Информационные процессы, Том 19, № 3, 2019, стр. 284–312.

Когнитивные аспекты Скользящего контроля и Деревьев решений

Е.А.Ващенко, М.А.Витушко, В.С.Переверзев-Орлов

Институт проблем передачи информации им. А.А. Харкевича, Российская академия наук, Москва, Россия

Поступила в редколлегию 31.05.2019

Аннотация—В статье представлена идея использования процедуры Скользящего Контроля (cross-validation) в качестве когнитивного инструмента, направленного на возможность смысловой интерпретации результатов машинного обучения, также позволяющего преодолеть некоторые проблемы, возникающие при использовании Скользящего Контроля (СК). Рассматриваются проблемы применения СК, связанные с трудностью получения устойчивых оценок в условиях нестационарности порождающих исследуемые данные процессов, а также некоторые нестандартные возможности, основанные на использовании процедуры СК. Цель - найти дополнительные способы оценивания результатов обучения, позволяющие целенаправленно улучшать построенные модели. На примере конкретных данных показаны приемы, позволяющие выявить неоднородность формируемых распознающих моделей, провести сопоставление их альтернативных вариантов, дать смысловую интерпретацию моделей, уточнить классификацию и сформировать гипотезы о не выявленных ранее подклассах.

КЛЮЧЕВЫЕ СЛОВА: машинное обучение, интерпретируемость, скользящий контроль, достаточность выборки, динамика моделей, устойчивость моделей, полнота данных, “расплетение” деревьев, информативные признаки деревьев.

1. ВВЕДЕНИЕ

При рассмотрении способов улучшения результатов применения методов машинного обучения немаловажно учитывать потенциал, связанный со смысловой интерпретацией получаемых моделей, а также с возможной корректировкой постановки задачи, ведущей к более адекватным результатам. Проблема интерпретируемости становится тем более актуальной, чем более сложные по своей внутренней структуре модели возникают в процессе развития методов машинного обучения. Всесторонние аспекты этой проблемы рассматриваются в работе [1].

Важность интерпретируемости видна в первую очередь в связи с двумя аспектами:

- пользователь модели склонен в большей степени доверять ей, если имеет возможность получить объяснение в какой-либо форме, оценить влияние различных факторов на принятие решения, сопоставить с собственным представлением;
- если пользователь способен получить понятную интерпретацию (например, оценить информативность используемых признаков), то это ему дает возможность содержательного анализа, направленного на улучшение качества модели (например, путем изменения состава используемых признаков) или даже корректировки постановки задачи (например, изменения классификации объектов на более адекватную).

Современные идеи в области интерпретации результатов машинного обучения изложены в статье [2]. Наиболее полное описание актуальных методов интерпретации и их классификация даны в книге [3].

Основные идеи современных методов интерпретации базируются на том, что целый ряд известных типов моделей (например, деревья решений или линейная регрессия) считаются интерпретируемыми по своей природе и если построить модель такого типа, "объясняющую" поведение рассматриваемой в качестве черного ящика модели, то анализ структуры объясняющей модели позволит дать интерпретацию. А наиболее универсальным приемом интерпретации является оценивание тем или иным способом важности используемых для принятия решений признаков. Так, одним из известных на сегодня методов является метод LIME [4], представляющий собой пример метода локальной интерпретации, т.е. объясняющего поведение модели на конкретных примерах. Методы же глобальной интерпретации направлены на объяснение поведения модели в целом. Примером известного способа глобальной интерпретации является подсчет важности признаков (feature importance) [5], основанный на оценке изменения ошибок модели при перемешивании значений признака в рамках обучающей выборки. Модельно-независимый, т.е. применимый к моделям произвольного типа вариант этого метода предложен в работе [6]. Примером метода интерпретации, представляющим унифицированный подход к объяснению результатов произвольных моделей, является SHAP (SHapley Additive exPlanations) [7]. Основная идея этого метода состоит в расчете вклада каждой переменной для всех возможных комбинаций других переменных в модели. Финальный вклад каждой переменной рассчитывается как средневзвешенное всех возможных вкладов. Сравнение и сопоставление нескольких различных техник интерпретации проводится в работе [8]. Упомянутые и им подобные методы реализованы и активно развиваются в рамках ряда программных библиотек, таких как LIME [9], SHAP [10], Skater [11], ELI5 [12].

Идея представляемой работы состоит в том, чтобы рассмотреть процедуру Скользящего Контроля (СК) в качестве когнитивного инструмента. С одной стороны - это поможет преодолеть проблемы СК, связанные с адекватностью получаемых оценок, опираясь на возможность смысловой интерпретации и объяснения получаемых результатов. С другой - мы можем использовать механизмы СК в качестве инструмента, обеспечивающего интерпретируемость и, в конечном итоге, - улучшение моделей.

Идея скользящего контроля возникла у М.Н.Вайнцвайга в конце 60-х годов. В 1969 году она в виде варианта теории была опубликована в [13] В.Л.Браиловским и А.Л.Лунцем, после чего практически сразу же стала одним из стандартов де-факто среди методов эмпирического оценивания качества алгоритмов обучения распознаванию образов по прецедентам, при полном игнорировании того факта, что в абсолютном большинстве практических задач, возникающих в этой области, этот подход не соответствует требованиям статистической устойчивости, независимости описаний и достаточности исследуемых экспериментальных данных [14].

На сайте MachineLearning.ru приводится исчерпывающее описание вариантов реализации СК с условиями, возможностями и проблемами их применения [15]. В основе процедуры СК лежит разбиение исходной выборки прецедентов на множество подмножеств, каждое из которых состоит из обучающей и экзаменационной частей, причем суммарный объем таких частей равен полной обучающей выборке, экзаменационные части не пересекаются, а их объединение образует полную выборку. На каждой из обучающих частей по некой стандартной процедуре с помощью как-то выбранного алгоритма обучения происходит формирование решающего правила, применяемого затем к соответствующей экзаменационной части. В качестве главного достоинства СК подчеркивается возможность получения с помощью этой технологии несмещенных оценок вероятности ошибок, выбора оптимальных алгоритмов и способов их применения, отбора информативных признаков и т.п. из стандартного набора пожеланий к обычным способам обучения распознаванию. Отмечаются также проблемы с оценками доверительных интервалов для получаемых оценок вероятности ошибок. В то же время, остаются недостаточно раскрытыми возможные проблемы, обусловленные нестационарностью данных.

Кроме того, все выводы делаются для вероятностей, а не частот. В реальных условиях выборки обычно малы, размерности пространств описаний велики, а классификации условны и данные не независимы. Эти обстоятельства заметно ограничивают возможности СК по оцениванию качества получаемых моделей.

В качестве цели описываемой работы рассматривалась задача найти дополнительные приемы, опирающиеся на процедуру СК, но позволяющие не просто оценить качество получаемых моделей, а также:

- обозначить границы применимости оценок СК;
- выдвинуть гипотезы о причинах недостаточного качества получаемых моделей;
- выявить смысловые структуры получаемых моделей, допускающие интерпретацию;
- обозначить пути улучшения качества моделей.

Рассматриваемые приемы можно отнести к глобальным модельно-зависимым способам интерпретации, однако, используемые подходы потенциально применимы ко многим типам моделей.

Полученные результаты позволяют потенциально рассматривать процедуру СК в качестве когнитивного инструмента исследования данных и задач, развивающего представления исследователей о решаемых ими проблемах.

С этой точки зрения, мы смогли выделить несколько моментов, с учетом которых применение СК позволяет обнаружить важное в несложных экспериментах. Учет этих особенностей дает возможность:

- оценить достаточность имеющихся данных и планируемых для использования алгоритмов обучения распознаванию для сколько-нибудь разумных решений стоящей практической задачи;
- построить простую и разумную оценку разброса получаемых решений в пределах имеющихся данных;
- выявить неоднородность формируемых при СК распознающих моделей;
- подвергнуть сомнению обоснованность использованной в решаемой задаче классификации данных, уточнить эту классификацию и сформировать гипотезы о не выявленных ранее подклассах.

Эти возможности рассматриваются на основе тестовых экспериментов, выполненных с помощью процедуры обучения распознаванию “Деревья решений” (Decision Trees) из Python-библиотеки “Scikit-learn” [17] и общедоступной базы данных “Pima Indians Diabetes Data Set”, собранной ранее в рамках исследования заболеваний диабетом в индейском поселении Pima, проводившегося в конце 80-х годов прошлого века [18].

Параметры и режим используемой процедуры и подробное описание выборки данных содержится в *Приложении А*.

2. ОЦЕНКА ДОСТАТОЧНОСТИ ВЫБОРКИ

Была проведена серия экспериментов с целью подтвердить связь между свойствами обучающей выборки и характером кривых обучения при сравнении таковых для оценок качества моделей, построенных на обучающей выборке со средними значениями для СК. В частности, проверялась возможность оценивать достаточность объема выборки. В этой серии в качестве переменной выступал объем рассматриваемой выборки, а зависимым от него считалось качество обучения в виде доли правильно распознаваемых объектов:

- на всей выборке;

– усредненное по всем обучающим частям СК;
– усредненное только по экзаменационным частям СК.

Это иллюстрируют графики на Рис. 1–2, цель которых – помочь оценить поведение “кривых обучения” (learning curve) [19, 20, 21] и степень соответствия этого поведения теоретическим ожиданиям в разных режимах обучения.

В первой части этой серии на упорядоченных по возрасту данных (Рис. 1) строились кривые при фиксированном объеме экзаменационных частей СК.

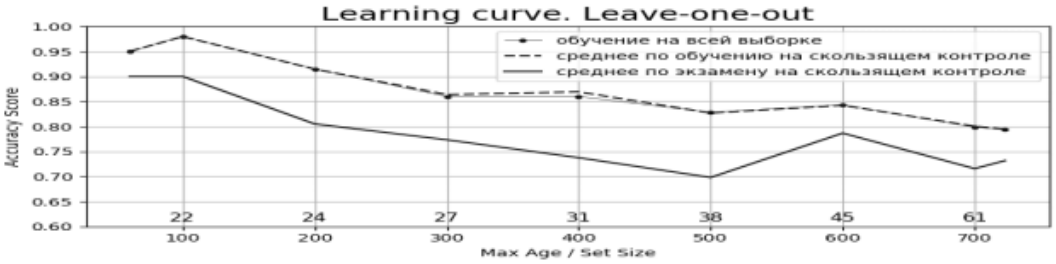


Рис. 1. Кривая обучения. Размер экзаменуемых частей СК (длина блока) = 1.

Видно, что по мере увеличения объема данных кривая качества экзамена в виде доли правильно распознаваемых объектов на экзаменационных долях в массивах СК в среднем монотонно приближается к результатам экзамена по всей обучающей выборке и усредненным результатам экзамена по обучающим массивам СК, оставаясь при этом заметно ниже. Даже при максимально доступных данных это различие не становится меньше 0.05 или 5%. В то же время для обучающих частей массивов СК и на полной выборке результаты практически совпадают по всему диапазону изменения объема данных. При этом видно, что, начиная с объема выборки, порядка 500 объектов из 724 вообще доступных, ситуация в целом стабилизируется, позволяя говорить об относительной достаточности этой выборки для формирования решающих правил такого рода на имеющихся данных, но из этого никак не следует, что в случае расширения набора данных эти результаты сохранились бы, поскольку источники данных в медицинских задачах обычно не обладает стационарностью ни в каком разумном смысле. При других объемах экзаменационных частей СК наблюдается похожая картина (см. *Приложение B*).

