МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

Предсказание медицинского диагноза на основе многомерных данных методами машинного обучения

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Бибичевой Татьяны Сергеевны

Научный руководитель:
Зав. кафедрой
к.ф.-м.н., доцент _____ М.В. Огнева
подпись, дата

Зав. кафедрой
к.ф.-м.н., доцент _____ М.В. Огнева
подпись, дата

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. В настоящее время методы машинного обучения применяются во многих областях: финансы, лингвистика, геология, физика, телекоммуникационная отрасль. Также они могут играть большую роль в развитии медицины, в том числе в диагностике заболеваний, прогнозировании состояния пациента и др.

Одной из важных областей применения машинного обучения в медицине является кардиология: заболевания сердечно-сосудистой системы являются главной причиной смертности в мире. Методы машинного обучения рекомендуются как инструмент доклинической диагностики, в работе [1] они рассматриваются для поддержки решения врача. Кроме этого, в [2] отмечается оптимизация диагностического процесса, уменьшение вероятности ошибки.

Скорость накопления разнородных данных о сердечно-сосудистых заболеваниях огромна. Выделяют четыре основных источника [3]:

1. Функциональные фенотипы (например, демография, эхокардиограмма, данные визуализации);
2. Молекулярные профили, полученные в клинических условиях;
3. Медицинские записи, содержащие результаты лабораторных анализов, заметки врачей, информацию о заболеваниях, лечении;
4. Научные публикации.

Существует необходимость в обработке и анализе таких массивов информации. В работе [4] подчеркивается, что ошибки в принятии решений человеком обусловлены предвзятостью и шумом (например, настроение, погода). Избежать этого может помочь включение инструментов искусственного интеллекта. Но врачи-кардиологи должны следить за всеми финальными решениями и управлением системы.

Цель выпускной квалификационной работы – решение задачи медицинской диагностики по прогнозированию исхода заболевания и диагноза.

Поставленная цель определила **следующие задачи**:

1 Изучить предметную область, провести анализ литературы.
2 Изучить способы представления и параметры многомерных данных медицинской диагностики.
3 Рассмотреть основные понятия машинного обучения.
4 Изучить методы машинного обучения для решения задач медицинской диагностики.
5 Рассмотреть способы решения задачи пропущенных данных
6 Использовать рассмотренные методы машинного обучения для прогнозирования исхода заболевания.
7 Использовать рассмотренные методы машинного обучения для прогнозирования диагноза.

Методологические основы методов машинного обучения представлены в работах Воронцова К.В. [5], Хайкина С. [6], применение методов машинного обучения в кардиологии и заполнения пропусков в данных рассматриваются в работах Johnson KW [7], Daghistani TA [8], Payrovnaziri SN [9], Shah SJ [10].

Теоретическая значимость магистерской работы заключается в исследовании разных вариантов заполнения пропусков в медицинских данных и предсказании исхода заболевания и диагноза методами машинного обучения.

Структура и объём работы. Магистерская работа состоит из введения, 5 разделов, заключения, списка использованных источников и 19 приложений. Общий объем работы – 121 страница, из них 59 страниц – основное содержание, включая 3 рисунка и 7 таблиц, список использованных источников информации – 55 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Проблемы медицинской диагностики сердечно-сосудистых заболеваний и методы их решения» посвящен рассмотрению

работ по изучению методов машинного обучения в медицине и заполнению пропусков в медицинских данных.

Например, в работе [10] рассмотрена возможность выявления фенотипически различных категорий сердечной недостаточности с сохраненной фракцией выброса. В рамках исследования изучены данные о 397 пациентах, проведены детальные клинические исследования и эхокардиографическое фенотипирование участников. Использовались статистические алгоритмы обучения, включая иерархический кластерный анализ фенотипических данных и кластеризацию со штрафной функцией. В результате был сделан вывод, что статистические алгоритмы обучения, применяемые к плотным фенотипическим данным, могут улучшить классификацию клинических синдромов с целью определения однородных подклассов пациентов.

Заполнение пропущенных данных изучается в работе [9] по прогнозированию смертности от острого инфаркта миокарда для раннего вмешательства и проведения необходимых процедур. Основной проблемой электронных медицинских данных отмечается их качество из-за неполноты. Задачей является их заполнение, а не удаление записей с недостающими значениями. Данные заполняются методами с использованием средних значений, уравнений, а также методами, основанными на моделировании с машинным обучением: метод k-ближайших соседей, случайный лес и нейронные сети. Лучшие результаты показали методы с нейросетью и случайным лесом.

