

УДК 004.8

НЕЧЁТКИЕ КЛАССИФИКАТОРЫ ДЛЯ ДИАГНОСТИКИ БОЛЕЗНИ ПАРКИНСОНА НА ОСНОВЕ СТАТИЧЕСКИХ РУКОПИСНЫХ ДАННЫХ

© И. А. Ходашинский, Ю. А. Шурыгин, К. С. Сарин,
М. Б. Бардамова, А. О. Слёзкин, М. О. Светлаков, Н. П. Корышев

*Томский государственный университет систем управления и радиоэлектроники,
634050, г. Томск, просп. Ленина, 40
E-mail: ilia.a.khodashinski@tusur.ru*

Диагностика болезни Паркинсона является дорогостоящей процедурой, включающей в себя транскраниальную сонографию и томографию головного мозга. В связи с этим актуальными являются простые и точные скрининговые методы диагностики. Рассматриваются вопросы анализа рукописных статичных рисунков спиралей и меандров методами машинного обучения для диагностики болезни Паркинсона на основе общедоступного набора данных HandPD. Построены нечёткие классификаторы с помощью оригинальных методов, способные по рисунку определять наличие или отсутствие болезни. Поскольку набор HandPD является несбалансированным, в работе применяются алгоритмы искусственной генерации выборок. Проведено статистическое сравнение точности применяемых моделей и методов, выполнено ранжирование признаков.

Ключевые слова: нейродегенеративные заболевания, машинное обучение, классификация, отбор признаков, нечёткие классификаторы, метаэвристические алгоритмы.

DOI: 10.15372/AUT20230308

Введение. Болезнь Паркинсона (БП) обусловлена недостатком дофамина, который является химическим веществом, вырабатываемым нейронами в чёрной субстанции среднего мозга. Когда эти нейроны начинают разрушаться, выработка дофамина снижается. Болезнь Паркинсона является неизлечимым заболеванием, и основная причина выхода из строя нейронов пока не известна. Однако при ранней диагностике правильный курс лечения может помочь облегчить тяжесть протекания заболевания [1].

Общемировая распространённость БП составляет 0,3 %. С возрастом число больных существенно увеличивается приблизительно от 1 % в группе лиц старше 60 лет и до 4 % после 75. Наблюдается тенденция к омоложению этой болезни: в настоящее время ею болеет каждый десятый человек до 50 лет и каждый двадцатый до 40 лет. Эта болезнь приводит к инвалидности пациентов, снижая качество жизни не только самого больного, но и его близких [2].

Искусственный интеллект и машинное обучение нашли применение в компьютерной диагностике, мониторинге и лечении нейродегенеративных двигательных расстройств паркинсонического типа. Эти задачи нетривиальны из-за высокой межпредметной вариативности и сходства клинических проявлений различных нейродегенеративных расстройств на ранних стадиях. Такие методы часто используются для диагностики заболеваний на основе данных, собранных с различных носителей. Кроме того, машинное обучение используется в дистанционном мониторинге и обнаружении тяжести БП, симптомов и реакции на терапию в режиме реального времени [3].

Болезнь Паркинсона неизлечима, однако, если её диагностировать на ранних стадиях, жизнь пациента можно улучшить с помощью физических упражнений и соответствующего лечения. Ранними симптомами у большинства пациентов с БП являются двигатель-

ные расстройства, такие как тремор, постуральная нестабильность, брадикинезия и ригидность. Изменения в головном мозге, вызванные нейродегенерацией, — атрофия мозга, потеря нейронов, синаптическая дисфункция и т. д. — приводят к дисфункции двигательной системы, а также к нарушениям выполнения ранее усвоенных двигательных навыков. По этой причине ключевую роль в ранней диагностике нейродегенеративных заболеваний (НЗ) может играть анализ почерка.

Почерк пациентов с БП характеризуется отсутствием беглости, медлительностью, резкой сменой направления и микрографией. Болезнь приводит к нарушению способности пациентов контролировать скорость и амплитуду движений, координировать движения пальцев и запястий [4]. На почерк человека влияет множество факторов, таких как грамотность и знание языка, поэтому в рамках диагностики БП предпочтительнее вместо написания предложений использовать для анализа рисование и/или обведение простых фигур, например спирали или меандра. Для интерпретации полученных рисунков требуются специалисты, но, если таковых нет, достаточно применить инструмент для компьютерной диагностики БП, обученный определять состояние здоровья испытуемого [1].

Системы искусственного интеллекта могут обрабатывать и анализировать большое количество данных, полученных из историй болезни и медицинских процедур. Это могло бы обеспечить более эффективные стратегии как для диагностики, так и для лечения заболеваний. Однако наиболее эффективные (с точки зрения точности) системы такого типа построены по принципу «чёрного ящика», который не предусматривает описания чётких правил получения результата. Отсутствие возможности интерпретации результата и объяснения работы модели препятствует принятию этих систем врачами-клиницистами и их внедрению в медицинскую практику [5].

Нечёткие системы, основанные на правилах, способны выражать знания в человеко-ориентированной форме благодаря принятию парадигмы, позволяющей использовать термины естественного языка. Такая возможность позволяет предоставить пользователям удобочитаемые объяснения, а также может продемонстрировать детали, касающиеся процесса вывода результата [6].

Цель работы заключается в исследовании методов построения нечётких классификаторов для диагностики болезни Паркинсона по статическим рукописным данным.

Научный вклад работы представлен следующими результатами:

- 1) впервые для диагностирования БП использован нечёткий классификатор с оригинальными алгоритмами построения, позволяющий на естественном языке интерпретировать полученный результат;
- 2) выделены информативные признаки в наборе рукописных данных для диагностики БП, впервые проведено ранжирование таких признаков;
- 3) для несбалансированных медицинских рукописных данных пациентов с БП применены алгоритмы оверсэмплинга, впервые статистически доказано, что добавление синтетических данных не влияет на распределение рангов признаков.

