НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2023 Том 23 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2023 Vol. 23 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2023-23-2-364-373
УДК 004.8

Мониторинг состояния здоровья населения по возрастным группам

Николай Александрович Игнатьев[1]✉, Мехрбону Акром кизи Рахимова[2]

[1,2] Национальный университет Узбекистана имени Мирзо Улугбека, Ташкент, 100174, Узбекистан

[1] n_ignatev@rambler.ru✉, https://orcid.org/0000-0002-7150-5837
[2] mehribonu@gmail.com, https://orcid.org/0000-0001-5849-3395

Аннотация

Рассмотрена многокритериальная методика отбора информативных наборов разнотипных признаков для количественной оценки состояния здоровья населения по 14 возрастным группам. Для сравнения выборок из двух классов (групп) сформировано унифицированное описание объектов по двум градациям номинальных признаков. Полученное описание использовано для синтеза латентных признаков и вычисления значений меры компактности объектов классов на числовой оси. Преобразование количественных признаков в градации номинальных реализовано по критерию поиска минимального покрытия их значений непересекающимися интервалами. Значения границ интервалов и их число определено рекурсивным алгоритмом с учетом принадлежности объектов к классам. Отмечено важное свойство преобразования — инвариантность к масштабам измерений. Предложена формула для вычисления функции принадлежности объектов классов по каждой градации признака. Значения функции применены при унификации описаний объектов и вычислении показателя устойчивости признака вне зависимости от его шкалы измерений. Унификация описаний по двум градациям не меняет показателя устойчивости, но увеличивает вклад каждой градации в разделение объектов классов. Ранжирование признаков по отношению к их устойчивости применено как для отдельных выборок, так и на множестве определяемых выборок. Результаты ранжирования по множеству выборок использованы для поиска закономерностей по отдельным признакам и формирования из них наборов для вычисления значений латентных признаков объектов. Множество из 13 выборок данных из представителей двух классов сформировано следующим образом. Первый класс представлен объектами младшей возрастной группы, второй — объектами разных возрастных групп. Определен набор из семи разнотипных признаков. По каждой из 13 выборок вычислены значения латентных признаков на этом наборе и меры компактности объектов классов на числовой оси. Получена монотонно неубывающая последовательность значений мер компактности выборок данных, инвариантных к порядку старшинства возрастных групп. Свойство монотонности значений последовательности согласуется с эмпирическими оценками состояния здоровья в процессе старения населения.

Ключевые слова

нелинейные преобразования, функции принадлежности, ранжирование признаков, обобщенные оценки объектов

Благодарности

Работа выполнено в рамках плана научных исследований кафедры «Искусственный интеллект» Национального университета Узбекистана.

Ссылка для цитирования: Игнатьев Н.А., Рахимова М.А. Мониторинг состояния здоровья населения по возрастным группам // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 2. С. 364–373. doi: 10.17586/2226-1494-2023-23-2-364-373

Monitoring the health status of the population by age groups

Nikolay A. Ignatev[1]✉, Mekhrbonu A. Rakhimova[2]

[1,2] National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100174, Uzbekistan

[1] n_ignatev@rambler.ru✉, https://orcid.org/0000-0002-7150-5837
[2] mehribonu@gmail.com, https://orcid.org/0000-0001-5849-3395

Abstract

A multi-criteria method for selecting informative sets of different features for a quantitative assessment of the population's health status in 14 age groups is considered. To compare samples from two classes (groups), it is proposed

to form a unified description of objects according to two gradations of nominal features. The unified description is used to synthesize latent features and calculate the values of the compactness measure of class objects on the numerical axis. The transformation of quantitative features into nominal gradations is implemented according to the search criterion for the minimum coverage of their values by non-overlapping intervals. The values of the boundaries of the intervals and their number are determined by a recursive algorithm considering the objects belonging to classes. An important property of the transformation is the invariance to measurement scales. A formula is proposed for calculating the membership function of class objects for each feature gradation. Function values are used to unify object descriptions and calculate the stability index of a feature, regardless of its measurement scale. The unification of descriptions by two gradations does not change the stability index but increases the contribution of each gradation to the separation of class objects. The ranking of features about their stability was used both for individual samples and for a set of defined samples. The results of ranking over a set of samples were used to search for patterns in individual features and to form sets from them to calculate the values of latent features of objects. A set of thirteen data samples from representatives of two classes was formed as follows. The first class was represented by objects of the younger age group, and the second class — by objects of different age groups. A set of seven different types of features has been identified. For each of 13 samples, the values of latent features on this set and measures of compactness of class objects on the numerical axis were calculated. A monotonically non-decreasing sequence of values of measures of compactness of data samples that are invariant to the order of precedence of age groups is obtained. The property of monotonicity of sequence values is consistent with empirical estimates of the health state in the process of population aging.

Keywords
nonlinear transformations, membership functions, ranging of features, generalized estimates of objects

Acknowledgements
The work was carried out within the framework of the scientific research plan of the Department of Artificial Intelligence of the National University of Uzbekistan.

For citation: Ignatev N.A., Rakhimova M.A. Monitoring the health status of the population by age groups. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 2, pp. 364–373 (in Russian). doi: 10.17586/2226-1494-2023-23-2-364-373

Введение

Анализ состояния здоровья населения — важное направление социальной политики государств и мира в целом. Для получения заключения о состоянии здоровья человека необходим сбор и анализ экспериментальных медицинских данных. Сбор данных, как правило, производится по утвержденным государственным программам и стандартам либо по заказам коммерческих фирм или научно-исследовательских организаций. Как показывает новейшая история пандемии COVID-19, сбор данных о пациентах — очень дорогой и не всегда технически реализуемый процесс.

