МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра дискретной математики и информационных технологий

ОТБОР ПРИЗНАКОВ В ЗАДАЧЕ ПОНИЖЕНИЯ РАЗМЕРНОСТИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Козынченко Вячеслава Сергеевича

Научный руководитель
доцент, к. ф.-м. н. _____ И. Д. Сагаева

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2025

ВВЕДЕНИЕ

На данный момент машинное обучение внедряется во все сферы деятельности человека. По мере развития технологий сбора данных увеличиваются и объёмы имеющихся данных. Для того чтобы извлечь пользу из такой ситуации, необходимо правильно подходить к задачам анализа данных и понижения размерности.

Модели машинного обучения легко поддаются модификациям и настройке. Разнообразие архитектур нейронных сетей позволяет решить широкий спектр задач, используя различные подходы. Производительность и точность моделей превосходят оценку эксперта. Таким образом, используя машинное обучение, можно сократить затраты на экспертов, а также увеличить скорость вычисления результата.

Не все имеющиеся данные могут быть полезны при работе с моделями машинного обучения. Это может приводить к ухудшению точности модели и её эффективности. Отбор признаков может уменьшить количество информации, передаваемой на вход модели, оставляя только необходимые для получения наиболее точного результата характеристики и сокращения времени, затрачиваемого на обучение [1].

Сахарный диабет является заболеванием, затрагивающим миллионы людей, и требует не только внимания, но и быстрого реагирования для предотвращения его развития. Критичность данного заболевания заключается не только в необходимом контроле за состоянием здоровья, но и развитием множества осложнений, включая сердечно-сосудистые заболевания. В такой ситуации на помощь приходит машинное обучение, благодаря которому можно получить рекомендации по обследованию.

Для того, чтобы получить наиболее точные результаты для группы риска, необходимо решить задачу понижения размерности с помощью одновременного использования нескольких алгоритмов отбора признаков. Также важно предоставить интерпретируемые результаты машинного обучения для повышения доверия от медицинских специалистов и дальнейшего развития стратегий профилактики сахарного диабета с внедрением таких моделей. Данный подход сочетает в себе высокую точность предсказаний и прозрачность работы модели, что является критически важным для внедрения алгоритмов машинного обучения в медицинские задачи.

Целью данной работы является решение задачи понижения размерности для классификации наличия сахарного диабета у человека.

Для этого были поставлены следующие задачи:

- обзор существующих решений задачи понижения размерности;
- поиск набора данных для проведения экспериментов;
- реализация алгоритмов дерева решений и случайного леса;
- реализация алгоритмов отбора признаков;
- отбор признаков для моделей машинного обучения;
- визуализация вклада отобранных признаков в предсказание модели.

Структура и объем работы. Для решения поставленных задач выполнена выпускная квалификационная работа, которая включает в себя введение, 7 основных глав, заключение, список использованных источников из 42 наименований и 10 приложений. Работа изложена на 80 страницах, содержит 43 рисунка, одну таблицу.

Первая глава имеет название «Машинное обучение» и содержит информацию об основных понятиях данной сферы. Также в главе были описаны модели машинного обучения, которые были разработаны в ходе выполнения работы.

Вторая глава имеет название «Алгоритмы выбора признаков» и содержит информацию о работе алгоритмов выбора признаков.

Третья глава имеет название «Обзор существующих решений» и содержит информацию о существующих решениях задачи понижения размерности. В главе было рассмотрено полноценное приложение с пользовательским интерфейсом и библиотеки на языке программирования Python

Четвертая глава имеет название «Используемые средства» и содержит перечисление использованных в работе инструментов.

Пятая глава имеет название «Реализация алгоритмов машинного обучения» и содержит подробное описание реализации моделей машинного обучения.

Шестая глава имеет название «Реализация алгоритмов отбора признаков» и содержит подробное описание реализации алгоритмов отбора признаков.