Ближкие работы. Диагностика и мониторинг нейродегенеративных заболеваний с использованием неинвазивных процедур и компьютерных технологий в настоящее время является актуальной темой научных исследований. Структурная магниторезонансная томография считается надёжным подходом к оценке признаков нейродегенерации. Однако одной нейровизуализации недостаточно для понимания НЗ [7]. Визуализация головного мозга — это мощный инструмент диагностики заболеваний, однако она является дорогостоящей, труднодоступной и непригодной для проведения непрерывного мониторинга прогрессирования заболевания [8].

В последние годы возрос интерес к применению биометрических методов в сфере медицины. Анализ почерка позволяет получить ценные сведения о начинающихся и прогрессирующих нарушениях, вызванных НЗ.

Болезнь Паркинсона характеризуется двигательной дисфункцией, часто заканчивающейся потерей подвижности. Авторы [9] считают, что понимание прогрессирования и хронологии моторной дегенерации необходимо для надлежащей поддержки пациентов на каждой стадии заболевания. Двигательная активность измеряется во время рутинных клинических обследований, которые часто оценивают как мелкие, так и крупные мышечные функции. Машинное обучение можно использовать для оценки состояния пациентов при рисовании простых фигур, написании слов или подобных задач. Анализ результатов выполнения таких заданий всё чаще используется в ранней диагностике БП, поскольку у людей с заболеванием почерк больше не является автоматизированной функцией и выполняется более сегментированным, последовательным образом, с многочисленными паузами. Внедрение методов машинного обучения в анализ письменных заданий может помочь классифицировать людей с БП и служить диагностическим инструментом [9].

К настоящему времени собраны и опубликованы наборы данных, которые содержат образцы рукописных данных пациентов с НЗ и контрольной группы (КГ) людей без этого заболевания. Далее приведены примеры таких наборов данных. База данных PaHaW состоит из множества образцов почерка 37 пациентов с БП и 38 человек из КГ [10]. Набор данных ParkinsonHW содержит информацию о 62 пациентах с БП и 15 человек из КГ [11]. Набор данных HandPD включает в себя сигналы оцифрованных текстов и рисунков пациентов с БП и КГ. В дальнейшем в этот набор помимо статических признаков были включены ещё и динамические. Расширенная версия, которая содержит данные о 66 человек (35 пациентов КГ и 31 пациент с БП), получила название NewHandPD.

Использование для исследований набора данных HandPD, содержащего только статические признаки, авторы [12] обосновывают тем, что: 1) анализ старых записей или рисунков пациента (доступных только на бумаге) может быть полезен для восстановления истории болезни пациента или для определения даты начала заболевания; 2) некоторым пациентам, особенно пожилым, неудобно писать на графическом планшете, и это может привести к ошибкам в оценке их состояния. В [12] предложен эволюционный подход к автоматической диагностике БП с помощью анализа рукописных данных с использованием декартового генетического программирования для построения понятной модели классификации.

Исследование [5] является продолжением работы [12]. В [5] при работе с набором данных NewHandPD создан классификатор на основе декартова генетического программирования, в котором ориентированный ациклический граф используется для представления вычислительной структуры эволюционировавших функциональных выражений. Каждый испытуемый был представлен вектором из 72 признаков, вычисленных на спиральных и мандрах, нарисованных этим испытуемым. Алгебраические манипуляции со значениями признаков были использованы в [5] для построения системы объяснения. Однако сделано уточнение, поскольку генетическое программирование включает в себя алгебраические манипуляции со значениями признаков, то эти манипуляции не имеют никакого смысла на поведенческом уровне.

Исследования набора данных HandPD [13] показали, что точность классификации смещается в сторону класса большинства, когда модели обучаются на несбалансированных данных. Наблюдались высокий уровень чувствительности и низкий уровень специфичности, поскольку класс пациентов составлял большинство, а класс здоровых — меньшинство. Чтобы снизить предвзятость в построенных моделях, был применён метод случайной заниженной выборки (random undersampling method). В [13] использовались классификатор AdaBoost и метод отбора признаков по критерию χ^2 .

В [14] использовали набор данных NewHandPD, метод главных компонент для извлечения признаков и следующие алгоритмы классификации: логистическая регрессия,

XGBoost и метод опорных векторов с линейным ядром. Набор данных HandPD исследовался с использованием свёрточной нейронной сети, параметры которой настраивались метаэвристическими алгоритмами, в результате чего получена достаточно хорошая точность. В [15] объединены несколько наборов рукописных данных и применены алгоритмы глубокого обучения свёрточных нейронных сетей различной архитектуры. Была достигнута точность классификации, близкая к 100 %. Однако используемые классификаторы не дают информации ни о том, как принимается решение, ни о том, как влияет каждый признак на результат классификации.

Набор данных для построения моделей диагностики БП. Набор данных HandPD был собран в работе [16]. Он состоит из изображений рукописных рисунков 92 человек, среди которых 74 страдают болезнью Паркинсона, а 18 не подвержены данному заболеванию. Таким образом, 80,44 % набора данных составляют пациенты, а 19,56 % — контрольная группа. Каждому участнику требовалось за четыре попытки обвести на бумажном листе шаблоны спиралей и меандров (рис. 1).

Всего было получено 736 изображений. Они были оцифрованы для последующей предварительной обработки и извлечения признаков [16]. Предварительная обработка заключалась в извлечении из оцифрованного изображения шаблона (ST) и фигуры (HT), нарисованной участником эксперимента. Для этого были объединены некоторые классические методы обработки изображений, такие как фильтры размытия и математическая морфология с процессом выделения контуров фигур. На рис. 2 представлены результаты извлечения ST и HT из изображения задания эксперимента.

Далее полученные изображения подверглись обработке алгоритмом прореживания из [17] с целью уменьшения толщины линии рисунка до одного пиксела и получения каркаса. Из каждой полученной пары каркасов (шаблонного и рукописного) извлекались по девять числовых признаков, которые описывают статистическое различие между ними.