Во второй части экспериментов этой серии (Рис. 2) исследовалось поведение кривых качества древовидных решателей в зависимости от объема выборки при ее разбиении на 40 массивов для СК с выделением 1/40 объема данных в каждом из получаемых так массивов в качестве экзаменационных при сохранении неизменными значений остальных параметров, установленных в первой части этого эксперимента.

Общий ход “кривых обучения” очень похож на получаемое в экспериментах первой части (Рис. 1), повторяя в целом выводы, сделанные там. Важно заметить, что “кривые обучения”, полученные в этих двух сериях, демонстрируют высокую степень согласованности при отсутствии тенденции к их слиянию, что явно свидетельствует о том, что исходная выборка нестационарна и/или имеет недостаточный объем.

Таким образом, эти результаты, с одной стороны, подтверждают зависимость оценок СК от объема выборки, что позволяет ориентироваться, например, на кривые обучения при определении достаточности этого объема [23, 24, 25]. С другой стороны, остается открытым вопрос о качестве оценок СК в ситуации малого объема либо нестационарности обучающей выборки [22]. Ниже рассматриваются возможности, позволяющие надеяться, что при более деталь-

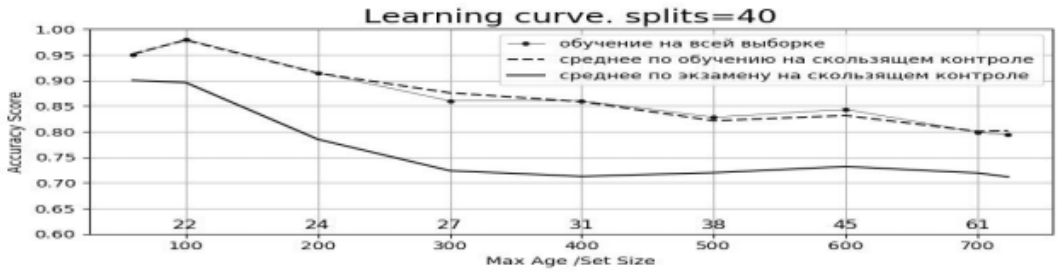


Рис. 2. Кривая обучения. 40 разбиений для формирования моделей и экзамена СК, 1/40 в каждом разбиении – для экзамена.

ном исследовании общее утверждение о нестационарности исследуемой выборки может быть дополнительно уточнено.

3. ОЦЕНКА РАЗБРОСА РЕЗУЛЬТАТОВ

Цель анализа данных – выявление в исследуемых наблюдениях таких закономерностей, которые позволяют предсказывать интересующие свойства новых наблюдений. Применительно к скользящему контролю обычно рассматриваются два вида таких закономерностей: оценка средней частоты правильных распознаваний и оценка доверительного интервала для средней частоты.

В этом разделе рассматриваются возможности СК для определения доверительного интервала для средней частоты правильных распознаваний с теми же оговорками относительно нестационарности источников данных с одним лишь уточнением, предполагающим, что изменения свойств процессов происходят сравнительно медленно. Это – достаточно серьезное ограничение, но оно позволяет локально экстраполировать оценки для средней частоты на новые данные допустимостью определенных вариаций частоты за пределами исходной области определения данных.

В ситуациях применения СК обычно приходится иметь дело с выборками данных сравнительно небольшого объема. В этих условиях как правило нет возможности оправданно применять гипотезы о допустимости параметрического оценивания и приходится использовать универсальные непараметрические методы [14, 15], основанные на использовании для формирования доверительных оценок разбиения множества анализируемых данных на множество подмножеств с последующим усреднением оценок по таким подмножествам. Удивительным образом это согласуется с базовой идеей самого скользящего контроля [13].

Основываясь на этом, мы провели серию экспериментов по определению доверительных интервалов на тех же данных о диагностике диабета, с которыми имели дело в предыдущем разделе, используя при этом ту же программу построения деревьев решений с теми же параметрами, но с некоторой доработкой ее, позволяющей расширить возможности работы с гистограммами средних частот по подвыборкам для СК.

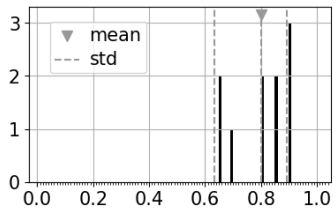
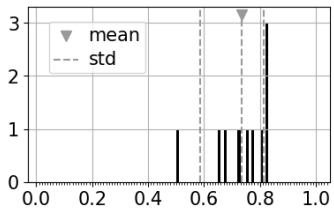
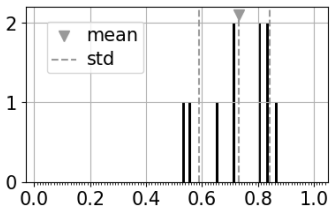
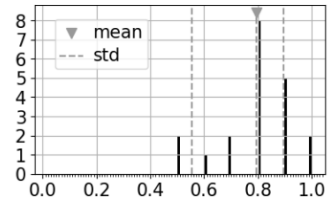
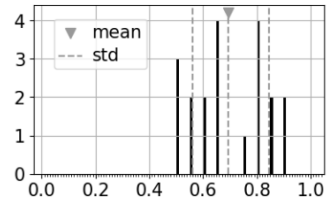
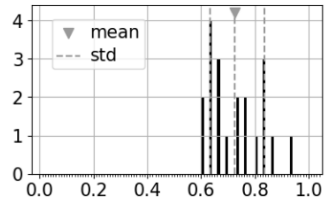
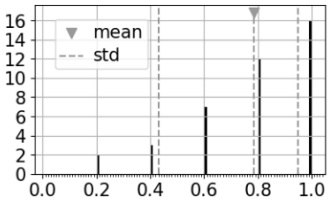
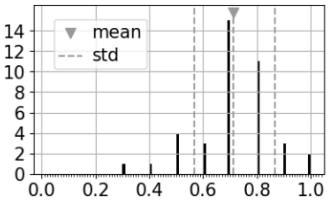
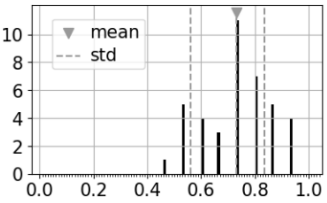
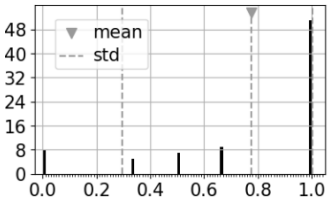
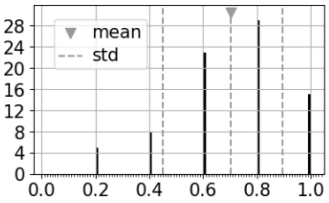
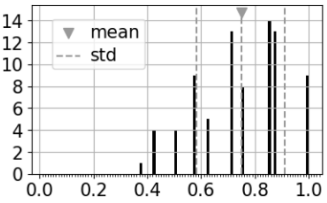
В первой части этих экспериментов формировались распределения долей правильно распознаваемых объектов, полученных на экзаменационных частях скользящего контроля при разных объёмах выборки, монотонно увеличиваемой с 200 до 400 и 600 объектов, и различном числе разбиений – на 10, 20, 40 и 80 блоков, включающих в каждом типе разбиений равные количества экзаменационных данных, в совокупности составляющих общую выборку.

В Таблице 1 приведены результаты этих экспериментов.

Видно, что все показанные здесь гистограммы весьма далеки от нормального распределения, предсказываемого теорией для множества оценок среднего по независимым подвыборкам.

Из этого можно заключить, что либо рассматриваемая выборка достаточно мала, либо это – следствие ее статистической неоднородности.

Таблица 1. Распределения долей правильно распознаваемых объектов, в зависимости от объема выборки и минимального числа объектов в узле.

Разбиения	Объем последовательно увеличивающейся выборки		
	200	**400**	**600**
10			
20			
40			
80			

В этой ситуации для оценок среднего качества формируемых решений следует использовать рекомендации из [15, 16] и считать доверительным интервалом с вероятностью $P = 2/(N + 1)$ выхода за него интервал определения соответствующей гистограммы из Таблицы 1, где N – число разбиений для СК. В частности, при N равном 40, вероятность выхода за соответствующие интервалы не превосходит 0.05. Заметим, что среднеквадратичное отклонение (std), приведенное на всех гистограммах Таблицы 1 оказывается существенно более жесткой характеристикой.

Можно заключить, что получаемые в таком случае доверительные интервалы очень широки и в данных обстоятельствах имеющиеся оценки качества обучения, основанные на СК, показывают недостаточно удовлетворительный с точки зрения решаемой задачи результат. В то же время, часть моделей демонстрируют хороший результат, что потенциально свидетельствует о возможности найти хорошие решения.

4. ТИПЫ МОДЕЛЕЙ

На основе сравнения поведения множества альтернативных моделей, получаемых в ходе СК, мы можем оценить качество каждой из них. Для этого сопоставим результаты модели на локальном экзамене соответствующей итерации СК и ее результаты на чистом экзамене, где используются заранее выделенные данные, не вошедшие ни в одно обучение. Таким образом мы попытаемся проверить гипотезу о том, что часть моделей обладает выраженной способностью к предсказанию результатов на чистом экзамене.

В следующей серии экспериментов, проводившейся на упорядоченных по возрасту данных, из которых первые 500 были включены в обучающую выборку, а оставшиеся 224 наблюдения – в независимый экзамен, иллюстрируемых Таблицей 2 и тоже связанных с размахом оценок при СК, было желание выяснить, как формируемые в процедуре СК решающие правила ведут себя не только на экзаменационных частях тех блоков данных, что использовались для построения этих правил, но и на их обучающих частях, используемых для формирования этих правил, а так же – на чистом экзамене.

С этой целью была проведена серия экспериментов на упорядоченных по возрасту данных, из которых первые 500 были включены в обучающую выборку, а оставшиеся 224 наблюдения – независимый экзамен. Результаты этих экспериментов приведены в Таблице 2. Они также связаны и с размахом оценок при СК.

В первой гистограмме показано распределение 40 моделей по качеству распознавания ими экзаменационных блоков в разбиениях, на которых эти модели формировались. Видно, что это распределение простирается по доле правильно распознаваемого от примерно 0.41 до 0.93. Это же составляет в данном случае и пятипроцентный доверительный интервал, как видим, достаточно широкий, чтобы основываясь на нем строить сколько-нибудь интересные предсказания.

Вторая гистограмма, столбцы которой совпадают с первыми по положению, показывает распределение ошибок распознавания 40 моделями. Она имеет заметный перекос в сторону увеличения доли ошибок, что особенно хорошо видно по объемам 33-процентных квантилей (**Q1 Vol**, **Q2 Vol** и **Q3 Vol** - на этих двух гистограммах справа) для моделей, показанных в первой гистограмме.

