НОВЫЙ МЕТОД ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ В НАБОРЕ ДАННЫХ НА ПРИМЕРЕ ИММУНОСИГНАТУР

DOI: 10.17691/stm2019.11.2.03
УДК 616–071:579
Поступила 3.05.2018 г.

© **А.А. Кошечкин,** лаборант Научно-образовательного центра компьютерных наук и технологий;
В.С. Андрющенко, лаборант Научно-образовательного центра компьютерных наук и технологий; программист Института прикладной математики и компьютерных наук;
А.В. Замятин, д.техн.н., зав. кафедрой теоретических основ информатики; директор Научно-образовательного центра компьютерных наук и технологий

Национальный исследовательский Томский государственный университет, проспект Ленина, 36, Томск, 634050

Технология иммуносигнатур основана на применении пептидных микрочипов, в которых пептиды со случайными аминокислотными последовательностями используются для обнаружения болезней в зависимости от изменений в профиле циркулирующих антител. Определение заболеваний производится при помощи алгоритмов классификации, обученных на редуцированной выборке иммуносигнатурных паттернов пациентов с известными диагнозами.

Цель исследования — разработка нового метода восстановления пропущенных значений применительно к данным иммуносигнатурного анализа, позволяющего сохранить качество классификации на достаточно высоком уровне.

Материалы и методы. В работе использовались данные иммуносигнатурного анализа, полученные с использованием пептидного микрочипа высокого разрешения, имеющего порядка десяти тысяч пептидных ячеек.

Произведена оценка применимости различных методов восстановления пропущенных значений, таких как простое восстановление, метод k-взвешенных ближайших соседей, в том числе методов, основанных на использовании машинного обучения: линейная регрессия, случайный лес, градиентный бустинг.

Результаты. В рамках исследования разработан метод восстановления пропущенных значений на основе градиентного бустинга. Принцип его работы заключается в последовательном обходе всех признаков и обучении на экземплярах, чьи значения присутствуют в признаке, с последующим уточнением отсутствующих значений признака. Такая операция повторяется, пока суммарная ошибка обучения по всем признакам продолжает уменьшаться или пока не будет достигнуто максимальное число итераций. В качестве метрики ошибки обучения используется среднеквадратичная ошибка.

Для оценки качества восстановления пропущенных значений в нашем исследовании применяются результаты классификации по данным после процедуры восстановления. Выявлена эффективность вариации предложенного в статье метода восстановления пропущенных значений, основанного на линейном градиентном бустинге, в условиях высокого содержания пропущенных значений по сравнению с рассматриваемыми аналогами. Результаты работы демонстрируют перспективность применения методов машинного обучения для восстановления пропущенных значений в данных иммуносигнатурного анализа.

Ключевые слова: ранняя диагностика заболеваний; иммуносигнатура; восстановление пропущенных значений в выборке; машинное обучение.

Как цитировать: Koshechkin A.A., Andryushchenko V.S., Zamyatin A.V. A new method to missing value imputation for immunosignature data. Sovremennye tehnologii v medicine 2019; 11(2): 19–24, https://doi.org/10.17691/stm2019.11.2.03

English

A New Method to Missing Value Imputation for Immunosignature Data

A.A. Koshechkin, Laboratory Assistant, Scientific and Educational Center of Computer Science and Technology;
V.S. Andryushchenko, Laboratory Assistant, Scientific and Educational Center of Computer Science and Technology; Programmer, Institute of Applied Mathematics and Computer Science;
A.V. Zamyatin, DSc, Head of the Department of Theoretical Foundations of Informatics; Director of the Scientific and Educational Center of Computer Science and Technology

National Research Tomsk State University, 36 Lenin Avenue, Tomsk, 634050, Russia

Для контактов: Кошечкин Александр Алексеевич, e-mail: kaa1994g@mail.ru

The immunosignature technology uses microarray chips of random amino acid sequence peptides to detect diseases based on the change in the profile of circulating antibodies. Diseases are detected using classification algorithms trained on a reduced sample of immunosignature patterns of patients with known diagnoses.

The aim of the study was to develop a new method of missing value imputation in immunosignature data, which allows maintaining sufficient accuracy of classification.

Materials and Methods. The study was carried out using immunosignature data obtained by utilizing a high-resolution peptide microarray chip with nearly ten thousand peptide cells.

The applicability of various missing value imputation methods such as simple imputation, weighted k-nearest neighbors and machine learning techniques (linear regression, random forest, gradient boosting) was evaluated.