Признак x_1 — среднеквадратичное значение различия между шаблонной и рукописной линиями:

$$x_1 = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_{HT}^i - r_{ST}^i)^2}.$$

Здесь n — количество случайно выбранных точек на линии рисунка (значение $n = 360$ в работе [16] было получено эмпирическим путём, так как данное количество точек выборки показало хороший компромисс между скоростью расчёта и точностью); r_{HT}^i — расстояние (радиус) от центра рисунка до точки i на рукописной линии; r_{ST}^i — расстояние от центра рисунка до точки i на шаблонной линии, $i = 1, \dots, n$.

Признаки x_2 и x_3 — максимальное и минимальное различия между r_{HT}^i и r_{ST}^i :

$$x_2 = \max(\{|r_{HT}^1 - r_{ST}^1|, \dots, |r_{HT}^n - r_{ST}^n|\}), \quad x_3 = \min(\{|r_{HT}^1 - r_{ST}^1|, \dots, |r_{HT}^n - r_{ST}^n|\}).$$

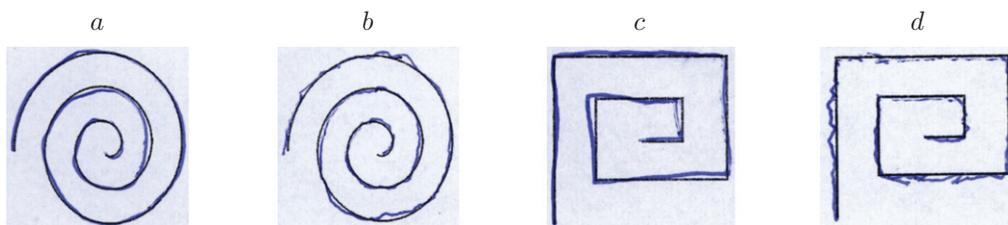


Рис. 1. Изображения спиралей и меандров участников эксперимента: a, c — рисунки участника из контрольной группы; b, d — рисунки пациента с БП [16]

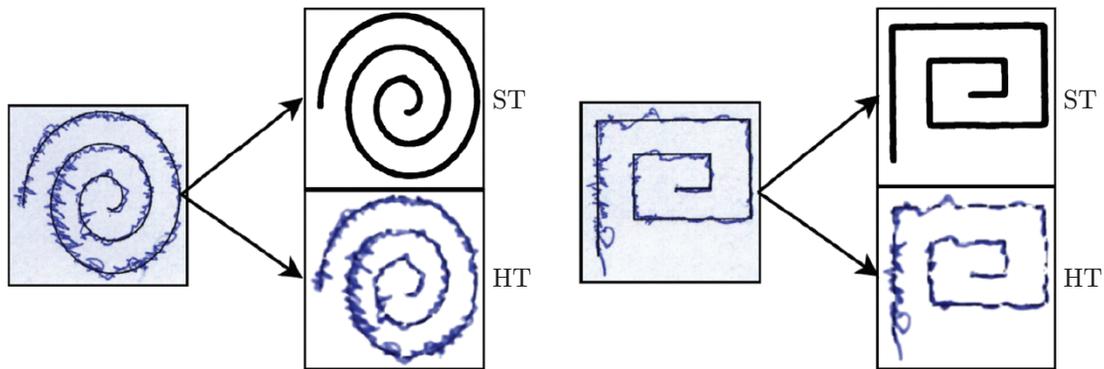


Рис. 2. Результаты извлечения шаблона (ST) и рукописной фигуры (HT) на этапе предварительной обработки [16]

Признак x_4 — стандартное отклонение между r_{HT}^i и r_{ST}^i .

Признак x_5 — средний относительный тремор, определяемый как средняя разница между радиусом точек рукописного образца и их ближайших соседей слева:

$$x_5 = \frac{1}{n-d} \sum_{i=d}^n |r_{HT}^i - r_{HT}^{i-d+1}|,$$

где d определяет позицию соседней точки слева, в работе [16] значение равнялось 10.

Признаки x_6 и x_7 — максимальное и минимальное значения радиуса HT, x_8 — стандартное отклонение радиуса HT.

Признак x_9 — количество раз, когда разница между радиусом ST и HT меняется с отрицательной на положительную и наоборот.

Описание эксперимента и оценок эффективности моделей диагностики. Для оценки предлагаемых методов построения модели диагностики БП применялась схема эксперимента, описанная в [16] как «Experiment 2». Эта схема предполагает использование всех признаков и 20-кратную перекрёстную проверку. Таблица наблюдения состоит из 18 признаков и 368 образцов. Первые 9 признаков — это признаки изображения спирали, следующие 9 — признаки изображения меандра. Образцом является набор из признаков спирали и меандра, нарисованных одним субъектом. Модель диагностики представляет собой классификатор, определяющий по входному образцу данных, относится ли субъект к больным или нет. В качестве оценок эффективности модели используются три показателя:

1) точность диагностирования больного:

$$acc_p = \frac{\hat{N}_p}{N_p},$$

где \hat{N}_p — число правильно определённых пациентов с БП, N_p — общее число пациентов с БП;

2) точность диагностирования участника из контрольной группы:

$$acc_c = \frac{\hat{N}_c}{N_c},$$

где \hat{N}_c — число правильно диагностированных контрольных участников, N_c — количество человек в контрольной группе;

3) средняя точность:

$$acc = \frac{acc_p + acc_c}{2}.$$

Задачей построения модели диагностики является получение наибольшей средней точности.

Методы и алгоритмы. Представим использованные для построения моделей диагностики БП алгоритмы и методы машинного обучения.