Седьмая глава имеет название «Проведение экспериментов» и содержит описание проведённых экспериментов с реализованными алгоритмами. Также в главе описан и проанализирован используемый набор данных.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Машинное обучение

Машинное обучение — это раздел искусственного интеллекта, направленный на создание систем, которые обучаются на основе входных данных программы. Такие системы позволяют улучшить производительность и снизить затраты на ручную работу экспертов [2].

Входные данные представляют собой векторы признаков, которые хранят значения, описывающие характеристики объекта. Для всех входных данных признак в позиции j всегда представляет одну и ту же характеристику.

Переобучение является одной из основных проблем обучения с учителем. Оно заключается в том, что модель вместо того, чтобы решать поставленную задачу, запоминает обучающие примеры. Одним из способов борьбы с переобучением является решение задачи уменьшения размерности.

Целью задачи уменьшения размерности является сокращение числа признаков набора данных. Вектор признаков может содержать слабо информативные или неинформативные характеристики. Наличие таких характеристик может понизить эффективность используемой модели.

Уменьшение размерности может быть осуществлено с помощью следующих методов.

— Выбор признаков. Такой метод упрощает вектор признаков, удаляя из него слабо информативные характеристики [3].

— Выделение признаков. Такой метод заключается в создании новых признаков на основе исходных, но при этом теряется репрезентативность данных.

Кросс-валидация — это метод оценки производительности и точности модели, заключающийся в разбиении доступной для обучения информации на обучающую и тестовую выборки. Данное разбиение позволяет снизить риск переобучения модели, а также недопустить оценивание на уже известной информации.

Дерево решений — инструмент принятия решений, использующийся в машинном обучении и анализе данных. Данный инструмент основан на структуре данных «дерево», где «листья» представляют собой конечный результат классификации. Перемещение по «узлам» структуры происходит по значениям признаков, находящихся в «ветвях» [4].

Случайный лес — это алгоритм машинного обучения, который является ансамблем деревьев решений. Он применяется в задачах классификации и регрессии, а также для обнаружения выбросов. Каждое дерево ансамбля может обладать низким качеством результатов, но за счет их большого количества общая оценка работы остается выше [5].

Алгоритмы выбора признаков

В работе были реализованы следующие алгоритмы выбора признаков [6].

— На основе фильтров. Данные алгоритмы не используют обучение и учитывают только релевантность между признаками. Таким образом, они обладают стабильностью и масштабируемостью, но могут игнорировать определённые информативные признаки, а также допускать несбалансированность.

— На основе оболочки. Такие алгоритмы используют результаты обучаемой модели для определения информативных признаков и позволяют облегчить анализ результатов. Данный подход порождает NP задачу, которую пытаются решить оптимизационные алгоритмы.

— Гибридные. Выбор признаков предполагает комбинацию вышеуказанных видов алгоритмов. Например, используя алгоритмы на основе фильтров, удалить наиболее дискриминационные признаки, а затем выделить самые информативные с помощью методов на основе оболочки.

Критерий хи-квадрат — статистический критерий, проверяющий наличие связи между признаком и целевой переменной. Этот критерий используется для изучения независимых категориальных признаков.

Метод Уэлфорда — метод, который используется для вычисления средних, дисперсий и других величин.

Исчерпывающий поиск признаков — жадный алгоритм выбора признаков, оценивающий каждое подмножество признаков и предоставляющий оптимальное решение на их основе. Данный алгоритм является значительно медленным для задач высокой размерности и подходит только для данных, обладающих малым набором признаков.

Последовательный прямой отбор — алгоритм выбора признаков, итеративно добавляющий признаки в изначально пустой набор. Данный алгоритм применим, когда осуществляется выбор минимального набора признаков для

решения задачи классификации или регрессии.

Последовательный плавающий прямой отбор — алгоритм, основанный на последовательном прямом отборе, который выполняет перерасчёт предыдущих итераций.