Results. Missing value imputation method based on gradient boosting has been developed in the framework of the study. Its operating principle implies iterating through all features (attributes) and training on examples (samples) whose values are present in the feature, followed by clarification of missing feature values. This process is repeated until the total training error for all features stops decreasing or until the maximum number of iterations is reached. The root mean squared error is employed as a training error metric.

To assess the quality of missing value imputation, classification results based on the data obtained after imputation procedure are used in our investigation.

The proposed missing value imputation algorithm based on linear gradient boosting proves to be effective under conditions of a high proportion of missing values as compared to other methods under consideration. The results of the investigation demonstrate the viability of using machine learning techniques for missing value imputation in immunosignature data.

Key words: early diagnosis of diseases; immunosignature; missing value imputation in the sample; machine learning.

Введение

На сегодняшний день в связи с бурным развитием современной техники для анализа доступны большие объемы различной информации. Однако зачастую возникает проблема репрезентативности данных, которая неизбежна при сборе и анализе этой информации. Репрезентативность в основном снижается вследствие пропущенных значений под влиянием шума и человеческого фактора при сборе информации. Пропущенные значения — одна из наиболее острых проблем для исследователей, ведь все методы анализа данных в такой ситуации некорректно работают или теряют свою эффективность. Поэтому восстановление пропущенных значений в выборке является одной из первостепенных задач при анализе данных [1].

В настоящее время многие виды рака потенциально излечимы, если диагностированы на ранней стадии заболевания. Для раннего выявления злокачественных опухолей необходимы простые, недорогие, минимально инвазивные, но при этом точные методы диагностики [2]. Технология иммуносигнатур (immunosignature) является одним из таких методов [3].

Данная технология основана на применении пептидных микрочипов, когда пептиды со случайными аминокислотными последовательностями используются для обнаружения болезней в зависимости от изменений в профиле циркулирующих антител [4]. Данные пептиды служат частичным или полным подобием эпитопов антигена. Использование множества пептидов, представляющих вероятные аминокислотные последовательности белков, делает возможным определение связывающего партнера для многих антител, даже если точное совпадение для эпитопа отсутствует [5]. Полученные данные могут помочь медицинскому работнику в постановке окончательного диагноза. Однако в процессе их получения могут возникнуть ошибки, вызванные как технологическими особенностями оборудования, так и последующей оцифровкой данных. Из-за этого теряется репрезентативность данных и некоторые значения светимостей пептидов становятся недоступными. Этим обусловлена актуальность решения задачи адекватного восстановления пропущенных значений.

Целью данного исследования явилась оценка существующих методов восстановления пропущенных значений в выборке и разработка эффективного метода обработки таких значений.

Для достижения цели необходимо решить следующие задачи:

провести анализ существующих методов восстановления пропущенных значений, основанных на различном математическом аппарате;

на основе полученных знаний предложить наиболее подходящий метод восстановления пропущенных значений, применимый к данным иммуносигнатур;

сравнить эффективность применения существующих и предложенного в рамках исследования методов восстановления пропущенных значений к иммуносигнатурным данным.

Материалы и методы

Материалы, полученные с помощью анализа иммуносигнатур, представляют собой набор значений интенсивности флюоресценции пептидов, где названия пептидов являются столбцами (признаками), а метки классов (экземпляры) — строками. В исследовании использовался набор данных из публичного хранилища биомедицинских данных (идентификатор GSE52580), полученный посредством оцифровки дан-

ных иммуносигнатур [6, 7]. Набор данных не имеет пропущенных значений, что позволяет управлять характером и количеством пропущенных значений в нашей выборке. Набор данных имеет следующие характеристики:

количество экземпляров — 240;
количество признаков — 9781;
количество классов — 6.

Количество экземпляров каждого класса одинаково, поэтому набор данных представляет собой сбалансированную выборку. В табл. 1 показан фрагмент набора данных с известными классами болезней и пептидами**.** Названия пептидов представлены в виде последовательности аминокислот.

Таблица 1

Фрагмент пептидной матрицы

Метка класса	Название пептида	
	CSGYNSFAMKANYIFNG	CSGSNYYDWWFRIAVMITI
Рак мозга	5.27889752	9.15952333
Рак молочной железы	0.89180777	0.89329176
Рак пищевода	0.88392227	1.12217693
Множественная миелома	0.82533253	0.93682348
Рак поджелудочной железы	0.96485786	1.02698893
Контрольный класс (здоровые лица)	0.85648045	0.84041385

Рассматриваемые методы восстановления пропущенных значений. На сегодняшний день существует большое количество различных методов восстановления пропущенных значений, которые применимы к решению данной проблемы [8–10], поэтому для анализа целесообразно выбрать методы, основанные на принципиально различном математическом аппарате. Рассмотрим выбранные методы подробнее.

