К ним относят логистическую регрессию, скрытые марковские цепочки, деревья решений и случайные леса. Каждый из этих методов имеет свои преимущества перед аналогичными методами решения задач прогнозирования. Тем не менее стоит отметить, что сегодня в практике применения методов машинного обучения встречаются примеры эффективности различных подходов, поэтому перед постановкой прогноза необходимо сравнить эффективность алгоритмов для рассматриваемого набора данных.

1. Обзор исследований

Исследованиями медицинских данных с целью прогнозирования, классификации и автоматизации внутренних процессов занимаются по всему миру.

Авторский коллектив из Национального исследовательского университета Высшей школы экономики в публикации [9] рассматривает вопросы применения современных облачных технологий при хранении и обработке кардиологической информации. В частности, в работе [10] исследователя Е.Ю. Зиминой рассматриваются способы решения проблемы диагностики состояния здоровья сердца пациента с применением методов классификации Data Mining при обработке кардиологических данных. Кластерный анализ проводился на основе поиска схожих форм спектров Фурье, полученных путем моделирования работы сердца при использовании разложения Ферми – Пласта – Улана.

Авторы исследования [11] отмечают перспективность метода анализа больших данных (Big Data) при оценке качественных и количественных показателей фармакотерапии пациентов с артериальной гипертензией. В рамках публикации [12] выполнен обзор методов и систем интеллектуального анализа медицинских данных, а также предложена архитектура и программная платформа по анализу разнородных источников структурированных и неструктурированных данных.

Диссертационное исследование [13] И.В. Степаняна посвящено разработке теоретических и методических аспектов риск-менеджмента с применением биоинформационных технологий для прогнозирования нарушения здоровья работников. Для проведения кластерного анализа в работе использовалась бионическая самоорганизующаяся сеть Кохонена.

Авторы исследования [14] задаются вопросом применения методов машинного обучения для улучшения прогнозирования риска сердечно-сосудистых заболеваний, на основе обработки массивов клинических данных Clinical Practice Research Datalink (CPRD). Экспериментальные исследования показывают улучшение точности прогнозирования. Учёный Shankar M. Krishnan из Технологического института Уэнтворт (США) в работе [15] отмечает, что использование аналитики в сфере здравоохранения вместе с эффективной организацией, оптимизацией и анализом больших данных обеспечивает быстрое и точное диагностирование, а также снижение количества предотвратимых ошибок.

В публикации [16] производится оценка глобального управления рисками сердечно-сосудистых заболеваний в клинической практике среди врачей, разделённых на группы в соответствии с использованием обычной либо электронной поддержки для сбора и регистрации клинических данных.

Таким образом, обзор исследований показал, что использование технологий машинного обучения при обработке кардиологических данных с целью решения проблемы диагностики – один из наиболее актуальных вопросов в настоящий момент.

Данное исследование направлено на построение модели управления потоком пациентов с сердечно-сосудистыми заболеваниями (ССЗ) на основе прогноза обращений в ближайший месяц или 3 месяца за медицинской помощью в поликлинику по поводу ССЗ при анализе электронных персонифицированных карт пациентов.

Определение прогноза на предмет обращения в поликлинику с заболеваниями сердца осуществлено методом логистической регрессии, алгоритмом построения деревьев решений ID3 и методом обучения ансамбля – случайные леса. В рамках экспериментального исследования проведена оценка эффективности применения рассмотренных методов для прогнозирования на основе анализа ROC-кривой и метрики AUC.

2. Постановка задачи

Рассмотрим базу данных Территориального фонда ОМС по обращениям пациентов в поликлиники с ССЗ, которая содержит набор пациентов $U = \{u_1, u_2, ..., u_k\}$ и записей, характеризующих отдельно взятые посещения $C = \{c_1, c_2, ..., c_p\}$. Исходный набор данных представлен таблицей и каждый столбец соответствует одной из следующих характеристик записи посещения:

1. ID – код пациента;
2. MO – код медицинской организации;
3. CODE_MP – код медицинской помощи;
4. DATE_IN – дата прихода в медицинское учреждение;
5. DATE_OUT – дата выхода из медицинского учреждения;
6. DLITELN – длительность лечения;
7. MKB – код болезни по МКБ;
8. POL – 1 – муж., 2 – жен.;
9. VOZRAST – возраст текущий пациента;
10. ATER – обнаружен ли атеросклероз при приеме пациента;
11. ISHEM – обнаружена ли ишемия при приеме пациента;
12. GIPER – обнаружена ли гипертония при приеме пациента;
13. STENOK – обнаружена ли стенокардия при приеме пациента;
14. INF_MIOK – обнаружен ли инфаркт миокарда при приеме пациента.

