Рахматова А.Ю.[1], Косаченко А.И.[1], Москалева А.С.[1], Львова О.А.[1,2], Сергеева М.В.[3], Сергеев А.П.[1,3]

Точность методов Случайный лес и Многослойный персептрон в задаче прогнозирования исходов детских ишемических инсультов

1 - Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, 2 - ФГБОУ ВО Уральский государственный медицинский университет Минздрава РФ, 3 - Институт промышленной экологии Уральского отделения Российской академии наук, г. Екатеринбург

Rakhmatova A.Yu., Kosachenko A.I., Moskaleva A.S., Lvova O.A., Sergeeva M.V., Sergeev A.P.

Accuracy of Random forest method and Multilayer perceptron method in predicting of the outcomes in pediatric ischemic stroke

Резюме

Проведено сравнение точности предсказания двух методов машинного обучения: Случайный лес (СЛ) и Многослойный персептрон (МСП) в задаче прогнозирования исходов «инвалидность» и «отсутствие инвалидности» детских ишемических инсультов (ИИ). Выборка представляет собой серию клинических случаев: 172 ребенка с ишемическим инсультом, доказанным по данным МРТ головного мозга. В качестве предикторов были использованы данные генетических исследований на носительство полиморфизмов 8 генов тромбофильного спектра: FGB:-455G>A, F2:20210G>A, F5:1691G>A, F7:10976G>A, F13:103G>T, ITGA2:807C>T, ITGB3:1565T>C, PAI-1:-675 5G>4G, и 4 генов фолатного цикла: MTHFR:677C>T, MTHFR:1298A>C, MTRR:66A>G, MTR:2756A>G. МСП продемонстрировал более высокие показатели правильных распознаний исходов, чем случайный лес (0,88 против 0,67 соответственно).

Ключевые слова: детский ишемический инсульт, генетический полиморфизм, случайный лес, многослойный персептрон, прогноз исхода

Summary

Authors made the comparison between two methods (random forest and multilayer perceptron) to forecast the outcome of the pediatric ischemic stroke. Two options of the outcome were estimated: disability and the absence of disability. Case series included 172 patients data base, all patients had MRI confirmation of stroke and signed concern form. Eight thrombophilic genes polymorphisms: FGB:-455G>A, F2:20210G>A, F5:1691G>A, F7:10976G>A, F13:103G>T, ITGA2:807C>T, ITGB3:1565T>C, PAI-1:-675 5G>4G, and four genes polymorphisms of folic acid enzymes: MTHFR:677C>T, MTHFR:1298A>C, MTRR:66A>G, MTR:2756A>G were established as feasible predictors. Multilayer perceptron method showed higher rates of correct recognition of the outcomes, than random forest method (0,88 vs 0,67).

Key words: pediatric ischemic stroke, genes polymorphism, random forest, multilayer perceptron, prognosis of outcome

Введение

Детский ишемический инсульт (ИИ) - заболевание, приводящее к формированию тяжелых исходов: моторному и интеллектуальному дефициту, эпилепсии и другим дефицитарным состояниям. В настоящее время методы прогнозирования исходов детских ИИ на основе анализа наличия генетических полиморфизмов разработаны, на наш взгляд, недостаточно. Предикторы неблагоприятного прогноза заболевания мало изучены; роль полиморфизмов генов в процессе формирования тяжелого неврологического дефицита до сих пор остается невыясненной [1, 2]. В то же время прогнозирование исходов детских ИИ на как можно более ранних стадиях болезни остается важной задачей, поскольку позволяет рационально и эффективно выстраивать терапевтическую тактику на всех этапах болезни. При решении задачи прогнозирования одной из ключевых трудностей является поиск оптимального метода машинного обучения. За последние десятилетия было разработано большое число таких методов, что значительно затрудняет их выбор [3].

Среди множества методов машинного обучения методы Случайный лес (СЛ) и Искусственная нейронная сеть (ИНС) показывают лучшие результаты, причем СЛ по некоторым данным, занимает лидирующую позицию [3, 4, 5].

Случайный лес (Random Forest) – метод, основанный на построении ансамбля деревьев решений, каждое из которых строится по выборке, полученной из исходной с помощью процедуры изъятия с возвращением – метод бэггинга (bagging, bootstrap aggregation). Помимо бэггинга, для построения каждого из деревьев ансамбля применяется метод случайных подпространств (random subspace method, RSM). Построение деревьев происходит на разных подмножествах - используется $\sqrt{p}$ предикторов, где p – общее число предикторов [6].