Число исследований, связанных с применением методов машинного обучения в кардиологии, быстро растет. Однако проблема предсказания диагноза сердечно-сосудистой системы не решена, в открытом доступе готовых программных продуктов нет. Поэтому необходимо рассматривать применение методов машинного обучения к диагностике заболеваний сердечно-сосудистой системы.

Также и исследования на тему пропусков в данных активно ведутся, но однозначного решения таких задач не существует. В каждом конкретном случае подбирается свой набор методов и осуществляется настройка параметров, при этом результат не всегда гарантирован.

Второй раздел «Постановка задачи машинного обучения» посвящен теоретическому описанию методов машинного обучения, применяемых в работе для решения задачи прогнозирования с помощью методов классификации:

1. Метод k-ближайших соседей
2. Метод опорных векторов
3. Случайный лес
4. Градиентный бустинг
5. Многослойный персептрон

Также в этом разделе рассматриваются методы регуляризации гребневая регрессия, лассо-регрессия, эластичная сеть для заполнения пропущенных данных.

Далее рассматривается реализация этих методов в библиотеке Python scikit-learn. Для каждого метода указаны возможные настраиваемые гиперпараметры.

Для реализации метода k-ближайших соседей в библиотеке scikit-learn реализован классификатор KNeighborsClassifier, для наивного байесовского классификатора - 3 вида классификаторов: классы GaussianNB, BernoulliNB и MultinomialNB, метод опорных векторов представлен в классе svm.SVC, случайный лес рассматривается в классе RandomForestClassifier, градиентный бустинг – в классе GradientBoostingClassifier, многослойный персептрон для классификации реализован в классе MLPClassifier.

Гребневая регрессия представлена в библиотеке классом Ridge, Лассо-регрессия реализована в классе Lasso, эластичная сеть реализуется в ElasticNet.

Рассмотренные методы имеют свои плюсы и минусы, каждый может использоваться в задачах медицинской диагностики. Но для их использования нужно точно понимать алгоритмы, грамотно подбирать параметры и осуществлять обучение.

Третий раздел «Восстановление пропусков в данных» посвящен различным подходам к заполнению пропусков в данных:

1. Игнорирование объектов с пропущенными значениями
2. Замена специальным значением
3. Замена самым частым или средним значением
4. Повторение результата последнего наблюдения
5. Метод k-ближайших соседей (KNN)
6. Восстановление с помощью методов машинного обучения

Каждый из вариантов работы с пропусками может быть использован в разных ситуациях, например, повторение результата последнего наблюдения может быть эффективно при заполнении пропусков во временных рядах, когда последующие значения, скорее всего, сильно взаимосвязаны с предыдущими.

Проблема наличия большого количества пропусков в таблице может быть связана с прекращением действия исследуемой процедуры; с невозможностью заполнения в данный момент; длительностью исследования или большим количеством параметров. Проблема пропуска данных является одной из самых серьезных в случае медицинских данных и может стать причиной смещения результатов и ошибок в выводах.

Четвертый раздел «Подбор параметров» посвящен настройке гиперпараметров методов машинного обучения. Для этого рассматриваются поиск по сетке и случайный поиск по сетке:

1. Поиск по сетке. На входе дается модель и возможные значения параметров для модели. Для всех возможных сочетаний значений параметров считается ошибка и выбирается наилучший набор.

2. Случайный поиск по сетке. На входе дается модель и возможные значения параметров для модели. Также производится поиск параметров для модели с лучшим результатом, но, в отличие от поиска по сетке, проверяются не все сочетания значений параметров.

В этом разделе приводится их реализация в Python. Для автоматического перебора в библиотеке scikit-learn существует класс GridSearchCV, случайный поиск реализован в библиотеке scikit-learn в классе RandomizedSearchCV.