При анализе медицинских данных преследуются разные цели: упорядочить по важности факторы риска сердечно-сосудистых заболеваний; определить генетическую предрасположенность к болезням людей, проживающих на определенной территории; изучить реакцию иммунной системы на вакцинацию. Проблема получения информации из экспериментальных данных может быть связана с комбинаторной сложностью алгоритмов для анализа, использованием разных масштабов и типов шкал измерений, наличием пропусков в данных. Чаще всего для доказательства эффективности используемой методики лечения (профилактики здоровья) выбирают деление данных на экспериментальную и контрольную выборки.

Классические методы прикладной статистики не всегда являются эффективными для изучения состояния здоровья населения, поскольку для принятия решения по результатам анализа данных чаще всего за основу берутся усредненные показатели. Из-за принципа усреднения статистические методы существенно ограничивают возможности для проверки различных гипотез с целью поиска закономерностей в данных.

Альтернативным инструментом для поиска закономерностей являются методы интеллектуального анализа данных (ИАД). Применение данных методов существенно увеличивает возможности доказательства истинности гипотез, выдвигаемых экспертами-врачами. Источником нового знания, как правило, становятся неожиданные и практически полезные результаты, полученные методами ИАД.

Общепринятой методики для анализа количественных и качественных (номинальных) медицинских данных не существует. При объяснении результатов анализа часто пользуются отношением «показатель больше (меньше) нормы». Как правило, указывается интервал, в границах которого находятся значения нормы. Существует альтернативная точка зрения, что у каждого человека есть своя «норма здоровья» [1], которая не обязательно совпадает с официально принятыми показателями для оценки состояния здоровья. Для обоснования этой точки зрения необходима проверка гипотезы о существовании нескольких интервалов, в границах которых находятся значения нормы для отдельных групп людей. Например, представителей мировой элиты по отдельным видам спорта.

В работе [2] рассмотрена проблема адекватного описания данных для решения задач медицинской диагностики. Использование методов селекции и преобразования признаков позволяет сократить время обработки, повысить качество классификации и возможности интерпретации полученных решений. Отмечено, что применение линейных методов преобразования признаков не всегда является эффективным на данных, демонстрирующих нелинейность. Для описания модели данных предложено использовать комбинацию известных методов выбора размерностей и оценки качества обучаемой классификации.

В [3] представлен обзор исследований связи отдельных показателей подгрупп пациентов и значений исхода их лечения, а также установлено существование двух задач для анализа таких связей. Решение первой задачи заключается в проверке гипотез на известных подгруппах, второй — к выявлению подгрупп и их оценке.

Медицинская диагностика состояния здоровья населения чаще всего апеллирует к качественным методам оценки. За распространенным диагнозом «практически здоров» нередко скрываются три группы людей: абсолютно здоровые; находящиеся в состоянии адаптивного напряжения; имеющие высокий риск болезни или признаки предболезненных состояний. С методологической точки зрения важным допущением является то, что количественная мера состояния здоровья оказывается латентным показателем, формируемым из значений номинальных и количественных признаков.

Традиционным в медицинской практике считается деление людей на группы по возрасту. Такое деление учитывается при выборе процедур лечения, дозировке лекарств, противопоказаниях на прием медицинских препаратов и ограничениях на объемы физической нагрузки.

Решение проблемы разработки и обоснования меры состояния здоровья связано с отбором информативных признаков и разработкой процедур сравнения различных групп населения по этим признакам. Состав информативных наборов признаков, используемый для мониторинга здоровья, не является уникальным и может зависеть от гендерной принадлежности, возраста, времени и географии проживания, уровня образования, особенностей культуры питания и т. д.

Эффективность принимаемых решений по мониторингу состояния здоровья людей во многом зависит от наличия пополняемых баз медицинских данных и извлечения из них полезных знаний методами ИАД. Примерами полезных знаний могут быть изменение уровня резистенции медицинских препаратов, влияние постковидного синдрома на состояние здоровья, генетическая предрасположенность к заболеваниям или занятиям отдельными видами спорта.

Отбор информативных признаков через решение многокритериальной задачи

Обозначим типичные проблемы, связанные с отбором информативных признаков: обоснование выбора критерия отбора; зависимость числа и состава набора признаков от эвристик, используемых для реализации алгоритмов; плохая интерпретируемость результатов отбора.

Решение перечисленных проблем предлагается рассматривать как многокритериальную задачу. Одно из средств, используемых при решении данной задачи — формирование наборов исходных признаков. На основе исходных признаков выполнен синтез латентных признаков, рассмотренных в качестве метапонятий, которые являются обобщением наборов исходных признаков и позволяют выносить суждения о сходстве и различии анализируемых групп людей (классов объектов) через их попарное сравнение.

В работе [4] рассмотрена задача отбора информативных наборов разнотипных признаков на основе значений их устойчивости по парам из l ($l > 2$) непересекающихся классов объектов. При отборе использованы правила алгоритма иерархической агломеративной группировки признаков для синтеза из них латентных показателей по методу вычисления обобщенных оценок объектов. Во время применения принципа динамического программирования, при реализации алгоритма, нет гарантии сходства составов наборов информативных признаков для всех пар классов. Для случая с мониторингом состояния здоровья на определяемых выборках выполнен поиск набора информативных признаков с учетом следующих требований: множество пар из непересекающихся классов объектов формируются относительно одного указанного класса (каждой паре определено значение (номер) в порядковой шкале); значения латентного признака объектов по паре классов вычислены по информативному набору, в результате которого определена мера компактности как произведение внутриклассового сходства и межклассового различия; порядок следования значения меры компактности по каждой паре классов соответствует порядку следования ее (пары) номера.

Отметим, что существование набора признаков, отвечающего данным требованиям, рассмотрено как гипотеза. Для проверки гипотезы предложена методика, согласно которой поиск информативного набора состоит из следующих шагов:

— разбиение значений количественных признаков на непересекающиеся интервалы;
— нелинейное преобразование значений разнотипных признаков в описании объектов в {1, 2};
— ранжирование признаков по отношению к их устойчивости;
— вычисление обобщенных оценок (значений латентного признака) объектов по наборам признаков, сформированных на основе их рангов;
— анализ значений меры компактности классов по обобщенным оценкам объектов.

