УДК 004.85:519.25

СЕЛЕКЦИЯ ОБЪЕКТОВ ПО ОТНОШЕНИЯМ С БЛИЖАЙШИМИ ОБЪЕКТАМИ ПРОТИВОПОЛОЖНОГО КЛАССА И СОХРАНЕНИЕ ЛОКАЛЬНОЙ СТРУКТУРЫ ОБУЧАЮЩЕЙ ВЫБОРКИ

Рамазонов Ш.Ш.

ramazonov377@gmail.com

Национальный университет Узбекистана имени Мирзо Улугбека,
100174, Узбекистан, г. Ташкент, ул. Университетская, 4

Рассматривается задача селекции объектов обучающей выборки и анализа того, как селекция влияет на локальную структуру данных. Предлагается алгоритм ENRS (Enemy-Neighbor Relation Selection), в котором решение об удалении объекта принимается на основе его отношений с ближайшими объектами противоположного класса и с объектами своего класса, в сферы влияния которых он попадает. Для описания локальной структуры выборки используется признак, формируемый по составу k ближайших соседей объекта, и показатель устойчивости этого признака; совокупность значений показателя при различных k представляет выборку как метаобъект в унифицированном пространстве. Предложен способ оценки сохранения локальной структуры, основанный на сравнении редуцированной выборки со случайными подвыборками того же объёма и классового состава; взаимное сходство редуцированных выборок оценивается с помощью корреляционного анализа их профилей устойчивости и метода главных компонент. Вычислительные эксперименты проведены на наборе данных German Credit и пяти других наборах репозитория UCI, выполнено сравнение с методами селекции CNN, ENN и Tomek Links. Установлено, что на наборах German Credit и Pima выборки, полученные алгоритмом ENRS, отличаются от случайных подвыборок меньшей однородностью окрестностей объектов по классам, а степень этого отличия зависит от параметра алгоритма; точность классификации на выборках ENRS не превышает точности на случайных подвыборках того же объёма. Предложенный подход позволяет отделить изменения локальной структуры, вызванные отбором объектов, от изменений, обусловленных уменьшением объёма выборки, и применим для анализа любых методов селекции.

Ключевые слова: селекция объектов, ближайший объект противоположного класса, k ближайших соседей, локальная структура выборки, устойчивость признака, метаобъект, случайные подвыборки

Цитирование: *Рамазонов Ш.Ш.* Селекция объектов по отношениям с ближайшими объектами противоположного класса и сохранение локальной структуры обучающей выборки // Проблемы вычислительной и прикладной математики. – 2026. – № 5(76). – С. 46-59.

DOI: 10.71310/pcam.5_76.2026.04

1 Введение

Качество обучающей выборки во многом определяет качество моделей распознавания образов и машинного обучения. Даже при одном и том же классификаторе точность и устойчивость решения зависят от того, содержит ли обучающее множество избыточные, шумовые или пограничные объекты. Поэтому селекция объектов

(instance selection, prototype selection) — выбор подмножества обучающей выборки, на котором классификатор работает не хуже, чем на всей выборке, — давно является самостоятельным направлением исследований [1–6].

Классические методы делятся на конденсирующие, которые оставляют объекты вблизи границ классов и удаляют внутренние (CNN [1], IB-алгоритмы [4]), редактирующие, которые удаляют шумовые и пограничные объекты (ENN [2], Tomek Links [3]), и гибридные (DROP [5], ICF [7]). Систематические обзоры и сравнения этих методов приведены в [6–8]. Как показывают обзоры [6, 8], большинство подходов оценивается по двум критериям: степени редукции и точности классификации на редуцированной выборке. Развитие направления связано с методами на основе локальных множеств [9, 10], методами линейной сложности для больших данных [11], селекцией для SVM [12] и многокритериальными эволюционными методами [13].

Гораздо меньше изучен вопрос о том, как селекция меняет локальную структуру данных — соотношение классов в окрестностях объектов. Две выборки одинакового объёма и с одинаковой точностью классификации могут иметь разную локальную организацию. Кроме того, при оценке любого метода селекции нужно отделять эффект собственно отбора объектов от эффекта уменьшения объёма выборки: часть изменений структурных характеристик возникает уже при случайном прореживании.

В работах [14–16] предложен аппарат описания обучающей выборки через устойчивость признаков, сформированных по отношениям объектов, и понятие выборки как метаобъекта в унифицированном признаковом пространстве. В настоящей работе этот аппарат используется для анализа редуцированных выборок. Вклад работы состоит в следующем:

1. предложен алгоритм селекции ENRS, дано его формальное описание в виде псевдокода, доказана завершимость и получена оценка сложности;
2. предложен протокол оценки сохранения локальной структуры, в котором профиль устойчивости редуцированной выборки сравнивается с распределением профилей случайных подвыборок того же объёма и классового состава, а также с профилями выборок, полученных методами CNN, ENN и Tomek Links; приводятся оценки статистической значимости;
3. выполнено экспериментальное сравнение ENRS с указанными методами и со случайной селекцией по степени редукции и точности классификации на шести наборах данных.

Признак, построенный по составу k ближайших соседей, показатель устойчивости (3) и понятие метаобъекта заимствованы из [14, 15] и в работе не претендуют на новизну.

2 Связь с предшествующими работами

В [14] обучающее множество строится через цензурирование выборки: выделяются шумовые и граничные объекты, а отбор эталонов управляется коэффициентом регуляризации $\lambda \in (0, 1)$, определяющим относительный отступ. Основной результат — сокращённое множество эталонов и повышение качества распознавания. В [15] близость структур отношений объектов сравнивается на разных наборах латентных признаков при фиксированном числе объектов; меры близости строятся по граничным объектам. В отличие от [14], в настоящей работе разнообразие редуцированных выборок задаётся параметром q алгоритма ENRS, а главным предметом анализа является не точность классификации, а изменение локальной структуры после се-

лекции. В отличие от [15], сравниваются выборки разного объёма, но в одном и том же признаковом пространстве; переход к профилям устойчивости $\{g_k\}$ позволяет сравнивать такие выборки как векторы одинаковой размерности.

