Применение методов машинного обучения в поиске статистических паттернов для диагностики обсессивно-компульсивного расстройства

© *В.А. Юнусов, С.А. Демин*

Казанский федеральный университет, Казань, Россия
E-mail: valentin.yunusov@gmail.com

Поступило в Редакцию 5 мая 2023 г.
В окончательной редакции 29 июня 2023 г.
Принято к публикации 30 октября 2023 г.

Одной из актуальных задач современных наук о данных является определение диагностических критериев психических расстройств. Эта задача усложняется наличием множества биофизических параметров, часть из которых может быть избыточной. Нами применяются методы отбора признаков для диагностики обсессивно-компульсивного расстройства. С помощью методов машинного обучения на первом этапе решена задача классификации для исходного набора признаков, а на втором этапе — отбора подмножеств наиболее значимых диагностических признаков для волонтеров, у которых существенно проявлялись симптомы указанного расстройства, и представителей контрольной группы.

Ключевые слова: живые системы, обсессивно-компульсивное расстройство, биомедицинские данные, методы машинного обучения, отбор признаков.

DOI: 10.61011/PJTF.2023.23.56840.13A

Определение диагностических критериев патологических изменений в функционировании мозга человека, например, при неврологических заболеваниях и психиатрических расстройствах является важной проблемой современных наук о данных и биофизики. Одним из распространенных психических отклонений является обсессивно-компульсивное расстройство (ОКР). Для ОКР характерно наличие обсессий и компульсий. Обсессии — навязчивые, повторяющиеся и неприятные мысли, побуждения, вызывающие тревогу. Компульсии — повторяющиеся действия или умственные ритуалы, которые выполняются для уменьшения стресса, вызванного навязчивыми идеями.

Для определения диагностических критериев данного заболевания широко используются методы статистического анализа сигналов электроэнцефалограмм (ЭЭГ) и/или магнитоэнцефалограмм и др. [1,2]. Фиксация большого числа экспериментальных данных функциональной активности мозга человека способствовала активному развитию методов машинного обучения для решения нейрофизиологических и биофизических задач [3–5]. Методы машинного обучения позволяют обнаружить скрытые закономерности, автоматизировать и ускорить процессы классификации и отбора признаков в исходных биомедицинских данных. Для решения подобных задач разрабатываются программные пакеты и библиотеки на различных языках программирования.

В настоящей работе используется программный пакет Weka, предназначенный для предобработки и анализа данных, в том числе методами машинного обучения [6]. Для проведения процедуры отбора значимых признаков в данном пакете используется комбинация методов поиска и средств оценки значимости атрибутов (признаков). Метод поиска используется в пространстве признаков для нахождения подходящего подмножества признаков. Оценщик признаков — это метод, с помощью которого каждая функция оценивается в контексте целевой переменной.

В данном исследовании мы применяем методы отбора подмножества атрибутов: CfsSubsetEval и CorrelationAttributeEval. CfsSubsetEval оценивает значимость подмножества атрибутов, рассматривая индивидуальную прогностическую способность каждой функции, а также степень избыточности. В итоге получается подмножество признаков только с сильной корреляцией с целевым классом [7]. CorrelationAttributeEval для получения значимого подмножества характеристик оценивает коэффициент корреляции Пирсона относительно целевого класса для каждой переменной [7].

Экспериментальные данные были получены ранее в результате международного сотрудничества с Голдсмитским колледжем Лондонского университета. Файлы данных представляли собой сигналы ЭЭГ для двух групп испытуемых: 15 испытуемых, у которых значительно проявлялись признаки обсессивно-компульсивных расстройств, и 15 человек, у которых эти признаки проявлялись незначительно (условно контрольная группа). В дополнение для всех испытуемых была проведена оценка согласно OCI-R (обновленный обсессивно-компульсивный опросник) — заполняемой анкете, оценивающей показатели ОКР по шести областям симптомов (далее наименования приведены в сокращенном виде: например, мытье рук — преувеличенный страх загрязнения и т.д.): мытье, проверка, упорядочивание, одержимость, накопление и умственная нейтрализация [8]. ЭЭГ-записи фиксировались при трех условиях: фаза чтения, фаза визуализации и фаза подавления. В первой фазе испытуемые повторяли вслух определенное пред-