Простое восстановление данных (simple imputation) — один из самых простых и известных методов восстановления пропущенных значений [11]. Он заключается в замене пропущенных значений признака медианой, средним или модой, вычисленными по присутствующим значениям признака. Преимущество данного метода — в быстром восстановлении пропущенных значений. Однако если таких значений достаточно много, простое восстановление приведет к существенному искажению результатов анализа данных.

Метод k-взвешенных ближайших соседей (weighted k nearest neighbors algorithm) — простой и эффективный метод восстановления пропущенных значений, основывающийся на гипотезе о том, что если экземпляры близки в пространстве измеренных признаков, то из этого следует их близость по неизмеренным признакам [12]. Расстояние между двумя экземплярами и вычисляется по присутствующим значениям признаков. Для вычисления пропущенных значений используются средневзвешенные значения соседей [13].

Восстановить пропущенные значения в конкретном признаке можно, предсказав его значения по другим признакам с помощью различных методов машинного обучения. Суть данного подхода заключается в последовательном объявлении каждого признака в качестве целевой переменной, далее происходит обучение на экземплярах, не имеющих пропущенных значений в целевой переменной, с последующим предсказанием пропущенных значений целевой переменной. Так как среди признаков, по которым производится обучение, имеются пропущенные значения, то необходимо их изначально заменить с помощью одного из простейших методов восстановления, а затем уточнить с помощью одного из методов машинного обучения. Рассмотрим несколько наиболее популярных методов машинного обучения подробнее.

Линейная регрессия (linear regression) — метод машинного обучения, заключающийся в построении уравнения с линейной функцией зависимости целевого признака от одного или нескольких других признаков [14]. Для двух признаков Y_1 и Y_2 можно построить уравнение зависимости одного признака от другого:

$$Y_2=aY_1+b,$$

используя значения, известные для каждого признака, и восстановить пропущенные значения с помощью полученного регрессионного уравнения по имеющимся значениям. Данный метод является эффективным только при наличии определенного уровня линейной зависимости между признаками.

Случайный лес (random forest) — метод машинного обучения, основанный на множестве деревьев принятия решений, объединенных в ансамбль [15]. Данный метод является универсальным решением и эффективно работает как с непрерывными, так и с категориальными признаками [16]. Для категориальных признаков предсказанное значение определяется по принципу мажоритарного голосования каждого отдельного дерева в ансамбле. В свою очередь для числовых признаков предсказанное значение определяется как среднее между ответами каждого дерева в ансамбле.

Предложенный метод восстановления пропущенных значений. Разработанный в рамках исследования метод восстановления пропущенных значений основан на использовании такого метода машинного обучения, как градиентный бустинг.

Градиентный бустинг (gradient boosting) — метод машинного обучения, основанный на построении линейной комбинации простых алгоритмов путем изменения веса входных данных [17, 18]. Каждый простой алгоритм (линейный классификатор или дерево принятия решений) строится таким образом, чтобы придавать больший вес и оказывать предпочтение

ранее некорректно предсказанным значениям. Длина линейной комбинации простых алгоритмов равна числу раундов модели.

Поскольку существует возможность в рамках градиентного бустинга построить линейную комбинацию различных алгоритмов, мы фактически можем использовать два различных метода — *линейный градиентный бустинг и градиентный бустинг деревьев*.

В рамках восстановления пропущенных значений применение градиентного бустинга организовано следующим образом. Выполняется восстановление пропущенных значений во всем наборе данных с помощью простого восстановления (в данной работе пропущенные значения в каждом признаке заменяются на медиану, вычисленную по присутствующим значениям признака). Далее последовательно для каждого признака происходит обучение на экземплярах, чьи значения присутствуют в данном признаке, с последующим уточнением отсутствующих значений признака. Такая операция повторяется, пока суммарная ошибка обучения по всем признакам продолжает уменьшаться или пока не будет достигнуто максимальное число итераций. В качестве метрики ошибки обучения используется среднеквадратичная ошибка [19].

Сравнение эффективности работы методов восстановления пропущенных значений. Основной целью применения технологии иммуносигнатур является поддержка принятия решений в постановке диагноза, что в терминологии анализа данных сводится к задаче классификации. В связи с этим для оценки качества восстановления пропущенных значений в нашем исследовании используются результаты классификации по данным после процедуры восстановления.