На третьей гистограмме показано распределение результатов правильного распознавания сформированными 40 моделями тех обучающих частей разбиений для СК, на которых они строились. Видно сильное смещение среднего в сторону увеличения качества распознавания (примерно на 0.2), сопровождаемое резким сужением самой гистограммы, обусловленным тем, что данные, на которых она строилась, соответствуют обучающим частям массивов для СК.

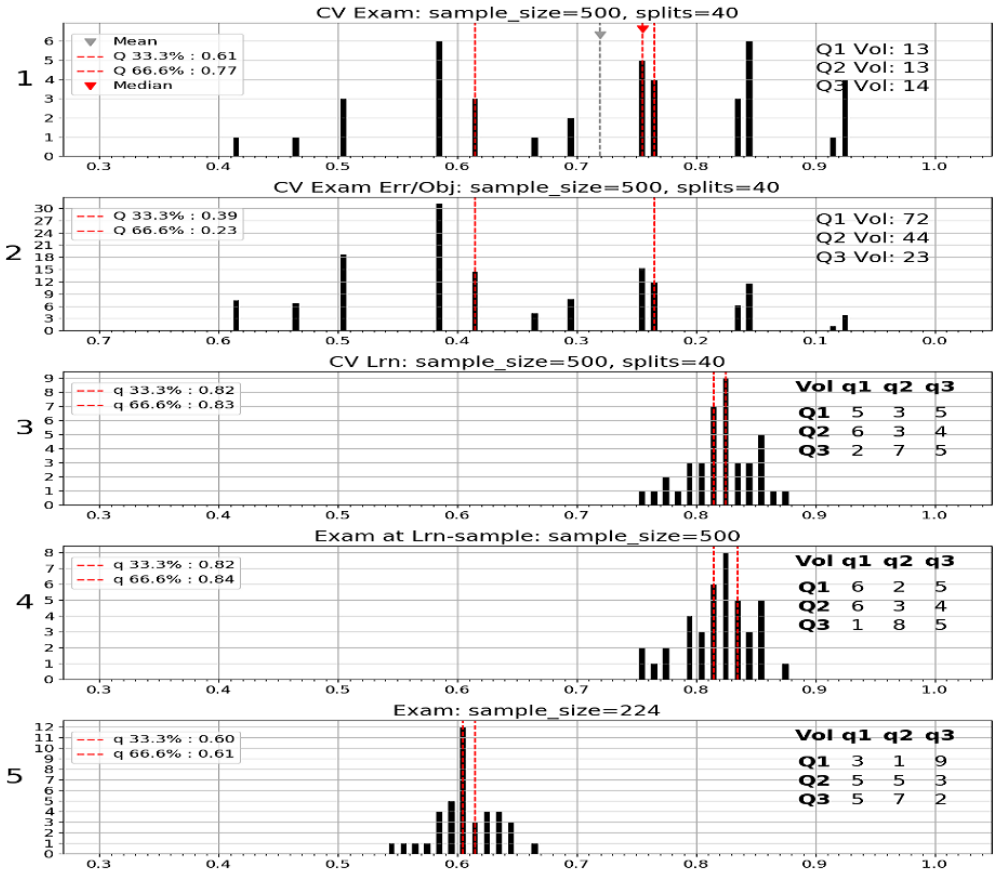
На четвертой гистограмме показано распределение результатов правильного распознавания 40 сформированными в режиме СК моделями полной обучающей выборки. Как и следовало ожидать, она мало отличается от третьей гистограммы.

Наконец, на пятой гистограмме показано распределение результатов правильного распознавания 40 моделями экзаменационного массива новых данных из 224 объектов, вообще не включавшихся в процесс формирования этих моделей. Эта гистограмма в целом похожа на предыдущие две, но сильно (примерно на 0.22) смещена в сторону ухудшения качества распознавания. При этом следует заметить, что интервалы определения последней гистограммы и предшествующих двух не пересекаются, но не выходят за пределы доверительного интервала, определяемого первой гистограммой.

Приведенные в Таблице 2. гистограммы являются обобщенным результатом эксперимента по применению процедуры СК к рассматриваемым данным и процедуре формирования решающих правил – моделей. В то же время из них явно следует, что сами полученные здесь

модели достаточно различны. Чтобы выявить эти различия, было проанализировано поведение моделей каждого из 3-х квантилей первой гистограммы на обучающей и экзаменационной выборках. Подробности анализа работы каждой из моделей приведены в *Приложении C*, а на гистограммах 3–5 Таблицы 2 приведены данные по совместному распределению объемов квантилей **Q**i для исходной гистограммы под номером 1 в трех квантилях **q**j для каждой из гистограмм 3–5.

Таблица 2. Распределения 40 моделей по качеству распознавания и ошибкам распознавания при различных условиях проведения экзамена.



Результаты анализа позволяют по-разному сгруппировать 40 моделей, сформированных в этом эксперименте. Но наиболее интересными представляются варианты, в которых выделяемые модели обладают выраженной способностью к предсказанию результатов обучения на чистом экзамене. Как видим, такого рода групп оказалось несколько, причем наиболее многочисленными являются две, модели в которых, сформированные в первом **Q1** и третьем **Q3** квантилях обучающей выборки, начинают эффективно работать в квантилях **q3** и **q1** экзаменационной выборки, выявляя встречные тенденции по изменению каких-то подмножеств свойств в данных экзаменационной выборки. При этом модели среднего квантиля вполне равномерно рассеваются по всем трем квантилям экзаменационных данных.

Этим формулируется когнитивная по сути задача о том, что именно в этих моделях как детекторов свойств данных придает им такое свойство в данной задаче.

По существу, такого рода задачи могут и должны решаться при активном участии проблемного специалиста, которым в данном случае является диабетолог, располагающий существенно большей информацией об объектах данного исследования, полностью отсутствующей в предоставляемой базе прецедентных данных. Только он, руководствуясь подсказками, содержащимися в моделях и особенностях их поведения на данных, мог бы сформулировать какие-то гипотезы, потенциально несущие возможность расширения области его профессиональных знаний. Но ни такого специалиста, ни существенно большей врачебной информации в нашем распоряжении нет, и поэтому приходится ограничиться только лишь тем, что связано с конкретной структурой отобранных правил-моделей, оставаясь в рамках возможностей, предоставляемых методами анализа и обучения распознаванию. Это тоже дает некоторые возможности для развития знаний в области анализа данных.

5. К ВОЗМОЖНОСТИ АНАЛИЗА ДИНАМИКИ МОДЕЛЕЙ

Рассмотрим теперь возможности сравнения разного типа моделей с точки зрения разницы их поведения на экзамене и на обучении. Цель в том, чтобы найти формальные способы выделения предпочтительных моделей, а также их осмысления путем выделения наиболее информативных базовых признаков.

Формально, задача определения различий типов моделей ничем не отличается от других задач обучения распознаванию. Проблема, однако, заключается в том, что распознаваемые объекты в данном случае являются бинарными нерегулярными графами в виде разного вида деревьев решений, а не векторами свойств, как обычно бывает в такого рода задачах.

Выход был найден в преобразовании деревьев решений в приемлемую для обычных методов обучения распознаванию форму. Были рассмотрены две возможности такого рода:

1. “Расплетение” деревьев на составляющие их ветви.
2. Отображение деревьев в некие интегральные функционалы исходных признаков, зависящие от того, как близко к корню дерева находится узел с рассматриваемым признаком, как часто соответствующий признак встречается в используемых моделях и т.д.

Обе эти возможности детально рассматриваются далее.

5.1. “Расплетение” деревьев на составляющие их ветви

Число возможных разных вариантов обработки деревом распознаваемых описаний конкретного случая равно числу его терминальных узлов (выходов). Каждый такой узел в какой-то мере соответствует одному из рассматриваемых классов. Множество цепочек узлов от терминальных к вершине, с приписанными им признаками и порогами описывает множество сценариев обработки. Каждое такое описание есть множество условий (значение признака меньше либо равно или больше заданного порога), одновременное выполнение которых приводит к реализации данного сценария. Для любого признака, входящего в цепочку один или более раз, набор связанных с ним условий можно всегда представить в виде двойного условия, заданного интервалом из двух порогов, один из которых, в частности, может быть $+\infty$ или $-\infty$. Описание любой цепочки можно при необходимости формально дополнить условиями для не вошедших в цепочку признаков с порогами $+\infty$ и $-\infty$ или кодом “неизвестно”. Таким образом, для каждой цепочки получаем описание фиксированной длины. Выход такой цепочки (ветки) не зависит от порядка фиксируемых в ней узлов-признаков, поэтому все они могут быть приведены к одному формату – последовательности признаков в исходных данных, например, чтобы каждую из них можно было бы рассматривать строкой матрицы “объекты-признаки”, обычно необходимой для работы методов обучения распознаванию образов.

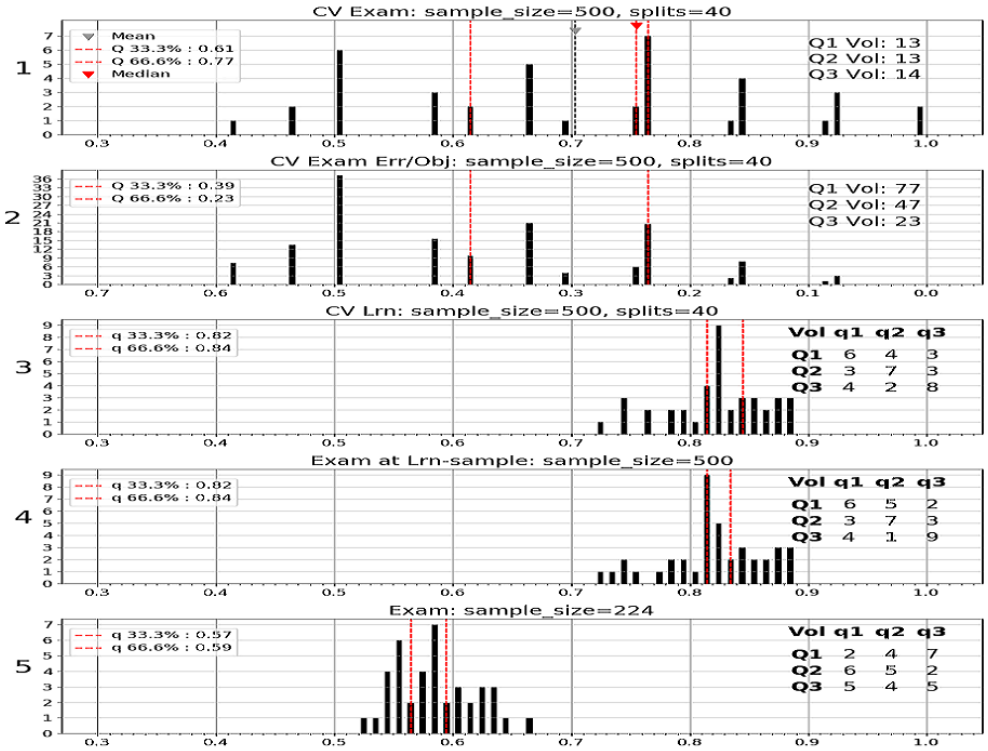
Приписав каждому такому описанию тип использующей его модели и соответствующий класс, можно рассматривать задачу классификации цепочек по типам интересующих моделей.