Исправление дисбаланса данных. Результат создания классификатора зависит не только от эффективности алгоритмов его построения, но и от того, достаточно ли данных для обучения классификатора. При наличии дисбаланса в данных процесс обучения затрудняется, так как классические алгоритмы построения отдают приоритет классам с наибольшим представительством. Для исправления такой ситуации применяются инструменты, позволяющие сбалансировать число образцов: алгоритмы удаления экземпляров наибольших классов (андерсэмплинг), алгоритмы добавления искусственных образцов данных (оверсэмплинг), а также их разнообразные комбинации [18]. Поскольку сокращение числа образцов неактуально для данных малого общего объёма, были опробованы четыре алгоритма генерации экземпляров наименьшего класса.

Простейший алгоритм случайного добавления данных (Random over-sampling, ROS) на каждом шаге генерирует дубликат экземпляра положительного класса, выбранного случайным образом, до тех пор, пока количество образцов не сравняется. Несмотря на простоту метода, его применения бывает достаточно для данных с небольшим дисбалансом и без существенного перекрытия классов [18]. Наиболее известным алгоритмом добавления данных и одним из самых распространённых инструментов преодоления дисбаланса данных в целом является SMOTE (Synthetic Minority Oversampling Technique) [19]. В нём генерация нового синтетического образца положительного класса происходит путём осуществления интерполяции между некоторым экземпляром положительного класса и случайно выбранным экземпляром этого же класса среди k ближайших соседей. Количество соседей k задаётся пользователем. Оригинальный SMOTE не пользуется информацией о распределении классов, что может привести к синтезу экземпляров, перекрывающих области отрицательного (наибольшего) класса.

Для исправления недостатков этого алгоритма разработан ряд модификаций. Например, пограничный SMOTE (Borderline-SMOTE, bSMOTE) генерирует новые образцы только для тех экземпляров, среди ближайших соседей которых есть, как минимум, один экземпляр отрицательного класса [20]. Ещё одним потомком SMOTE является алгоритм ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning). Генерация новых образцов в ADASYN происходит в первую очередь для самых сложных экземпляров: сложность экземпляра тем выше, чем больше среди его k ближайших соседей образцов противоположного класса. Количество генерируемых экземпляров для рассматриваемого экземпляра пропорционально его сложности [21].

Отбор признаков. Поиск подмножества признаков, позволяющего при сокращении размерности данных достичь точности, как минимум, сопоставимой с полным множеством признаков, является важным этапом построения классификатора. Оценка информативности каждого отдельного признака или их подмножеств считается полезным средством анализа предметной области, так как вносит ясность о степени важности переменных при определении вывода модели.

Для проведения отбора признаков в данной работе применяются два подхода: алгоритмы фильтрации и алгоритмы обёртки. В качестве фильтров использованы инструменты из открытой библиотеки scikit-learn [22], отбирающие признаки на основе критерия χ^2 , взаимной информации (mutual_info) и дисперсионного анализа (f_classif). Алгоритмы обёртки представлены алгоритмом рекурсивного исключения признаков (Recursive Feature Elimination, RFE) из библиотеки scikit-learn и оригинальными бинарными метаэвристиками. Достоинством обёрток является способность подбирать набор признаков под конкретный классификатор. Они оперируют векторами вида $\mathbf{P} = (p_1, p_2, \dots, p_n)$, где p_i — бинарное значение, отражающее участие i -го признака в процессе классификации (единица) или его игнорирование (ноль), $i = 1, \dots, n$, где n — количество признаков в исходном наборе. Бинарные метаэвристики в течение заданного количества итераций осуществляют поиск вектора \mathbf{P} , максимально улучшающего целевую функцию. После окончания итераций выдаётся лучшее найденное решение [23].

Нечёткие классификаторы, предусматривающие работу с неполным набором признаков, состоят из базы продукционных правил, оценивающих значение признаков объекта $\mathbf{x} = (x_1, x_2, \dots, x_n)$ нечёткими терминами:

$$R_j : \text{ ЕСЛИ } p_1 \wedge x_1 = T_{1j}, \text{ И } p_2 \wedge x_2 = T_{2j}, \text{ И } \dots, \text{ И } p_n \wedge x_n = T_{nj}, \text{ ТО } \text{class} = c_j,$$

где T_{ij} — нечёткое множество (терм), характеризующее i -й признак в правиле R_j , $j = 1, \dots, R$ (R — количество правил); c_j — индекс класса j -го правила. Запись $p_i \wedge x_i$ вводится для указания использовать ($p_i = 1$) или игнорировать ($p_i = 0$) i -й признак в процессе нечёткого вывода.

Нечёткий вывод, т. е. определение для некоторого входного вектора признаков \mathbf{x} наиболее подходящего ему класса, заключается в выполнении следующих шагов. На первом осуществляется расчёт степени принадлежности вектора \mathbf{x} к перечню термов каждого правила. Для j -го правила степень принадлежности $\mu_j(\mathbf{x})$ определяется формулой

$$\mu_j(\mathbf{x}) = \prod_{i=1}^n \mu_{T_{ij}}(x_i),$$

где $\mu_{T_{ij}}(x_i)$ — степень принадлежности признака x_i нечёткому терму T_{ij} . На втором шаге вычисляются степени уверенности в принадлежности входного вектора \mathbf{x} к каждому имеющемуся классу. Для k -го класса ($k = 1, \dots, M$, где M — число классов) степень уверенности $\beta_k(\mathbf{x})$ рассчитывается путём сложения степени принадлежности объекта к правилам, консеквент которых соответствует k -му классу:

$$\beta_k(\mathbf{x}) = \sum_{(j|c_j=k)} \mu_j(\mathbf{x}).$$

Третий шаг заключается в определении итогового вывода путём нахождения класса, к которому \mathbf{x} получил максимальную уверенность в принадлежности:

$$\text{class} = \arg \max_{k=1, \dots, M} \{\beta_k(\mathbf{x})\}.$$

Методика построения нечёткого классификатора алгоритмом экстремальных значений признаков классов и метаэвристикой для настройки параметров термов. Данная методика включает три этапа. Первый заключается в создании структуры классификатора алгоритмом экстремальных значений признаков классов (АЭПК) [23]. Алгоритм разбивает обучающие данные по классам, определяет для каждого признака внутри класса

экстремумы и строит между ними покрывающий терм. Таким образом получается база правил минимально возможного объёма, поскольку на один класс приходится ровно одно правило, в котором каждый признак представлен одним термом. Второй этап — отбор признаков, который проводится с помощью бинарной метаэвристики в том случае, если до построения классификатора не был использован алгоритм фильтрации. Третий этап необходим для уточнения термов с целью улучшения точности. Для подбора их параметров применяются непрерывные версии метаэвристик: алгоритм птиц (Bird Swarm Optimization, BSO) [24], алгоритм ласточек (Swallow Swarm Optimization, SSO) [25], гравитационный алгоритм (Gravitational Search Algorithm, GSA) [23].