Алгоритм ENRS использует понятия, близкие к понятию локального множества [7, 9]: локальное множество объекта — это объекты его класса, лежащие ближе ближайшего объекта противоположного класса. На этом понятии основаны методы ICF [7] и LSSm, LSBo [9]. В ENRS используются те же сферы, но иной порядок просмотра объектов и иное правило удаления (разд. 3).

3 Постановка задачи и алгоритм ENRS

Пусть задана обучающая выборка $E_0 = \{S_1, \ldots, S_m\}$, разбитая на два непересекающихся класса K_1 и K_2, и функция расстояния $\rho(S_i, S_j)$. Требуется построить редуцированные выборки $D \subset E_0$ при разных значениях управляющего параметра q и оценить, насколько они сохраняют локальную структуру выборки E_0.

Для каждого объекта $S_i \in E_0$ из класса K_t определим:

- $r(S_i) = \min\limits_{S_j \notin K_t} \rho(S_i, S_j)$ — расстояние до ближайшего объекта противоположного класса (радиус сферы влияния);
- $c(S_i) = |\{S_j \in K_t \setminus \{S_i\} : \rho(S_i, S_j) < r(S_i)\}|$ — число объектов своего класса внутри сферы влияния;
- $z(S_i, q)$ — расстояние от S_i до его q-го ближайшего соседа (любого класса, сам объект не учитывается);
- $d(S_i) = \dfrac{|z(S_i, q) - r(S_i)|}{z(S_i, q) + r(S_i)}$, при $z(S_i, q) + r(S_i) = 0$ полагается $d(S_i) = 0$.

Величина $d(S_i) \in [0, 1]$ мала, если q-й сосед объекта находится примерно на том же расстоянии, что и ближайший объект противоположного класса. Параметр q используется только в $z(S_i, q)$; для числа соседей при построении признака в разд. 4 используется отдельное обозначение k.

Объекты просматриваются в порядке возрастания $d(S_i)$. Для очередного объекта S_i из класса K_t строится множество

$$L(S_i) = \{S_j \in D \cap K_t : S_j \neq S_i,\ \rho(S_j, S_i) < r(S_j),\ c(S_j) > 1\}, \tag{1}$$

т. е. объектов своего класса, в сферу влияния которых попадает S_i и в сфере которых, кроме S_i, есть ещё хотя бы один объект своего класса. Правила алгоритма таковы:

- если $|L(S_i)| = 0$, объект S_i не лежит ни в одной «непустой» сфере своего класса и сохраняется;
- если $|L(S_i)| = 1$, объект S_i представлен единственной сферой и удаляется как избыточный;
- если $|L(S_i)| > 1$, объект S_i лежит на пересечении нескольких сфер и сохраняется как их общий представитель; счётчики $c(S_j)$ всех $S_j \in L(S_i)$ уменьшаются на единицу, а объекты S_j, у которых после этого $c(S_j) = 1$, удаляются: их сфера уже представлена объектом S_i.

Каждый объект рассматривается в качестве кандидата не более одного раза. Формально это описано в алгоритме 1.

В исходной формулировке алгоритма объект после просмотра «блокировался» заменой $d(S_i)$ на $-d(S_i)$, после чего очередь упорядочивалась заново. Поскольку r, z и d вычисляются один раз по исходной выборке E_0 и в ходе работы не меняются, повторное упорядочивание даёт тот же порядок, и алгоритм 1 ему эквивален-

Алгоритм 1 ENRS

Вход: выборка $E_0 = \{S_1, \ldots, S_m\}$ с метками классов, функция ρ, параметр q
Выход: редуцированная выборка D и множество удалённых объектов R

вычислить $r(S_i)$, $c(S_i)$, $z(S_i, q)$, $d(S_i)$ для всех $S_i \in E_0$
$D \leftarrow E_0$, $R \leftarrow \varnothing$
упорядочить E_0 по возрастанию $d(S_i)$ (при равенстве — по возрастанию индекса): $S_{\pi(1)}, \ldots, S_{\pi(m)}$
для $l = 1, \ldots, m$
  $S_i \leftarrow S_{\pi(l)}$
  если $S_i \notin D$ **то**
    перейти к следующему l
  конец если
  построить $L(S_i)$ по формуле (1)
  если $|L(S_i)| = 1$ **то**
    $D \leftarrow D \setminus \{S_i\}$, $R \leftarrow R \cup \{S_i\}$
  иначе если $|L(S_i)| > 1$ **то**
    для всех $S_j \in L(S_i)$
      $c(S_j) \leftarrow c(S_j) - 1$
      если $c(S_j) = 1$ **то**
        $D \leftarrow D \setminus \{S_j\}$, $R \leftarrow R \cup \{S_j\}$
      конец если
    конец для
  конец если
конец для
вернуть D, R

тен. Явный однократный проход устраняет и неоднозначность исходной записи при $d(S_i) = 0$, когда замена знака не блокирует объект.

Утверждение 1. *Алгоритм 1 завершается не более чем за m итераций внешнего цикла. При заданной матрице расстояний его трудоёмкость составляет $O(m^2)$ операций и $O(m^2)$ памяти; с учётом вычисления матрицы расстояний для p признаков — $O(m^2 p)$.*

Доказательство. Внешний цикл перебирает фиксированную перестановку π объектов E_0, поэтому число его итераций равно m, и каждый объект рассматривается как кандидат не более одного раза. Вычисление r и c требует $O(m^2)$ операций, $z(S_i, q)$ — $O(m)$ на объект при поиске q-й порядковой статистики, сортировка по d — $O(m \log m)$. На каждой итерации построение $L(S_i)$ и коррекция счётчиков требуют $O(m)$ операций, что в сумме даёт $O(m^2)$. □

Следует подчеркнуть две особенности алгоритма. Во-первых, радиусы r, расстояния z и величины d не пересчитываются после удалений: они описывают геометрию исходной выборки, а счётчики c корректируются только по правилу $|L| > 1$. Это упрощение сохраняет квадратичную сложность. Во-вторых, параметр q влияет на результат только через порядок просмотра объектов. Поэтому степень редукции не является монотонной функцией q и непосредственно этим параметром не управляется (см. табл. 1); параметр q определяет главным образом состав редуцированной выборки.