На основе представленного исходного набора данных достаточно сложно оценивать прогноз прихода пациента с сердечно-сосудистыми заболеваниями через фиксированный промежуток времени и выбрать стратегии управления потоками пациентов. В связи с этим необходимо провести подготовку данных.

В необработанных данных, содержащихся в БД, сложно учитывать важный критерий для прогнозирования – время. Для его учета можно искусственно создать признаки, агрегирующие некоторые показатели пациентов за некоторый промежуток времени. Например, это может быть количество посещений за последние полгода. Признаки, идентифицирующие визит и сведения о пациенте, останутся неизменными. В настоящем исследовании собрана статистика по количеству посещений каждой болезни и классу болезни (по МКБ) за последние 3 и 6 месяцев.

Отметим, что для более точного прогноза нужно исключить пациентов с отсутствующей электронной медицинской картой и пациентов, для которых невозможно проверить прогноз. В связи с этим информация о первых и последних записях в БД (с интервалом времени 6 мес.) не включена в исходный набор.

Таким образом, в результате статистической обработки данных выделены следующие дополнительные признаки для прогнозирования обращений пациентов в поликлинику с заболеваниями сердца:

1. МКБ – код болезни по МКБ-10 не детализированный;
2. COUNT_CODE_MP – количества посещений по каждому типу учреждения за последние полгода;

3. BOOL_CODE_MP – было ли посещение по каждому типу учреждения за последние 3 месяца;

4. ATER_6M – сколько раз обнаружен атеросклероз при приеме пациента за последние 6 месяцев;

5. ATER_3M – был ли обнаружен атеросклероз при приеме пациента за последние 3 месяца;

6. ISHEM_6M – сколько раз обнаружена ишемия при приеме пациента за последние 6 месяцев;

7. ISHEM_3M – была ли обнаружена ишемия при приеме пациента за последние 3 месяца;

8. GIPER_6M – сколько раз обнаружена гипертония при приеме пациента за последние 6 месяцев;

9. GIPER_3M – была ли обнаружена гипертония при приеме пациента за последние 3 месяца;

10. STENOK_6M – сколько раз обнаружена стенокардия при приеме пациента за последние 6 месяцев;

11. STENOK_3M – была ли обнаружена стенокардия при приеме пациента за последние 3 месяца;

12. INF_MIOK_6M – сколько раз обнаружен инфаркт миокарда при приеме пациента за последние 6 месяцев;

13. INF_MIOK_3M – был ли обнаружен инфаркт миокарда при приеме пациента за последние 3 месяца;

14. MKB_CLASS_COUNT_6M – сумма количеств обращений с разными классами болезней по МКБ за последние 6 месяцев;

15. IS_AGAIN_1M – придет ли пациент в ближайший месяц после DATE_OUT или нет;

16. MKB_CLASS_BOOL_3M – было ли обращение с разными классами болезней по МКБ за последние 3 месяца;

17. MKB_CLASS_BOOL_1M – было ли обращение с разными классами болезней по МКБ за последний месяц.

Поле MKB_CLASS_BOOL_1M и MKB_CLASS_BOOL_3M отвечает за повторное посещение пациента, при 0 – посещения нет, при 1 – посещение есть.

Данные поля получены простым агрегированием с кодированием one-hot и bug-of-words.

Данное исследование направлено на построение прогноза для пациента на предмет того, обратится ли он в ближайший месяц или 3 месяца в поликлинику с сердечно-сосудистыми заболеваниями на основе анализа персонифицированных карт. В связи с этим выходным полем для обучения классификатора является и MKB_HEART_BOOL_1M, и MKB_HEART_BOOL_3M.

Пусть функция $Y = f(C_i)$ описывает определенный классификатор, который получает вектор характеристик посещений C_i пациента u_i.

Функция $f(C_i)$ определяет некоторое значение $Y \in \{0;1\}$, обратится ли пациент в ближайшие 3 месяца в поликлинику с сердечно-сосудистыми заболеваниями.

При определенных условиях значение Y может быть преобразовано в вероятность того, что пациент обратится в поликлинику. В этом случае действует условие монотонности, означающее, что чем выше значение Y, тем выше вероятность его прихода. Необходимо найти такие параметры, при которых классификатор будет давать наилучшие вероятностные оценки с точки зрения выбранных метрик.