Последовательный обратный отбор — алгоритм выбора признаков, итеративно удаляющий признаки из набора всех признаков. Данный алгоритм применим, когда осуществляется выбор максимального набора признаков для решения задачи классификации или регрессии.

Последовательный плавающий обратный отбор — алгоритм, основанный на последовательном обратном отборе, выполняющий перерасчёт предыдущих итераций.

Рекурсивное удаление признаков — алгоритм поиска, рекурсивно выбирающий уменьшающийся набор признаков. Признаки в данном случае ранжируются по очерёдности их удаления из набора.

Регуляризация Лассо — метод регуляризации, используемый для оценки линейной регрессии и решения задачи понижения размерности. Он уменьшает веса нерелевантных признаков вплоть до 0. Это позволяет избежать переобучения и упростить модель, так как коэффициенты, соответствующие нерелевантным признакам, будут исключены в процессе обучения.

Обзор существующих решений

Были рассмотрены следующие существующие решения задачи понижения размерности с помощью отбора признаков:

— FeatureSelect;
— Sklearn;
— Feature-engine.

Используемые средства

Jupyter Notebook — это веб-приложение с открытым исходным кодом, которое позволяет специалистам по обработке и анализу данных создавать и обмениваться документами, содержащими «живой» код на языке Python, а также уравнения и другие мультимедийные ресурсы.

Подготовка данных осуществляется с помощью библиотеки Pandas.

Для разбиения данных и создания наборов признаков используются сле-

дующие библиотеки:

— itertools — встроенный модуль Python, предоставляющий набор бесконечных и конечных итераторов, а также комбинаторные генераторы.

— Sklearn — это библиотека для машинного обучения в Python. Она предоставляет набор методов, с помощью которых можно производить оценку точности построенной модели и решать задачу уменьшения размерности.

Matplotlib — это библиотека построения графиков, доступная для языка программирования Python в качестве компонента NumPy, используемого числовой обработки больших данных.

Imblearn — это библиотека, предназначенная для балансирования классов данных перед разбиением набора на обучающую и проверяющую выборки.

SHAP — фреймворк, вычисляющий SHAP значения для моделей машинного обучения и способы визуализации вклада признаков.

graphviz — пакет, упрощающий построение и визуализацию графов.

Joblib — пакет для инициализации параллельных вычислений в Jupyter Notebook.

Реализация алгоритмов машинного обучения

Для реализации алгоритма дерева решений был реализован класс DecisionTree, который имеет следующее параметры:

— максимальная высота;

— минимальное число образцов для разбиения.

Данная реализация использует энтропию для поиска наилучшего разделения выборки.

Для визуализации дерева решений была реализована функция, выполняющий прямой обход дерева. Для инициализации визуализации используется конструкция digraph библиотеки graphviz. Функция создаёт ветки и узлы, которые представлены абстрактной грамматикой.

Для реализации алгоритма случайного леса был реализован класс RandomForest, который имеет следующее параметры:

— количество деревьев;

— максимальная высота дерева;

— размер выборки для каждого дерева;

— минимальное число образцов для разбиения.

Реализация алгоритмов выбора признаков

Для реализации метода Уэлфорда был реализован класс VT, параметром которого является пороговое значение вариативности.

Был реализован класс KBest, осуществляющий выбор k лучших признаков по значению теста с минимальным значением p, которое отвечает за вероятность ошибки полученного значения. Данный алгоритм принимает следующие параметры:

— критерий для отбора признаков, например, хи-квадрат;
— необходимое количество признаков.

Для реализации исчерпывающего поиска признаков был реализован класс ESF, который принимает следующие параметры:

— модель машинного обучения;
— минимальное количество признаков в результирующем наборе;
— максимальное количество признаков в результирующем наборе;
— функция для оценки наборов данных, например, F1-мера.

Для реализации последовательного прямого отбора был реализован класс SFS, параметры которого представлены следующим образом:

— модель машинного обучения;
— количество признаков;
— флаг плавающей реализации;
— функция для оценки наборов данных.