Методика построения нечёткого классификатора алгоритмом k -средних и метаэвристическим алгоритмом. В данной методике на первом шаге для генерации базы правил нечёткого классификатора применяется гибридный алгоритм кластеризации данных на основе алгоритма k -средних и метаэвристического алгоритма китов (Whale Optimization Algorithm, WOA) [26]. Максимальный размер базы правил, который мог быть получен на каждой отдельной выборке данных, был ограничен 10 правилами. Второй шаг методики — выбор оптимального набора признаков — осуществлялся с помощью бинарной версии WOA в режиме обёртки. На последнем шаге осуществляется оптимизация параметров термов непрерывной версией WOA.

Генетические нечёткие системы. Нечёткие системы, созданные с помощью генетического алгоритма, получили название «генетические нечёткие системы» (Genetic Fuzzy System, GFS). В представленной работе был применён генетический алгоритм для генерации нечётких классификаторов согласно подходу [27]. В этих работах использовалась многокритериальная оптимизация для обучения, обеспечивающая компромисс между интерпретируемостью и точностью получаемых классификаторов.

Хромосома (индивид) в популяции представляет отдельный нечёткий классификатор. Каждый ген хромосомы кодирует функцию принадлежности определённого правила целочисленным значением; количество возможных вариантов этого значения зависит от нечёткого разбиения (грануляции) пространства изменения переменной. Например, если пространство разбито на три нечётких множества, которые ассоциируются с лингвистическими гранулами (термами) Маленький, Средний и Большой, то значение гена может быть выбрано из множества $\{0, 1, 2, 3\}$, где 0 обозначает отсутствие функции принадлежности, остальные значения кодируют соответствующие функции принадлежности.

В данной работе используется многокритериальный генетический алгоритм NSGA II (Nondominated Sorting Genetic Algorithm II) [28]. В качестве критериев оптимизации выступают: ошибка классификации, количество термов во всей базе правил, количество нечётких правил. Все вышеописанные критерии должны быть минимизированы. Алгоритм возвращает популяцию классификаторов, находящихся на Парето-фронте данных критериев. На результирующем Парето-фронте выбирается решение для классификатора с наименьшей ошибкой классификации.

Эксперимент и обсуждение. Для оценки эффективности нечётких классификаторов и методов их построения было проведено несколько экспериментов, направленных на достижение лучшей точности классификации и на определение информативных признаков. Первый эксперимент был проведён без применения алгоритмов оверсэмплинга. Во втором эксперименте на этапе предварительной обработки использовались четыре алгоритма оверсэмплинга для решения проблемы дисбаланса классов, поскольку набор данных был создан на основе выполненных заданий 74 пациентов с БП и только 18 здоровых. Для сохранения возможности сравнения с результатами оригинальной публикации [16] и работ [4, 13], использующих базу HandPD и аналогичный набор признаков в предлагаемой работе, применялся метод двадцатикратной перекрёстной проверки. Таблицы с полным описанием результатов экспериментов представлены в [29]. В табл. 1 приведены варианты

Таблица 1

Перечень вариантов построения нечёткого классификатора

№	Обозначение	Генерация	R	Отбор признаков	Настройка термов
1	extr- χ^2 -sso	АЭПК	2	χ^2	SSO
2	extr-bin-sso	АЭПК	2	бин. SSO	SSO
3	extr- χ^2 -gsa	АЭПК	2	χ^2	GSA
4	extr-bin-gsa	АЭПК	2	бин. GSA	GSA
5	extr- χ^2 -bso	АЭПК	2	χ^2	BSO
6	extr-bin-bso	АЭПК	2	бин. SSO	BSO
7	kmeans-woa	k-ср.-WOA	2–10	бин. WOA	WOA
8	GFS	GFS	3–8	—	—

Таблица 2

Значения точности *acc* в %

№	Алгоритм	Normal	ADASYN	bSMOTE	SMOTE	ROS
1	extr- χ^2 -sso	66,55	66,61	67,49	65,28	66,42
2	extr-bin-sso	59,09	63,40	64,72	64,69	63,06
3	extr- χ^2 -gsa	64,13	68,37	68,09	68,70	68,04
4	extr-bin-gsa	71,91	62,11	66,50	64,04	65,52
5	extr- χ^2 -bso	70,00	70,22	66,82	67,96	69,40
6	extr-bin-bso	69,87	61,92	67,50	66,02	70,89
7	kmeans-woa	58,70	62,43	61,48	64,45	61,81
8	GFS	67,80	69,09	65,73	66,48	67,06

построения нечётких классификаторов, опробованные в экспериментах. В качестве термов использовались функции Гаусса.