Если удается найти определяющее заданный тип правило, то оно будет описываться через диапазоны значений тех же исходных признаков и может быть на этой основе легко интерпретировано по смыслу.

Однако, обычно для методов обучения распознаванию образов элементами матриц “объекты-признаки” являются скаляры, а в данном случае с “Деревьями решений” элементами матрицы оказываются двумерные вектора, в качестве координат которых выступают границы интервалов по признакам в узлах, направляющих потоки данных по исходящим из узлов ветвям. Для такого варианта представления матриц “объекты-признаки” было решено заменить исходные числовые признаки на наборы бинарных, соответствующих разным по величине и положению на шкале признака интервалов, позволяющих с достаточной точностью описать те интервалы исходных признаков, которые формируются на них используемой процедурой обучения распознаванию. Размерность пространства первичных описаний при этом резко возрастает, но зато сами описания становятся бинарными “симптомами”, характеризующими попадание величины исходного признака исследуемого объекта в некий интервал его шкалы типа “нормальная температура”, “учащенный пульс” и т.п. Соответственно, и приведенная к бинаризованному виду ветка дерева решений становится неким “синдромом” как строго выполняемой конъюнкцией входящих в ветку узлов-симптомов, а дерево в целом становится сборкой по ИЛИ множества входящих в него синдромов веток. Такое представление можно рассматривать как частный случай рассматриваемых М.Вайнцвайгом “размытых” или “нестрогих” конъюнкций и дизъюнкций, использованных им при разработке программы “Кора” [26].

Подробно преобразование исходных признаков в бинарные описано в *Приложении D*.

Таблица 3. Аналог Таблицы 2, но для бинарных признаков полученных из 20 квантилей исходных признаков.



В Таблице 3, аналогичной Таблице 2, но построенной в эксперименте по формированию в режиме скользящего контроля для бинарных признаков, приведены результаты экспериментов по 40 моделям.

В целом, результаты, приведенные в Таблице 3 для бинаризованных признаков, близки показанному в Таблице 2 для исходных числовых признаков. Различия их, насколько можно судить, обусловлены двумя причинами:

– случайными отличиями в разбиениях исходной выборки на подвыборки для скользящего контроля (СК) и
– неточностями соответствия выбранного способа бинаризации тем реальным интервалам признаков, которые были определены процедурой "Деревья решений" при решении данной задачи обучения распознаванию.

Однако, с точки зрения цели данной работы эти различия совершенно несущественны, так как цель в данном случае – показать, какую пользу можно извлечь, изучая тонкие различия в поведении моделей принятия решений. На этом основании было решено, что примененная схема бинаризации адекватна исходным числовым признакам и может быть использована для поиска содержательных различий между моделями, которые улучшали и ухудшали результаты своей работы при переходе от обучения к экзамену.

В соответствии с Таблицей 4 переходов для срабатывания моделей из квантилей **Q**j обучающей выборки (первая строка Таблицы 3) в квантили **q**i экзаменационной выборки (последняя строка Таблицы 3), характеризующей 9 типов или классов поведения моделей, класс 7 был выделен как класс, связанный с максимальным улучшением результатов при переходе к экзаменационным данным, так как в нем модели из первого квантиля с наибольшей долей ошибок для обучения переходили с наименьшими долями ошибок в третий квантиль экзамена, а класс 3 – как альтернатива 7-ому, соответствующая переходу моделей из третьего квантиля для обучения в первый квантиль для экзамена, демонстрируя при этом заметное ухудшение результатов.

Таблица 4. Типы динамики моделей при переходе от квантилей их распределения по качеству на обучающей выборке **Q**j к квантилям распределения на экзамене **q**i.

	q1	**q2**	**q3**
Q1	1	4	7
Q2	2	5	8
Q3	3	6	9

Сформированные так новые классы моделей были приписаны бинарным описаниям составляющих их веток, что позволило получить стандартную матрицу "объекты-признаки" для последующего использования методов обучения распознаванию.

Для решения задачи поиска содержательных различий между моделями 2-х выделенных классов были отобраны следующие симптомы-бинарные признаки: gluc_1_10, mass_7_12 и bloo_10_8. Они образуют трехсимптомный одноуровневый синдром, обеспечивающий отбор моделей с наиболее заметным улучшением результата распознавания экзаменационных данных в классе 7 по отношению к классу 3 (Рис. 3).

5.2. Отображение деревьев решений в веса исходных признаков

Для оценки значимости (весов) отдельных признаков в рамках каждой модели типа дерева принятия решений был применен критерий, основанный на положении признака в структуре

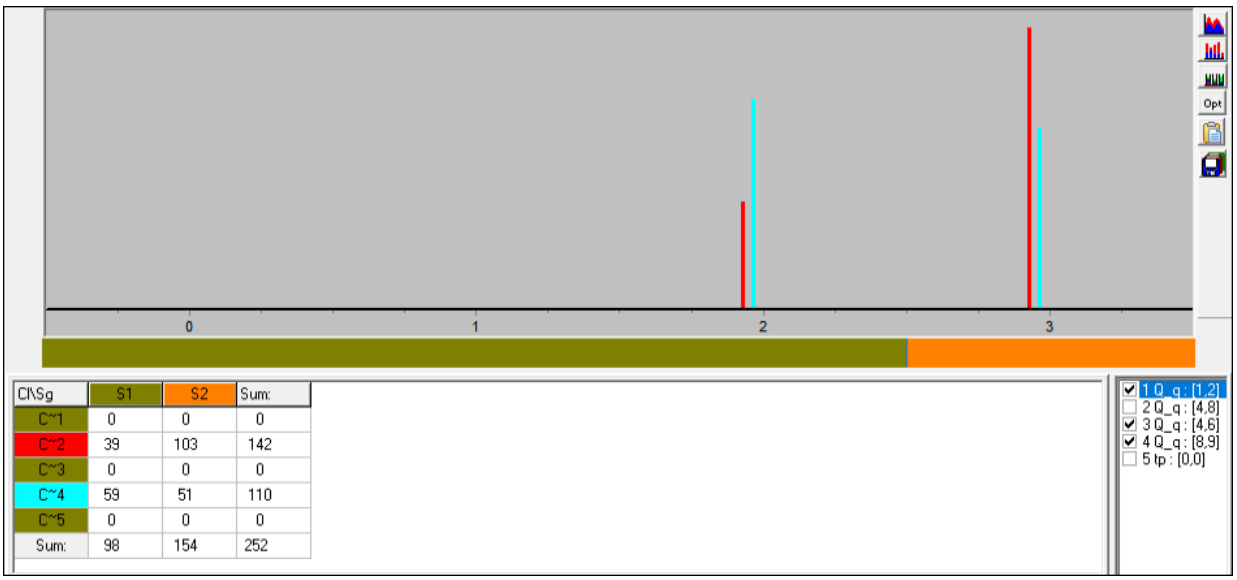


Рис. 3. Гистограмма для результата выбора трехэлементного набора “симптомов”, наиболее информативных в совокупности при отделения моделей, характеризующихся наиболее заметным улучшением результата распознавания экзаменационных данных в классе 7. Оранжевый интервал по оси X и оранжевый столбец в таблице под гистограммой соответствуют выполнению трех условий. Красные столбцы гистограммы и красная строка в таблице соответствуют случаям наличия диабета, а голубые столбцы гистограммы и голубая строка в таблице соответствуют пациентам без диабета. Столбцы гистограммы отображены в процентах от объемов соответствующих классов (39 и 103 составляют 27% и 73% соответственно от 142 - красные столбцы гистограммы; 59 и 51 составляют 54% и 46% соответственно от 110 - голубые столбцы гистограммы).

дерева. Вес признака рассчитывался следующим образом:

$$W_f = \sum_{n=1}^{N} w_n \ ,$$

где w_n – вес отдельного узла дерева, рассчитывавшийся одним из трех способов: как $\frac{1}{d_n}$, 1 или $\sqrt{d_n}$; N – число вхождений признака в качестве узла дерева; d_n – уровень узла в дереве соответствующего (n-го) вхождения (1 соответствует вершине дерева).

Понятно, что такого рода подход может иметь множество модификаций и вопрос лишь в том, какая из них в наибольшей степени соответствует смыслу решаемой задачи.

В *Приложении E* приведены различные таблицы с результатами оценивания значимости исходных данных в соответствии с приведенной формулой взвешивания применительно ко всем 40 моделям и средние по моделям значимости исходных признаков для трех вариантов взвешивания

Видно, что такого рода признаки оказываются резко неоднородными, причем наиболее значимыми с отрывом от других в 2 и более раз при этом оказываются “glucose”, “blood”, “mass” и “pedigree”.

“Профили” значимости исходных признаков в этих трех вариантах достаточно похожи, что можно рассматривать в качестве некого свидетельства того, что способ взвешивания узлов деревьев решений мало влияет на оценку, но это только лишь усложняет переход к тому, что и как надо было бы делать для достижения желаемого понимания того, какие новые знания в проблеме, решаемой “деревьями”, может реально получить человек.

Задача, решаемая в данном разделе, связана с тем, что характеризует поведение тех моделей из 40, построенных в процедуре скользящего контроля, которые улучшили свое качество при переходе от их экзамена на обучающей выборке к экзамену на данных, не входивших в обучение. Таких моделей оказалось 9, в *Приложении E* для них приведены номера и значимости исходных признаков, формируемые ими в варианте с убыванием веса по мере удаления от входа в дерево (с глубиной), а также усредненные результаты по этим 9 моделям для трех вариантов вычисления веса узла в оценке значимости приписанного ему признака. В Таблице 5 приведено сравнение результатов для 9 и 40 моделей при всех вариантах расчета весов. Для каждого варианта третьей строкой приведены значения меры сходства значимостей случаев 9 и 40 моделей: $D_f = \frac{\left|\overline{W_f}_{[40]} - \overline{W_f}_{[9]}\right|}{\overline{W_f}_{[40]} + \overline{W_f}_{[9]}}$, где $\overline{W_f}_{[40]}$ и $\overline{W_f}_{[9]}$ - средние значимости признака по 40 и 9 моделям соответственно.

Таблица 5. Средние значимости признаков для 9 и 40 моделей и значения меры различия между ними при разных способах расчета веса узлов, в зависимости от глубины.

Вес узла	моделей	pregnan	glucose	blood	mass	pedigree	age
$w_n = 1/d_n$	40	0.48	1.82	0.68	1.78	1.42	0.49
	9	0.66	2.00	0.69	1.75	1.44	0.49
	D_f	**0.158**	**0.047**	0.007	0.010	0.007	0.000
$w_n = 1$	40	1.82	3.52	3.8	6.95	6.1	2.12
	9	2.56	4.44	3.89	6.56	6.33	2.22
	D_f	**0.169**	**0.116**	0.012	0.029	0.019	0.023
$w_n = \sqrt{d_n}$	40	3.7	5.79	9.16	14.66	13.09	4.48
	9	5.24	7.98	9.33	13.47	13.75	4.85
	D_f	**0.172**	**0.159**	0.009	0.042	0.025	0.040

Видно, что для 9 моделей вид профилей значимости немного изменился, сделав их более контрастными по сравнению с профилями, посчитанными для всех 40 моделей.