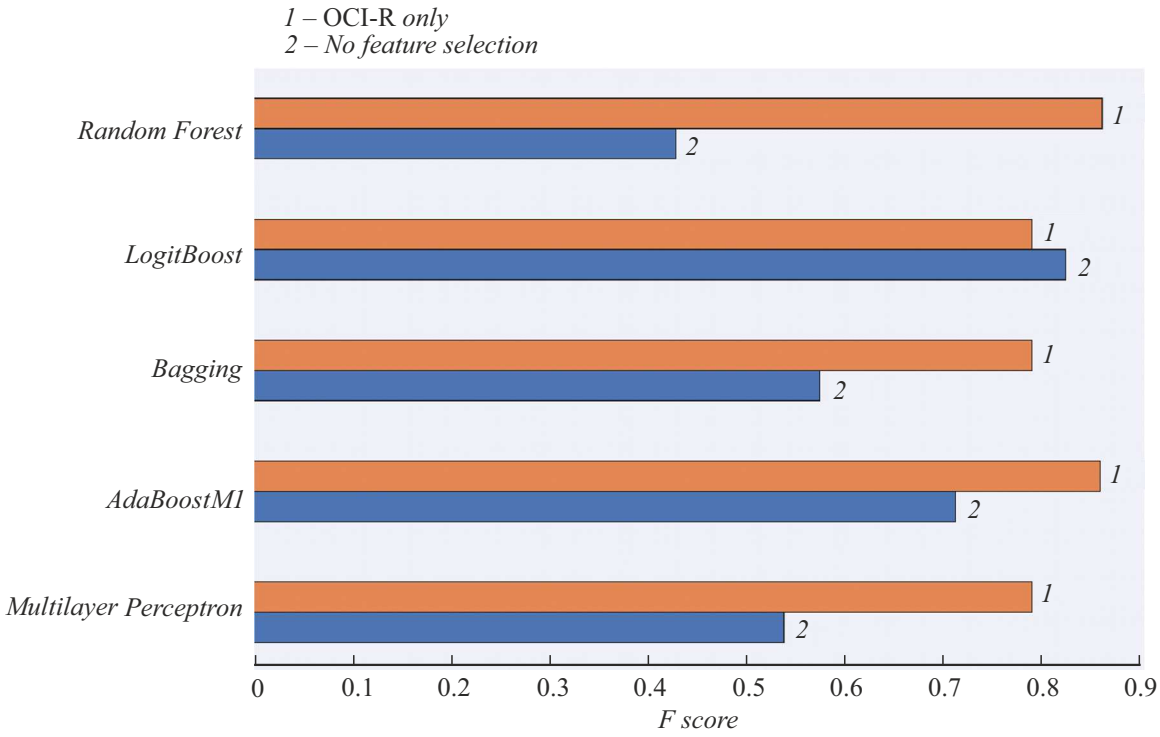

Рис. 1. *F*-мера классификаторов без применения методов отбора признаков для полного набора статистических параметров и подмножества параметров OCI-R.

ложение с описанием некоторого события, используя в нем имя друга или члена семьи. Во второй фазе участники исследования визуализировали сказанное ими событие в течение 1 min. В последней фазе в течение 1 min испытуемые должны были думать о чем угодно, кроме этого события. Биоэлектрическая активность с разных областей коры головного мозга фиксировалась электродами, расположенными в соответствии с расширенной международной схемой размещения „10–20%“ [9].