В результате получаем задачу прогнозирования, которую будем решать с помощью метода логистической регрессии, алгоритма построения деревьев решений ID3 и метода обучения ансамбля – случайные леса.

3. Интеллектуальные методы анализа данных

Метод логистической регрессии

Логистическая регрессия – это тип обобщенной линейной модели (GLM), которая использует логистическую функцию для прогнозирования бинарной характеристики на основе любого вида независимых входных параметров.

Коэффициенты алгоритма логистической регрессии должны оцениваться на основе обучающей выборки с использованием метода оценки максимального правдоподобия, который является

самым распространенным алгоритмом обучения, используемым различными алгоритмами машинного обучения.

Основная идея метода максимального правдоподобия для логистической регрессии состоит в том, что алгоритм ищет значения для коэффициентов логистической функции, которые сводят к минимуму ошибку в вероятностях, прогнозируемых моделью, по значениям в данных.

Алгоритм построения решающего дерева ID3

Алгоритм ID3 строит дерево решений по принципу сверху вниз. В алгоритме ID3 реализована одна из разновидностей «жадного» поиска в пространстве всех возможных деревьев: он добавляет поддерево к текущему дереву и продолжает поиск, не делая возвратов. Благодаря такому подходу алгоритм становится очень эффективным. При этом, однако, он сильно зависит от процедуры выбора очередного свойства для тестирования.

Можно считать, что каждое свойство объекта вносит в решение задачи классификации какой-то объем новой информации и сокращает неопределенность.

В общем случае в теории информации энтропия вычисляется по формуле

$$I(P) = -\sum_{i=1}^{n} p_i \log p_i. \quad (1)$$

Алгоритм ID3 проводит выбор определенного свойства на роль корня текущего поддерева, основываясь на количестве информации, получаемой в результате его проверки: корнем поддерева выбирается то свойство, которое дает при проверке наибольшую информацию (больше всего сокращает неопределенность).

Случайный лес (Random Forest)

Случайный лес – это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме.

- Выбирается подвыборка обучающей выборки размера samplesize – по ней строится дерево (для каждого дерева – своя подвыборка).
- Для построения каждого расщепления в дереве просматриваем max_features случайных признаков (для каждого нового расщепления – свои случайные признаки).
- Выбираем наилучший признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Ясно, что такая схема построения соответствует главному принципу ансамблирования (построению алгоритма машинного обучения на базе нескольких, в данном случае решающих деревьев): базовые алгоритмы должны быть разнообразными (поэтому каждое дерево строится на своей обучающей выборке и при выборе расщеплений присутствует элемент случайности).

4. Вычислительные эксперименты

Вычислительные эксперименты, выполненные в работе, проводились на массиве электронных персонифицированных данных о медицинских услугах в Территориальном фонде обязательного медицинского страхования (ТФОМС) и медицинском информационно-аналитическом центре г. Оренбурга.

Набор данных содержит информацию о посещениях пациентов, дополненную статистическими характеристиками, определенными выше, с отметкой о том, обратится ли пациент в медицинское учреждение с сердечно-сосудистыми заболеваниями в течение одного или трех месяцев.

Анализ решающего дерева

Стоит отметить, что преимуществом алгоритма построения решающего дерева ID3 является простота представления и интерпретируемость результатов. В связи с этим для подтверждения того, что построенный прогноз может соответствует реальному анамнезу, проанализируем построенное дерево решений (рис. 1), выделив основные правила.