В табл. 2 приведены результаты построения классификаторов на исходных данных (Normal) и с использованием алгоритмов генерации экземпляров наименьшего класса. Для сравнения результатов приведён ряд аналогичных классификаторов из работ, в которых эксперимент был построен в полном соответствии с [16], где набор данных HandPD и был представлен впервые. В [16] были достигнуты следующие значения точности (%): оптимальный лес — 57,54, байесовский классификатор — 59,20, метод опорных векторов (SVM) — 66,72. В [13] классификаторы построены с помощью каскадной модели обучения, которая включает статистический критерий χ^2 , применяющийся для отбора признаков, и один из следующих алгоритмов классификации: наивный байесовский классификатор (GNB) с точностью 68,97, деревья решений (DT) — 65,09, линейный дискриминантный анализ (LDA) — 74,00, классификатор k -ближайших соседей (KNN) — 63,85, SVM — 58,65 и адаптивный бустинг (AdaBoost) — 76,44. Кроме этого, для балансировки обучающих данных использовалось случайное удаление экземпляров наибольшего класса. Классификатор, построенный в работе [4] с помощью методов генетического программирования (GP), достиг точности 76,60.

Методы глубокого обучения для обработки графических изображений HandPD также применялись при построении классификаторов диагностики БП. В работе [30] применялась непрерывная свёрточная сеть, результаты которой достигли 85,7 % общей точности. Здесь используется общая точность, т. е. процентное количество правильно классифицированных участников эксперимента, а не средняя точность по классам.

На исходных данных пять нечётких классификаторов из восьми превзошли по точности лучший результат [16]. Использование алгоритмов оверсэмплинга позволило повысить

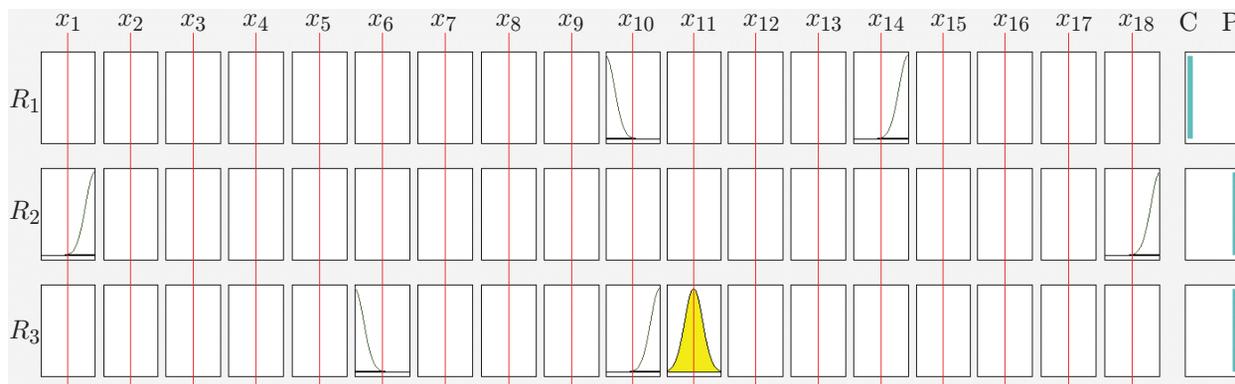


Рис. 3. База знаний нечёткого классификатора диагностики БП

точность классификации, при этом только один нечёткий классификатор показал меньшую точность по сравнению с лучшим результатом [16]. Классификаторы на основе каскадной модели обучения (χ^2 -LDA, χ^2 -AdaBoost) и генетического программирования показали лучшие по точности результаты по сравнению с предлагаемыми в данной работе нечёткими классификаторами. Однако при сравнении важно учитывать не только точность, но и объяснимость/интерпретируемость модели, особенно в медицине. Правила из базы знаний нечётких классификаторов могут быть легко интерпретированы, а полученный результат объяснён конечному пользователю. Интерпретация результата классификаторов линейного дискриминантного анализа (χ^2 -LDA) и адаптивного бустинга (χ^2 -AdaBoost) весьма затруднительна. Что касается генетического программирования, то считается, что данные модели интерпретируемы, но уровень их интерпретации ниже, чем у нечётких систем. В подтверждении этого далее приводятся базы знаний для диагностики БП классификаторов обоих типов (оба классификатора основываются на правилах).

На рис. 3 представлена одна из полученных в ходе эксперимента база знаний нечёткого классификатора, состоящая из трёх правил. Последний столбец на рисунке определяет метку класса правила: Р — наличие болезни Паркинсона, С — отсутствие болезни Паркинсона.

Извлечённые правила интерпретируются следующим образом.

Правило 1. ЕСЛИ среднее квадратичное значение различия между шаблонной и рукописной линиями на рисунке меандра является Маленьким И средний относительный тремор на рисунке меандра является Большим, ТО болезнь Паркинсона отсутствует.

Правило 2. ЕСЛИ среднее квадратичное значение различия между шаблонной и рукописной линиями на рисунке спирали является Большим И количество изменений знака различия между радиусами шаблонной и рукописной линиями на рисунке меандра является Большим, ТО болезнь Паркинсона присутствует.

Правило 3. ЕСЛИ максимальный радиус рукописной линии спирали является Маленьким И среднее квадратичное значение различия между шаблонной и рукописной линиями на рисунке меандра является Большим И максимальное различие между шаблонной и рукописной линиями на рисунке меандра является Средним, ТО болезнь Паркинсона присутствует.

Правило для классификатора GP, построенного на основе генетического программирования, выглядит следующим образом [12]:

ЕСЛИ $(2 \times x_1 + x_5 - x_7) \leq (x_1 + x_4)$, ТО болезнь Паркинсона отсутствует, ИНАЧЕ болезнь Паркинсона присутствует.

Здесь x_1, x_4, x_5, x_7 — вышеописанные признаки рисунка меандра.