Видно также, что для 9 моделей признак mass имеет меньшую информативность, а признаки blood, pedigree и age показывают незначительно более высокую информативность по отношению к 40 моделям.

Для признаков pregnan и glucose в случае 9 моделей видим более высокую информативность и значительно более высокие соответствующие значения D_f (отмечены зеленым в Таблице 5). Для остальных признаков (значения D_f отмечены красным) информативность меняется незначительно при переходе от 40 к 9 моделям. Эти особенности сохраняются для всех типов взвешивания.

Последнее обстоятельство позволяет заключить, что так формируемые характеристики обладают некоторой специфичностью, которая могла бы навести на размышления эндокринолога – специалиста в этой области, если бы таковой имелся. Но и в этом случае только лишь указание на то, что некое подмножество исходных признаков обладает повышенной информативностью с точки зрения понимания того, что способствует усилению прогностического эффекта этих моделей, все равно было бы недостаточно из-за отсутствия информации о том, как и в каких сочетаниях эти признаки желаемое свойство более явно описывают.

Заметим также, что три признака, в целом показывающие более высокую информативность при рассмотренных тут способах взвешивания – “glucose” (gluc_1_10), “mass” (mass_7_12) и “blood” (bloo_10_8), вошли в число наиболее информативных признаков в эксперименте с “синдромным” “расплетанием” моделей на ветви, когда решалась задача понять, почему некоторые из моделей улучшают свою предсказательную способность при переходе от обучающих данных к экзаменационным при отображении исходных описаний в бинарные их представления.

Таким образом, здесь в полной мере проявляется эффект обмена простоты на сложность (возникающей при переходе к интегральным функционалам), что открывает богатые возможности для развития этих подходов.

Тем не менее, идея выхода на интегральные функционалы более глубока и потенциально более интересна, чем это может показаться на первый взгляд на основе только лишь полученного здесь результата, так как ее можно рассматривать как попытку перенесения методов разложения функций в ряды на проблему интерпретирования сетевых структур, разнообразие, сложность и обилие которых во множестве практически важных задач нарастает с большой скоростью.

В завершение этого раздела необходимо еще раз отметить, что возможности, рассмотренные здесь, отнюдь не исчерпывают подходы к решению давно уже возникшей проблемы получения содержательных интерпретаций "черных ящиков", к которым можно отнести большинство решающих правил или моделей в сетевом их представлении. Одной из ярких демонстраций этого в последнее время явилось отторжение врачами решений, формируемых для медицины системой "Ватсон" фирмы IBM [27] в существующем ныне ее виде. Проблема эта критически важна для организации эффективного взаимодействия "интеллектуальных" компьютерных систем с человеком. И хотя она возникла вместе с кибернетикой в виде "черных ящиков", а несколько позже была переосмыслена в эру "Экспертных систем", пока по сути в ней нет существенного продвижения, а наблюдаемый прогресс в области компьютерных и интеллектуальных технологий с каждым днем делает эту проблему все актуальнее.

6. ОЦЕНКА СТРУКТУРНОЙ ПОЛНОТЫ ДАННЫХ. УТОЧНЕНИЕ КЛАССИФИКАЦИИ

Последней из рассматриваемых здесь является гипотеза о неполноте и неточности классификации, используемой в решаемой задаче. Предлагается способ уточнения исходной классификации данных, основанный на результатах многократного итеративного проведения СК, сопровождающегося выделением новых подклассов, исходно не заданных в постановке задачи.

В самом деле, ошибки, получаемые в результате обучения распознаванию, могут свидетельствовать не только о несовершенстве используемого метода или неадекватности его решаемой задаче, но и о том также, что в силу априорно существующих неполноты знаний о природе решаемой задачи и зашумленности оценок свойств изучаемых объектов неполнота знаний может распространяться на классификацию объектов. Таким образом, и исходная классификация может и должна в таких случаях рассматриваться как объект уточнения [28].

Поскольку исходная классификация подвергается сомнению и уточнению, ориентироваться в решении этой проблемы следует на методы, которые не опираются или опираются в малой мере на априорно заданную классификацию исходных данных, на которых решается задача обучению распознаванию. Условно можно считать, что это – методы кластеризации данных (Unsupervised learning) по какой-то приписываемой им мере сходства[29].

В данном случае, ориентируясь на специфику рассматриваемой в данной работе процедуры Скользящего контроля, мы модифицировали его для данной задачи, остановившись на подходе, состоящем из последовательных циклически повторяющихся фаз, что позволило нам, опираясь на исходную априорно неточную классификацию данных, выдвинуть гипотезы относительно улучшения этой классификации, подтвердить которые можно было бы только при наличии эндокринологов, имевших непосредственное отношение к сбору тех данных, с которыми мы экспериментировали в этой работе. Алгоритм разработанной процедуры состоит из следующих пяти фаз:

1. В первой фазе для полученных в обучении моделей строится "Матрица ошибок". Если она удовлетворяет требованиям качества, предъявляемым к решению задачи обучения распо-

знаванию, то задача считается решенной, в противном случае совершается переход ко второй фазе.
2. Во второй фазе ошибки, выявленные при построении распознающих моделей в обучении, выделяются в новый класс, после чего совершается переход к третьей фазе.
3. В третьей фазе для полученного так расширенного набора классов строятся новые модели.
4. В четвертой фазе на обучающих данных производится экзамен расширенного набора моделей, в результате которого могут быть получены ошибки двух типов – "ложные тревоги" и "отказы от распознавания".
5. В пятой фазе "ложные тревоги" включаются в те классы, за которые они были распознаны, а "отказы от распознавания" включаются в новый класс, после чего совершается переход к первой фазе.

В Таблице 6 приведена матрица ошибок для исходной двухклассовой задачи, а в Таблице 7 показан результат моделирования на скорректированных ошибками трех классах, получающийся после первой итерации в соответствии с описанным алгоритмом.

В последних трех столбцах каждой из таблиц 6, 7 приводятся количества объектов в исходной их классификации, говорящие о сильном перемешивании исходных классов в каждом из выделяемых подклассов. При этом столбец vol соответствует общему количеству объектов в выделяемом подклассе, true1 соответствует исходному классу "Нет диабета", а столбец true2 соответствует исходному классу "Диабет".

Таблица 6. Матрица ошибок для исходных двух классов.

clss\ress	1	2	**vol**	**true1**	**true2**
iniClass 1	270	96	366	270	43
iniClass 2	43	91	134	96	91
summ	313	187	500	366	134

Таблица 7. Результат моделирования на скорректированных ошибками трех классах.

clss\ress	1	2	3	**vol**	**true1**	**true2**
iniClass 1	180	9	47	236	204	16
iniClass 2	4	61	31	96	35	68
iniClass 3	36	33	99	168	127	50
summ	220	103	177	500	366	134

В соответствии с Таблицей 6 видно, что степень перемешанности исходных классов при используемых данных и методе построения решающих правил весьма велика, что может быть объяснено в данном случае главным образом недостатком существенных для решения задачи признаков в исходных данных. Но мы здесь исходим из того, чем реально располагаем, и стремимся лишь к тому, чтобы анализ этих данных позволил бы сформировать сильные гипотезы в отношении структуры анализируемых данных, опираясь на элементы которой – подклассы, проблемный специалист, привлекая дополнительные сведения об объектах распознавания и свои профессиональные знания, смог бы дать содержательную интерпретацию формируемым так подклассам.[1] Понятно, что более тонкая классификация данных позволяет существенно упростить для специалиста проблему интерпретации выделяемых подклассов.

[1] Здесь стоит заметить, что в процессе работы над текстом этой статьи мы обнаружили ссылку на работу большого коллектива шведских и финских медиков "Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables", опубликованную в журнале "The Lancet, Diabetes & Endocrinology"[30]. Они, исследуя очень большую базу данных, содержащую данные о примерно 15000

После 10 итераций такого типа получен результат, приведенный ниже в Таблице 8 для 12 подклассов, из которой следует, что при 12 подклассах Матрица ошибок оказывается практически диагональной. Матрицы ошибок для всех итераций в поиске подклассов вынесены в Приложение F.

Таблица 8. Результат моделирования на двенадцати классах.

clss\ress	1	2	3	4	5	6	7	8	9	10	11	12	**vol**	**true1**	**true2**
iniClass 1	104	0	0	0	0	0	0	0	0	0	0	0	104	102	2
iniClass 2	0	28	0	0	1	0	0	0	0	0	0	1	30	2	28
iniClass 3	0	0	55	0	2	0	0	0	0	0	0	0	57	47	13
iniClass 4	0	0	0	36	1	0	0	0	0	0	0	2	39	31	9
iniClass 5	0	0	2	0	29	0	0	1	0	0	0	4	36	24	14
iniClass 6	0	0	0	0	1	36	0	0	0	0	0	1	38	28	8
iniClass 7	0	0	0	0	0	0	31	0	0	0	1	0	32	18	13
iniClass 8	0	0	0	0	0	0	0	22	0	0	0	2	24	22	4
iniClass 9	0	0	2	0	1	0	0	0	31	0	3	0	37	26	9
iniClass 10	0	0	0	0	0	0	0	0	0	22	0	0	22	18	7
iniClass 11	0	0	0	0	0	0	0	0	0	0	51	0	51	37	19
iniClass 12	0	2	1	4	3	0	0	3	4	3	1	9	30	11	8
summ	104	30	60	40	38	36	31	26	35	25	56	19	500	366	134

Скорее всего, показанная здесь степень диагонализации в реальной задаче могла бы оказаться избыточной, если бы изучением осмысленности выделяемых подклассов с медицинской точки зрения занимался бы диабетолог-эндокринолог, но при отсутствии такового мы можем ориентироваться лишь на какие-то формальные критерии, на роль наилучшего из которых вполне может претендовать “степень диагонализации Матрицы ошибок” при не слишком малых объемах подклассов, составляющих ее.

Подводя итог, сказанному здесь, следует отметить, что рассматриваемая нами процедура последовательной кластеризации принадлежит к особому классу кластерных процедур, в которых, в отличие от широко используемых, в качестве меры сходства кластеризуемых объектов применяется не Евклидово расстояние в “Пространстве признаков” или какой-то иной вариант такого рода меры из числа широко используемых в математике, а “распознавательская” мера сходства, основанная на голосовании решателей, характеризуемая критериями близости распознаваемого к тем классам, на которых происходит обучение распознаванию[31]. Такой подход снимает многие не объясняемые проблемы обычной кластеризации, перекладывая их решение на возможности выбираемого для решения задачи распознавания метода. В частности, так существенно повышается осмысленность получаемых формально кластеров.

7. ЗАКЛЮЧЕНИЕ

На основании ряда вышеописанных экспериментов мы приходим к заключению, что процедура СК может быть использована в качестве когнитивного инструмента исследования данных и задач, развивающего представления исследователей о решаемых ими проблемах в области обучения распознаванию.