Этот набор данных уже был применен ранее в исследованиях [20], где установлено, что алгоритм случайный лес показывает высокие результаты классификации, в связи с чем его использование в работе в качестве классификатора является вполне целесообразным. В задачах классификации для оценки качества применяются различные метрики, а их выбор и анализ — непременная часть любого исследования [21]. Учитывая, что набор данных представляет собой сбалансированную выборку, в нашей работе мы применили метрику классификации «доля правильных ответов» (accuracy).

Набор данных обладает большим количеством признаков, что негативно сказывается на времени проведения вычислений. В связи с этим для экономии времени из этого набора с помощью t-критерия Стьюдента отобраны только 120 самых информативных признаков [22].

Для проведения сравнения эффективности рассматриваемых методов необходимо сформировать выборку с параметрами, варьируемыми в широких пределах. Это целесообразно реализовать путем создания пропущенных значений искусственно. При многоступенчатом анализе пептидного микрочипа возможно появление пропущенных значений на любой из стадий, без какой-либо закономерности. Ввиду этого невозможно подобрать метод создания пропущенных значений, имитирующих все возможные сценарии. В связи с этим в данном исследовании использовался метод полностью случайного создания пропущенных значений (missing completely at random) [23, 24].

Для сравнения рассматриваемых методов необходимо выполнить следующие шаги:

1) создать несколько наборов данных с различным количеством пропущенных значений;
2) восстановить пропущенные значения в образованных наборах данных поочередно каждым из рассматриваемых методов;
3) провести классификацию по данным каждого восстановленного набора;
4) повторить шаги 1–3 тридцать раз;
5) вычислить средние значения точности классификации, проведенной по восстановленным наборам, для каждого из рассматриваемых методов при различном количестве пропущенных значений;
6) оценить полученные результаты.

Вся работа выполнена с использованием языка программирования R и программных библиотек из публичного репозитория CRAN (табл. 2).

Таблица 2

Используемые пакеты и их параметры

Методы	Пакет/параметры настройки
Простое восстановление	caret/medianImpute
Метод k-взвешенных ближайших соседей	wNNSel
Линейная регрессия	ice/norm.predict
Случайный лес	missForest/100 деревьев, максимум итераций — 10
Линейный градиентный бустинг	xgboost/максимум итераций — 10, максимум раундов — 100, eta = 0.3
Градиентный бустинг деревьев	

Результаты

Необходимо отметить, что по результатам исследования выявлена неспособность методов линейной регрессии и k-взвешенных ближайших соседей обрабатывать наборы данных, доля пропущенных значений в которых превышает 0.7 и 0.85 соответственно. В связи с этим для данных методов частично отсутствуют значения оценок эффективности их работы.

Исходя из результатов (рис. 1), использование методов машинного обучения для восстановления пропущенных значений является наиболее эффективным решением. Наилучшие результаты показали случайный лес и линейный градиентный бустинг.

В табл. 3 представлены максимальные и минимальные значения доли правильных ответов.

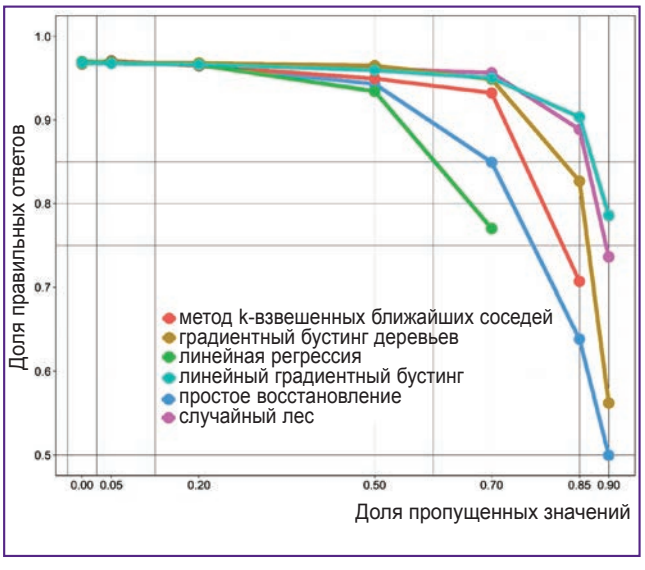

Рис. 1. Результаты классификации по данным после процедуры восстановления варьируемой доли пропущенных значений