Таблица 3

Значения рангов признаков

Признак	Спираль		Меандр	
	Исходный набор	Оверсэмплинг	Исходный набор	Оверсэмплинг
x_1	14,38	14,46	17	16,69
x_2	8,44	7,96	13,81	12,25
x_3	11,31	11,91	13,19	11,62
x_4	4,88	5,44	5,69	5,89
x_5	9,25	9,22	9,19	11,09
x_6	7,12	6,98	9,5	10,82
x_7	6,56	8,56	6,44	6,19
x_8	6,31	6,28	12,31	9,85
x_9	8,25	7,75	7,38	8,04

В силу малого объёма данных подтвердить на статистически значимом уровне увеличение точности при использовании алгоритмов оверсэмплинга не представляется возможным. Однако для алгоритмов $\text{extr-}\chi^2\text{-gsa}$ и kmeans-woa прирост точности в среднем составил 4,17 и 3,84 % соответственно. Все четыре алгоритма оверсэмплинга с точки зрения критерия Фридмана статистически неразличимы на уровне значимости 0,05 (p – value = 0,199). Средние ранги алгоритмов следующие: ADASYN — 3,00, SMOTE — 2,88, ROS — 2,38, bSMOTE — 1,75.

Для формирования системы диагностики БП по рукописным данным необходимо выявить наиболее информативные признаки в этих данных. Отбор признаков выполнялся методами фильтров и обёрток на оригинальном наборе данных и на наборах, сгенерированных алгоритмами оверсэмплинга. Значения степени важности признаков приведены в [29]. В табл. 3 указаны значения рангов, найденных в результате применения двухфакторного рангового дисперсионного анализа Фридмана для связанных выборок. Нулевая гипотеза — все признаки имеют одинаковые средние значения — отвергнута на уровне значимости 0,05 (p – value < 0,001).

Ранжирование признаков без оверсэмплинга:

$$x_1^M \succ x_1^C \succ x_2^M \succ x_3^M \succ x_8^M \succ x_3^C \succ x_6^M \succ x_5^C \succ x_5^M \succ \\ \succ x_2^C \succ x_9^C \succ x_9^M \succ x_6^C \succ x_7^C \succ x_7^M \succ x_8^C \succ x_4^M \succ x_4^C,$$

ранжирование признаков с оверсэмплингом:

$$x_1^M \succ x_1^C \succ x_2^M \succ x_3^C \succ x_3^M \succ x_5^M \succ x_6^M \succ x_8^M \succ x_5^C \succ \\ \succ x_7^C \succ x_9^M \succ x_2^C \succ x_9^C \succ x_6^C \succ x_8^C \succ x_7^M \succ x_4^M \succ x_4^C.$$

Сумма рангов признаков спирали равна 78,56, а признаков меандра — 92,44. Таким образом, по сумме общих рангов признаки меандра более информативны, чем признаки спирали. Наиболее информативным признаком является среднеквадратичное значение различия (x_1), наименее информативным признаком — стандартное отклонение (x_4).

Статистическое сравнение по критерию Уилкоксона значений рангов с оверсэмплингом и без оверсэмплинга позволило сделать вывод об их статистической неразличимости на уровне значимости 0,05 (p – value = 0,948). Отсюда можно сделать вывод о том, что алгоритмы оверсэмплинга не вносят статистически значимых искажений в ранжирование признаков.

Заключение. В представленной работе рассмотрены три методики построения нечётких классификаторов и их применение для диагностики болезни Паркинсона на основе статических рукописных данных на примере набора данных HandPD. Для решения проблемы несбалансированности данных в исходном наборе применены четыре алгоритма оверсэмплинга. Отбор информативных признаков в рукописных данных выполнен с применением методов обёртки и фильтрации. Для оценки эффективности предложенных методов использованы непараметрические критерии Фридмана и Уилкоксона. Сравнительный анализ точности представленных в работе классификаторов и моделей, описанных в [20], позволил сделать вывод о превосходстве нечётких классификаторов.

Новизна исследования заключается в следующих уникальных результатах.

1. Впервые для диагностирования болезни Паркинсона использован нечёткий классификатор с оригинальными алгоритмами построения, позволяющий на естественном языке интерпретировать полученный результат (выявлены закономерности в виде правил).

2. Выделены информативные признаки в общеизвестном наборе рукописных данных для диагностики болезни Паркинсона; впервые проведено ранжирование таких признаков. Практическое значение полученных закономерностей заключается в том, что рисование пациентами спиралей и меандров может быть выполнено в неавтоматизированном режиме; проведённое ранжирование указывает на признаки, на которые в первую очередь необходимо акцентировать внимание практикующего врача при анализе рисунков пациентов.

3. Для несбалансированных медицинских рукописных данных пациентов с БП применены алгоритмы оверсэмплинга; впервые статистически доказано, что обоснованное добавление синтетических данных не влияет на распределение рангов признаков, т. е. не меняет семантику исходных данных. Получено новое знание и предложен научный подход к проверке условно-реальных данных.

Новые выборки, полученные с применением алгоритмов оверсэмплинга и методов фильтров, могут быть использованы исследователями, работающими с набором данных HandPD, для апробации разрабатываемых ими классификаторов.

Финансирование. Работа выполнена при поддержке Российского научного фонда (проект № 22-21-00021).

СПИСОК ЛИТЕРАТУРЫ

1. **Lamba R., Gulati T., Al-Dhlan K. A., Jain A.** A systematic approach to diagnose Parkinson's disease through kinematic features extracted from handwritten drawings // *Journ. Reliab. Intell. Environ.* 2021. **7**, N 3. P. 253–262.
2. **Раздорская В. В., Воскресенская О. Н., Юдина Г. К.** Болезнь Паркинсона в России: распространённость и заболеваемость (обзор) // *Саратовский научно-медицинский журнал.* 2016. **12**, N 3. С. 379–384.
3. **Belič M., Bobič V., Atanasijević M. B. et al.** Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease — a review // *Clin. Neurol. Neurosurg.* 2019. **184**, N 9. P. 105442.
4. **Senatore R., Marcelli A.** A paradigm for emulating the early learning stage of handwriting: Performance comparison between healthy controls and Parkinson's disease patients in drawing loop shapes // *Hum. Mov. Sci.* 2019. **65**. P. 89–101.
5. **Parziale A., Senatore R., Della Cioppa A., Marcelli A.** Cartesian genetic programming for diagnosis of Parkinson disease through handwriting analysis: Performance vs. interpretability issues // *Artif. Intell. Med.* 2021. **111**. P. 101984.
6. **Casalino G., Castellano G., Castiello C., Mencar C.** Effect of fuzziness in fuzzy rule-based classifiers defined by strong fuzzy partitions and winner-takes-all inference // *Soft Comput.* 2022. **26**, N 14. P. 6519–6527.