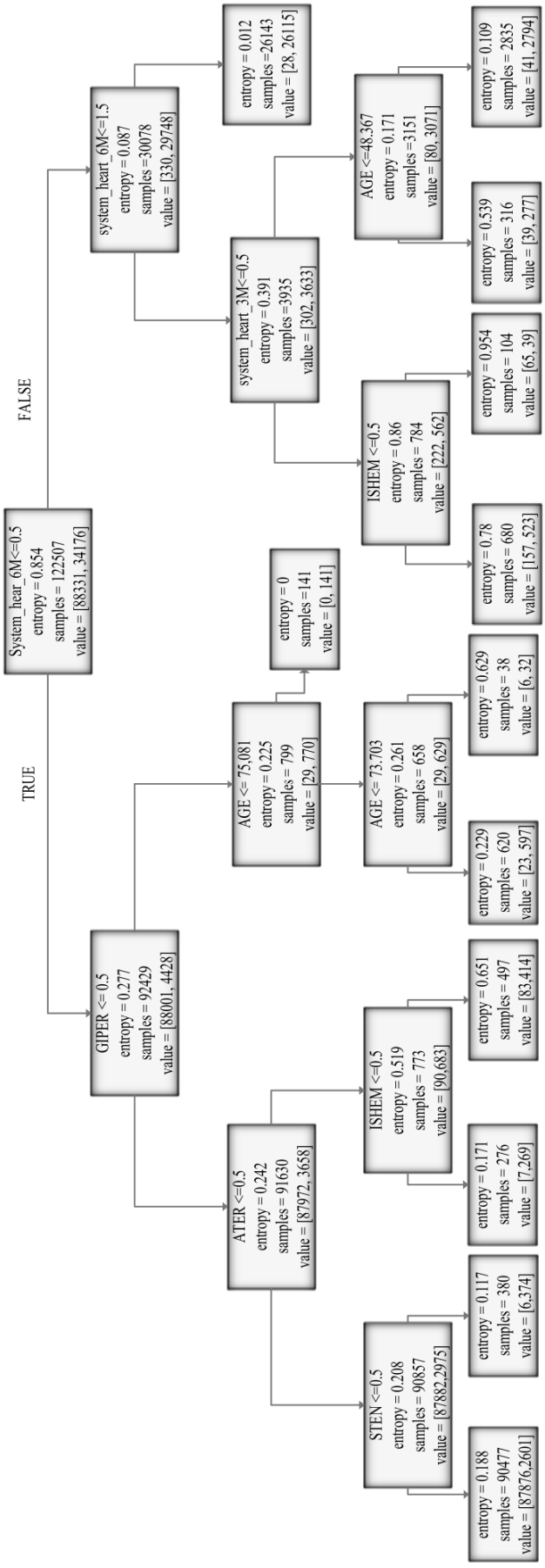

Рис. 1. Дерево решений для прогноза MKB_HEART_BOOL_3M на обучающем множестве
Fig. 1. MKB_HEART_BOOL_3M Prediction Solutions Tree on the Learning Set

В соответствии с построенным деревом решений можно выделить наиболее важные обобщающие правила:

ЕСЛИ пациент не обращался последние 6 мес. с ССЗ
ТО
  ЕСЛИ пациент не имеет гипертонии, атеросклероза и стенокардии
  ТО не придет в ближайшие 3 мес. с ССЗ
  ИНАЧЕ
    ЕСЛИ пациент имеет гипертонию / атеросклероз / стенокардию
    ТО придет ближайшие 3 мес. с ССЗ
ИНАЧЕ
ЕСЛИ пациент приходил последние 6 мес. больше 1 раза с ССЗ
  ЕСЛИ пациент не приходил последние 3 мес. с ССЗ
  ТО
    ЕСЛИ пациент имеет ишемию
    ТО придет ближайшие 3 мес. с ССЗ
    ИНАЧЕ не придет ближайшие 3 мес. с ССЗ
  ИНАЧЕ
    ЕСЛИ возраст больше 48
    ТО придет ближайшие 3 мес. с ССЗ
    ИНАЧЕ не придет ближайшие 3 мес. с ССЗ
ИНАЧЕ пациент придет ближайшие 3 мес. с ССЗ.

В связи с тем, что построенное дерево решений с достаточной точностью проводит классификацию обучающего множества по выходному признаку, то можно говорить об адекватности разработаннной модели.

Сравнительный анализ эффективности алгоритмов

В рамках данного исследования проведена оценка эффективности применения рассмотренных методов для прогнозирования посещений пациентов на основе анализа ROC-кривой и метрики AUC.

Прогноз строился отдельно для обращений в ближайший месяц (рис. 2) и 3 месяца (рис. 3) в поликлинику с сердечно-сосудистыми заболеваниями на основе анализа электронных персонифицированных карт.