К возможностям такого рода, базирующимся на применении этой процедуры, следует отнести предложенные методы смысловой интерпретации моделей, получаемых в результате машинного обучения (в т.ч. в условиях малого объема выборки, либо ее неоднородности), а также способы уточнения как самих моделей, так и постановки задачи.

В частности, на основе исследованных примеров можно сказать, что:

больных диабетом, установили, что диагноз “диабет 2-го типа” на самом деле должен быть представлен пятью разными диагнозами. По мнению медиков, это воспринимается как очень серьезное достижение.

– Подтверждается связь между свойствами обучающей выборки и характером кривых обучения при сравнении таковых для оценок качества моделей, построенных на обучающей выборке со средними значениями для скользящего контроля.
– В ситуации, когда формальная оценка разброса результатов СК дает слишком широкие доверительные интервалы по качеству моделей (в условиях малого объема выборки, либо ее неоднородности), применение предложенного метода исследования динамики свойств моделей позволяет потенциально выделить подмножество моделей, обладающих выраженной способностью к предсказанию результатов на чистом экзамене.
– Модифицированная итеративная процедура СК позволяет уточнять исходную классификацию данных, выдвигая гипотезы о дополнительных смысловых кластерах и, как следствие, – получать более качественное решение задачи в целом. Применительно к рассматривавшейся в качестве экспериментальной базы с описанием больных диабетом, данная возможность подтверждается исследованием [30], согласно которому диабет второго типа на самом деле объединяет пять совершенно разных видов диабета, требующих самостоятельного лечения.

ПРИЛОЖЕНИЕ

А. ОПИСАНИЕ ВЫБОРКИ ДАННЫХ И ПАРАМЕТРЫ МЕТОДА “ДЕРЕВЬЯ РЕШЕНИЙ” ИЗ БИБЛИОТЕКИ “SCIKIT-LEARN”

Имена признаков выборки:

1. “pregnan”: число беременностей / Number of times pregnant
2. “glucose”: концентрация глюкозы в плазме через 2 часа при пероральном тестировании на толерантность к глюкозе / Plasma glucose concentration after 2 hours in an oral glucose tolerance test
3. “blood/pr”: диастолическое (нижнее) давление крови / Diastolic blood pressure ($mmHg$)
4. “mass/ind”: индекс массы тела / Body mass index ($(weight\ in\ kg)/(height\ in\ m)^2$)
5. “pedigree”: Diabetes pedigree function which provide some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient
6. “age”: возраст обследуемого / Age ($years$)
7. Triceps skin fold thickness (mm)
8. 2-Hour serum insulin (muU/ml)

В этих данных оказалось большое число пропусков измерений, наблюдаемых в основном в признаках “Triceps skin fold thickness” и “2-Hour serum insulin”, а также в небольшом числе россыпью встречающихся почти во всех остальных признаках. Было решено, что для целей данного исследования нет необходимости сохранять в исходных данных эти два признака и те строки-объекты, в которых хоть раз встречается пропуск измерений. В результате размерность задачи снизилась до 6, а объем выборки сократился до 723 наблюдений.

В серии предварительных экспериментов в соответствии с параметрами исследуемого массива данных были выбраны и зафиксированы следующие параметры программы, определяющие условия ветвления в узлах дерева и его глубину, а также – критерий ветвления:

class_weight=‘balanced’ (взвешивание объектов обратно пропорционально объемам классов)

criterion=‘entropy’ (критерий минимизации энтропии)

max_depth=7 (максимальная глубина дерева)

min_samples_leaf=3 (минимальное число объектов в узле дерева)

min_impurity_decrease=0.007 (минимальное уменьшение нечистоты деления (impurity) в узле)

Для остальных параметров использовались значения по умолчанию [17].

В. ДЕТАЛИЗАЦИЯ ЭКСПЕРИМЕНТОВ ПО ОПРЕДЕЛЕНИЮ ЗАВИСИМОСТИ ЭФФЕКТА ОБУЧЕНИЯ ОТ ОБЪЕМА ОБУЧАЮЩЕЙ ВЫБОРКИ.

В серии, состоящей из пяти экспериментов на упорядоченных по возрасту данных менялся объем экзаменационных частей СК, составляя последовательно 1, 3, 7, 15 и 30 объектов.

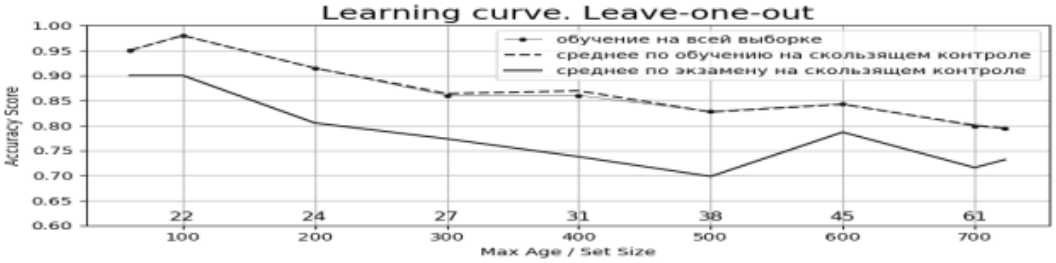


Рис. 4. Кривая обучения. Размер экзаменуемых частей СК (длина блока) = 1.

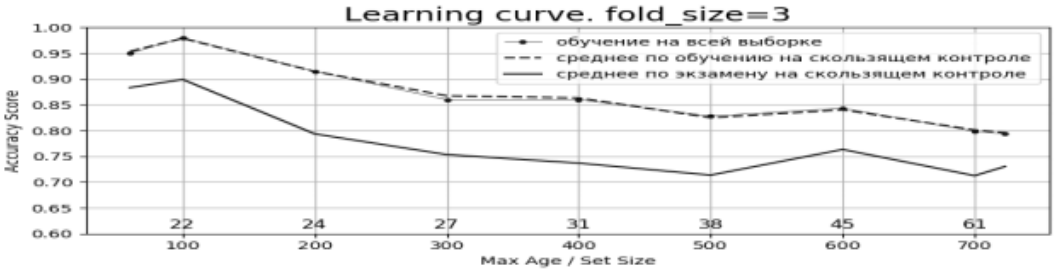


Рис. 5. Кривая обучения. Размер экзаменуемых частей СК (длина блока) = 3.

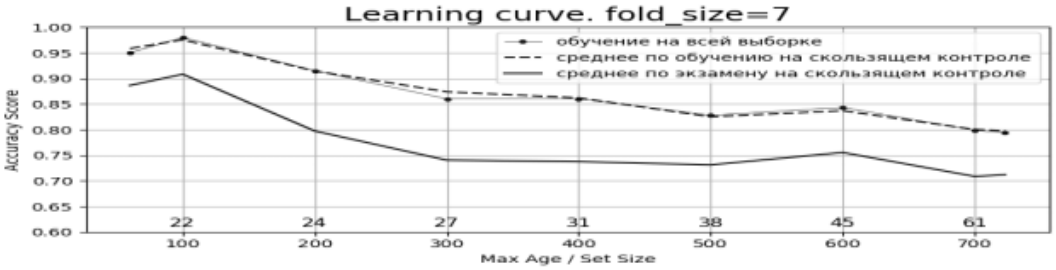


Рис. 6. Кривая обучения. Размер экзаменуемых частей СК (длина блока) = 7.

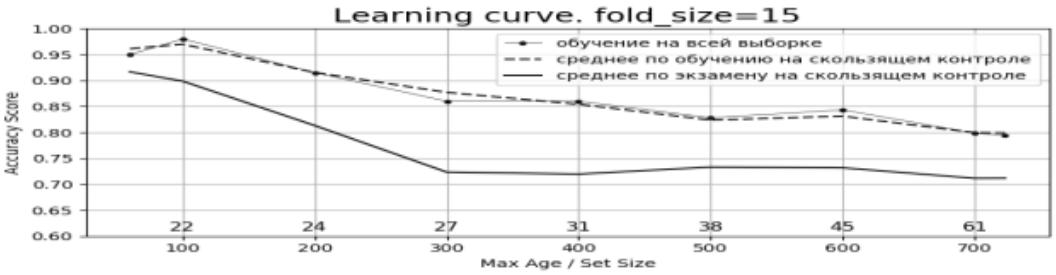


Рис. 7. Кривая обучения. Размер экзаменуемых частей СК (длина блока) = 15.

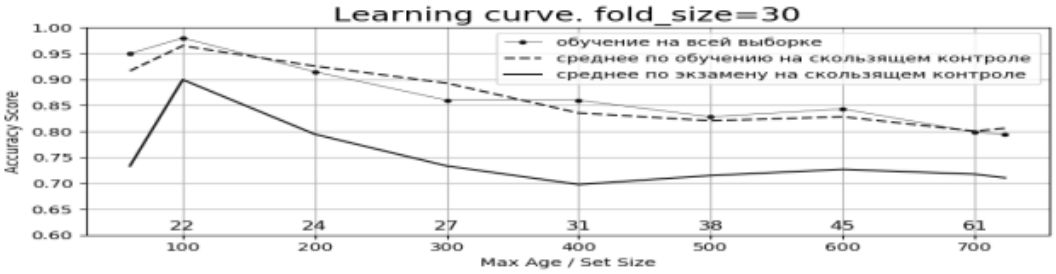


Рис. 8. Кривая обучения. Размер экзаменуемых частей СК (длина блока) = 30.

C. АНАЛИЗ РАБОТЫ МОДЕЛЕЙ ИЗ РАЗНЫХ КВАНТИЛЕЙ НА ОБУЧАЮЩЕЙ И ЭКЗАМЕНАЦИОННОЙ ВЫБОРКАХ.

В Таблицах 9a–9c, приведенных ниже, показано, как каждая из сформированных моделей работает на обучающей и экзаменационной выборках. А чтобы удобнее это было видеть, исследуемые модели, сведенные в Таблицы 9a–9c, раскрашены с учетом их распределения по 33-процентным квантилям в соответствии с гистограммой 1 в Таблице 2.

Таблица 9a. Модели первого квантиля Таблицы 2.

Models ID	Lrn Size	True, Lrn %	False, Lrn %	True, Exm %	False, Exm %	Exm q1: A q2: B q3: C
2	488	41.67	58.33	63.39	36.61	C
12	487	46.15	53.85	62.05	37.95	C
25	488	50.0	50.0	63.39	36.61	C
27	488	50.0	50.0	64.73	35.27	C
33	488	50.0	50.0	59.38	40.62	A
20	488	58.33	41.67	63.39	36.61	C
21	488	58.33	41.67	64.73	35.27	C
28	488	58.33	41.67	54.91	45.09	A
35	488	58.33	41.67	62.95	37.05	C
36	488	58.33	41.67	58.48	41.52	A
37	488	58.33	41.67	61.61	38.39	C
5	487	61.54	38.46	62.05	37.95	C
10	487	61.54	38.46	60.27	39.73	B
Mean	487.77	54.68	45.32	61.64	38.36	

Таблица 9b. Модели второго квантиля Таблицы 2.