7. **Oxtoby N. P., Alexander D. C.** Imaging plus X: Multimodal models of neurodegenerative disease // *Curr. Opin. Neurol.* 2017. **30**, N 4. P. 371–379.
8. **Tautan A.-M., Ionescu B., Santarnecchi E.** Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques // *Artif. Intell. Med.* 2021. **117**. P. 102081.
9. **Myszczyńska M. A., Ojamies P. N., Lacoste A. M. B. et al.** Applications of machine learning to diagnosis and treatment of neurodegenerative diseases // *Nat. Rev. Neurol.* 2020. **16**. P. 440–456.
10. **Drotár P., Mekyska J., Rektorová I. et al.** A new modality for quantitative evaluation of Parkinson's disease: In-air movement // *Proc. of the 13th IEEE Int. Conf. on BioInformatics and BioEngineering*. Chania, Greece, 10-13 Nov., 2013. P. 1–4.
11. **Isenkul M., Sakar B., Kursun O.** Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease // *Proc. of the 2nd Int. Conf. on E-Health and TeleMedicine*. Istanbul, Turkey, 22-24 May, 2014. P. 171–175.
12. **Senatore R., Della Cioppa A., Marcelli A.** Automatic diagnosis of neurodegenerative diseases: An evolutionary approach for facing the interpretability problem // *Information*. 2019. **10**. P. 30.
13. **Ali L., Zhu C., Golilarz N. A. et al.** Reliable Parkinson's disease detection by analyzing handwritten drawings: Construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model // *IEEE Access*. 2019. **7**. P. 116480–116489.
14. **Pereira C. R., Weber S. A. T., Hook C. et al.** Deep learning-aided Parkinson's disease diagnosis from hand-written dynamics // *Proc. of the 29th SIBGRAPI Conf. on Graphics, Patterns and Images*. Sao Paulo, Brazil, 04-07 Oct., 2016. P. 340–346.
15. **Kamran I., Naz S., Razzak I., Imran M.** Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease // *Future Gener. Comput. Syst.* 2021. **117**. P. 234–244.
16. **Pereira C. R., Pereira D. R., Silva F. A. et al.** A new computer vision-based approach to aid the diagnosis of Parkinson's disease // *Comput. Methods Programs Biomed.* 2016. **136**. P. 79–88.
17. **Zhang T. Y., Suen C. Y.** A fast parallel algorithm for thinning digital patterns // *Commun. ACM*. 1984. **27**, N 3. P. 236–239.
18. **Бардамова М. Б., Ходашинский И. А.** Формирование структуры нечёткого классификатора комбинацией алгоритма экстремумов классов и алгоритма «прыгающих лягушек» для несбалансированных данных с двумя классами // *Автометрия*. 2021. **57**, № 4. С. 54–64. DOI: 10.15372/AUT20210407.
19. **Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P.** SMOTE: synthetic minority over-sampling technique // *Journ. Artif. Intell. Res.* 2002. **16**. P. 321–357.
20. **Han H., Wang W. Y., Mao B. H.** Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning // *Lect. Notes Comput. Sci.* 2005. **3644**. P. 878–887.
21. **He H., Bai Y., Garcia E. A., Li S.** ADASYN: adaptive synthetic sampling approach for imbalanced learning // *Proc. of the 5th IEEE Int. Joint Conf. on Neural Networks*. Sydney, Australia, 1-8 June, 2008. P. 1322–1328.
22. **Machine Learning in Python (scikit-learn).** API Reference. Feature Selection. URL: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection (дата обращения: 10.02.2022).
23. **Bardamova M., Konev A., Hodashinsky I., Shelupanov A.** A fuzzy classifier with feature selection based on the gravitational search algorithm // *Symmetry*. 2018. **10**, N 11. P. 609.

24. **Светлаков М. О.** Отбор признаков для нечёткого классификатора с использованием мета-эвристического алгоритма “bird swarm” // Сб. избр. стат. научн. сессии ТУСУР. Ч. 2. Томск: В-Спектр, 2021. С. 230–233.
25. **Hodashinsky I., Sarin K., Shelupanov A., Slezkin A.** Feature selection based on swallow swarm optimization for fuzzy classification // Symmetry. 2019. **11**, N 11. P. 1423.
26. **Koryshev N., Hodashinsky I., Shelupanov A.** Building a fuzzy classifier based on whale optimization algorithm to detect network intrusions // Symmetry. 2021. **13**, N 7. P. 1211.
27. **Nishihara A., Masuyama N., Nojima Y., Ishibuchi H.** Michigan-style fuzzy genetics-based machine learning for class imbalance data // Journ. Japan Soc. Fuzzy Theory Intell. Inform. 2021. **33**, N 1. P. 525–530.
28. **Deb K., Agrawal S., Pratap A., Meyarivan T.** A fast and elitist multiobjective genetic algorithm: NSGA-II // IEEE Trans. Evol. Comput. 2002. **6**, N 2. P. 182–197.
29. **Нечёткие классификаторы для диагностики болезни Паркинсона на основе статических рукописных данных. Результаты.** Mendeley Data. URL: <https://data.mendeley.com/datasets/xj7cpx6z9z/> (дата обращения: 25.07.2022).
30. **Li Z., Yang J., Wang Y. et al.** Early diagnosis of Parkinson’s disease using continuous convolution network: Handwriting recognition based on off-line hand drawing without template // Journ. Biomed. Inform. 2022. **130**. P. 104085.

Поступила в редакцию 17.10.2022

После доработки 21.11.2022

Принята к публикации 29.12.2022