Формирование информативных наборов признаков

Теоретической основой при выборе признакового пространства для описания объектов классов использована гипотеза о компактности. Существуют несколько мер компактности, оценивающих отношения объектов по их описаниям: на числовой оси; в пространстве размерности два и выше.

При выборе мер компактности для построения информационных моделей в медицине необходимо учитывать: инвариантность к масштабам измерений количественных признаков; многообразие способов интерпретации наборов разнотипных признаков для принятия решений.

Числовая ось рассмотрена в качестве универсальной шкалы для анализа отношений между объектами. Исследовать отношения между объектами, описываемых набором «сырых» признаков, возможно через синтез значений латентного признака по данному набору.

Для описания объектов в работе [4] применены два способа преобразования количественных признаков в градации номинальной шкалы с использованием разбиения их значений на непересекающиеся интервалы. Для первого способа градациями выбраны номера непересекающихся интервалов. Описания объектов, полученные по первому способу, стали исходными данными для второго способа. По каждой градации признака вычислены частоты встречаемости и значения функции принадлежности объектов к одному из двух классов. На основе значений функции принадлежности сформировано новое описание объектов в виде бинарной таблицы.

Преобразование в {1, 2} градаций признаков (при числе несовпадающих значений больше числа классов) путем замены их значений на значения функции принадлежности является нелинейным и неинвариантным к порядку следования. Изменение порядка следования может привести к корректному (без ошибок) распознаванию объектов обучающей выборки по одному признаку. В работе [5] выполнено сравнение двух способов описания объектов на данных больных лейкемией. Эффективность второго (нелинейного) способа преобразования перед первым показана в увеличении точности распознавания по обобщенным оценкам объектов. В работе [6] для доказательства использован стохастический алгоритм вычисления обобщенных оценок объектов в разнотипном признаковом пространстве. Значение линейной проекции объекта на числовую ось определено как сумма проекций по количественным и номинальным признакам. Единство результатов (значений обобщенных оценок) алгоритм гарантирует только на наборах номинальных признаков.

Ранжирование — один из способов предобработки данных для формирования информативных наборов признаков в задачах классификации. Значения рангов зависят от выбора показателя для упорядочения. В работе [7] в качестве такого показателя предложено использовать устойчивость признака.

В результате применения ранжирования было значительно сокращено число наборов признаков, используемых для вычисления обобщенных оценок объектов. Получена возможность проанализировать сходство (различие) между группами как по обобщенным оценкам, так и по признакам, используемым для их синтеза. Примером может служить исследование описаний четырех групп больных с бессимптомной, легкой, среднетяжелой и тяжелой формами COVID-19. Выполнено попарное сравнение групп с целью: анализа и объяснения степени различий значений разнотипных признаков между группами; поиска закономерностей по множеству индексированных показателей (обобщенных оценок объектов) состояния здоровья относительно указанной группы.

Функциональная схема отбора информативных признаков по возрастным группам

Данные медицинских обследований[1] по 14 возрастным группам (20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, 85 лет и старше) населения Южной Кореи за период с 2002 по 2018 год послужили основой для демонстрации методики отбора информативных признаков. По данным за 2018 год сформированы 14 групп населения $G_0, \ldots, G_{13}$ мужского пола, индексы которых упорядочены по отношению старшинства возраста. Объекты каждой группы описаны набором из 23 разнотипных (17 количественных и 6 качественных) признаков. Требуется обосновать отбор информативных признаков и вычисление значений обобщенных оценок объектов на их основе для сравнения самой младшей по возрасту группы G_0 с группами $G_1, \ldots, G_{13}$. Функциональная схема отбора информативных признаков приведена на рис. 1.

Поиск закономерностей по описаниям объектов из групп $G_0, \ldots, G_{13}$

В качестве инструмента для поиска закономерностей на выборках данных использованы методы ИАД

[1] [Электронный ресурс]. Режим доступа: https://www.data.go.kr/dataset/15007122/fileData.do (дата обращения: 19.12.2022).

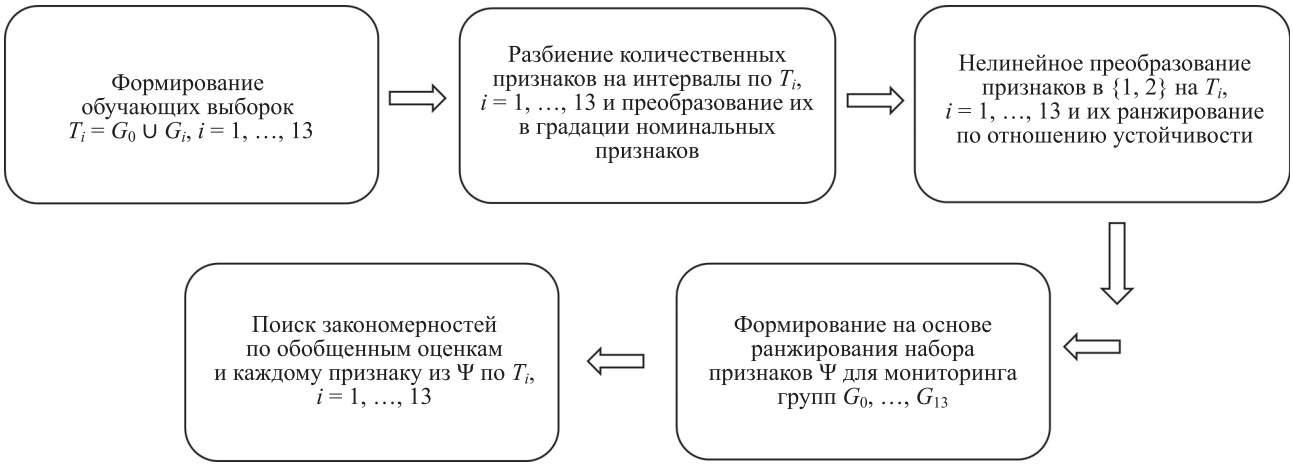

Рис. 1. Функциональная схема отбора информативных признаков

Fig. 1. Functional scheme for selecting informative features

[8]. Для упрощения записи математической символики при изложении методов ИАД множество объектов из объединения групп $T_i = G_0 \cup G_i$, $i = 1, \ldots, 13$ обозначим как обучающую выборку $E_0 = \{S_1, \ldots, S_m\}$, $m = |T_i|$, разделенную на два непересекающихся класса $K_1(G_0)$, $K_2(G_i)$. Считается, что объекты классов описываются набором разнотипных признаков $X(n) = (x_1, \ldots, x_n)$.