Проводимое нами исследование включало два этапа. На первом этапе для каждой записи ЭЭГ был рассчитан набор статистических показателей: параметры Хьорта (активность, сложность и мобильность), мощность α-, β-, θ- и δ-активности коры головного мозга, анализ колебаний без тренда (DFA), фрактальная размерность Хигучи, сложность Лемпеля–Зива, фрактальная размерность Петросяна и выборочная энтропия. Для полученных параметров посредством пяти методов машинного обучения, реализованных в программном пакете Weka, решалась задача классификации ЭЭГ-записей испытуемых по группам: с низкой и высокой степенью проявления симптомов ОКР. Было проведено сравнение эффективности методов машинного обучения для распределения испытуемых с различным проявлением ОКР на две группы. Точность классификации оценивалась с помощью *F*-меры (гармоническое среднее между точностью и полнотой классификатора) и AUC ROC — параметр, описывающий взаимосвязь между чувствительностью модели (доля истинно положительных примеров) и ее специфичностью (описываемой в отношении долей ложноположительных результатов). Нами установлено, что для большинства методов *F*-мера и AUC ROC принимают большие значения при учете только OCI-R параметров (рис. 1). Максимальная *F*-мера (0.856) и AUC ROC (0.946) достигались методом Random Forest. Вследствие небольшого объема рассматриваемого набора данных деление выборки осуществлялось с помощью стратифицированной кросс-валидации с делением на пять блоков, при применении которой эмулируется наличие тестовой выборки.

На втором этапе методами отбора признаков CorrelationAttributeEval и CfsSubsetEval определялись характеристики, которые вносили наиболее значимый вклад в классификацию. Отбор подмножеств значимых признаков выполнялся для возможного повышения точности классификаторов, так как избыточные (или шумовые) переменные могли искажать прогнозируемую величину.

Для метода CorrelationAttributeEval в подмножество параметров входили те из них, для которых коэффициент корреляции Пирсона, рассчитанный для целевого класса, был достаточно высоким (табл. 1). Для метода CfsSubsetEval исходя из ранее указанных механизмов оценки значимости атрибутов в искомое подмножество вошли сложность Хьорта для электрода O_2, δ-активность для электрода PO_8, θ-активность для электрода O_z, β-активность для электродов CP_3 и AF_8, а также по-