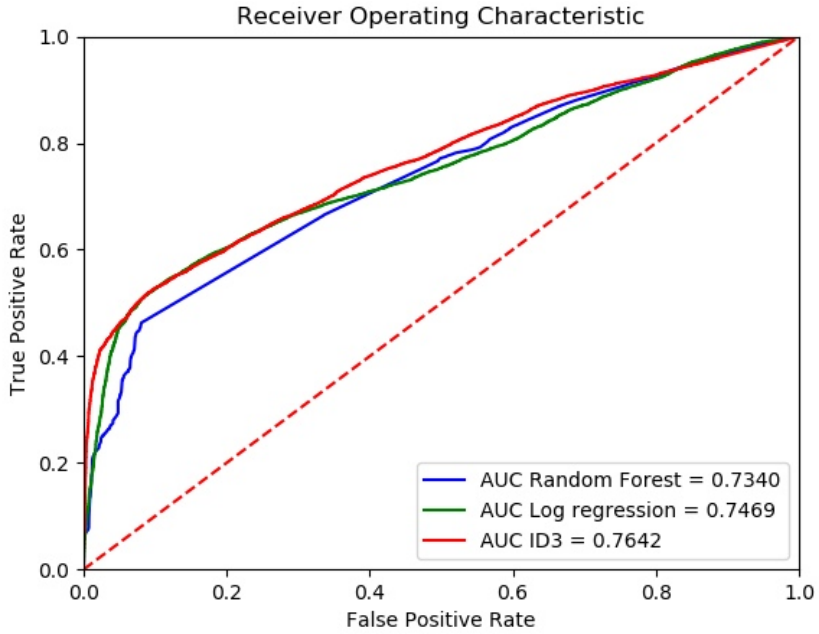

Рис. 2. ROC-кривая для MKB_HEART_BOOL_1M
Fig. 2. ROC curve for MKB_HEART_BOOL_1M

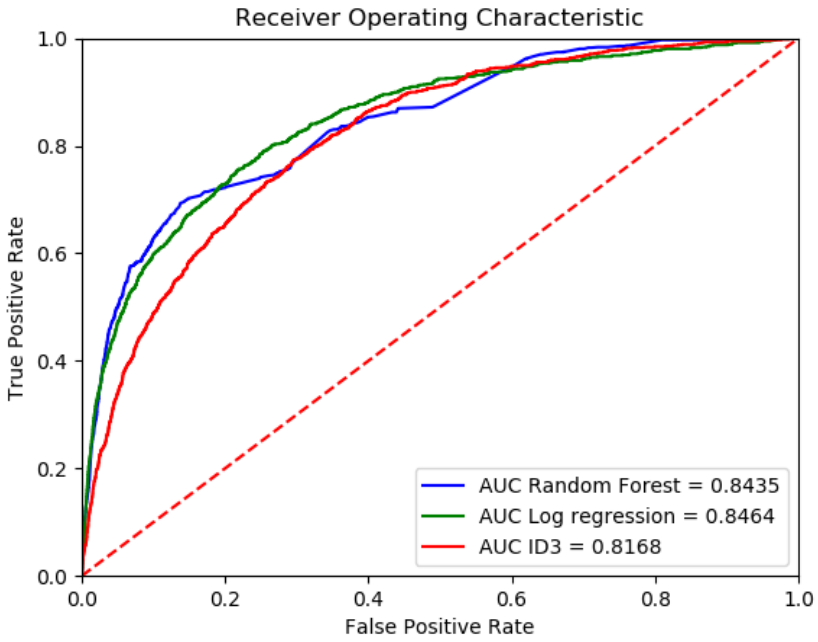

Рис. 3. ROC-кривая для MKB_HEART_BOOL_3M
Fig. 3. ROC curve for MKB_HEART_BOOL_3M

Стоит отметить, что, согласно метрике AUC, алгоритм ID3 показал лучшие результаты прогнозирования (AUC ID3 = 0,7642) на тестовом множестве для анализа обращений в ближайший месяц. Однако для прогнозирования на более долгий срок (3 месяца) более высокую точность показал метод построения логистической регрессии (AUC Log. Regression = 0,8464).

Кроме того, обратим внимание, что точность прогноза посещений на ближайший месяц (0,75) ниже точности на ближайшие 3 месяца (0,84). Это связанно с тем, что обострение сердечно-сосудистых заболеваний происходит достаточно редко при регулярных обследованиях пациентов, и при увеличении рассматриваемого периода для прогноза точность должна увеличиваться.

Заключение

В рамках данного исследования построена модель управления потоком пациентов с сердечно-сосудистыми заболеваниями на основе прогноза обращения в ближайший месяц или 3 месяца за медицинской помощью в поликлинику по поводу ССЗ при анализе электронных персонифицированных карт пациентов.