Искусственная нейронная сеть (Artificial Neuron Network) – последовательность функций, преобразующих входные факторы в выходные. ИНС является вычислительной системой с большим числом параллельно функционирующих простых процессоров (нейронов) и множеством связей. Искусственная нейронная сеть не программируется в обычном смысле этого слова, а обучается. Обучение сети состоит в подстройке весовых коэффициентов каждого искусственного нейрона. В настоящей работе рассматривается Многослойный персептрон (МСП) – один из многочисленных вариантов ИНС.

Цель исследования – сравнить показатели прогностической точности методов Случайный лес и Многослойный персептрон в задаче предсказания исходов детского ИИ: «инвалидность» и «отсутствие инвалидности» по данным генетического профиля детей.

Материалы и методы

Выборка: 172 ребенка с доказанным ИИ в соответствии с критериями включения-исключения.

Критерии включения: возраст от 0 до 15 лет; диагноз острого нарушения мозгового кровообращения по ишемическому типу (I63.0-I64.9, G45.0-45.9 по МКБ-10) [8], подтвержденный по клиническим данным, результатам компьютерной томографии (КТ) и/или МРТ головного мозга; письменное информированное согласие родителей или их законных представителей на участие в исследовании.

Критерии исключения: отказ пациентов и/или их родителей от обследования; отсутствие точного диагноза острых нарушений мозгового кровообращения (дети на этапе дифференциальной диагностики); внутричерепные кровоизлияния любой этиологии; симметричные перивентрикулярные ишемические очаги, лейкомаляции и пери-, интравентрикулярные кровоизлияния как морфологический субстрат перинатального поражения центральной нервной системы (ЦНС); возраст дебюта инсульта старше 15 лет.

Состав выборки: «девочки без инвалидности» - 30 человек (0,17), «девочки с инвалидностью» - 31 человек (0,18), «мальчики с инвалидностью» - 51 человек (0,30), «мальчики без инвалидности» - 60 человек (0,35).

У всех детей была проведена оценка носительства полиморфизмов 12 генов: 8 тромбофилии (FGB:-455G>A, F2:20210G>A, F5:1691G>A, F7:10976G>A, F13:103G>T, ITGA2:807C>T, ITGB3:1565T>C, PAI-1:-675 5G>4G) и 4 фолатного цикла (MTHFR:677C>T, MTHFR:1298A>C, MTRR:66A>G, MTR:2756A>G) в образцах крови методом полимеразной цепной реакции.

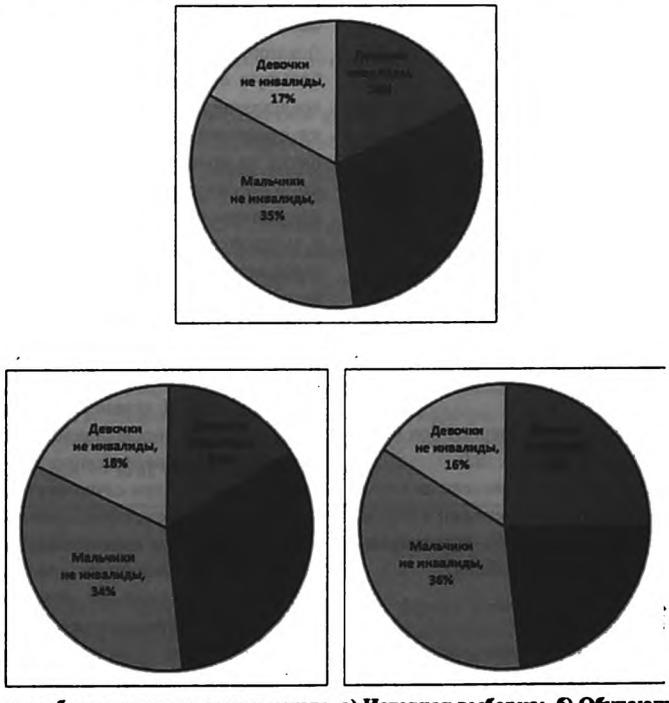

Рисунок 1. Состав выборок с учетом пола и исхода. а) Исходная выборка; б) Обучающая подвыборка; в) Тестовая подвыборка

Таблица 1. Состав выборок

Выборка	Объем	Девочки не инвалиды, доли	Девочки инвалиды, доли	Мальчики не инвалиды, доли	Мальчики инвалиды, доли
Исходная	172	0,17	0,18	0,35	0,30
Обучающая	120	0,18	0,16	0,34	0,32
Тестовая	52	0,16	0,25	0,36	0,23

Таблица 2. Критерии качества моделей на тестовой выборке

Метод	Точность, %	Чувствительность, %	Специфичность, %
Случайный лес (СЛ)	67	43	76
Многослойный персептрон (МСП)	88	73	90

Варианты исходов (инвалидность или отсутствие инвалидности) были оценены спустя не менее 4 лет после ИИ. Критерии установления инвалидности оценивали в соответствии с приказом от 29 сентября 2014 г. № 664н [9] (наличие справки серии МСЭ).