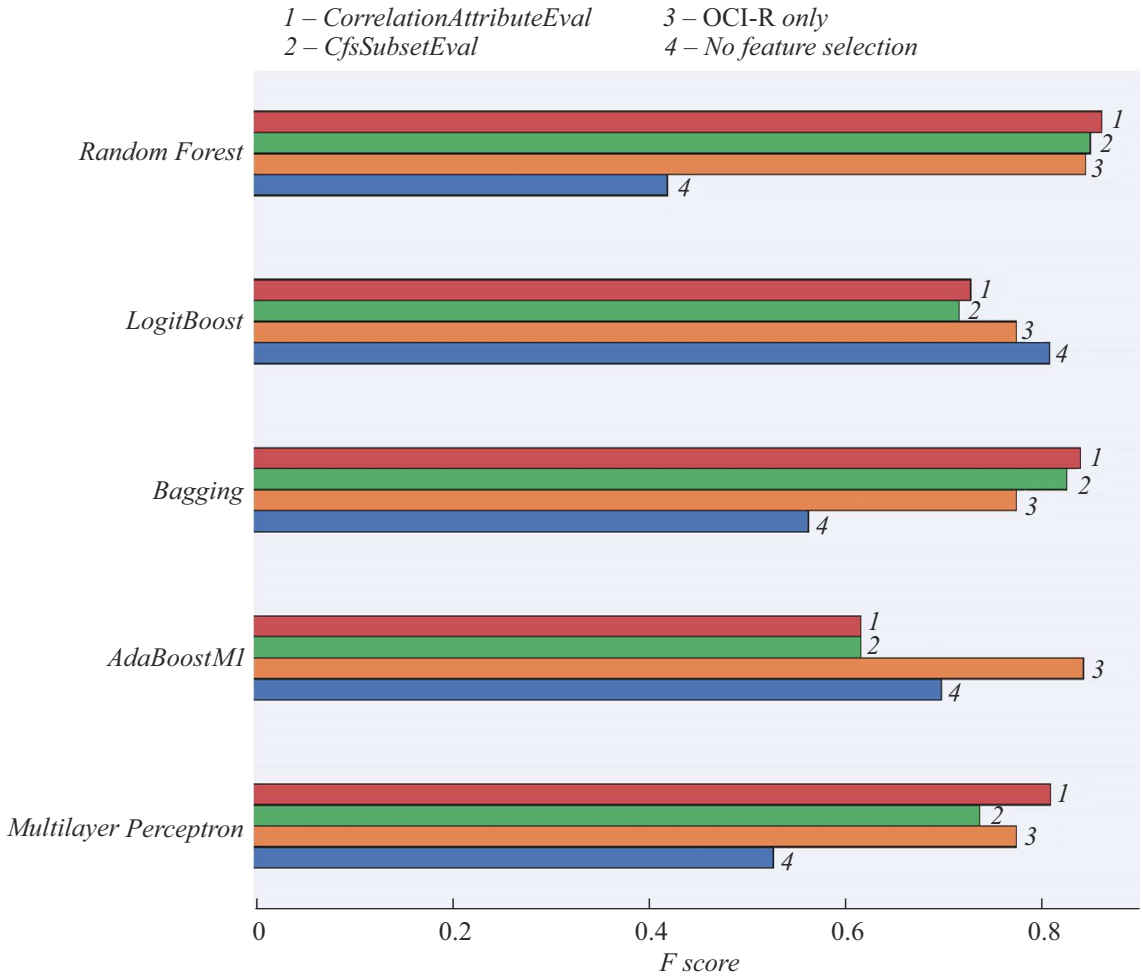

Рис. 2. *F*-мера для классификаторов с использованием различных методов выбора признаков.

Таблица 1. Характеристики с наибольшим коэффициентом корреляции Пирсона для целевого класса (оценщик CorrelationAttributeEval)

Характеристика	Коэффициент Пирсона
Проверка	0.733
Одержимость	0.685
Упорядочивание	0.628
δ-активность для электрода PO_8	0.585
β-активность для электрода O_2	0.541
Умственная нейтрализация	0.536
Мытье	0.533
DFA для электрода O_z	0.504

казатели OCI-R: проверка, упорядочивание, умственная нейтрализация и одержимость.

После отбора признаков была проведена повторная классификация для полученных подмножеств (рис. 2). Сравнение *F*-метрики и метрики AUC ROC классификаторов с отбором признаков и без него показывает, что для одних методов точность повышается, а для других снижается (табл. 2, 3). В целом *F*-мера и AUC ROC без применения методов отбора признаков для полного набора статистических параметров ниже, чем для подмножеств выделяемых признаков. Для метода Random Forest *F*-мера и мера AUC ROC классификатора максимальны для всех подмножеств; наибольших значений 0.873 и 0.969 соответственно они достигают при использовании метода атрибуции CorrelationAttributeEval. Следует отметить, что при уменьшении числа параметров в методе Random Forest точность растет. Однако в данном случае существенно, что при использовании методов отбора признаков точность метода выросла не только относительно полного набора характеристик, а, кроме того, еще и относительно набора характеристик OCI-R, содержащего меньшее количество признаков.

В настоящей работе мы применили методы машинного обучения в сочетании с методами выбора признаков, включенными в программный пакет Weka. Мы рассчитали набор статистических признаков для сигналов ЭЭГ людей с ОКР и представителей контрольной группы. Была выполнена задача классификации признаков, а

Таблица 2. *F*-мера классификатора для различных методов отбора признаков

Классификатор	*F*-мера			
	Без отбора признаков	Только характеристики OCI-R	CorrelationAttributeEval	CfsSubsetEval
Random Forest	0.426	0.856	0.873	0.861
LogitBoost	0.819	0.785	0.738	0.726
Bagging	0.571	0.785	0.851	0.837
AdaBoostM1	0.708	0.854	0.625	0.625
Multilayer Perceptron	0.535	0.785	0.82	0.747

Таблица 3. AUC ROC-метрика классификатора для различных методов отбора признаков

Классифиатор	AUC ROC			
	Без отбора признаков	Только характеристики OCI-R	CorrelationAttributeEval	CfsSubsetEval
Random Forest	0.617	0.946	0.969	0.935
LogitBoost	0.811	0.809	0.936	0.8
Bagging	0.648	0.924	0.926	0.909
AdaBoostM1	0.77	0.688	0.849	0.688
Multilayer Perceptron	0.577	0.883	0.898	0.803

также отбор наиболее значимых признаков для двух групп испытуемых.