4 Признак и показатель устойчивости

Пусть V — рассматриваемая выборка (исходная или редуцированная), $h = |V|$, а $|K_1^V|$ и $|K_2^V|$ — число объектов каждого класса в V. Для каждого $S \in V$ найдём k ближайших соседей в V (сам объект не учитывается) и обозначим через $i(S) \in \{0, 1, \ldots, k\}$ число объектов класса K_1 среди них. Значение признака $\mu(S) = i(S) + 1 \in \{1, \ldots, k+1\}$; признак является порядковым. Пусть $d_{1k}(\mu)$ и $d_{2k}(\mu)$ — число объектов классов K_1 и K_2 со значением признака μ, $n(\mu) = d_{1k}(\mu) + d_{2k}(\mu)$, $M_k = \{\mu : n(\mu) > 0\}$. Для $\mu \in M_k$ функция принадлежности градации к классу K_1 с учётом численности классов вычисляется по формуле [14, 15]

$$f_k(\mu) = \frac{d_{1k}(\mu)/|K_1^V|}{d_{1k}(\mu)/|K_1^V| + d_{2k}(\mu)/|K_2^V|}, \tag{2}$$

а устойчивость признака — по формуле

$$g_k = \frac{1}{h} \sum_{\mu \in M_k} n(\mu)\, \max\bigl\{ f_k(\mu),\, 1 - f_k(\mu) \bigr\}. \tag{3}$$

Градации с $n(\mu) = 0$ в (3) вклада не дают, поэтому $f_k(\mu)$ для них не определяется. Величина $g_k \in [0{,}5;\, 1]$ тем больше, чем лучше значения признака разделяют классы, т. е. чем более однородны по классам окрестности объектов. Вектор $G(V) = (g_3, \ldots, g_{32})$ будем называть профилем устойчивости выборки V; он описывает V как метаобъект в унифицированном 30-мерном пространстве независимо от объёма выборки и исходного набора признаков.

Для одной выборки объёма h порядок случайной погрешности g_k можно оценить как $\sqrt{g_k(1-g_k)/h}$; при $h \approx 550$ и $g_k \approx 0{,}65$ это около 0,02. Кроме того, g_k зависит от объёма и классового состава выборки. Поэтому сравнение «сырых» значений g_k разных выборок недостаточно, и в работе используется следующий протокол.

1. Для редуцированной выборки V строится $B = 200$ случайных подвыборок E_0 с тем же числом объектов каждого класса, что и в V, и вычисляются их профили.
2. Для каждого k находится 95 %-й интервал значений g_k на случайных подвыборках; для среднего $\bar{g}(V) = \frac{1}{30}\sum_k g_k$ вычисляется эмпирический уровень значимости $p = (1 + \#\{b : \bar{g}_b \leqslant \bar{g}(V)\})/(B+1)$.
3. Профили сравниваются коэффициентами корреляции Пирсона, вычисленными как по исходным профилям, так и после удаления линейного тренда по k, поскольку общий рост g_k с увеличением k даёт положительную корреляцию любых профилей.
4. Метод главных компонент [17] применяется к профилям для наглядного представления взаимного расположения метаобъектов и не рассматривается как независимое подтверждение корреляционного анализа, поскольку строится по тем же данным.

5 Вычислительный эксперимент

Данные и предобработка

Основной набор данных — Statlog (German Credit Data) [18]: 1000 объектов, 700 в классе K_1 («хороший заёмщик») и 300 в классе K_2, 20 признаков, из которых 7

количественных и 13 категориальных. Для сравнения методов селекции использованы ещё пять наборов из репозитория UCI: Pima Indians Diabetes [19] (768 объектов, 8 признаков), Ionosphere [20] (351, 33; постоянный второй признак исключён), Sonar [21] (208, 60), Breast Cancer Wisconsin (Diagnostic, WDBC) [22] (569, 30) и Banknote Authentication [23] (1372, 4).

Расстояние между объектами — гетерогенная метрика HEOM [24]: количественные признаки нормируются на размах, для категориальных используется расстояние 0 при совпадении значений и 1 иначе, а итоговое расстояние равно корню из суммы квадратов покомпонентных расстояний. Для наборов только с количественными признаками HEOM совпадает с евклидовым расстоянием после нормировки признаков на $[0, 1]$. При равенстве расстояний соседи упорядочиваются по индексу. Вычисления выполнены на Python с использованием NumPy, SciPy и scikit-learn [25]; реализации CNN, ENN ($k = 3$) и Tomek Links (удаляются оба объекта каждой связи) выполнены по [1–3].

Редуцированные выборки German Credit

Алгоритм ENRS применялся ко всей выборке E_0 при $q \in \{7, 32, 57, 82, 107\}$; полученные выборки обозначены $V_7, \ldots, V_{107}$. Их объёмы и состав приведены в табл. 1. ENRS сохраняет 54–60 % объектов и сильнее сокращает класс K_1 (до 45–53 %), чем класс K_2 (75 %). Наибольшая редукция достигается при $q = 7$; при $q \geqslant 32$ объёмы почти совпадают, а выборки сильно пересекаются (коэффициент Жаккара между $V_{32}, \ldots, V_{107}$ равен 0,92–0,98, между V_7 и остальными — 0,77–0,78). Время работы ENRS при готовой матрице расстояний составляет около 0,02 с. При $q = 7$ из 916 просмотренных объектов 162 удалены по правилу $|L| = 1$, 299 — как объекты S_j по правилу $|L| > 1$, для 319 объектов $L = \varnothing$.