Исходные деперсонализованные данные были предоставлены ДГКБ №9 и представляют собой таблицу MS Excel, содержащую различные анамнестические данные о пациенте с перенесенным ИИ и его биологических родителях. На основе этой таблицы была построена база данных в MS Access, в которой содержалась информация по полу, и возрасту. Кроме того, в базу были внесены данные об оценке носительства 12 вышеуказанных полиморфизмов генов. Отсутствие полиморфизма кодировалось нулем, наличие полиморфизма в одной из аллелей генов – единицей, наличие полиморфизма в двух аллелях – двойкой.

Из исходной выборки (Рисунок 1а) были сгенерированы обучающая (70%) и тестовая (30%) подвыборки. Для генерации случайных номеров записей был создан код VBA, который генерировал два массива: в первом содержались номера записей базы по порядку (N записей), второй был пуст. Следующим шагом была генерация случайного числа из набора номеров базы. Число перемещалось в пустой массив, в исходном оставались все записи, кроме сгенерированного числа (n-1 записей). Путем 172 итераций был получен массив со случайными номерами записей без повторений. После этого были вызваны записи с соответствующими номерами, которые впоследствии были разделены в заданном процентном соотношении для создания обучающей (Рисунок 1б) и тестовой (Рисунок 1в) подвыборок (Таблица 1).

Случайный лес и Многослойный персептрон были реализованы в приложении STATISTICA 12.0.

Оценка качества моделей была проведена на тестовой выборке. Для каждого метода критериями качества моделей выступали точность (precision), чувствительность (sensitivity) и специфичность (specificity).

Результаты и обсуждение

Число деревьев в методе случайного леса выбиралось исходя из величины коэффициента неправильной классификации. Как видно из Рисунка 2, коэффициент неправильной классификации для тестовой выборки минимален при ансамбле из 62 деревьев решений.

На Рисунке 3 представлена полученная важность предикторов для метода Случайный лес. Как видно из рисунка, наибольшую прогностическую ценность представляют 4 гена: ITGA2:807C>T, MTRR:66A>G, PAI-1:-675 5G>4G, MTR:2756A>G.

Архитектура МСП (MLP 38-10-2 с экспоненциальной функцией активации) была выбрана на основе максимизации показателей точности ее предсказания на тестовой выборке.

В таблице 2 приведены показатели качества моделей, которые демонстрируют значительное преимущество ИНС типа МСП перед Случайным лесом по всем трем выбранным показателям качества. Чувствительность метода Случайного леса (43%) делает его непригодным для целей прогноза исхода «инвалидность»: в рассматриваемом случае метод Случайный лес может быть использован только как «специфичный» классификатор. Искусственная нейронная сеть МСП имеет большую специфичность, что делает ее пригодной для предсказания исходов обоих типов. Оба метода обладают относительно низкой чувствительностью по сравнению с их специфичностью. Это может говорить о том, что в рассматриваемой выборке отсутствуют «сильные» предикторы инвалидизации. Однако способность МСП с высокой вероятностью правильно отклонять прогноз «инвалид» указывает на наличие в этой выборке «сильных» предикторов восстановления утраченных в результате ИИ функций.

Заключение

Метод МСП лучше справляется с задачей классификации исхода «инвалидность» или «отсутствие инвалидности» после перенесенного ИИ в детском возрасте в сравнении с методом случайного леса по всем показателям. Таким образом, искусственная нейронная сеть типа МСП, может быть использована в качестве и чувствительного, и специфичного классификатора, в то время как метод Случайный лес можно использовать только в качестве классификатора специфичного, несмотря на то, что исследования [3, 4] говорят о лидирующей позиции Случайного леса по сравнению с множеством других методов машинного обучения, в частности, искусственными нейронными сетями.■