Определение прогноза осуществлено логистической регрессией, алгоритмом построения деревьев решений ID3 и методом обучения ансамбля – случайные леса.

Построенные модели показали хороший результат, так как имели высокую обобщающую способность и точность. В рамках экспериментального исследования проведена оценка эффективности применения рассмотренных методов для прогнозирования обращений пациентов в поликлиники на основе анализа ROC-кривой и метрики AUC.

Каждый из рассмотренных методов имеет свои преимущества перед аналогичными методами решения задач прогнозирования. Тем не менее стоит отметить, что для короткого временного периода прогнозирования (1 месяц) более высокие результаты показал алгоритм ID3 построения решающих деревьев, а при увеличении рассматриваемого периода до 3 месяцев наилучшие результаты показал метод логистической регрессии.

Предложенный подход к прогнозированию обращений пациентов позволяет повысить качество управления клинико-организационной системой здравоохранения при оказании медицинской помощи, а также спланировать объем и количество отдельных медицинских услуг.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-07-01065, а также гранта Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации (НШ-2502.2020.9).

Литература

1. Windmann, S. Big Data Analysis of Manufacturing Processes / S. Windmann, A. Maier, O. Niggemann, C. Frey // Journal of Physics: Conference Series. – 2015. – DOI: 10.1088/1742-6596/659/1/012055

2. Kim, G. Big Data Applications in the Government Sector: A Comparative Analysis among Leading Countries / G. Kim, J. Chung // Communications of the ACM. – 2014. – Vol. 57. – P. 78–85. DOI: 10.1145/2500873

3. Alshura, M. Big Data in Marketing Arena. Big Opportunity, Big Challenge, and Research Trends: An Integrated View / M. Alshura, A. Zabadi, M. Abughazaleh // MANAGEMENT AND ECONOMICS REVIEW. – 2018. – Vol. 3. DOI: 10.24818/mer/2018.06-06

4. Beam, A. Big Data and Machine Learning in Health Care / A. Beam, I. Kohane // JAMA. – 2018. – Vol. 319. DOI: 10.1001/jama.2017.18391

5. Patient-Level Effectiveness Prediction Modeling for Glioblastoma Using Classification Trees / T. Geldof, N. Damme, I. Huys, W. Dyck // Frontiers in Pharmacology. – 2020. – Vol. 10. DOI: 10.3389/fphar.2019.01665

6. Blasiak, A. CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence / A. Blasiak, J. Khong, T. Kee // SLAS TECHNOLOGY: Translating Life Sciences Innovation. – 2019. – DOI: 10.1177/2472630319890316

7. Lin, R. Diseases and Health Monitoring Big Data: A Survey. / R. Lin, Z. Chronic, H. Wang // IEEE Reviews in Biomedical Engineering. – 2018. DOI: 10.1109/RBME.2018.2829704

8. Leung, K. Application of Big Data in Decision Making for Emergency Healthcare Management / K. Leung, A. Stevenson // International Journal of Re-search and Engineering. – 2018. – Vol. 5. – P. 311–314. DOI: 10.21276/ijre.2018.5.2.2

9. Зимина, Е.Ю. Применение облачных технологий в задачах математического анализа кардиологической информации / Е.Ю. Зимина, М.А. Новопашин, А.В. Шмид // Сборник трудов IV международной конференции и молодежной школы «Информационные технологии и нанотехнологии» (ИТНТ-2018). – Самара, 2018. – С. 2282–2287.

10. Зимина, Е.Ю. Кластерный анализ кардиологических данных / Е.Ю. Зимина // Статистика и Экономика. – 2018. – Т. 15, № 2. – С. 30–37. DOI: 10.21686/2500-3925-2018-2-30-37

11. Бурыкин, И.М. Перспективность метода анализа больших данных (big data) для оценки качества и эффективности фармакотерапии пациентов с артериальной гипертензией / И.М. Бурыкин, Г.Н. Алеева, Р.Х. Хафизьянова // Современные технологии в медицине. – 2017. – № 9. – С. 194. DOI: 10.17691/stm2017.9.4.24

12. Баранов, А.А. Методы и средства комплексного интеллектуального анализа медицинских данных / А.А. Баранов // Труды Института системного анализа Российской академии наук. – 2015. – Т. 65, № 2. – С. 81–93.

13. Степанян, И.В. Научно-методические основы и биоинформационные технологии управления профессиональными рисками в медицине труда: дис. … д-ра биол. наук / И.В. Степанян. – М., 2012.