Полное исходное пространство характеристик избыточно для исследуемой задачи. Было установлено, что показатели, определяемые в OCI-R-методике, вносят значительный вклад в работу моделей машинного обучения. Кроме того, статистические параметры биоэлектрических сигналов затылочной области коры головного мозга также значительным образом влияют на проводимый отбор признаков. Наилучшего результата удалось достичь при помощи метода отбора признаков CorrelationAttributeEval и классификатора Random Forest, значение *F*-метрики для которых составило 0.873, а значение метрики AUC ROC было равно 0.969.

Дальнейшая верификация полученных результатов подразумевает вовлечение большего числа волонтеров с разным уровнем проявления ОКР-симптомов, а также расширение опросников, измерителей и шкал для выявления и мониторинга ОКР. В дальнейших исследованиях при увеличении рассматриваемого набора данных для более объективного анализа можно будет применить метод случайного смешивания. Применение методов машинного обучения с подкреплением методами отбора признаков при статистической обработке сигналов биоэлектрической активности мозга человека будет способствовать автоматизированному поиску диагностических критериев психиатрических расстройств, нейродегенеративных и неврологических заболеваний [10], а также повышению точности и ускорению постановки диагнозов.

Финансирование работы

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета („Приоритет-2030“).

Соблюдение этических стандартов

Все работы, входящие в состав исследования с участием людей, выполнены в соответствии с этическими стандартами национального Комитета по научной этике, а также с Хельсинкской декларацией 1964 г. и ее последующими дополнениями или с аналогичными этическими стандартами. Информированное добровольное согласие было получено от каждого участника исследования.

Конфликт интересов

Авторы заявляют, что у них нет конфликта интересов.

Список литературы

[1] S.A. Demin, R.M. Yulmetyev, O.Yu. Panischev, P. Hänggi, Physica A, **387** (8-9), 2100 (2008). DOI: 10.1016/j.physa.2007.12.003

[2] V.A. Yunusov, S.A. Demin, O.Yu. Panischev, N.Y. Demina, J. Phys.: Conf. Ser., **2103** (1), 012044 (2022). DOI: 10.1088/1742-6596/2103/1/012044
[3] K. Hilbert, T. Jacobi, S.L. Kunas, B. Elsner, B. Reuter, U. Leuken, N. Kathmann, Psychother. Res., **31** (1), 52 (2021). DOI: 10.1080/10503307.2020.1839140
[4] F. Ferreri, A. Bourla, C.S. Peretti, T. Segawa, N. Jaafari, S. Mouchabac, J. Med. Internet Res., **6** (12), e11643 (2019). DOI: 10.2196/11643
[5] M. Hoexter, E. Miguel, J. Diniz, R. Shavitt, G. Busatto, J. Sato, J. Affect. Disord., **150** (3), 1213 (2013). DOI: 10.1016/j.jad.2013.05.041
[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, ACM SIGKDD Explor. Newsl., **11** (1), 10 (2008). DOI: 10.1145/1656274.1656278
[7] M. Hall, *Correlation-based feature subset selection for machine learning*, Ph.D. thesis (University of Waikato, Hamilton, New Zealand, 1999).
[8] E.B. Foa, J.D. Huppert, S. Leiberg, R. Langner, R. Kichic, G. Hajcak, P.M. Salkovskis, Psychol. Assess., **14** (4), 485 (2002). DOI: 10.1037/1040-3590.14.4.485
[9] R. Jones, J. Bhattacharya, J. Behav. Addict., **1** (3), 96 (2012). DOI: 10.1556/JBA.1.2012.005
[10] С.А. Демин, О.Ю. Панищев, С.Ф. Тимашев, Р.Р. Латыпов, Изв. РАН. Сер. физ., **84** (11), 1569 (2020). DOI: 10.31857/S0367676520110083 [S.A. Demin, O.Yu. Panischev, S.F. Timashev, R.R. Latypov, Bull. Russ. Acad. Sci. Phys., **84** (11), 1349 (2020). DOI: 10.3103/S1062873820110088].