Разбиение количественных признаков на непересекающиеся интервалы

Пусть для значений количественного признака $x_c \in X(n)$ в описании объектов E_0 построена упорядоченная по неубыванию последовательность

$$r_1, \ldots, r_v, \ldots, r_m. \quad (1)$$

При разбиении выражения (1) на непересекающиеся интервалы их число считается неизвестным. Определено условие разбиения, что в границах каждого интервала частота встречаемости значений признака из описаний объектов класса K_t больше чем в K_{3-t}, $t = 1,2$.

В работе [6] предложен критерий для разбиения (1) на множество из p_c, ($p_c \geq 2$) непересекающихся интервалов $\{[r_u; r_v]^i\}$, $1 \leq u$, $u \leq v \leq m$, $i = 1, ..., p_c$. Значения данных в границах интервала $[r_u; r_v]^i$ могут использоваться методами ИАД как градация номинального признака. Считается, что множество чисел, идентифицирующих p_c градаций номинального признака, всегда можно взаимно-однозначно отобразить в множество $\{1, ..., p_c\}$.

Пусть $d_{tc}(u, v)$, $d_{3-t,c}(u, v)$ — количество представителей классов K_t, K_{3-t} в интервале $[r_u; r_v]^i$, $i \in \{1, ..., p_c\}$. Для рекурсивной процедуры выбора значений r_u, r_v используем критерий

$$\left| \frac{d_{tc}(u, v)}{|K_t|} - \frac{d_{3-t,c}(u, v)}{|K_{3-t}|} \right| \to \max. \quad (2)$$

Границы первого интервала $[r_u; r_v]^1$ последовательности (1) вычислим по максимуму критерия (2). Аналогичным образом определим границы для $[r_u; r_v]^q$, $q > 1$ на значениях (1), не вошедших в последовательность $[r_u; r_v]^1, \ldots, [r_u; r_v]^{q-1}$. Критерием останова процедуры служит покрытие всех значений (1) непересекающимися интервалами. Пространство из номинальных признаков, в формировании которого использовано разбиение на интервалы по критерию (2), назовем «сырым».

Вычисление значений функции принадлежности и нелинейные преобразования признаков

В зависимости от шкалы измерений признака $x_c \in X(n)$ через $d_{tc}(\mu)$ ($d_{3-t,c}(\mu)$), $t = 1,2$ обозначим число значений объектов в границах интервала $[r_u; r_v]^\mu$ или объектов, описываемых градацией $\mu \in \{1, ..., p_c\}$, из класса K_t, (K_{3-t}). Суть нелинейных преобразований признаков сводится к замене их исходных значений на значения функции принадлежности объектов к классам. Рассчитаем значение функции принадлежности $f_c(\mu)$ к классу K_1:

$$f_c(\mu) = \frac{d_{1c}(\mu)/|K_1|}{d_{1c}(\mu)/|K_1| + d_{2c}(\mu)/|K_2|}. \quad (3)$$

Определим границу между объектами классов по (3) для $x_c \in X(n)$:

$$\Gamma_c = (s1 + s2)/2, \quad (4)$$

где $s2 = \max\{f_c(\mu) | 0,5 - f_c(\mu) > 0, \ \mu = 1, \ldots, p_c\}$ и $s1 = \min\{f_c(\mu) | 1 - f_c(\mu) < 0,5, \ \mu = 1, ..., p_c\}$. Полученное значение (4) возможно использовать для классификации объектов и для их описания в новом (бинарном) признаковом пространстве. Найдем преобразование номинального признака x_c по значениям $\mu \in \{1, ..., p_c\}$, $c = 1, \ldots, n$ в градации из $\{1, 2\}$ для объекта $S_i = (x_{i1}, \ldots, x_{in})$:

$$x_{ic}^* = \begin{cases} 1, x_{ic} = \mu, f_c(\mu) < \Gamma_c, \\ 2, x_{ic} = \mu, f_c(\mu) > \Gamma_c. \end{cases} \quad (5)$$

Важной характеристикой для анализа данных, определяемой с помощью значений функции принадлежности (3), является устойчивость признака. Вычислим устойчивость признака $x_c \in X(n)$ по множеству значений градаций $\mu \in \{1, ..., p_c\}$:

$$U(c) = \frac{1}{m} \sum_{r=1}^{m} \begin{cases} f_c(\mu), x_{rc} = \mu, f_c(\mu) > 0,5, \\ 1 - f_c(\mu), x_{rc} = \mu, f_c(\mu) < 0,5, \\ 0, x_{rc} = \mu, f_c(\mu) = 0,5. \end{cases} \quad (6)$$

Замена исходных значений признака $X(n)$ объектов на значение (3) в идеале при $U(c) = 1$ может привести к корректному (без ошибок) разделению объектов E_0 на классы. Рассмотрим данное утверждение на примере. Пусть при использовании критерия (2) на (1) получено разбиение на p_c, ($p_c > 2$) интервалов, в границах каждого из которых представлены объекты только одного класса. Тогда значения функции принадлежности по (3) для всех объектов из K_1 будут равны 1, для K_2 — 0, граница (4) между классами $\Gamma_c = 0,5$.