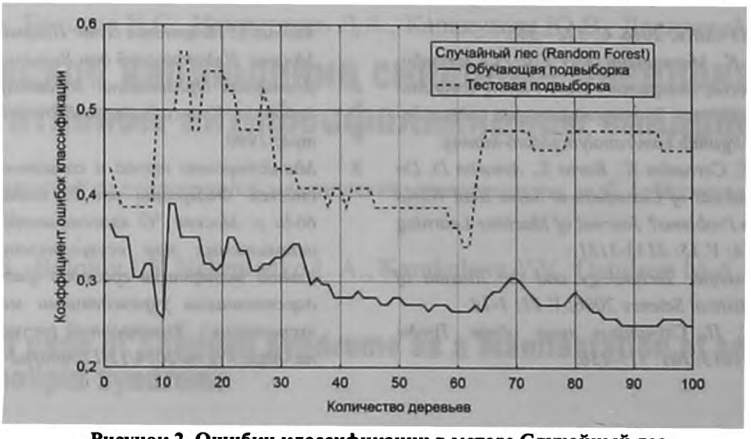

Рисунок 2. Ошибки классификации в методе Случайный лес

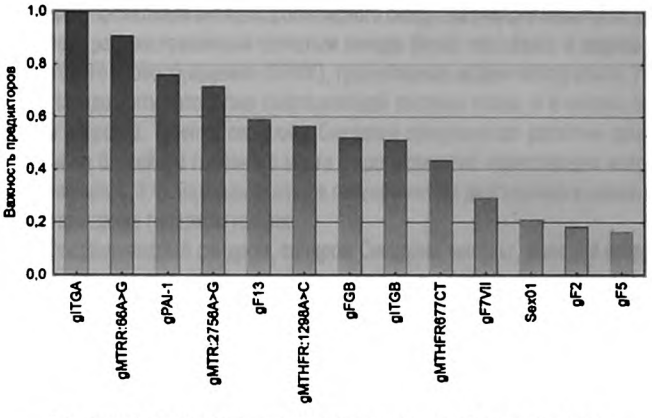

Рисунок 3. Важность предикторов в методе Случайный лес

Рахматова А.Ю. *– студент кафедры технической физики ФТИ Уральского федерального университета имени первого Президента России Б.Н. Ельцина г. Екатеринбург.* ***Косаченко А.И.*** *– студент кафедры технической физики ФТИ Уральского Уральского федерального университета имени первого Президента России Б.Н. Ельцина г. Екатеринбург.****Москалева А.С.*** *- студент кафедры технической физики ФТИ Уральского Уральского федерального университета имени первого Президента России Б.Н. Ельцина г. Екатеринбург.****Сергеева М.В.*** *– научный сотрудник Лаборатории устойчивого развития территорий Института промышленной экологии Уральского отделения Российской академии наук, г. Екатеринбург.* ***Сергеев А.П.*** *– к.ф.-м.н., заведующий Лаборатории физики и экологии Института промышленной экологии Уральского отделения Российской академии наук, г. Екатеринбург; доцент Департамента информационных технологий и автоматики ИРИТ-РТФ Уральского федерального университета имени первого Президента России Б.Н. Ельцина г. Екатеринбург. Автор, ответственный за переписку: Львова О.А. - к.м.н., ведущий научный сотрудник лаборатории мозга и нейрокогнитивного развития Уральского федерального университета имени первого Президента России Б.Н. Ельцина г. Екатеринбург; доцент кафедры психиатрии ФГБОУ ВО Уральский государственный медицинский университет Минздрава РФ, г. Екатеринбург; адрес для переписки: 620014, г. Екатеринбург, ул. Репина, д. 3, телефон +7 9222 093259, olvova@bk.ru*

Литература:

1. *Львова О.А., Сергеев А.П., Рахматова А.Ю. Изучение влияния полиморфизмов прокоагулянтных генов на исход ишемического инсульта у детей методом логистической регрессии // Актуальные вопросы современной медицинской науки и здравоохранения: Материалы II Международной (72 Всероссийской) научно-практической конференции молодых ученых и студентов г. Екатеринбург: Изд-во УГМУ; 2017. С. 445-449.*
2. *Инсульты у детей / В.П. Зыков, И.Б. Комарова, Е.Н. Дьяконова и др. / В кн.: Федеральное руководство по детской неврологии; под ред. профессора Гузевой*

В.И. М.: ООО «МК»; 2016. С. 323 - 360.

3. *Шитиков В. К., Мастицкий С. Э. (2017) Классификация, регрессия, алгоритмы Data Mining с использованием R. 351 с. - Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining.*
4. *Delgado M.F., Cernadas E., Barro S., Amorim D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 2014; V. 15: 3133-3181.*
5. *Hand D. Classifier Technology and the Illusion of Progress. Statistical Science 2006; V. 21: 1-14.*
6. *Чистяков С. П. Случайные леса: обзор. Труды КарНЦ РАН 2013; №1: 117-136.*
7. *Хайкин С. Нейронные сети: Полный курс / Хайкин С. Москва: Издательский дом Вильямс; 2006. 1104 с.*
8. *Всемирная Организация Здравоохранения Международная Классификация Болезней 10-го пересмотра; 1990.*
9. *Министерства труда и социальной защиты Российской Федерации от 29 сентября 2014 г. N 664н г. Москва "О классификациях и критериях, используемых при осуществлении медико-социальной экспертизы граждан федеральными государственными учреждениями медико-социальной экспертизы". Электронный ресурс, режим доступа: https://rg.ru/2014/12/12/mintrud-dok.html.*