

Оценка текущего состояния беременной на различных сроках гестации

В. М. Гушин

Кафедра биотехнических систем

*Санкт-Петербургский государственный
электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)*

vmguschin@stud.etu.ru

И. П. Корнеева

Кафедра биотехнических систем

*Санкт-Петербургский государственный
электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)*

korneeva.inkorn@gmail.com

Аннотация. В работе рассматривается вопрос комплексной оценки состояния здоровья беременных пациенток в рамках амбулаторного или удаленного мониторинга. Реализация алгоритмов классификации текущего состояния построена на основе различных модификаций деревьев решений с учетом срока гестации.

Ключевые слова: телемедицина; удаленный мониторинг состояния здоровья; анализ данных

I. ВВЕДЕНИЕ

Разработки в сфере мониторинговых систем для нужд здравоохранения увеличиваются количественно и качественно наряду с развитием технологий. Одной из важнейших задач в этой области на сегодняшний день является мониторинг состояния здоровья беременных женщин с целью обеспечения их безопасности вплоть до родов. В том числе и удаленный мониторинг, построенный на принципах телемедицины, который становится предпочтительнее во многих ситуациях, когда женщинам становится все труднее регулярно посещать лечебное учреждение, а риски увеличиваются и необходимо быстро реагировать на возможное изменение состояния здоровья. Однако количество анализов и соответствующих показателей увеличивается, а средства для их обработки и интерпретации, которые могли бы оптимизировать работу врача, становятся все более востребованными. Кроме того, как правило, в таких исследованиях не учитываются определенные факторы, такие как сроки беременности, влияющие на референтные значения показателей, что в некоторых ситуациях может сыграть решающую роль [1].

Развитие аппаратной части мониторинговых систем позволило собирать большое количество различного вида данных о пациентах. Такими же стремительными темпами происходит развитие методов для обработки этих данных, поэтому значительно чаще в медицине стали использовать современные средства интеллектуальной обработки данных, к которым относятся методы машинного обучения, нейросети и другие подобные модели. Медицинские данные могут быть сложными и содержать множество параметров, которые необходимо учитывать при анализе. Средства интеллектуальной обработки данных помогают автоматизировать процесс анализа данных, что может ускорить процесс принятия решений и улучшить качество медицинской помощи.

Машинное обучение (МО) активно используется в медицине для различных задач по работе с большим количеством данных. К ним относится диагностика заболеваний, анализ медицинских изображений, лабораторные данные и результатов клинических тестов для раннего выявления патологий, а также прогнозирование исходов лечения (в том числе исходов беременности). Кроме того, с помощью методов машинного обучения можно провести анализ генетических данных, анамнез и других факторов для разработки индивидуальных планов лечения, что повышает эффективность терапии и снижает риск побочных эффектов. МО позволяет анализировать большие объемы данных из электронных медицинских карт, генетических исследований и других источников для выявления скрытых закономерностей и тенденций. Помимо прочего использование методов машинного обучения часто необходимо для автоматизации рутинных задач, чтобы направить возможности медицинского персонала на решение более важных вопросов.

Таким образом, учитывая вышеуказанные условия, можно сделать вывод о том, что для решения задачи комплексной оценки состояния требуется аналитическая система, учитывающая также срок гестации для повышения точности анализа, которая может быть интегрирована в мониторинговую систему. Такая система в будущем может стать основой для системы поддержки принятия решения врача в клинической практике, а также для прогностической модели для предупреждения возможных патологий на ранних этапах.

Целью данного исследования является разработка системы для оценки текущего состояния беременной на основе комплекса показателей, полученных в результате мониторинга, на различных сроках гестации.

Для достижения цели исследования необходимо выполнить следующие задачи:

- выбрать модель для обработки данных;
- выполнить подготовку данных для обучения модели и ее дальнейшего тестирования;
- обучить и протестировать модель;
- оценить работу модели.