Для графической иллюстрации смысла нелинейного преобразования на рис. 2 показано разбиение количественного признака $x_c \in X(n)$ на четыре интервала (обозначены *1*, *2*, *3*, *4*), каждому из которых соответствует класс (K_1 или K_2) с максимальной частотой встречаемости по (2) и значение функции принадлежности $f_c(\mu)$, $\mu \in \{1, ..., 4\}$ (3) из $\{0,3, 0,63, 0,42, 0,88\}$. Граница между классами (4) $\Gamma_c = 0,525$.

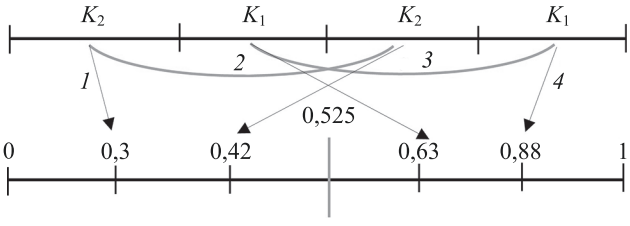

Рис. 2. Нелинейное преобразование признака по функции принадлежности

Fig. 2. Nonlinear transformation of a feature by a membership function

Вычисление весов номинальных признаков и обобщенных оценок объектов на их основе

Обозначим через g_{1c}^j, g_{2c}^j — количество значений градации $j \in \{1, \ldots, p_c\}$ признака $x_c \in X(n)$ в описании объектов классов K_1 и K_2. Определим межклассовое различие по признаку x_c:

$$\lambda_c = 1 - \frac{\sum_{j=1}^{p_c} g_{1c}^j g_{2c}^j}{|K_1||K_2|}. \quad (7)$$

Степень однородности (мера внутриклассового сходства) β_c значений градаций признака по классам K_1, K_2 вычислим по формулам:

$$D_{dc} = \begin{cases} (|K_d| - l_{dc} + 1)(|K_d| - l_{dc}), p_c > 2, \\ |K_d|(|K_d| - 1), p_c \le 2, \end{cases}$$

$$\beta_c = \begin{cases} \dfrac{\sum_{j=1}^{p_c} g_{1c}^j (g_{1c}^j - 1) + g_{2c}^j (g_{2c}^j - 1)}{D_{1c} + D_{2c}}, D_{1c} + D_{2c} > 0, \\ 0, \qquad D_{1c} + D_{2c} = 0, \end{cases} \quad (8)$$

где l_{dc} — число градаций признака x_c в описании объектов из K_d, $d = 1,2$.

С помощью выражений (7) и (8) вес признака $x_c \in X(n)$ в номинальной шкале определим как произведение внутриклассового сходства и межклассового различия

$$\omega_c = \beta_c \lambda_c. \quad (9)$$

Множество допустимых значений весов признаков, вычисленных по формуле (9), лежит в интервале [0; 1].

Для определения обобщенных оценок объектов [4] на E_0 используем вклады градаций признаков. Получим вклад градации $j \in \{1, \ldots, p_c\}$ признака $x_c \in X(n)$ в виде:

$$\eta_c(j) = \omega_c \left(\frac{\alpha_{cj}^1}{|K_1|} - \frac{\alpha_{cj}^2}{|K_2|} \right), \quad (10)$$

где α_{cj}^1, α_{cj}^2 — количество значений градации j признака x_c соответственно в классах K_1 и K_2; ω_c — вес признака x_c по (9).

Вычислим обобщенную оценку объекта $S_r \in E_0$ по описанию в номинальной шкале измерений $S_r = (a_{r1}, \ldots, a_{rn})$ на наборе $X(n)$ и вкладам (10):

$$Z(S_r) = \sum_{i=1}^{n} \eta_i(a_{ri}). \quad (11)$$

В табл. 1 приведены результаты расчетов устойчивости (6) и весов «сырых» и бинарных признаков

Таблица 1. Устойчивость признаков и их веса в «сыром» и бинарном пространствах

Table 1. Stability of features and their weights in raw and binary spaces

Название признака (число градаций «сырых» признаков)	Веса признаков в пространстве		Устойчивость по (6)
	«сыром»	бинарном	
Код города (17)	0,1369	0,2621	0,5776
Рост (2)	0,2533	0,2533	0,5839
Вес тела (4)	0,2611	0,2596	0,5549
Окружность талии (3)	0,2921	0,2914	0,6209
Зрение (слева) (3)	0,2323	0,2296	0,5662
Зрение (справа) (4)	0,2495	0,2306	0,5386
Слух (слева) (2)	0,0067	0,0067	0,5039
Слух (справа) (2)	0,0100	0,0100	0,5016
Систолическое артериальное давление (4)	0,2511	0,2465	0,5547
Диастолическое артериальное давление (2)	0,2819	0,2819	0,6378
Сахар в крови до еды (натощак) (2)	0,3301	0,3301	0,6574
Общий холестерол (4)	0,3491	0,3447	0,6735
Триглицерид (2)	0,3753	0,3753	0,7233
HDL холестерол (2)	0,2989	0,2989	0,6246
LDL холестерол (3)	0,3094	0,3040	0,6641
Гемоглобин (9)	0,2513	0,2499	0,6104
Белок в моче (5)	0,1231	0,1180	0,5293
Сывороточный креатинин (3)	0,2551	0,2530	0,5467
AST (3)	0,2667	0,2699	0,6240
ALT (3)	0,2866	0,2923	0,6567
Гамма GTP (4)	0,3394	0,3333	0,7034
Курение (3)	0,2644	0,2518	0,5862
Потребление алкоголя (2)	0,2113	0,2113	0,5105

по выражению (9) на примере выборки $T_4 = G_0 \cup G_4$ (группы пациентов с возрастами 20–24 и 40–44 лет). Равенством двух подмножеств объектов выборки E_0, для градаций признака $x_c \in X(n)$ которых в «сыром» и бинарном (5) пространствах выполняется одно из неравенств $f_c(\mu) > 0{,}5$ или $f_c(\mu) < 0{,}5$, объясняется малая различимость (в шестом или седьмом знаке) значений устойчивости (6).