Для реализации последовательного обратного отбора был реализован класс SBS, параметры которого представлены следующим образом:

— модель машинного обучения;
— количество признаков;
— флаг плавающей реализации;
— функция для оценки наборов данных.

Для реализации алгоритма, рекурсивно удаляющего менее важные признаки из набора, был реализован класс RFE, который принимает следующие параметры:

— модель машинного обучения;
— способ, с помощью которого будет извлекаться важность признаков;
— количество признаков, удаляемых за итерацию;
— итоговое число признаков;

Гибридные алгоритмы представляют собой последовательное выполнение алгоритмов выбора признаков. Для реализации таких алгоритмов был создан класс FSPipeline, который принимает список алгоритмов.

Проведение экспериментов

Для решения задачи с использованием реализованных алгоритмов машинного обучения был выбран датасет с сайта Kaggle, содержащий ответы на телефонный опрос, связанный со здоровьем. Признаки представляют собой либо вопросы, непосредственно заданные участникам, либо вычисленные значения, основанные на ответах отдельных участников опроса.

В наборе данных преобладает количество образцов негативного класса (отсутствие сахарного диабета). Это может привести к следующим последствиям:

— потеря производительности при подборе признаков;
— переобучение доминирующего класса;
— усложнение использования метрик.

Для балансирования классов набора данных был использован метод NearMiss, который предоставляет выбор важных образцов из доминирующего класса, используя среднее расстояние до ближайших образцов меньшего класса, и выбирает среди них меньшее.

Для нормализации числовых данных используется StandardScaler от Sklearn, который индивидуально масштабирует каждый признак набора данных до единичной дисперсии.

Для оценки подбираемых признаков была выбрана F-мера, вычисляемая по формуле 1.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{1}$$

Проведение эксперимента с алгоритмами на основе оболочки показало, что оптимальным алгоритмом выбора признаков для задачи классификации наличия сахарного диабета является последовательный прямой плавающий алгоритм выбора признаков. Данный алгоритм продемонстрировал схожие результаты с исчерпывающим алгоритмом и показал меньшие временные затраты на выполнение.

Было построено 3 гибридных алгоритма выбора признаков. Основой дан-

ных алгоритмов является выбор k лучших категориальных признаков по значению критерия хи-квадрат. После фильтрации выполняются следующие алгоритмы:

— последовательный плавающий прямой отбор на основе моделей дерева решений с неограниченной высотой и значением минимального количества образцов для разбиения, установленного в 2;
— построение случайного леса с числом деревьев 100 и максимальной высотой дерева 6;
— логистическая регрессия с регуляризацией Лассо.

Лучший набор признаков, подобранный для дерева решений представлен следующим списком:

— физическое состояние;
— психологическое состояние;
— уровень дохода;
— общее состояние здоровья;
— уровень образования;
— наличие сердечно-сосудистых заболеваний;
— трудности при передвижении;
— высокая физическая активность;
— случай инсульта;
— чрезмерное употребление алкоголя.

SHAP-значения для i признака вычисляются для каждого образца из набора данных на всех возможных комбинациях признаков, и сумма полученных значений представляет важность i признака [7].

Были вычислены SHAP-значения и построен график вклада каждого признака в предсказание модели. Наибольший вклад оказывают психологическое и физическое состояния респондента. Чем больше дней человек испытывает физическую боль или низкое эмоциональное состояние, тем выше риск наличия сахарного диабета. Также вклад вносит индекс массы тела, чем больше его значение, тем выше риск наличия заболевания. Наименьший вклад вносят образование, высокий уровень холестерина, пол и возраст участника опроса. Данные наблюдения подтверждаются медицинскими исследованиями [8, 9].

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были рассмотрены и реализованы алгоритмы машинного обучения и отбора признаков. Также в алгоритмы отбора признаков были добавлены конструкции для параллельного выполнения и визуализации результатов работы.