Таблица 1 Редуцированные выборки German Credit и их средняя устойчивость

Выборка	h	$\lvert K_1^V\rvert/\lvert K_2^V\rvert$	Доля, %	$\bar{g}$	$\bar{g}$ случ. подвыб.	p
E_0	1000	700/300	100	0,681	—	—
V_7	539	315/224	53,9	0,619	$0{,}677 \pm 0{,}013$	0,005
V_{32}	583	359/224	58,3	0,646	$0{,}678 \pm 0{,}013$	0,020
V_{57}	596	369/227	59,6	0,654	$0{,}678 \pm 0{,}012$	0,030
V_{82}	596	369/227	59,6	0,657	$0{,}676 \pm 0{,}013$	0,080
V_{107}	593	369/224	59,3	0,655	$0{,}678 \pm 0{,}013$	0,025
CNN	487	296/191	48,7	0,566	—	—
ENN	726	606/120	72,6	0,859	—	—
Tomek Links	898	649/249	89,8	0,710	—	—

Примечание. $\bar{g}$ — среднее значение g_k по $k = 3, \ldots, 32$; для случайных подвыборок того же объёма и состава указаны среднее и стандартное отклонение по $B = 200$ повторениям; p — эмпирический уровень значимости того, что $\bar{g}$ случайной подвыборки не превосходит $\bar{g}$ выборки ENRS.

Профили устойчивости

Значения g_k для отдельных k приведены в табл. 2, полные профили — на рис. 1. Для всех выборок, кроме CNN и ENN, g_k растёт с увеличением k (коэффициент корреляции с k от 0,59 для V_{57} до 0,84 для V_7). Профили случайных подвыборок близки к профилю E_0: уменьшение объёма почти вдвое само по себе не меняет сред-

нюю устойчивость (табл. 1), а разброс профилей случайных подвыборок составляет около 0,015 при каждом k.

Таблица 2 Значения устойчивости g_k на исходной и редуцированных выборках German Credit

Выборка	k							
	3	5	8	10	15	20	25	32
E_0	0,642	0,654	0,668	0,678	0,689	0,692	0,683	0,696
V_7	0,568	0,601	0,607	0,604	0,605	0,627	0,629	0,633
V_{32}	0,586	0,623	0,644	0,645	0,637	0,639	0,653	0,662
V_{57}	0,600	0,636	0,647	0,657	0,657	0,644	0,664	0,674
V_{82}	0,601	0,640	0,657	0,671	0,667	0,653	0,670	0,676
V_{107}	0,593	0,634	0,660	0,668	0,657	0,657	0,667	0,668
Случ. ($h = 539$)	0,645	0,660	0,673	0,674	0,675	0,681	0,683	0,689
CNN	0,655	0,617	0,547	0,534	0,546	0,548	0,565	0,587
ENN	0,915	0,891	0,874	0,847	0,862	0,856	0,858	0,857
Tomek Links	0,702	0,701	0,704	0,702	0,703	0,715	0,714	0,718

Примечание. Строка «Случ.» — среднее по 200 случайным подвыборкам объёма и состава V_7; стандартное отклонение при каждом k равно 0,015–0,019.

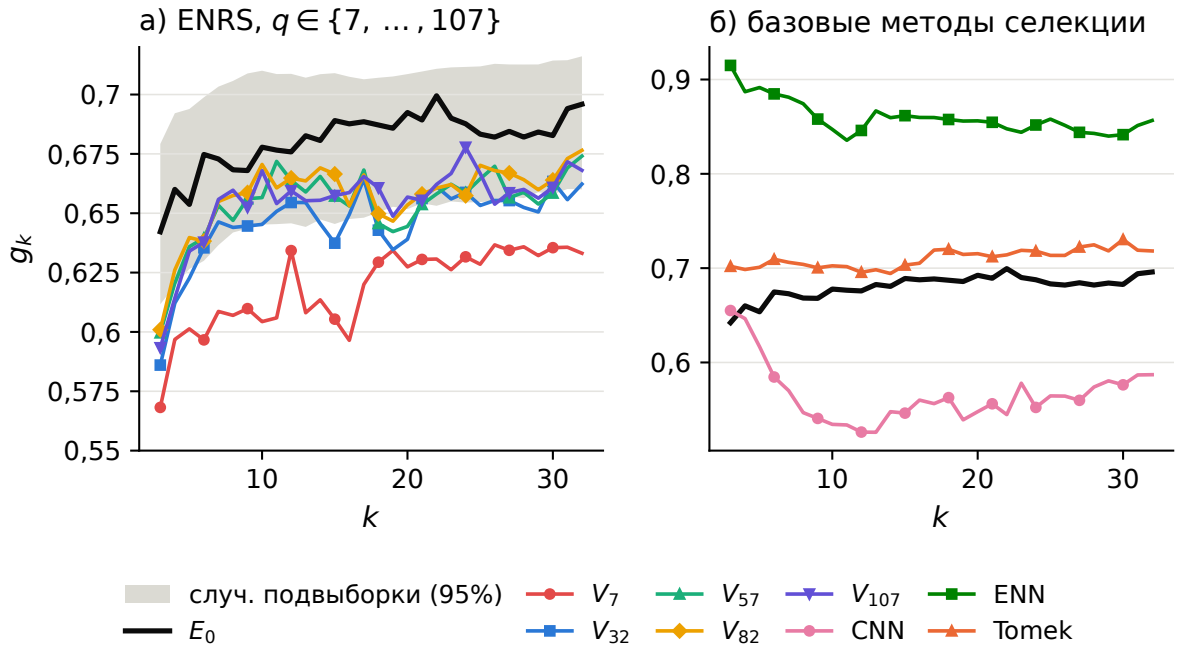


Рис. 1 Профили устойчивости g_k выборок German Credit: а) исходная выборка E_0, выборки ENRS $V_7, \ldots, V_{107}$ и 95 %-я полоса значений g_k для случайных подвыборок того же объёма и состава; б) E_0 и выборки, полученные методами CNN, ENN и Tomek Links.