14. Weng, S. Can Machine-learning improve cardiovascular risk prediction using routine clinical data / S. Weng, J. Reps, J. Kai. // PLoS ONE. – 2017. – Vol. 12. DOI: 10.1371/journal.pone.0174944

15. Krishnan, S. Application of Analytics to Big Data in Healthcare / S. Krishnan // 32nd Southern Biomedical Engineering Conference. – 2016. – P. 156–157. DOI: 10.1109/SBEC.2016.88

16. Tocci, G. Use of Electronic Support for Implementing Global Cardiovascular Risk Management / G. Tocci, A. Ferrucci, P. Guida // High HEART Press Cardiovasc Prev. – 2010. – Vol. 17. – P. 37–47. DOI: 10.2165/11311750-000000000-00000

Болодурина Ирина Павловна, д-р техн. наук, профессор, заведующий кафедрой прикладной математики, Оренбургский государственный университет, г. Оренбург; prmat@mail.osu.ru.

Назаров Александр Михайлович, канд. мед. наук, заведующий отделением реанимации и интенсивной терапии, Оренбургская областная клиническая больница, г. Оренбург; ookbmedis@mail.ru.

Кича Дмитрий Иванович, д-р мед. наук, профессор, заведующий кафедрой организации здравоохранения, лекарственного обеспечения, медицинских технологий и гигиены, Российский университет дружбы народов, г. Москва; kichad@yandex.ru.

Забродина Любовь Сергеевна, ассистент кафедры прикладной математики, Оренбургский государственный университет, г. Оренбург; zabrodina97@inbox.ru.

Жигалов Артур Юрьевич, ведущий программист, Оренбургский государственный университет, г. Оренбург; leroy137.artur@gmail.com.

Поступила в редакцию 27 февраля 2020 г.

DOI: 10.14529/ctcr200210

DEVELOPMENT OF A MODEL FOR CONTROL THE FLOW OF PATIENTS WITH CARDIOVASCULAR DISEASES USING DATA MINING METHODS

I.P. Bolodurina*[1]*, prmat@mail.osu.ru,
A.M. Nazarov*[2]*, ookbmedis@mail.ru,
D.I. Kicha*[3]*, kichad@yandex.ru,
L.S. Zabrodina*[1]*, zabrodina97@inbox.ru,
A.Yu. Zhigalov*[1]*, leroy137.artur@gmail.com

[1] Orenburg State University, Orenburg, Russian Federation,
[2] Orenburg Regional Clinical Hospital, Orenburg, Russian Federation,
[3] Peoples' Friendship University of Russia, Moscow, Russian Federation

Introduction. Currently, the development of Big Data technologies and methods of big data mining has opened up the possibility of investigating the timeliness, availability and effectiveness of therapy when processing all available information about the treatment practice. Personalized and preventive medicine methods based on remote monitoring of patients and intelligent analysis of similar treatment practices will lead to significant cost savings and improved quality of life. One of the most effective methods of studying patient data and their electronic medical records is machine learning methods.

Aim. This study is aimed at building a model for managing the flow of patients with cardiovascular diseases based on the analysis of personalized patient data maps.

Materials and methods. The forecast for treatment of patients with heart diseases was determined using the method of logistic regression, the algorithm for building ID3 decision trees, and the method of training the ensemble – random forests. As part of the experimental study, the effectiveness of the methods considered for forecasting was evaluated based on the analysis of the ROC curve and the AUC metric.

Results. Experiments on an array of electronic personalized data about medical services in the territorial Fund of compulsory medical insurance (TFOMS) and the medical information and analytical center of Orenburg showed that for short-term forecasting for 1 month, the ID3 algorithm for constructing decision trees showed better results, and when the period under consideration was increased to 3 months, the method of logistic regression was more effective.

Conclusion. The proposed approach to predicting patient requests allows us to improve the quality of management of the clinical and organizational health care system in the provision of medical care, as well as to plan the volume and number of individual medical services.

Keywords: logistic regression, decision trees, random forest, cardiovascular diseases, learning algorithms.