Заметим, что возможное максимальное значение весов и устойчивости признаков равно 1. Из полученных результатов видно значительное отклонение величин от максимума, что указывает на плохую разделимость объектов групп.

Обобщенные оценки объектов по выражению (11) на разных наборах признаков в «сыром» и бинарном пространствах рассмотрим как латентные признаки. О целесообразности использования данных оценок можно судить по компактности их значений на числовой оси.

Пусть на выборке T_i, $i = 1, \ldots, 13$ по значениям обобщенных оценок объектов $Z(S_1), \ldots, Z(S_m)$, $m = |T_i|$ построена упорядоченная по неубыванию последовательность

$$\delta_1, \ldots, \delta_j, \ldots, \delta_m. \qquad (12)$$

Разделим последовательность (12) на два непересекающихся интервала $[\pi_1; \pi_2]$, $(\pi_2; \pi_3]$ по значению критерия [5]:

$$\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2}(u_i^d - 1)u_i^d}{\sum_{i=1}^{2}|K_i|(|K_i| - 1)} \right) \left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2}u_i^d(|K_{3-i}| - u_{3-i}^d)}{2|K_1||K_2|} \right) \rightarrow \rightarrow \max_{\pi_1 < \pi_2 < \pi_3}, \qquad (13)$$

где $\pi_1 = \delta_1$, $\pi_2 = \delta_j$, $\pi_3 = \delta_m$, u_1, u_1^1, u_1^2 (u_2^1, u_2^2) — количество значений обобщенных оценок объектов из классов $K_1(G_0)$, $K_2(G_i)$ в интервалах $[\pi_1; \pi_2]$ и $(\pi_2; \pi_3]$.

Определим границу между классами:

$$\Theta = (\pi_2 + b)/2, \qquad (14)$$

где b — ближайшее к π_2 значение из $(\pi_2; \pi_3]$. Если в границах каждого из интервалов $[\pi_1; \pi_2]$, $(\pi_2; \pi_3]$ содержатся все оценки объектов только одного класса, критерий (13) равен 1. Значение (13) меньше 1 соответствует точности распознавания на T_i (разделению на классы по (14)) меньше 100 %.

Результаты распознавания по обобщенным оценкам объектов T_4 на наборе из 23 признаков (табл. 1) с использованием значений, полученных из выражений (13) и (14), приведены в табл. 2.

Из табл. 2 видно, что точность распознавания по бинарным признакам выше, чем по «сырым». Выводы об изменении точности распознавания (увеличивается, уменьшается) на разных наборах признаков на выборке T_4 предложено выполнить по выражению (13) с учетом результатов, полученных в табл. 2.

Формирование наборов признаков на основе их ранжирования

Для отбора информативных наборов признаков выполним их ранжирование по устойчивости (6). Процедуру отбора осуществим по значениям обобщенных оценок объектов на наборах признаков, сформированных по результатам ранжирования. Исходя из ограниченных возможностей экспертов для анализа и интерпретации данных в сложно организованных системах, количество признаков, используемых для синтеза обобщенных оценок, ограничено магическим числом 7. В табл. 3 приведены результаты экспериментов на выборке T_4 по вычислению обобщенных оценок объектов, синтезированным по трем наборам признаков с максимальными значениями устойчивости (6).

При эксперименте на обобщенных оценках объектов по набору из семи бинарных признаков с самыми низкими показателями устойчивости из табл. 1 получено значение критерия (13), равное 0,2700. На аналогичном по мощности наборе (табл. 3) значение (13) равно 0,4215. Вычисления обобщенных оценок объектов на наборах бинарных признаков по (11) и границы между классами (14) рассмотрены как реализация метамодели по ансамблю базовых алгоритмов распознавания (9]. Значение (4) является параметром базового алгоритма.

Пусть Π_{ic} — значение ранга признака $x_c \in X(n)$, полученное по выборке $T_i = G_0 \cup G_i$, $i = 1, \ldots, 13$. Рассчитаем ранг признака x_c по 13 выборкам данных:

$$R_c = \sum_{i=1}^{13}\Pi_{ic}/13. \qquad (15)$$

Выполним исследование наборов из упорядоченной по (15) последовательности признаков. Первые 10 признаков приведены в табл. 4.

Таблица 2. Результаты распознавания по обобщенным оценкам объектов на выборке T_4

Table 2. Results of recognition based on generalized estimates of objects on the T_4 set

Показатели	Пространство	
	«сырое»	бинарное
Границы интервалов	[–1,2465; –0,0432], (–0,0432; 1,2363]	[–1,3221; –0,0916], (–0,0916; 1,3371]
Значение критерия (13)	0,4532	0,4669
Граница между классами (14)	–0,0429	–0,0901
Число ошибок (точность распознавания)	91(79,59 %)	87(80,49 %)

Таблица 3. Разбиение на интервалы обобщенных оценок объектов по «сырым» и бинарным признакам
Table 3. Splitting into intervals of generalized estimates of objects by raw and binary features

Число признаков в наборе	Обобщенные оценки объектов по признакам:			
	«сырым»		бинарным	
	границы интервалов	значение критерия (13)	границы интервалов	значение критерия (13)
7	[–0,7762, 0,0803], (0,0803, 0,7744]	0,4186	[–0,7780; 0,0785], (0,0785; 0,7780]	0,4215
6	[–0,6990, –0,0384], (–0,0384, 0,6972]	0,4163	[–0,7008; –0,0400], (–0,0400; 0,7008]	0,4163
5	[–0,6341, –0,1171], (–0,1171, 0,6386]	0,4183	[–0,6351; –0,0964], (–0,0964; 0,6351]	0,4222

Таблица 4. Упорядоченный по (15) набор из 10 признаков
Table 4. Ordered by (15) sets of 10 features

Название признака		Ранг признака (15)
Антропометрические показатели	Рост	6,4615
	окружность талии	8,0769
	зрение (слева)	8,7692
	зрение (справа)	9,4615
Состояние здоровья	сахар в крови до еды (натощак)	4,7692
	Триглицерид	6,0000
	гамма GTP	6,4615
	курение	7,7692
	общий холестерол	9,1538
	AST	10,0769

Тенденция на уменьшение значений признака «Рост» с повышением возраста — известная закономерность. По этой причине значения признака не представляют интереса для анализа и далее не рассмотрены. Для исследования закономерностей по другим признакам из табл. 4 использованы значения их устойчивости (6) на выборках T_i, $i = 1, \ldots, 13$. Графики изменения показателей устойчивости 7 признаков (табл. 4) по возрастным группам показаны на рис. 3.