Все выборки ENRS имеют меньшую устойчивость, чем случайные подвыборки того же объёма и состава. Для V_7 значения g_k лежат ниже 95 %-го интервала случайных подвыборок при всех 30 значениях k, для V_{32} — при 16, V_{57} — при 10, V_{107} — при 6, V_{82} — при 2. Различие средних значений значимо на уровне 0,05 для всех

выборок, кроме V_{82} (табл. 1). Таким образом, ENRS, как и CNN, удаляет преимущественно объекты, окрестности которых однородны по классам, и сохраняет объекты зон пересечения классов; ENN и Tomek Links, наоборот, удаляют пограничные объекты и повышают g_k. Наибольшее изменение локальной структуры даёт выборка V_7, которая одновременно является и наименьшей по объёму, однако отличие её объёма от остальных выборок ENRS (539 против 583–596) мало по сравнению с отличием её профиля: случайные подвыборки объёма 539 сохраняют профиль E_0 (табл. 2). Следовательно, отличие V_7 вызвано составом отобранных объектов, а не степенью редукции.

Корреляционный анализ и метод главных компонент

Коэффициенты корреляции между профилями выборок ENRS приведены в табл. 3. Без удаления тренда все корреляции значимы ($p < 0{,}001$) и велики (0,64–0,95), что в значительной мере объясняется общим ростом g_k с увеличением k. После удаления линейного тренда корреляции между V_{32}, V_{57}, V_{82}, V_{107} остаются высокими (0,79–0,93, $p < 10^{-6}$), а корреляции V_7 с остальными снижаются до 0,34–0,45 (p от 0,012 до 0,065). Следует учитывать, что соседние значения профиля не независимы (автокорреляция остатков с лагом 1 составляет 0,6–0,7 для $V_{32}, \ldots, V_{107}$), поэтому приведённые p-значения являются заниженными. Высокое сходство профилей $V_{32}, \ldots, V_{107}$ согласуется и с тем, что сами эти выборки на 92–98 % совпадают по составу.

Таблица 3 Коэффициенты корреляции между профилями устойчивости выборок ENRS: над диагональю — по исходным профилям, под диагональю — после удаления линейного тренда по k

	V_7	V_{32}	V_{57}	V_{82}	V_{107}
V_7	1	0,745	0,642	0,675	0,704
V_{32}	0,451*	1	0,913	0,891	0,909
V_{57}	0,341	0,866*	1	0,952	0,867
V_{82}	0,365*	0,817*	0,926*	1	0,901
V_{107}	0,421*	0,845*	0,792*	0,839*	1

Примечание. Все коэффициенты над диагональю значимы при $p < 0{,}001$; знаком * отмечены коэффициенты под диагональю, значимые при $p < 0{,}05$ (без поправки на автокорреляцию).

На рис. 2 показано расположение метаобъектов в пространстве первых двух главных компонент, построенных по профилям всех выборок, включая 1000 случайных подвыборок. Первая компонента объясняет 73,5 % дисперсии и практически совпадает с общим уровнем устойчивости $\bar{g}$, вторая — 7,1 %. Выборки $V_{32}, \ldots, V_{107}$ расположены на краю облака случайных подвыборок, V_7 — за его пределами, между ними и CNN; ENN и Tomek Links смещены в противоположную сторону.

Результаты на других наборах данных

В табл. 4 тот же анализ выполнен для всех шести наборов данных ($B = 100$). Снижение устойчивости по сравнению со случайными подвыборками значимо для Pima при всех q, для German Credit при всех q, кроме $q = 82$, и для WDBC при $q = 7$. Для Ionosphere и Sonar различия не значимы, а для Banknote значения g_k близки к 1 у всех выборок, и показатель не различает их. Следовательно, характер

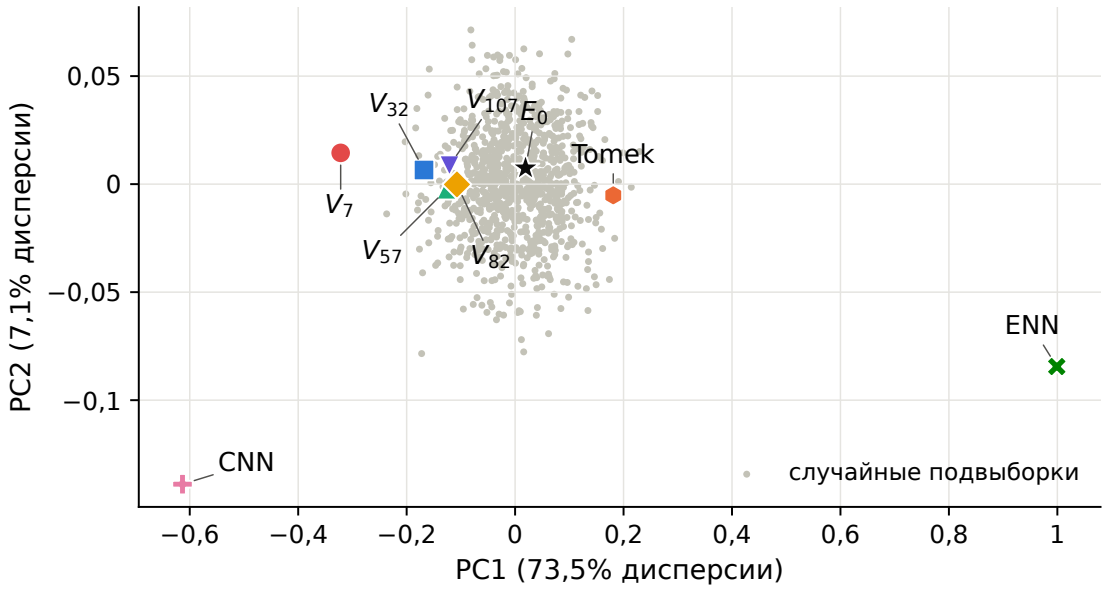


Рис. 2 Метаобъекты German Credit в пространстве первых двух главных компонент профилей устойчивости; серые точки — 1000 случайных подвыборок того же объёма и состава, что и $V_7, \ldots, V_{107}$.

изменения структуры зависит от данных, и вывод о снижении устойчивости не может быть перенесён на все наборы без проверки.