Пятый раздел «Решение задачи прогнозирования» посвящен заполнению пропущенных данных,

В работе использовались обезличенные данные о приеме пациентов с различными заболеваниями сердечно-сосудистой системы за 2014-2018 гг. в саратовских больницах, заполненные врачами. Размер – 150000 записей. Для каждой записи оценивается >100 параметров. Основные параметры для каждого пациента:

1. Пол (Мужской/Женский - 1/0);
2. Возраст;
3. Жизненный статус (Жив/Мёртв/Нет данных - 1/0/null);
4. Рост, вес;
5. Дата выписки-дата поступления (в часах);
6. Основной диагноз при выписке (код диагноза в соответствии со справочником «Международная статистическая классификация болезней и проблем, связанных со здоровьем, 10-го пересмотра»);
7. Различные клинические состояния (например, ишемический инсульт, тампонада сердца, развитие жизнеопасных нарушений ритма; Да/Нет – 1/0);
8. Наличие различных состояний в анамнезе (инфаркт миокарда, хроническое легочное заболевание и др.; Да/Нет – 1/0);

9. Исследования крови (уровень глюкозы плазмы крови, гемоглобин, лейкоциты и др.);
10. Назначенное лечение (например, бета-блокаторы (Имеется/Отсутствует/Нет данных – 1/0/null));

Для заполнения пропущенных данных используются метод случайного леса для классификации и задачи регрессии, метод гребневая регрессия, лассо-регрессия, эластичная сеть. Для всех методов и признаков подобраны гиперпараметры с помощью поиска по сетке.

Для решения задачи бинарной классификации – исхода заболевания (жизненный статус) использовались метод k-ближайших соседей, случайный лес, градиентный бустинг, метод опорных векторов, наивный байесовский классификатор, многослойный персептрон, реализованные в библиотеке scikit-learn языка Python. Для всех методов подобраны гиперпараметры с помощью поиска по сетке.

В результате классификации предварительно необработанных данных ошибка составила 16-19%.

После заполнения пропущенных данных с помощью случайного леса ошибка классификации для разных методов составила 5-8%.

После заполнения пропущенных данных определенными вручную значениями и удаления строк с пропущенными значениями ошибка классификации составила 8-10%.

При заполнении пропущенных данных с использованием случайного леса для классификации и гребневой регрессии ошибка классификации составила 9-13%.

При заполнении пропущенных данных с использованием случайного леса для классификации и лассо-регрессии ошибка классификации составила 7-14%.

При заполнении пропущенных данных с использованием случайного леса для классификации и эластичной сети ошибка классификации составила 6-10%.

Сравнение результатов приведено в таблице 1.

Таблица 1 – Сравнение результатов

Методы классификации	Ошибка для необработанных данных, %	Ошибка после обработки данных, заполнение, случайный лес, %	Ошибка после обработки данных, заполнение, удаление пропусков, %	Ошибка после обработки данных, гребневая, случайный лес, %	Ошибка после обработки данных, лассо, случайный лес, %	Ошибка после обработки данных, эластичная сеть, случайный лес, %
Метод k-ближайших соседей	17%	8%	10%	10%	13%	10%
Случайный лес	16%	4%	8%	10%	7%	6%
Градиентный бустинг	17%	5%	9%	13%	11%	8%
Метод опорных векторов	17%	7%	10%	9%	9%	10%
Наивный байесовский классификатор	19%	7%	9%	11%	14%	8%
Многослойный персептрон	9%	4%	6%	8%	10%	6%

По полученным данным можно сделать вывод, что обработка медицинских данных с помощью случайного леса для классификации и регрессии больше остальных улучшила результаты классификации для всех методов, но разница с простым методом удаления пропусков небольшая. Это может быть связано с разницей в заполнении данных, проведении исследований для более и менее здоровых пациентов (возможно, для больных и здоровых пациентов проводилось разное количество исследований, т.е. оценивалось большее или меньшее количество признаков).

Для получения более определенных результатов необходима дополнительная информация о комбинациях исследований, лечения.

Для решения задачи мультиклассовой классификации – постановки диагноза были исследованы результаты предсказания исхода заболевания. Наиболее качественным оказался метод заполнения случайным лесом для классификации и регресии. Для предсказания диагноза использовались методы: метод k-ближайших соседей, случайный лес, градиентный бустинг, метод опорных векторов, наивный байесовский классификатор, многослойный персептрон.

Сравнение результатов приведено в таблице 2.