Models ID	Lrn Size	True, Lrn %	False, Lrn %	True, Exm %	False, Exm %	Exm q1: A q2: B q3: C
14	487	61.54	38.46	57.59	42.41	A
23	488	66.67	33.33	56.7	43.3	A
4	487	69.23	30.77	60.71	39.29	B
15	487	69.23	30.77	60.71	39.29	B
30	488	75.0	25.0	60.27	39.73	B
31	488	75.0	25.0	55.8	44.2	A
34	488	75.0	25.0	60.71	39.29	B
38	488	75.0	25.0	61.61	38.39	C
39	488	75.0	25.0	60.71	39.29	B
2	487	76.92	23.08	58.48	41.52	A
13	487	76.92	23.08	59.82	40.18	A
16	487	76.92	23.08	63.39	36.61	C
17	487	76.92	23.08	62.05	37.95	C
Mean	487.46	73.02	26.97	59.92	40.11	

Каждая из Таблиц 9a–9c характеризует поведение моделей, попавших в соответствующие 33.3-процентные квантили распределения 40 моделей по их качеству распознавания экзаменационных блоков в эксперименте со скользящим контролем на первых 500 наблюдениях из упорядоченного по возрасту массива данных.

В первых их столбцах указаны номера моделей в порядке формирования их процедурой построения древовидных решателей, во вторых – реальный объем обучающих блоков в каждом из циклов СК, в третьих и четвертых – проценты правильного распознавания и ошибок на обучающих блоках в циклах СК, в пятых и шестых – проценты правильного распознавания и ошибок этими же моделями 224 наблюдений, вообще не входивших в процедуру СК, соответ-

Таблица 9c. Модели третьего квантиля Таблицы 2.

Models ID	Lrn Size	True, Lrn %	False, Lrn %	True, Exm %	False, Exm %	Exm q1: A q2: B q3: C
24	488	83.33	16.67	59.82	40.18	A
26	488	83.33	16.67	60.71	39.29	B
29	488	83.33	16.67	66.07	33.93	C
0	487	84.62	15.38	59.82	40.18	A
6	487	84.62	15.38	59.82	40.18	A
7	487	84.62	15.38	61.16	38.84	B
9	487	84.62	15.38	58.04	41.96	A
18	487	84.62	15.38	60.71	39.29	B
19	487	84.62	15.38	64.29	35.71	C
22	488	91.67	8.33	60.27	39.73	B
1	487	92.31	7.69	58.04	41.96	A
3	487	92.31	7.69	60.71	39.29	B
8	487	92.31	7.69	60.27	39.73	B
11	487	92.31	7.69	60.71	39.29	B
Mean	487.29	87.07	12.93	60.71	39.25	

ствующих максимальному возрасту и рассматриваемых как “чистый” экзаменационный набор, а в седьмом – классификационный индекс, более детально описываемый позже.

В нижних строках Таблиц 9a–9c приведены средние значения оцениваемых параметров выделенных групп моделей. В соответствии с ними средние для правильно распознаваемого и ошибок СК в фазе обучения показывают, как и должно было бы быть, наихудшие результаты для моделей первого квантиля. При переходе к моделям второго и третьего квантилей эти результаты монотонно улучшаются. В то же время на чистом экзамене средние значения оценок для правильно и ошибочно распознанных наблюдений почти не различаются, имея при этом незначительный разброс, подтверждающий показанное на пятой гистограмме Таблицы 2.

Для каждой из моделей было определено, в какой из 33.3-процентных квантилей для экзаменационной выборки переходит каждая из моделей в Таблицах 9a–9c. Для облегчения последующего анализа квантилям для экзаменационной выборки в порядке возрастания их номеров присвоены условные коды A, B и C, а результаты такого рода классификации размещены в Таблицах 9a–9c в последнем их седьмом столбце, причем для облегчения визуального анализа строки этих таблиц окрашены в красный, зеленый и синий цвета соответственно для кодов A, B и C.

D. ПРЕОБРАЗОВАНИЕ ИСХОДНЫХ ПРИЗНАКОВ В БИНАРНЫЕ.

Было решено в качестве основы для бинаризации использовать квантильное разбиение исходных числовых признаков на интервалы с последующим их укрупнением путем объединения соседних. Такая схема позволяет в принципе с желаемой точностью воспроизвести исходные разбиения числовых переменных, обеспечиваемое "Деревьями решений", однако ее недостатком является квадратичный рост получаемого так числа бинарных признаков. Поэтому в этом случае приходится искать баланс между потенциальной точностью и размерностью получаемого так признакового пространства.

В серии экспериментов по оценке качества получаемых так новых описаний, проводимых в режиме скользящего контроля все для тех же "Деревьев решений", с которыми здесь проводились и другие исследования, было установлено, что, начиная с 20 квантилей для каждой исходной числовой переменной, дальнейшее увеличение числа квантилей не приводит к увеличению качества, а размерность получаемого так бинарного признакового пространства несколько превосходит 1000, что в принципе может породить некоторые проблемы при последующем анализе. Тем не менее, мы остановились на этом варианте.

В результате получилась бинарная матрица, размерность которой по признакам превосходила 1250 при 854 объектах в выборке (из них 224 в классах 7 и 3) и она была сильно разреженной. Объектом в данном случае является одна из веток, на которые расплетались 40 моделей, формировавшихся в обучении для СК-экспериментов. Т.е. в среднем каждое из 40 деревьев содержало порядка 20 терминальных узлов-веток.

В Таблице 10, показаны интервалы исходных признаков при формировании представляющих их бинарных признаков.

Таблица 10. Границы квантилей исходных признаков.

features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pregnan	0	0	0	1	1	1	1	2	2	2.35	3	3.65	4	5	5	6	7	8	9	10	17
glucose	44	80	86.3	91	95	99.8	102	106	109	112	117	122	125	129	135	142	147	158	168	181	199
blood/pr	24	52	58	60	62	64	66	68	70	70	72	74	75	76	78	80	82	85	88	91.7	122
mass/ind	18.2	22.3	24	25.2	26.2	27.5	28.5	29.6	30.4	31.2	32.4	33.1	33.8	34.6	35.5	36.6	37.8	39.3	41.4	44.5	67.1
pedigree	0.08	0.14	0.17	0.2	0.22	0.24	0.26	0.28	0.31	0.34	0.38	0.42	0.46	0.51	0.56	0.63	0.69	0.75	0.88	1.14	2.42
age	21	21	22	22	23	24	25	26	27	28	29	31	33	36	38	41	43	46	51	58	81

Строка "features" содержит номера базовых квантилей для всех рассматриваемых признаков, а каждая из следующих границы соответствующих признаков (см. *Приложение А*).

В соответствии с Таблицей 10, отобранные признаки соответствуют следующим интервалам исходных:

- gluc_1_10: glucose (Plasma glucose concentration at 2 hours in an oral glucose tolerance test) меньше 112;
- mass_7_12: mass/ind (Body mass index) в диапазоне 28.5-39.3;
- bloo_10_8: blood/pr (Diastolic blood pressure) в диапазоне 70-82.

Е. ЗНАЧИМОСТИ ПРИЗНАКОВ.

Таблица 11. Значимости признаков для 40 моделей, полученных на скользящем контроле по упорядоченным по возрасту 500 объектам обучающей выборки. Вес убывает с глубиной.

N	pregnan	glucose	blood	mass	pedigree	age
0	0.33	1.92	0.87	1.77	1.43	0.5
1	0.33	1.75	0.7	1.79	1.43	0.64
2	0.33	1.75	0.56	1.38	1.48	0.7
3	0.48	1.75	0.25	1.9	1.48	0.25
4	0.48	1.75	0.73	1.79	1.83	0.25
5	0.33	1.89	0.62	1.67	1.65	0.45
6	0.48	1.75	0.25	1.9	1.88	0.25
7	0.33	2.06	0.87	1.63	1.18	0.5
8	0.33	1.75	0.73	1.77	1.18	0.64
9	0.39	1.95	0.73	1.99	1.74	0.53
10	0.33	1.75	0.54	2.1	0.87	0.5
11	0.48	1.75	0.73	1.64	1.8	0.42
12	0.78	1.81	0.34	2.13	1.31	0.64
13	0.93	1.75	0.14	2.05	1.48	0.42
14	0.33	1.5	0.79	1.91	1.33	0.5
15	0.93	1.75	0.64	1.91	1.39	0.64
16	0.33	1.95	0.54	2.04	1.06	0.5
17	0.48	2.06	1.04	1.28	1.48	0.5
18	0.33	2.06	0.58	1.43	1.63	0.64
19	0.33	1.7	0.76	1.93	0.93	0.5
20	0.33	1.7	0.62	2.24	1.49	0.5
21	0.7	2.48	0.93	1.03	1.13	0.25
22	0.78	1.75	0.81	1.71	1.54	0.84
23	0.5	1.95	0.39	2.05	1.28	0.5
24	0.33	1.75	1.04	1.65	1.4	0.5
25	0.7	1.95	0.74	1.62	1.93	0.64
26	0.48	1.75	1.04	1.65	1.01	0.5
27	0.98	2.45	0.62	1.99	1.63	0.33
28	0.5	1.75	0.56	1.65	1.26	0.5
29	0.37	2.09	0.73	1.68	1.99	0.33
30	0.33	1.75	0.79	1.9	1.4	0.5
31	0.33	1.5	1.29	1.74	1.77	0.5
32	0.73	2.04	1.18	1.33	1.2	0.7
33	0.33	1.64	0.31	2.04	1.68	0.45
34	0.33	1.89	1.01	1.82	1.26	0.5
35	0.81	1.89	0.45	2.22	1.25	0.49
36	0.48	1.5	0.17	1.59	1.27	0.31
37	0.53	1.75	0.75	1.5	1.38	0.39
38	0.33	1.5	0.57	1.74	1.17	0.5
39	0.33	1.5	0.7	1.94	1.43	0.5
mean	**0.48**	**1.82**	**0.68**	**1.78**	**1.42**	**0.49**

Таблица 12. Значимости признаков для 9 моделей, улучшающих прогноз. Вес убывает с глубиной ($w_n = 1/d_n$).

N	pregnan	glucose	blood	mass	pedigree	age
5	0.33	1.89	0.62	1.67	1.65	0.45
12	0.78	1.81	0.34	2.13	1.31	0.64
20	0.33	1.70	0.62	2.24	1.49	0.50
21	0.70	2.48	0.93	1.03	1.13	0.25
25	0.70	1.95	0.74	1.62	1.93	0.64
27	0.98	2.45	0.62	1.99	1.63	0.33
32	0.73	2.04	1.18	1.33	1.20	0.70
35	0.81	1.89	0.45	2.22	1.25	0.49
37	0.53	1.75	0.75	1.50	1.38	0.39
mean	**0.66**	**2.00**	**0.69**	**1.75**	**1.44**	**0.49**

F. ОСНОВНЫЕ РЕЗУЛЬТАТЫ 10 ИТЕРАЦИЙ ПО ВЫЯВЛЕНИЮ ПОДКЛАССОВ.

Таблицами 13b–13d иллюстрируются детали преобразования исходной двухклассовой классификации в классификацию с добавлением к ней нового класса.

В Таблицах 14–20 опущены детали, содержащиеся в Таблицах 13b–13d и для краткости приводятся только результаты серий уточнения классификаций, начиная с четырех классов на входе, позволяющие видеть, как применение исследуемой процедуры выделения подклассов приводит к диагонализации "Матрицы ошибок" в системе уточняемых подклассов.