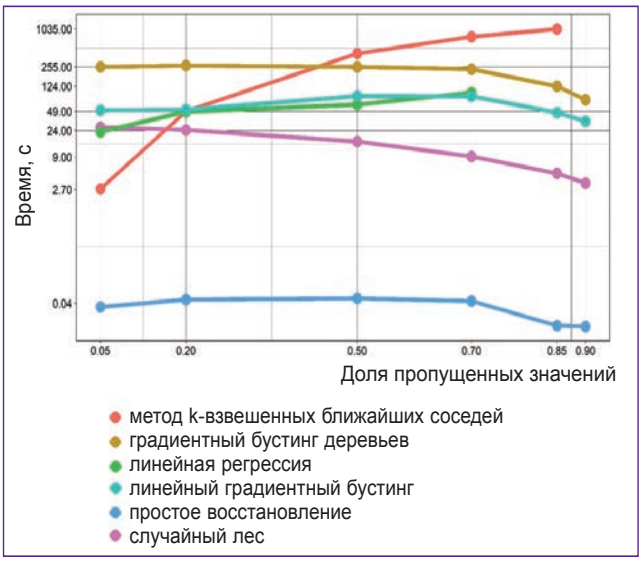

Рис. 2. Время работы методов с различной долей пропущенных значений в логарифмической шкале

Таблица 3

Минимальные и максимальные значения результатов классификации

Методы и модели		Доля пропущенных значений					
		0.05	0.20	0.50	0.70	0.85	0.90
Простое восстановление	max	0.974	0.973	0.956	0.872	0.668	0.553
	min	0.965	0.957	0.930	0.822	0.608	0.432
Метод k-взвешенных ближайших соседей	max	0.977	0.971	0.961	0.951	0.743	NA
	min	0.966	0.955	0.937	0.909	0.657	NA
Линейная регрессия	max	0.974	0.971	0.947	0.806	NA	NA
	min	0.961	0.960	0.922	0.721	NA	NA
Случайный лес	max	0.973	0.973	0.971	0.972	0.921	0.811
	min	0.964	0.960	0.952	0.938	0.846	0.621
Линейный градиентный бустинг	max	0.974	0.972	0.970	0.960	0.930	0.823
	min	0.963	0.957	0.950	0.931	0.877	0.691
Градиентный бустинг деревьев	max	0.975	0.975	0.974	0.973	0.861	0.694
	min	0.964	0.969	0.956	0.933	0.736	0.444

Примечание: NA (англ. *not available*) — отсутствующее значение.

При доле пропущенных значений вплоть до 0.7 такие методы, как случайный лес, линейный градиентный бустинг и градиентный бустинг деревьев, сохраняют одинаковые показатели разброса, что говорит об их высокой эффективности.

На основании данных табл. 3 и рис. 1 можно сделать вывод, что при высоком содержании пропущенных значений в наборе данных случайный лес и градиентный бустинг деревьев показывают резкое ухудшение результатов. В связи с этим линейный градиентный бустинг является в этой ситуации более предпочтительным. В то же время при низком содержании пропущенных значений различий между методами не наблюдается.

Исходя из значений времени работы методов (рис. 2) можно выделить две особенности.

Первая — время работы алгоритмов восстановления пропущенных значений, основанных на методах машинного обучения, уменьшается с увеличением количества пропущенных значений. Это происходит ввиду того, что при увеличении количества отсутствующих значений признака уменьшается размер обучающей выборки, а следовательно, и время обучения алгоритма.

Вторая — при малом количестве пропущенных значений метод k-взвешенных ближайших соседей является предпочтительным ввиду незначительных временны́х затрат на восстановление пропущенных значений.

Заключение

По результатам исследования выявлена эффективность предложенного в статье метода восстановления пропущенных значений, основанного на линейном градиентном бустинге, в условиях высокого содержания пропущенных значений по сравнению с рассматриваемыми аналогами. В свою очередь метод k-взвешенных ближайших соседей является наиболее предпочтительным при низком содержании пропущенных значений ввиду незначительных временны́х затрат на обработку набора данных и сравнимой с более сложными методами эффективности работы.

Полученные знания являются основой для будущих исследований с последующим созданием программных пакетов для предварительной обработки данных пептидных микрочипов.

Финансирование исследования и конфликт интересов. Исследование не финансировалось какими-либо источниками, и конфликты интересов, связанные с данным исследованием, отсутствуют.