Таблица 2 – Сравнение результатов

Методы классификации	Метод k-ближайших соседей	Случайный лес	Градиентный бустинг	Метод опорных векторов	Наивный байесовский классификатор	Многослойный персептрон
Ошибка классификации	37%	25%	26%	30%	33%	32%

Таким образом, наиболее эффективными оказались методы случайный лес, градиентный бустинг. Более плохие результаты для многослойного персептрона могут быть связаны с недостаточным количеством данных для классификации.

ЗАКЛЮЧЕНИЕ

В проделанной работе были решены все поставленные задачи, что позволило рассмотреть уже проведенные исследования по применению методов машинного обучения при заболеваниях сердечно-сосудистой системы, работы по изучению пропусков в медицинских данных и их заполнению, исследовать форму представления данных о пациентах, а также методы машинного обучения, которые применимы для прогнозирования

заболеваний сердечно-сосудистой системы.

Для борьбы с проблемой пропусков в данных проведены эксперименты с такими методами, как удаление пропусков, восстановление пропущенных данных с использованием случайного леса для классификации и регрессии, гребневой регрессии, лассо-регрессии, эластичной сети. Для всех методов подобраны гиперпараметры с использованием поиска по сетке. Показано, что восстановление пропусков с помощью данных методов улучшает результаты классификации на 7-12%.

Для решения задачи прогнозирования исхода заболевания и диагноза были рассмотрены и реализованы методы классификации: метод k-ближайших соседей, случайный лес, градиентный бустинг, метод опорных векторов, наивный байесовский классификатор, многослойный персептрон. В результате для решения задачи классификации на медицинских данных лучшими стали случайный лес, градиентный бустинг, нейронные сети, которые показали результат на 4-5% лучше других методов для бинарной классификации и 7-9% для предсказания диагноза.

Отдельные части магистерской работы были представлены на конференции:

Научно-практическая конференция студентов факультета компьютерных наук и информационных технологий 2020 года: доклад «Анализ и обработка медицинских данных для решения задачи классификации».

Основные источники информации:

1. Yoruk U. Automatic Renal Segmentation for MR Urography Using 3D-GrabCut and Random Forests / U Yoruk, BA Hargreaves, SS Vasanawala // International Society for Magnetic Resonance in Medicine. – 2017. – Vol. 79, No. 3. – P. 1696–1707.

2. Boughorbel S. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric / S Boughorbel, F Jarray, M El-Anbari // PLoS ONE. – 2017. – Vol. 12, No. 6.– 17 p.

3. Perez RV. Utility of artificial intelligence in cardiology / RV. Perez // HealthManagement. – 2018. – Vol. 18, No. 1. – P. 50–52.

4. Subhi J Al'Aref. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging / J Al'Aref Subhi, K Anchouche, G Singh, PJ Slomka, KK Kolli, A Kumar, M Pandey,G Maliakal, AR Rosendael // European Heart Journal. – 2019. – Vol. 40, No. 24. – P. 1975–1986.

5. Воронцов К.В. Математические методы обучения по прецедентам. / К.В. Воронцов. – 2007. – 141 с.

6. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. М.: Издательский дом «Вильямс», 2006. 1104 с.

7. Johnson KW. Artificial Intelligence in Cardiology / KW Johnson, JT Soto, BS Glicksberg, K Shameer, R Miotto, M Ali, E Ashley, JT Dudley // JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY. – 2018. – Vol. 71, No. 23. – P. 2668-2679.

8. Daghistani TA. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach / TA Daghistani, R Elshawi, SSakr, AM Ahmed,A Al-Thwayee,MH Al-Mallah // Int J Cardiol. – 2019. – Vol.288. – P.140-147.

9. Payrovnaziri SN. The Impact of Missing Value Imputation on the Interpretations of Predictive Models: A Case Study on One-year Mortality Prediction in ICU Patients with Acute Myocardial Infarction / SN Payrovnaziri, A Xing, S Salman, X Liu, J Bian, Z He // Explainable Artificial Intelligence in Medicine for Mortality Prediction for ICU Patients. – 2020. – 10 p.

10. Shah SJ. Phenomapping for Novel Classification of Heart Failure with Preserved Ejection Fraction / SJ Shah, DH Katz, S Selvaraj, MA Burke, CW Yancy, M Gheorghiade, RO Bonow, CC Huang, RC Deo // Circulation. – 2015. – Vol.131, No. 3. – P. 269–279.