II. МАТЕРИАЛЫ И МЕТОДЫ

Для анализа данных, полученных в ходе мониторинга, предлагается использовать методы машинного обучения. Основной целью анализа является классификация данных, то есть их разделение на группы. В большинстве диагностических исследований в области медицины, данные классифицируются на два основных класса: «Больные» и «Здоровые». Однако в зависимости от специфики заболевания могут использоваться и другие категории. В случае, когда речь идет о беременных, корректнее выделять классы «Норма» и «Патология». Кроме того, предлагается расширить диагностический потенциал системы и выделить дополнительный класс «Внимание», так как важно обнаружить ранние нарушения в динамике показателей [2].

А. Описание метода

Для стандартной бинарной классификации эффективным инструментом является дерево решений, представляющее собой систему алгоритмов, которая разделяет информацию на основе заранее определенных признаков, выявленных в обучающей выборке. Дерево решений представляет собой структуру, начинающуюся с корневого узла, который соответствует начальному состоянию алгоритма, и продолжающуюся через промежуточные узлы до конечных узлов (листьев), представляющих собой результаты классификации для каждой ветви. В каждом узле, на основе обучающей выборки и соответствующих меток классов, осуществляется рекурсивное разделение пространства объектов таким образом, чтобы выборки с одинаковыми или схожими метками концентрировались в одной группе. Затем, с использованием функции потерь, происходит расчет и минимизация ошибки классификации. В результате формируется правило, определяющее направление движения данных из тестовой выборки к определенному узлу, что позволяет выявить все возможные варианты классификации анализируемой информации.

В качестве функции потерь в данном случае принята информационная энтропия. Этот критерий рассчитывается на основании логарифма вероятности принадлежности одному из классов (логарифмические потери). Таким образом, задача сводится к минимизации логарифмических потерь, иначе говоря, достижении однозначной принадлежности отдельно взятой выборки данных к одному из классов в каждом узле принятия решения при построении дерева.

Так как данные должны разделяться по трем классам, обучение проводится в два этапа. Сначала происходит выделение представителей класса «Норма», после чего применимо к оставшейся выборке формируются решающие правила для разделения данных на классы «Внимание» и «Патология». Чем равномернее распределение представителей обучающей выборки по классам, тем точнее проводится классификация.

В. Виды деревьев решений

Существует несколько основных типов деревьев решений, которые различаются в зависимости от используемого алгоритма обучения [3], однако есть также некоторые аналоги, адаптированные под конкретные задачи, которые отличаются большей

точностью, но и большей специфичностью. Среди основных выделяют:

- **ID3.** Алгоритм позволяет работать только с дискретной целевой переменной. Число потомков в узле неограниченно. Алгоритм не работает с пропущенными данными.
- **C4.5.** Алгоритм, который является преемником ID3 и преобразует обученные деревья в наборы правил, строящихся на условных конструкциях. Затем оценивается точность каждого правила, чтобы определить порядок, в котором они должны применяться.
- **C5.0.** Алгоритм, использующий меньше памяти и создающий меньшие наборы правил, чем C4.5, будучи при этом более точным.
- **CART.** Алгоритм решает задачи классификации и регрессии, так как позволяет использовать дискретную и непрерывную целевые переменные. CART строит деревья, в каждом узле которых только два потомка.

Почти все представленные методы подходят для выполнения задач, поставленных в рамках данного исследования, однако на практике было установлено, что наиболее универсальным, простым и точным является метод CART, поэтому предпочтение было отдано ему.

С. Специфика исследования

В данной работе был выделен основополагающий фактор – срок беременности. Клинические исследования и статистика за многие годы позволили выявить определенную закономерность динамики показателей лабораторного и инструментального анализов от времени у беременной. Принято разделять весь срок гестации на три примерно равных промежутка, в течение которых меняется не только сам плод, но и мать, организм которой испытывает серьезные нагрузки и постепенно готовится к родам [4]. Таким образом, хотя нормы диагностически значимых показателей меняются почти на всем протяжении беременности и, как правило, носят индивидуальный характер, прежде всего выводы делаются на основе статистики для текущего триместра.