Были построены 3 гибридных алгоритма выбора признаков. В их основе используется критерий хи-квадрат. Данные алгоритмы представлены следующим списком:

- последовательный прямой плавающий отбор признаков для дерева решений;
- случайный лес;
- логистическая регрессия с регуляризацией Лассо.

Проведение экспериментов показало, что алгоритмы выбора признаков склонны выбирать важными признаками:

- информацию об общем состоянии здоровья;
- индекс массы тела;
- социальную информацию;
- наличие сердечно-сосудистых заболеваний и их осложнений;
- информацию об образе жизни.

С помощью алгоритмов отбора признаков удалось уменьшить количество признаков в наборе данных в два раза, что улучшило производительность моделей. Лучшей моделью по результатам экспериментов является случайный лес.

Реализованные алгоритмы отбора признаков можно улучшить с помощью внедрения алгоритмов оптимизации и параллельных участков программного кода. Также данные алгоритмы могут быть улучшены с помощью критериев остановки, которые прекращают отбор по заданному условию, например, малый прирост метрики или переобучение.

Полученные в ходе экспериментов модели машинного обучения могут использоваться как в персональных целях для определения необходимости прохождения медицинского обследования, так и в специализированных учреждениях для уменьшения нагрузки на персонал.

Отдельные части магистерской работы были опубликованы в журнале, индексируемом в РИНЦ:

1 *Козынченко, В. С.* Алгоритмы выбора признаков на основе оболочки с исполь-

зованием деревьев решений [Электронный ресурс] / В.С Козынченко, И. Д. Сагаева // *Актуальные вопросы современной экономики*. — 2025. — Т.5. — URL: https://avse-journal.ru/file/file_242525258428.pdf (Дата обращения: 01.05.2025). — Загл. с экр. — Яз. рус.

Основные источники информации:

1 *Miao, J.* A survey on feature selection / J. Miao, L. Niu // *Procedia Computer Science*. — 2016. — Vol. 91. — Pp. 919–926.

2 *Вьюгин, В. В.* Математические основы машинного обучения и прогнозирования / В. В. Вьюгин. — М.: МЦНМО, 2022.

3 *Chen, R.-C.* Selecting critical features for data classification based on machine learning methods / R.-C. Chen, C. Dewi, S. Huang, R. Caraka // *Journal Of Big Data*. — 2020. — Vol. 7. — P. 26.

4 *Alharan, A.* Popular decision tree algorithms of data mining techniques: A review / A. Alharan, R. Alsagheer, A. Al-Haboobi // *International Journal of Computer Science and Mobile Computing*. — 2017. — Vol. 6. — Pp. 133–142.

5 *Mei, K.* Modeling of feature selection based on random forest algorithm and pearson correlation coefficient / K. Mei, M. Tan, Z. Yang, S. Shi // *Journal of Physics: Conference Series*. — 2022. — Vol. 2219, no. 1. — Pp. 12–46.

6 *Pushpam, C.* Research on feature selection using svm / C. Pushpam, G. Joseph // *International Journal of Recent Technology and Engineering (IJRTE)*. — 2019. — Vol. 8. — Pp. 7252–7256.

7 *Lundberg, S. M.* A unified approach to interpreting model predictions / S. M. Lundberg, S. Lee // Proceedings of the 31st International Conference on Neural Information Processing Systems. — New York: Curran Associates Inc., 2017. — Pp. 4768–4777.

8 *Garrett, C.* Diabetes and mental health / C. Garrett, A. Doherty // *Clinical medicine*. — 2014. — Vol. 14. — Pp. 669–672.

9 *Lysy, Z.* The impact of income on the incidence of diabetes: A population-based study / Z. Lysy, G. L. Booth, J. Luo, B. R. Shah, P. C. Austin, L. L. Lipscombe // *Diabetes Research and Clinical Practice*. — 2013. — Vol. 99. — Pp. 372–379.