Таблица 4 Средняя устойчивость $\bar{g}$ выборок ENRS и (в скобках) среднее $\bar{g}$ случайных подвыборок того же объёма и состава

Набор	E_0	$q = 7$	$q = 32$	$q = 57$	$q = 82$	$q = 107$
German	0,681	0,619 (0,676)*	0,646 (0,678)*	0,654 (0,676)*	0,657 (0,676)	0,655 (0,678)*
Pima	0,735	0,600 (0,732)*	0,681 (0,729)*	0,684 (0,733)*	0,690 (0,733)*	0,689 (0,733)*
Ionosphere	0,870	0,879 (0,834)	0,850 (0,840)	0,810 (0,832)	0,801 (0,831)	0,810 (0,832)
Sonar	0,759	0,685 (0,712)	0,727 (0,724)	0,707 (0,724)	0,697 (0,723)	0,697 (0,726)
WDBC	0,962	0,891 (0,960)*	0,956 (0,961)	0,959 (0,960)	0,959 (0,960)	0,959 (0,959)
Banknote	0,999	0,999 (0,998)	1,000 (0,995)	0,998 (0,998)	0,999 (0,998)	0,999 (0,997)

Примечание. * — $\bar{g}$ выборки ENRS меньше, чем у случайных подвыборок, при $p < 0{,}05$ ($B = 100$).

Сравнение методов селекции по точности классификации

Для оценки ENRS как метода селекции использовалась 10-кратная стратифицированная перекрёстная проверка [26], повторённая 3 раза. Селекция выполнялась только на обучающей части каждого разбиения (нормировка признаков — также по обучающей части), затем классификатор 1-NN [27] строился по редуцированной выборке и проверялся на контрольной части. Поскольку классы в ряде наборов несбалансированы, в табл. 5 приведена сбалансированная точность — среднее значение полноты по двум классам — и доля сохранённых объектов. Для каждой выборки ENRS строилась случайная подвыборка того же объёма и классового состава (столбцы «Случ.»). Различия ENRS и случайной селекции проверялись критерием Уилкоксона для связанных выборок по 30 разбиениям [28].

Таблица 5 Сбалансированная точность 1-NN (%, верхняя строка) и доля сохранённых объектов (%, нижняя строка); 3×10-кратная перекрёстная проверка

Набор	Вся выборка	$q = 7$		$q = 57$		CNN	ENN	Tomek
		ENRS	Случ.	ENRS	Случ.			
German	66,3	62,4	63,2	63,0	63,7	62,0	62,9	65,7
	100	53	53	59	59	50	72	90
Pima	65,7	62,9†	66,9	62,8†	65,9	61,6	70,0	67,6
	100	51	51	60	60	50	74	87
Ionosphere	82,1	80,1†	82,3	80,8	82,5	81,3	79,7	82,1
	100	55	55	47	47	23	86	100
Sonar	85,5	84,0	81,2	83,3	81,5	84,4	80,7	84,7
	100	46	46	56	56	31	84	99
WDBC	94,8	89,6†	93,5	93,9	95,0	92,6	95,7	95,2
	100	31	31	71	71	14	97	98
Banknote	99,9	96,0†	99,8	99,4†	99,8	99,7	99,9	99,9
	100	54	54	53	53	2	100	100

Примечание. † — точность ENRS значимо ниже точности случайной подвыборки того же объёма и состава (критерий Уилкоксона, $p < 0{,}05$). Стандартное отклонение по разбиениям составляет 3–10 процентных пунктов (для Banknote — не более 2,5).

Результаты (табл. 5) показывают, что при сопоставимой степени редукции ENRS не превосходит случайную селекцию. Значимое ухудшение по сравнению со случайной подвыборкой наблюдается для Pima, Ionosphere, WDBC и Banknote при $q = 7$ и для Pima и Banknote при $q = 57$; значимых улучшений нет. С учётом всех пяти значений q и шести наборов (30 пар средних значений) точность ENRS ниже точности случайной подвыборки в 25 случаях (критерий Уилкоксона, $p < 0{,}001$). Метод CNN при меньшей доле сохранённых объектов даёт сопоставимую точность. Tomek Links сохраняет почти все объекты и точность полной выборки; ENN повышает точность на наборах Pima и WDBC, но снижает её на German, Ionosphere и Sonar. Эти результаты согласуются со структурным анализом: ENRS, по-видимому, удаляет преимущественно объекты с однородными по классу окрестностями, но, в отличие от CNN, не гарантирует согласованности редуцированной выборки с исходной (правильной классификации всех объектов E_0 по редуцированной выборке), что может объяснять потерю точности.

6 Обсуждение и ограничения

Протокол сравнения со случайными подвыборками того же объёма и состава позволил разделить два эффекта, которые при прямом сравнении профилей смешиваются: изменение g_k из-за уменьшения объёма и изменение из-за отбора конкретных объектов. Для German Credit первый эффект пренебрежимо мал, и снижение устойчивости выборок ENRS следует отнести к отбору объектов. Без этой базовой линии различия профилей нельзя интерпретировать: разброс g_k между случайными подвыборками (около 0,015) сопоставим с различиями между самими выборками $V_{32}, \ldots, V_{107}$.

Работа имеет ограничения. Правила удаления в ENRS эвристичны; величины r, z, d не пересчитываются после удалений, а параметр q не управляет степенью редукции. Показатель g_k определён для двух классов и перестаёт различать выборки, когда классы хорошо разделимы (набор Banknote). Использован один классифика-

тор (1-NN) и одна метрика (HEOM). Вариант ENRS с пересчётом характеристик и правилом, гарантирующим согласованность выборки, а также обобщение на многоклассовый случай являются предметом дальнейших исследований.

7 Заключение

Предложен алгоритм селекции объектов ENRS, основанный на отношениях объектов с ближайшими объектами противоположного класса; дано его формальное описание, доказана завершимость и получена оценка сложности $O(m^2)$. Для анализа сохранения локальной структуры выборки предложен протокол, в котором профиль устойчивости g_k порядкового признака, построенного по составу k ближайших соседей, сравнивается с профилями случайных подвыборок того же объёма и классового состава, исходной выборки и выборок, полученных методами CNN, ENN и Tomek Links.