Для поиска закономерностей по набору из 7 признаков (рис. 3) использованы значения обобщенных оценок (11), вычисленные по 13 выборкам. Исследована связь между увеличением возраста относительно группы G_0 и значением критерия (13) по обобщенным оценкам. Результаты анализа связей по обобщенным оценкам приведены в табл. 5.

Как закономерность можно рассматривать наличие монотонной неубывающей последовательности по значениям (13) (табл. 5), инвариантной порядку старшинства возраста в группах $G_1, \ldots, G_{13}$. Основой для обнаружения закономерности служит формирование информативных наборов признаков по значениям их устойчивости (6).

Таблица 5. Анализ связей между возрастными группами по обобщенным оценкам
Table 5. Analysis of relationships between age groups according to generalized estimates

Возрастная группа	Значение критерия (13)	Возрастная группа	Значение критерия (13)
25–29	0,2849	60–64	0,5735
30–34	0,3204	65–69	0,6275
35–39	0,3602	70–74	0,6973
40–44	0,4191	75–79	0,7421
45–49	0,4609	80–84	0,8039
50–54	0,4745	85+	0,8993
55–59	0,5071	—	—

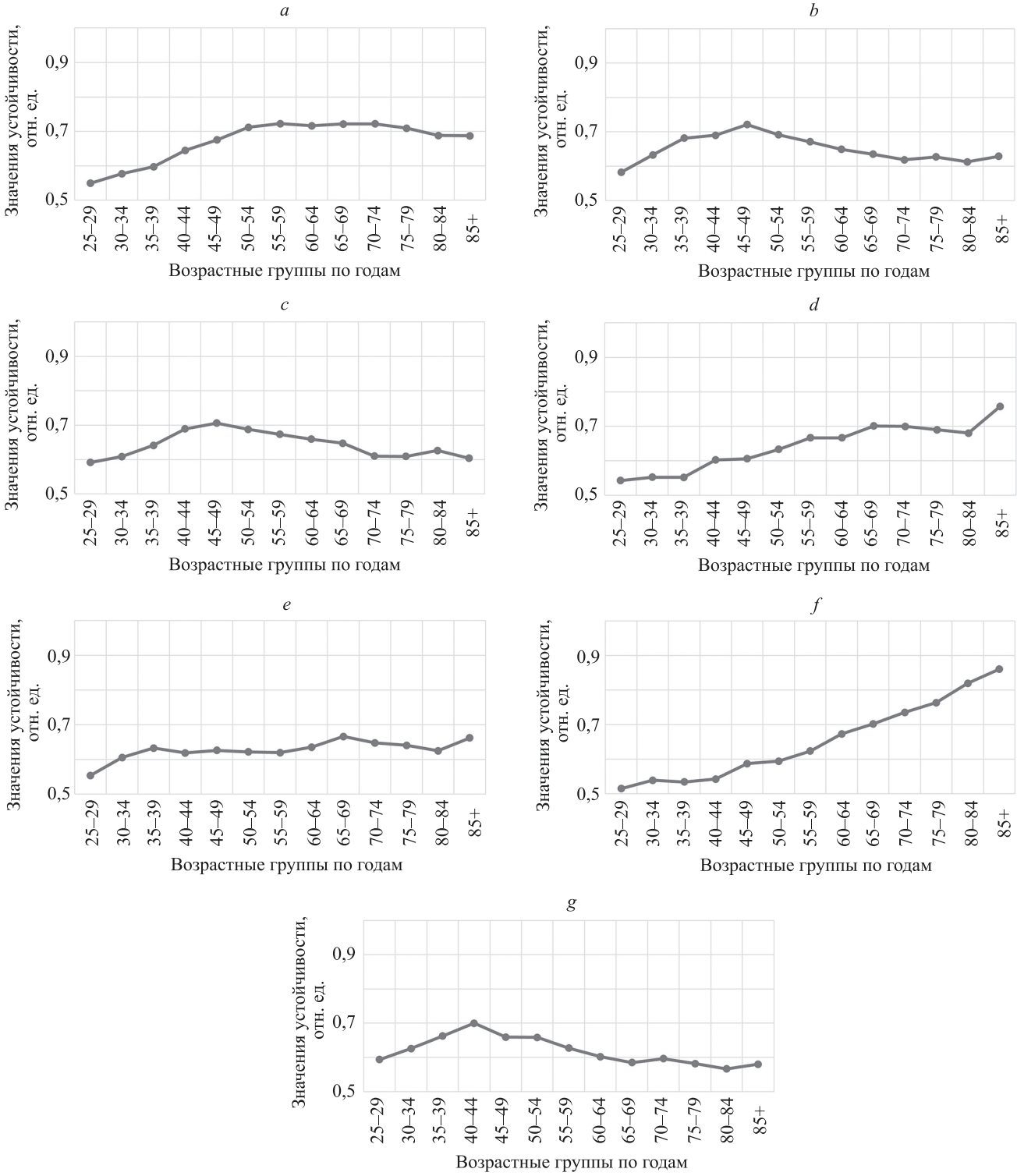

Рис. 3. Графики изменения показателей устойчивости (6) значений разнотипных признаков в интервале (0,5; 1] для различных возрастных групп: сахар в крови до еды (натощак) (*a*); триглицерид (*b*); гамма GTP (*c*); курение (*d*); окружность талии (*e*); зрение (слева) (*f*); общий холестерол (*g*)

Fig. 3. Graphs of changes in stability indicators (6) of different types of features values in the interval (0,5; 1] for different age groups: blood sugar before meal (in fasting state) (*a*); triglyceride (*b*); gamma GTP (*c*); smoking (*d*); waist circumference (*e*); eyesight (at the left) (*f*); total cholesterol (*g*)

Заключение

Предложена новая методика отбора и анализа информативных наборов разнотипных признаков с использованием нелинейных преобразований на основе функций принадлежности, процедур ранжирования по отношению устойчивости и вычисления обобщенных оценок объектов. Реализация методики является одним из путей автоматизации процесса формирования признакового пространства в информационных моделях для слабо структурированных предметных областей.