Отобранная модель дерева принятия решения изначально не учитывает такой параметр, как «срок беременности», поэтому его необходимо интегрировать в алгоритм при расчете решающих правил на этапе обучения модели. Учитывая, что модель способна выявлять скрытые паттерны, определение специфических характеристик и закономерностей можно заложить еще на этапе разметки данных, однако при использовании выборок малой размерности в процессе обучения наиболее вероятно расширение нормального диапазона, что может привести к ошибкам классификации.

Более выгодным с позиции повышения точности классификации является другой подход. Вместо использования одного дерева принятия решений предлагается обучить сразу три модели на данных, отобранных по каждому триместру. В случае использования ограниченного набора данных при обучении все еще есть вероятность возникновения ошибок на этапе апробации модели. Тем не менее, при таком подходе, модели не зависят друг от друга, что значительно снижает риски выбросов и позволяет

проводить более тонкую настройку системы в случае необходимости.

III. РЕЗУЛЬТАТЫ

Данные для проведения исследования были собраны и добровольно предоставлены самими пациентками, проходившими регулярное обследование, и представляют собой три выборки значений диагностически значимых показателей, среди которых результаты лабораторного анализа крови и мочи, пульс, температура тела, артериальное давление, а также информация о наличии осложнений (боль, тошнота, признаки ОРВИ).

Для того, чтобы начать работать с деревьями решений, необходимо было провести предварительную подготовку данных, которая включала в себя:

- Формирование единого шаблона для массива данных.
- Удаление пропусков и выбросов или замена на наиболее подходящие значения.
- Объединение в один массив.

После того, как все данные были обработаны и объединены, необходимо было провести разметку. Все представленные значения сравнивались с референтными и определялись в одну из трех категорий. Если значение

немного отклонялось от нормы (не более 20 %), оно относилось к категории класса «Внимание», что означало бы необходимость обратить внимание врача, у которого наблюдается пациентка, на возможные риски осложнений в случае бездействия. Если же отклонение оказывалось значительным, то значение попадало в категорию «Патология», что обозначало бы необходимость срочного вмешательства врача. Каждый из признаков (показателей) был взвешен в соответствии со степенью влияния на возникновение осложнений беременности. Разметка проводилась с учетом мнения эксперта.

Следующим шагом было разделение общей выборки на обучающую и тестовую. Так как количество данных, предоставленных для проведения исследования, сравнительно небольшое, а также необходимо было охватить в обучающем наборе данных весь период беременности, для него было выделено 70 % экземпляров данных, а остальное – для тестовой выборки.

При помощи средств scikit-learn на языке программирования Python были созданы модели деревьев решений, обучение которых проводилось с помощью подготовленного набора данных. Пример построения дерева решений представлен на рис. 1.