На наборе German Credit все выборки ENRS (54–60 % объектов) имеют меньшую устойчивость, чем случайные подвыборки, т. е. в них уменьшается однородность окрестностей по классам. Сильнее всего структура меняется при $q = 7$; это вызвано составом отобранных объектов, а не степенью редукции, которая у всех выборок близка. Выборки при $q \geqslant 32$ почти совпадают по составу и по профилям устойчивости. На шести наборах данных точность 1-NN на выборках ENRS не превышает точности на случайных подвыборках того же объёма, поэтому в нынешнем виде ENRS не может быть рекомендован как метод сокращения выборки для повышения качества классификации. Предложенный протокол применим к любому методу селекции и позволяет установить, сохраняет ли метод локальную структуру данных лучше или хуже, чем случайное сокращение выборки.

Декларации

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Доступность данных. Использованы общедоступные наборы данных репозитория UCI Machine Learning Repository: Statlog (German Credit Data), Pima Indians Diabetes, Ionosphere, Connectionist Bench (Sonar), Breast Cancer Wisconsin (Diagnostic) и Banknote Authentication. Программная реализация алгоритма ENRS и всех вычислительных экспериментов доступна у автора по запросу.

Литература

[1] *Hart P. E.* The condensed nearest neighbor rule // IEEE Transactions on Information Theory. — 1968. — Vol. 14, №3. — P. 515–516. — doi: https://doi.org/10.1109/TIT.1968.1054155.

[2] *Wilson D. L.* Asymptotic properties of nearest neighbor rules using edited data // IEEE Transactions on Systems, Man, and Cybernetics. — 1972. — Vol. 2, №3. — P. 408–421. — doi: https://doi.org/10.1109/TSMC.1972.4309137.

[3] *Tomek I.* Two modifications of CNN // IEEE Transactions on Systems, Man, and Cybernetics. — 1976. — Vol. 6, №11. — P. 769–772. — doi: https://doi.org/10.1109/TSMC.1976.4309452.

[4] *Aha D. W., Kibler D., Albert M. K.* Instance-based learning algorithms // Machine Learning. — 1991. — Vol. 6. — P. 37–66. — doi: https://doi.org/10.1023/A:1022689900470.

[5] *Wilson D. R., Martinez T. R.* Reduction techniques for instance-based learning algorithms // Machine Learning. — 2000. — Vol. 38. — P. 257–286. — doi: https://doi.org/10.1023/A:1007626913721.

[6] *García S., Derrac J., Cano J. R., Herrera F.* Prototype selection for nearest neighbor classification: taxonomy and empirical study // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2012. — Vol. 34, № 3. — P. 417–435. — doi: https://doi.org/10.1109/TPAMI.2011.142.

[7] *Brighton H., Mellish C.* Advances in instance selection for instance-based learning algorithms // Data Mining and Knowledge Discovery. — 2002. — Vol. 6, № 2. — P. 153–172. — doi: https://doi.org/10.1023/A:1014043630878.

[8] *Olvera-López J. A., Carrasco-Ochoa J. A., Martínez-Trinidad J. F., Kittler J.* A review of instance selection methods // Artificial Intelligence Review. — 2010. — Vol. 34, № 2. — P. 133–143. — doi: https://doi.org/10.1007/s10462-010-9165-y.

[9] *Leyva E., González A., Pérez R.* Three new instance selection methods based on local sets: a comparative study with several approaches from a bi-objective perspective // Pattern Recognition. — 2015. — Vol. 48, № 4. — P. 1523–1537. — doi: https://doi.org/10.1016/j.patcog.2014.10.001.

[10] *Arnaiz-González Á., Díez-Pastor J. F., Rodríguez J. J., García-Osorio C.* Local sets for multi-label instance selection // Applied Soft Computing. — 2018. — Vol. 68. — P. 651–666. — doi: https://doi.org/10.1016/j.asoc.2018.04.016.

[11] *Arnaiz-González Á., Díez-Pastor J. F., Rodríguez J. J., García-Osorio C.* Instance selection of linear complexity for big data // Knowledge-Based Systems. — 2016. — Vol. 107. — P. 83–95. — doi: https://doi.org/10.1016/j.knosys.2016.05.056.

[12] *Liu C., Wang W., Wang M. [et al.]* An efficient instance selection algorithm to reconstruct training set for support vector machine // Knowledge-Based Systems. — 2017. — Vol. 116. — P. 58–73. — doi: https://doi.org/10.1016/j.knosys.2016.10.031.

[13] *Jiménez F., Sánchez G., Palma J., Sciavicco G.* Three-objective constrained evolutionary instance selection for classification: wrapper and filter approaches // Engineering Applications of Artificial Intelligence. — 2022. — Vol. 107. — Art. 104531. — doi: https://doi.org/10.1016/j.engappai.2021.104531.

[14] *Игнатьев Н. А., Турсунмуротов Д. Х.* Цензурирование обучающих выборок с использованием регуляризации отношений связанности объектов классов // Научно-технический вестник информационных технологий, механики и оптики. — 2024. — Т. 24, № 2. — С. 322–329. — doi: https://doi.org/10.17586/2226-1494-2024-24-2-322-329.

[15] *Игнатьев Н. А., Акбаров Б. Х.* Оценка близости структур отношений объектов обучающей выборки на многообразиях наборов латентных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. — 2023. — № 65. — С. 69–78. — doi: https://doi.org/10.17223/19988605/65/7.

[16] *Игнатьев Н. А.* Выбор целевых признаков для классификации и кластерного анализа структур отношений объектов // International Journal of Open Information Technologies. — 2026. — Vol. 14, № 3.

[17] *Jolliffe I. T., Cadima J.* Principal component analysis: a review and recent developments // Philosophical Transactions of the Royal Society A. — 2016. — Vol. 374, № 2065. — Art. 20150202. — doi: https://doi.org/10.1098/rsta.2015.0202.

[18] *Hofmann H.* Statlog (German Credit Data) // UCI Machine Learning Repository. — 1994. — doi: https://doi.org/10.24432/C5NC77. — URL: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data (дата обращения: 08.05.2026).