Литература

1. Шумаков В.И., Новосельцев В.Н., Сахаров М.П., Штенголд Е.Ш. Моделирование физиологических систем организма. М.: Медицина, 1971. 352 с.
2. Кривенко М.П. Выбор модели данных в задачах медицинской диагностики // Информатика и ее применения. 2019. Т. 13. № 4. С. 27–29. https://doi.org/10.14357/19922264190404
3. Корепанова Н.В. Машинное обучение для оптимизации лечения в подгруппах пациентов // Искусственный интеллект и принятие решений. 2018. № 1. С. 53–65.
4. Игнатьев Н.А., Рахимова М.А. Формирование и анализ наборов информативных признаков объектов по парам классов // Искусственный интеллект и принятие решений. 2021. № 4. С. 18–26. https://doi.org/10.14357/20718594210402
5. Игнатьев Н.А., Згуральская Е.Н., Марковцева М.В. Поиск скрытых закономерностей, влияющих на общую выживаемость больных, методами интеллектуального анализа данных // Искусственный интеллект и принятие решений. 2020. № 3. С. 73–80. https://doi.org/10.14357/20718594200307
6. Игнатьев Н.А. Вычисление обобщенных показателей и интеллектуальный анализ данных // Автоматика и телемеханика. 2011. № 5. С. 183–190.
7. Згуральская Е.Н. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей // Известия Самарского научного центра Российской академии наук. 2018. Т. 20. № 4-3. С. 451–455.
8. Piatetsky-Shapiro G. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from «university» to «business» and «analytics» // Data Mining and Knowledge Discovery. 2007. V. 15. N 1. P. 99–105. https://doi.org/10.1007/s10618-006-0058-2
9. Joseph R. (2019, April 23). Ensemble methods: bagging, boosting and stacking. Understanding the key concepts of ensemble learning [Электронный ресурс]. URL: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (дата обращения: 25.12.2022).

References

1. Shumakov V.I., Novoseltcev V.N., Sakharov M.P., Shtengold E.Sh. *Simulation of The Body Physiological Systems*. Moscow, Medicina Publ., 1971, 352 p. (in Russian)
2. Krivenko M.P. Data model selection in medical diagnostic tasks. *Informatics and Applications*, 2019, vol. 13, no. 4, pp. 27–29. (in Russian). https://doi.org/10.14357/19922264190404
3. Korepanova N.V. Machine learning for treatment optimization in subgroups of patients. *Artificial Intelligence and Decision Making*, 2018, no. 1, pp. 53–65. (in Russian)
4. Ignatyev N.A., Rakhimova M.A. Formation and analysis of sets of informative features of objects by pairs of classes. *Artificial Intelligence and Decision Making*, 2021, no. 4, pp. 18–26. (in Russian). https://doi.org/10.14357/20718594210402
5. Ignatev N.A., Zguralskaya E.N., Markovtseva M.V. Searching for hidden patterns that affect the overall patient survival with data mining. *Scientific and Technical Information Processing*, 2021, vol. 48, no. 6, pp. 461–466. (in Russian). https://doi.org/10.3103/S014768822106006X
6. Ignat'ev N.A. Computing generalized parameters and data mining. *Automation and Remote Control*, 2011, vol. 72, no. 5, pp. 1068–1074. https://doi.org/10.1134/S0005117911050146
7. Zguralskaya E. Sustainability of dividing data in intervals in the problems of recognition and searching for hidden laws. *Izvestia of Samara Scientific Center of the Russian Academy of Sciences*, 2018, vol. 20, no. 4-3, pp. 451–455. (in Russian)
8. Piatetsky-Shapiro G. Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from «university» to «business» and «analytics». *Data Mining and Knowledge Discovery*, 2007, vol. 15, no. 1, pp. 99–105. https://doi.org/10.1007/s10618-006-0058-2
9. Joseph R. (2019, April 23). *Ensemble methods: bagging, boosting and stacking. Understanding the key concepts of ensemble learning*. Available at: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (accessed: 25.12.2022).

Авторы

Игнатьев Николай Александрович — доктор физико-математических наук, профессор, профессор, Национальный университет Узбекистана имени Мирзо Улугбека, Ташкент, 100174, Узбекистан, SC 39361638900, https://orcid.org/0000-0002-7150-5837, n_ignatev@rambler.ru

Рахимова Мехрбону Акром кизи — старший преподаватель, Национальный университет Узбекистана имени Мирзо Улугбека, Ташкент, 100174, Узбекистан, https://orcid.org/0000-0001-5849-3395, mehribonu@gmail.com

Authors

Nikolay A. Ignatev — D.Sc. (Physics & Mathematics), Full Professor, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100174, Uzbekistan, SC 39361638900, https://orcid.org/0000-0002-7150-5837, n_ignatev@rambler.ru

Mekhrbonu A. Rakhimova — Senior Lecturer, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100174, Uzbekistan, https://orcid.org/0000-0001-5849-3395, mehribonu@gmail.com

Статья поступила в редакцию 29.09.2022
Одобрена после рецензирования 09.02.2023
Принята к печати 28.03.2023

Received 29.09.2022
Approved after reviewing 09.02.2023
Accepted 28.03.2023