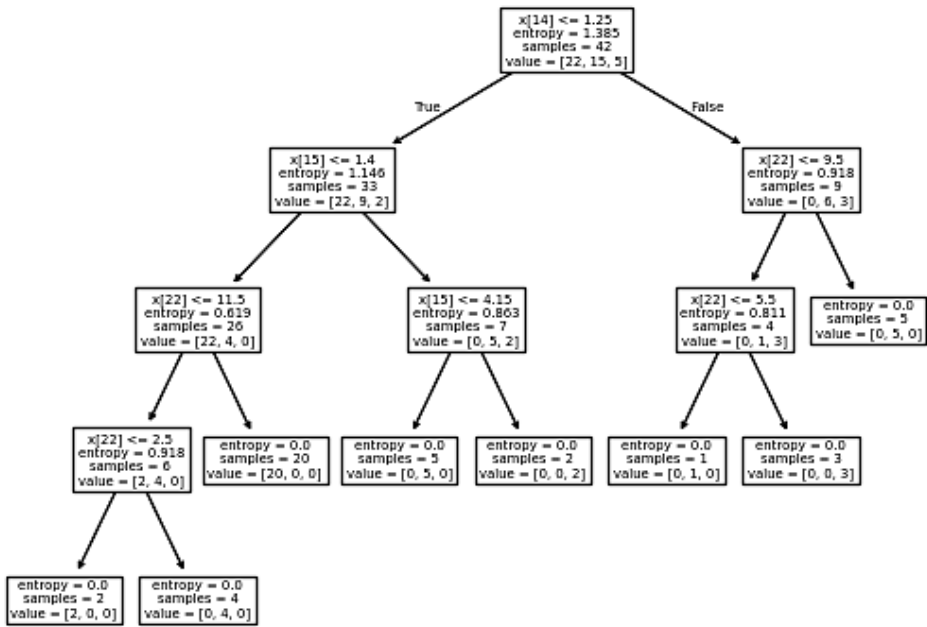


Рис. 1. Пример построения дерева решений

Обученные модели были применены к тестовой выборке данных. В табл. 1 представлены результаты: точность классификации составила 87,5 %. Также была построена матрица ошибок, по которой были определены дополнительные оценки алгоритма. Чувствительность составила всего примерно 16,7 %, что говорит о низкой точности выявления случаев, отличающихся от нормы. В это же время специфичность оказалась абсолютной в данном тесте, иначе говоря, все представители класса «Норма» были выявлены безошибочно. Тем не менее, несмотря на низкую специфичность, по модель верно определила единственный представленный в тестовой выборке

патологический случай, но не смогла распознать представителей класса «Внимание» и в большинстве случаев определяла их как нормальные. Это означает, что границы между этими двумя классами оказались размытыми, и неправильно были определены решающие правила для разделения на эти классы.

ТАБЛИЦА I.

Истинный класс	Результат классификации		
	«Норма»	«Внимание»	«Патология»
«Норма»	34	0	0
«Внимание»	4	0	1
«Патология»	0	0	1

IV. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Полученные результаты позволяют сделать вывод о работоспособности разработанной системы оценки состояния здоровья. Отобранные данные с относительно высокой точностью определяются классификатором в соответствии с размеченными классами. И хотя точность не достигла даже отметки 90 %, что считается сравнительно низким показателем при оценке методов машинного обучения стоит отметить, что эксперимент оказался успешным. Большей точности можно было бы добиться при наличии большего количества данных, чтобы провести более глубокое обучение модели. Кроме того, на точность и прочие показатели повлияла неравномерность обучающей выборки. Высокая специфичность и в то же время настолько низкая чувствительность объясняются тем, что подавляющая часть экземпляров данных относилась к классу «Норма». Вероятно, в будущем для проведения подобных экспериментов необходимо прибегнуть к генерации данных для выравнивания количества представителей каждого класса и увеличения размера выборки.

Также стоит учитывать, что, как правило, даже при наличии закономерностей, динамика показателей несет более сложный и индивидуальный характер, поэтому некоторые результаты классификации могут не соответствовать истине и снижать оценки работоспособности алгоритма.

В рамках исследования учитывался срок гестации как текущий триместр, в котором определены собственные нормальные диапазоны. Однако эти границы условны и изменения происходят в организме матери, конечно, гораздо чаще. Поэтому для оптимизации модели в будущем стоит учесть этот фактор.

V. ЗАКЛЮЧЕНИЕ

В результате проведения данного исследования были выполнены все задачи и достигнута поставленная цель, а также получены результаты, которые будут направлены на проведение новых исследований и улучшения аналитической системы для оценки текущего состояния здоровья беременной по комплексу диагностически значимых показателей.

СПИСОК ЛИТЕРАТУРЫ

- [1] Материнская смертность // Федеральная служба государственной статистики [Электронный ресурс]. URL: <https://rosstat.gov.ru/storga/mediabank/pUVvU4Rg/3-18.xls> (дата обращения: 01.03.2025).
- [2] Корнеева И.П., Крамарь К.А., Семенова Е.А., Сергеев А.М., Юлдашев З.М. Аппаратно-программный комплекс для удаленного мониторинга и контроля состояния беременных женщин // Информационно-управляющие системы, 2021, № 6, с.21–30.
- [3] Груздев А.В. "Прогнозное моделирование в IBM SPSS Statistics, R и Python. Метод деревьев решений и случайный лес" // ДМК Пресс, 2018, 644 с.
- [4] Анализы мочи во время беременности // Центр медицины плода [Электронный ресурс]. URL: https://mosplod.ru/topatients/articles/stati_9.html (дата обращения: 01.03.2025).