Литература/References

1. Padgett C.R., Skilbeck C.E., Summers M.J. Missing data: the importance and impact of missing data from clinical research. *Brain Impairment* 2014; 15(01): 1–9, https://doi.org/10.1017/brimp.2014.2.

2. Осипова Т.В., Рябых Т.П., Барышников А.Ю. Диагностические микрочипы: применение в онкологии. Российский биотерапевтический журнал 2006; 5(3): 72–81. Osipova T.V., Ryabykh T.P., Baryshnikov A.Yu. Diagnostic microchips: application in oncology. *Rossijskij bioterapevticeskij zurnal* 2006; 5(3): 72–81.

3. O'Donnell B., Maurer A., Papandreou-Suppappola A., Stafford P. Time-frequency analysis of peptide microarray data: application to brain cancer immunosignatures. *Cancer Inform* 2015; 14(Suppl 2): 219–233, https://doi.org/10.4137/cin.s17285.

4. Richer J., Johnston S.A., Stafford P. Epitope identification from fixed-complexity random-sequence peptide microarrays. *Mol Cell Proteomics* 2014; 14(1): 136–147, https://doi.org/10.1074/mcp.m114.043513.

5. Kukreja M., Johnston S.A., Stafford P. Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. *J Proteomics Bioinform* 2013; S6: 001, https://doi.org/10.4172/jpb.s6-001.

6. Stafford P., Halperin R., Legutki J.B., Magee D.M., Galgiani J., Johnston S.A. Physical characterization of the "immunosignaturing effect". *Mol Cell Proteomics* 2012; 11(4): M111.011593, https://doi.org/10.1074/mcp.m111.011593.

7. *National Center for Biotechnology Information Search database*. URL: https://www.ncbi.nlm.nih.gov/.

8. Efromovich S. Nonparametric regression with missing data. *Wiley Interdisciplinary Reviews: Computational Statistics* 2014; 6(4): 265–275, https://doi.org/10.1002/wics.1303.

9. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 2007; 16(3): 219–242, https://doi.org/10.1177/0962280206074463.

10. Zloba E., Yatskiv I. Statistical methods of reproducing of missed data. *Computer Modelling & New Technologies* 2002; 6(1): 51–61.

11. Žliobaitė I., Hollmén J. Optimizing regression models for data streams with missing values. *Machine Learning* 2014; 99(1): 47–73, https://doi.org/10.1007/s10994-014-5450-3.

12. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; 17(6): 520–525, https://doi.org/10.1093/bioinformatics/17.6.520.

13. Tutz G., Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis* 2015; 90: 84–99, https://doi.org/10.1016/j.csda.2015.04.009.

14. Little R.J.A. Regression with missing X's: a review. *J Am Stat Assoc* 1992; 87(420): 1227–1237, https://doi.org/10.2307/2290664.

15. Breiman L. Random forests. *Machine Learning* 2001; 45(1): 5–32.

16. Stekhoven D.J., Buhlmann P. MissForest — nonparametric missing value imputation for mixed-type data. *Bioinformatics* 2011; 28(1): 112–118, https://doi.org/10.1093/bioinformatics/btr597.

17. Natekin A., Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot* 2013; 7: 21, https://doi.org/10.3389/fnbot.2013.00021.

18. Friedman J.H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 2001; 29(5): 1189–1232, https://doi.org/10.1214/aos/1013203451.

19. Hyndman R.J., Koehler A.B. Another look at measures of forecast accuracy. *Int J Forecast* 2006; 22(4): 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

20. Andryushchenko V.S., Uglov A.S., Zamyatin A.V. Statistical classification of immunosignatures under significant reduction of the feature space dimensions for early diagnosis of diseases. *Sovremennye tehnologii v medicine* 2018; 10(3): 14–20, https://doi.org/10.17691/stm2018.10.3.2.

21. Arlot S., Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys* 2010; 4: 40–79, https://doi.org/10.1214/09-ss054.

22. Student. The probable error of a mean. *Biometrika* 1908; 6(1): 1–25, https://doi.org/10.2307/2331554.

23. Rubin D.B. Inference and missing data. *Biometrika* 1976; 63(3): 581, https://doi.org/10.2307/2335739.

24. Тихова Г.П. Пропуск данных в выборке: как решать проблему и как ее избежать. Регионарная анестезия и лечение острой боли 2016; 10(3): 205–209. Tikhova G.P. Data missing: how to solve and how to escape the problem. *Regionarnaya anesteziya i lechenie ostroy boli* 2016; 10(3): 205–209.