Таблица 13b. Результат группировки ошибок для формирования третьего класса.

clss\ress	1	2	3	**vol**	**true1**	**true2**
iniClass 1	270	0	0	270	270	43
iniClass 2	0	91	0	91	96	91
iniClass 3	43	96	0	139	0	0
summ	313	187	0	500	366	134

Таблица 13c. Результат промежуточного моделирования на трех классах.

clss\ress	1	2	3	**vol**	**true1**	**true2**
iniClass 1	193	10	67	270	218	18
iniClass 2	3	61	27	91	32	64
iniClass 3	40	25	74	139	116	52
summ	236	96	168	500	366	134

Таблица 13d. Скорректированные ошибками три класса для следующей итерации.

clss\ress	1	2	3	**vol**	**true1**	**true2**
iniClass 1	236	0	0	236	218	18
iniClass 2	0	96	0	96	32	64
iniClass 3	0	0	168	168	116	52
summ	236	96	168	500	366	134

Таблица 14. Результат моделирования на пяти классах.

clss\ress	1	2	3	4	5	**vol**	**true1**	**true2**
iniClass 1	151	1	0	2	11	165	158	3
iniClass 2	2	45	1	0	6	54	10	46
iniClass 3	1	0	112	0	11	124	106	35
iniClass 4	0	0	3	49	14	66	45	20
iniClass 5	7	10	25	14	35	91	47	30
summ	161	56	141	65	77	500	366	134

Таблица 15. Результат моделирования на шести классах.

clss\ress	1	2	3	4	5	6	**vol**	**true1**	**true2**
iniClass 1	140	0	1	0	2	11	154	145	3
iniClass 2	0	42	2	0	1	3	48	7	39
iniClass 3	2	0	104	4	5	11	126	90	25
iniClass 4	2	0	0	31	7	7	47	40	15
iniClass 5	0	0	1	9	26	15	51	35	26
iniClass 6	4	4	7	11	20	28	74	49	26
summ	148	46	115	55	61	75	500	366	134

Таблица 16. Результат моделирования на семи классах.

clss\ress	1	2	3	4	5	6	7	**vol**	**true1**	**true2**
iniClass 1	134	0	0	1	2	1	0	138	134	2
iniClass 2	0	38	0	1	2	0	0	41	3	37
iniClass 3	0	0	71	1	6	4	13	95	68	19
iniClass 4	0	0	0	32	1	1	4	38	39	14
iniClass 5	0	0	2	2	34	0	5	43	37	20
iniClass 6	0	1	3	1	0	30	8	43	38	13
iniClass 7	2	1	11	15	12	15	46	102	47	29
summ	136	40	87	53	57	51	76	500	366	134

Таблица 17. Результат моделирования на восьми классах.

clss\ress	1	2	3	4	5	6	7	8	**vol**	**true1**	**true2**
iniClass 1	123	0	0	0	0	0	0	4	127	123	2
iniClass 2	0	36	1	0	1	0	0	0	38	5	34
iniClass 3	0	1	43	1	0	0	9	11	65	59	19
iniClass 4	0	0	1	33	2	1	2	6	45	39	14
iniClass 5	0	0	0	0	35	0	1	2	38	32	17
iniClass 6	0	0	0	0	0	36	1	0	37	34	13
iniClass 7	0	0	5	2	1	0	21	10	39	40	22
iniClass 8	2	2	28	17	10	10	28	14	111	34	13
summ	125	39	78	53	49	47	62	47	500	366	134

Таблица 18. Результат моделирования на девяти классах.

clss\ress	1	2	3	4	5	6	7	8	9	**vol**	**true1**	**true2**
iniClass 1	120	0	0	0	0	0	0	1	1	122	118	2
iniClass 2	0	37	0	0	0	0	0	0	0	37	5	35
iniClass 3	0	0	44	0	0	0	1	0	8	53	46	15
iniClass 4	0	0	0	39	0	0	0	2	2	43	30	12
iniClass 5	0	1	0	0	41	0	0	0	0	42	30	12
iniClass 6	0	0	0	0	0	40	2	0	1	43	35	14
iniClass 7	0	0	2	0	0	2	29	5	1	39	26	19
iniClass 8	0	0	0	0	0	0	3	7	5	15	30	7
iniClass 9	0	2	15	3	1	7	10	22	46	106	46	18
summ	120	40	61	42	42	49	45	37	64	500	366	134

Таблица 19. Результат моделирования на десяти классах.

clss\ress	1	2	3	4	5	6	7	8	9	10	**vol**	**true1**	**true2**
iniClass 1	116	0	0	0	0	0	0	2	0	2	120	118	2
iniClass 2	0	36	2	0	0	0	0	0	0	0	38	3	35
iniClass 3	0	1	53	0	0	0	0	0	2	0	56	49	16
iniClass 4	0	0	0	42	0	0	0	0	0	0	42	30	12
iniClass 5	0	0	0	0	42	0	0	0	0	0	42	30	12
iniClass 6	0	0	0	0	0	41	1	1	0	0	43	30	11
iniClass 7	0	0	0	0	0	0	25	6	0	0	31	20	14
iniClass 8	0	0	0	0	0	0	2	12	1	4	19	29	7
iniClass 9	0	1	2	0	0	0	1	3	28	7	42	26	12
iniClass 10	4	0	8	0	0	0	5	12	7	31	67	31	13
summ	120	38	65	42	42	41	34	36	38	44	500	366	134

Таблица 20. Результат моделирования на одиннадцати классах.

clss\ress	1	2	3	4	5	6	7	8	9	10	11	**vol**	**true1**	**true2**
iniClass 1	104	0	0	0	0	0	0	0	0	0	10	114	102	2
iniClass 2	0	32	0	0	0	0	0	0	1	0	0	33	2	30
iniClass 3	0	0	57	0	0	0	0	0	0	0	3	60	45	12
iniClass 4	0	0	0	35	3	0	0	0	0	0	1	39	33	13
iniClass 5	0	0	0	3	33	0	0	0	0	0	1	37	26	12
iniClass 6	0	0	0	0	0	39	0	0	0	0	2	41	30	10
iniClass 7	0	0	0	0	0	0	27	0	0	0	0	27	18	14
iniClass 8	0	0	0	2	0	0	0	16	0	0	0	18	24	4
iniClass 9	0	0	0	0	0	0	0	0	30	0	0	30	24	14
iniClass 10	0	0	0	0	0	0	0	1	0	20	1	22	17	7
iniClass 11	0	0	0	6	2	1	5	11	7	4	43	79	45	16
summ	104	32	57	46	38	40	32	28	38	24	61	500	366	134

СПИСОК ЛИТЕРАТУРЫ

1. Miller T. *Explanation in Artificial Intelligence: Insights from the Social Sciences.* arXiv preprint. arXiv:1706.07269, 2017.
2. Hall P., Phan W., Ambati S.S. Ideas on interpreting machine learning. *O'Reilly Ideas*, 2017. URL: https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning
3. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box.* Leanpub, 2019 .
4. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
5. Breiman L. Random Forests. *Machine Learning*, 2001, vol. 45, no. 1, pp. 5–32.
6. Fisher A., Rudin C., Dominici F. *Model Class Reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective.* arXiv preprint. arXiv:1801.01489v1, 2018.
7. Lundberg S., Lee S. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, 2017.
8. Hall P., Gill N., Meng L. Testing machine learning explanation techniques. *O'Reilly Ideas*, 2018. URL: https://www.oreilly.com/ideas/testing-machine-learning-interpretability-techniques
9. LIME. URL: https://lime.readthedocs.io/
10. SHAP. URL: https://shap.readthedocs.io/
11. Skater. URL: https://oracle.github.io/Skater/
12. ELI5. URL: https://eli5.readthedocs.io/
13. А.Л. Лунц, В.Л. Браиловский. Об оценке признаков, полученных в статистических процедурах распознавания. Известия АН СССР, *Техническая кибернетика*, 1969, № 3
14. В.Н. Тутубалин. *Теория вероятностей в естествознании.* Математика, кибернетика. М.: Знание, 1972.
15. Скользящий контроль. URL: http://www.machinelearning.ru/wiki/index.php?title=CV
16. К.В. Воронцов. Комбинаторный подход к оценке качества обучаемых алгоритмов. *Математические вопросы кибернетики*, М.: Физматлит, 2004, № 13, стр. 5–36.
17. Scikit-learn. Decision Trees. URL: https://scikit-learn.org/stable/modules/tree
18. Pima Indians Diabetes Data Set. URL: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.
19. Validation curves: plotting scores to evaluate models. Learning curve. URL: http://scikit-learn.org/stable/modules/learning_curve.html#learning-curve
20. Perlich C. *Learning Curves in Machine Learning.* IBM Research Report. Yorktown Heights, 2009, no. RC24756.
21. Learning Curve. URL: http://www.ritchieng.com/machinelearning-learning-curve/
22. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 2018, vol. 180 Pt A, pp. 68–77.
23. Applying Machine Learning. URL: http://www.ritchieng.com/applying-machine-learning/
24. С. Гланц. *Медико-биологическая статистика.* М.: Практика, 1998. (Stanton A. Glantz. *Primer of Biostatistics.* McGraw-Hill, Health Professions Division, 1997.)
25. Gary M. Weiss and Alexander Battistin. Generating Well-Behaved Learning Curves: An Empirical Study. *Proceedings of the Tenth International Conference on Data Mining.* Las Vegas, NV, 2014, pp. 210–213.
26. Бонгард М.М. *Проблема узнавания.* М.: Физматгиз, 1967.

27. Strickland E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* 2019, vol. 56, no. 4, pp. 24–31.

28. В.Е. Левит, В.С. Переверзев-Орлов. *Структура и поле данных при распознавании образов.* М.: Наука, 1984.

29. Data Science Predicting The Future. URL: https://www.kdnuggets.com/2018/06/data-science-predicting-future.html

30. Ahlqvist E., Storm P., Käräjämäki A. et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet, Diabetes & Endocrinology.* May, 2018, vol. 6, no. 5, pp. 361–369.

31. Растригин Л.А., Эренштейн Р.Х. Метод коллективного распознавания. *Библиотека по автоматике*, М.: Энергоиздат, 1981, вып. 615.

Cognitive Aspects of Cross Validation and Decision Trees

E.A. Vashchenko, M.A. Vitushko, V.S. Pereverzev-Orlov

The article presents the idea of using the cross-validation procedure as a cognitive tool aimed at interpreting machine learning models. The problems of application of cross-validation are considered, which are caused by the difficulty of obtaining stable estimates in the conditions of nonstationarity of the processes generating the studied data. Some non-standard features based on the cross-validation procedure are also considered. We show techniques to identify the differences between models, to compare their alternatives, to give an interpretation of the models, to clarify the classification and to put forward hypotheses about unknown subclasses. The goal is to find additional ways of evaluating learning outcomes, allowing to improve the constructed models.

KEYWORDS: machine learning, interpreting, cross-validation, sampling sufficiency, model dynamics, model stability, data completeness, tree unweaving, informative features of trees.