References

1. Windmann S., Maier A., Niggemann O., Frey C. Big Data Analysis of Manufacturing Processes. *Journal of Physics: Conference Series*, 2015. DOI: 10.1088/1742-6596/659/1/012055

2. Kim G., Chung J. Big Data Applications in the Government Sector: A Comparative Analysis among Leading Countries. *Communications of the ACM*, 2014, vol. 57, pp. 78–85. DOI: 10.1145/2500873

3. Alshura M., Zabadi A., Abughazaleh M. Big Data in Marketing Arena. Big Opportunity, Big Challenge, and Research Trends: An Integrated View. *Management and economics review*, 2018, vol. 3, pp. 75–84. DOI: 10.24818/mer/2018.06-06

4. Beam A., Kohane I. Big Data and Machine Learning in Health Care. *JAMA*, 2018, vol. 319, pp. 1317–1318. DOI: 10.1001/jama.2017.18391

5. Geldof T., Damme N., Huys I., Dyck W. Patient-Level Effectiveness Prediction Modeling for Glioblastoma Using Classification Trees. *Frontiers in Pharmacology*, 2020, vol. 10. DOI: 10.3389/fphar.2019.01665

6. Blasiak A., Khong J., Kee T. CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence. *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, 2019, vol. 25, pp. 95–105. DOI: 10.1177/2472630319890316

7. Lin R., Chronic Z., Wang H. Diseases and Health Monitoring Big Data: A Survey. *IEEE Reviews in Biomedical Engineering*, 2018, vol. 11, pp. 275–288. DOI: 10.1109/RBME.2018.2829704

8. Leung K., Stevenson A. Application of Big Data in Decision Making for Emergency Healthcare Management. *International Journal of Re-search and Engineering*, 2018, vol. 5, pp. 311–314. DOI: 10.21276/ijre.2018.5.2.2

9. Zimina E., Novopashin M., Shmid A. Cloud technologies in the problems of mathematical analysis of cardiological information. IV International Conference on "Information Technology and Nanotechnology" (ITNT-2018), 2018, pp. 112–118. DOI: 10.18287/1613-0073-2018-2212-112-118

10. Zimina E.Yu. A Cluster Analysis of Cardiac Data. *Statistics and Economics*, 2018, vol. 15, no. 2, pp. 30–37. (in Russ.) DOI: 10.21686/2500-3925-2018-2-30-37

11. Burykin I.M., Aleeva G.N., Hafiz'yanova R.H. [Prospects of the Big Data Analysis Method for Evaluating the Quality and Effectiveness of Pharmacotherapy in Patients with Arterial Hypertension]. *Modern Technologies in Medicine*, 2017, vol. 9, no. 4, pp. 194–200. (in Russ.) DOI: 10.17691/stm2017.9.4.24

12. Baranov A.A. [Methods and Tools for Complex Medical Data Mining]. *Proc. of the Institute of System Analysis of the Russian Academy of Sciences*, 2015, vol. 65, no. 2, pp. 81–93. (in Russ.)

13. Stepanyan I.V. *Nauchno-metodicheskie osnovy i bioinformatsionnye tekhnologii upravleniya professional'nymi riskami v meditsine truda. Dis. doct. biol. nauk* [Scientific and Methodological Bases and Bioinformatic Technologies of Occupational Risk Management in Occupational Medicine. Thesis of Doct. of Biol. Sc.]. Moscow, 2012. 240 p.

14. Weng S., Reps J., Kai J. Can Machine-learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data. *PLoS ONE*, 2017, vol. 12. DOI: 10.1371/journal.pone.0174944

15. Krishnan, S. Application of Analytics to Big Data in Healthcare, *32nd Southern Biomedical Engineering Conference*, 2016, pp. 156–157. DOI: 10.1109/SBEC.2016.88

16. Tocci G., Ferrucci A., Guida P. Use of Electronic Support for Implementing Global Cardiovascular Risk Management. *High HEART Press Cardiovasc Prev.*, 2010, vol. 17, pp. 37–47. DOI: 10.2165/11311750-000000000-00000

Received 27 February 2020

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Разработка модели управления потоком пациентов с сердечно-сосудистыми заболеваниями методами интеллектуального анализа данных / И.П. Болодурина, А.М. Назаров, Д.И. Кича и др. // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2020. – Т. 20, № 2. – С. 105–115. DOI: 10.14529/ctcr200210

FOR CITATION

Bolodurina I.P., Nazarov A.M., Kicha D.I., Zabrodina L.S., Zhigalov A.Yu. Development of a Model for Control the Flow of Patients with Cardiovascular Diseases Using Data Mining Methods. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2020, vol. 20, no. 2, pp. 105–115. (in Russ.) DOI: 10.14529/ctcr200210