[19] *Smith J. W., Everhart J. E., Dickson W. C. [et al.]* Using the ADAP learning algorithm to forecast the onset of diabetes mellitus // Proceedings of the Annual Symposium on Computer Application in Medical Care. — 1988. — P. 261–265.

[20] *Sigillito V. G., Wing S. P., Hutton L. V., Baker K. B.* Classification of radar returns from the ionosphere using neural networks // Johns Hopkins APL Technical Digest. — 1989. — Vol. 10. — P. 262–266.

[21] *Gorman R. P., Sejnowski T. J.* Analysis of hidden units in a layered network trained to classify sonar targets // Neural Networks. — 1988. — Vol. 1, №1. — P. 75–89. — doi: https://doi.org/10.1016/0893-6080(88)90023-8.

[22] *Street W. N., Wolberg W. H., Mangasarian O. L.* Nuclear feature extraction for breast tumor diagnosis // Proceedings of SPIE. Biomedical Image Processing and Biomedical Visualization. — 1993. — Vol. 1905. — P. 861–870. — doi: https://doi.org/10.1117/12.148698.

[23] *Lohweg V.* Banknote Authentication // UCI Machine Learning Repository. — 2012. — doi: https://doi.org/10.24432/C55P57.

[24] *Wilson D. R., Martinez T. R.* Improved heterogeneous distance functions // Journal of Artificial Intelligence Research. — 1997. — Vol. 6. — P. 1–34. — doi: https://doi.org/10.1613/jair.346.

[25] *Pedregosa F., Varoquaux G., Gramfort A. [et al.]* Scikit-learn: machine learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.

[26] *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). — 1995. — Vol. 2. — P. 1137–1143.

[27] *Cover T. M., Hart P. E.* Nearest neighbor pattern classification // IEEE Transactions on Information Theory. — 1967. — Vol. 13, №1. — P. 21–27. — doi: https://doi.org/10.1109/TIT.1967.1053964.

[28] *Demšar J.* Statistical comparisons of classifiers over multiple data sets // Journal of Machine Learning Research. — 2006. — Vol. 7. — P. 1–30.

UDC 004.85:519.25

INSTANCE SELECTION BASED ON NEAREST-ENEMY RELATIONS AND PRESERVATION OF THE LOCAL STRUCTURE OF A TRAINING SAMPLE

Ramazonov Sh.Sh.

ramazonov377@gmail.com

National University of Uzbekistan named after Mirzo Ulugbek, 4, Universitetskaya str., Tashkent, 100174 Uzbekistan

The paper considers instance selection for a training sample and the analysis of how selection affects the local structure of the data. The ENRS (Enemy-Neighbor Relation Selection) algorithm is proposed, in which the decision to remove an object is based on its relations with the nearest objects of the opposite class and with same-class objects whose spheres of influence contain it. The local structure of a sample is described by a feature formed from the composition of the k nearest neighbours of an object and by the stability indicator of this feature; the set of indicator values for different k represents the sample as a meta-object in a unified space. A method for assessing the preservation of local structure is proposed, based on comparing a reduced sample with random subsamples of the same size and class composition. Computational experiments were carried out on the German Credit data set and five other UCI data sets, and ENRS was compared with the CNN,

ENN and Tomek Links selection methods. It is found that on the German Credit and Pima data sets the samples produced by ENRS differ from random subsamples by a lower class homogeneity of object neighbourhoods, that the extent of this difference depends on the algorithm parameter, and that the classification accuracy on ENRS samples does not exceed that on random subsamples of the same size. The proposed approach separates changes of local structure caused by the selection of objects from those caused by the reduction of sample size and can be applied to the analysis of any selection method.

Keywords: instance selection, nearest enemy, k nearest neighbours, local structure of a sample, feature stability, meta-object, random subsamples

Citation: Ramazonov Sh.Sh. 2026. Instance selection based on nearest-enemy relations and preservation of the local structure of a training sample. *Problems of Computational and Applied Mathematics.* 5(76): 46-59.

DOI: 10.71310/pcam.5_76.2026.04

№ 5(76) 2026

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

AiRI
RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 5(76) 2026

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Главный редактор:
Равшанов Н.

Заместители главного редактора:
Арипов М.М., Шадиметов Х.М., Ахмедов Д.Д.

Ответственный секретарь:
Убайдуллаев М.Ш.

Редакционный совет:
Азамов А.А., Алоев Р.Д., Амиргалиев Е.Н. (Казахстан), Арушанов М.Л., Абдуллозода С.Ф. (Таджикистан), Боборахимов Б.И., Бурнашев В.Ф., Джумаёзов У.З., Загребина С.А. (Россия), Задорин А.И. (Россия), Игнатьев Н.А., Ильин В.П. (Россия), Иманкулов Т.С. (Казахстан), Исмагилов И.И. (Россия), Кабанихин С.И. (Россия), Курбонов Н.М., Маматов Н.С., Мирзаев Н.М., Мурадов Ф.А., Назирова Э.Ш., Нормуродов Ч.Б., Нуралиев Ф.А., Нуралиев Ф.М., Опанасенко В.Н. (Украина), Расулмухамедов М.М., Садуллаева Ш.А., Старовойтов В.В. (Беларусь), Хаётов А.Р., Халджигитов А., Хамдамов Р.Х., Хужаев И.К., Хужаеров Б.Х., Эшмаматова Д.Б., Дустмуродова Ш.Ж., Чье Ен Ун (Россия), Шабозов М.Ш. (Таджикистан), Dimov I. (Болгария), Li Y. (США), Mascagni M. (США), Min A. (Германия), Singh M. (Южная Корея).

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

При перепечатке материалов ссылка на журнал обязательна.
За точность фактов и достоверность информации ответственность несут авторы.

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 71 263-41-98.
Э-почта: journals@airi.uz.
Веб-сайт: https://journals.airi.uz.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 28.10.2026 г.
Формат 60x84 1/8. Заказ № 6. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 5(76) 2026

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

Editor-in-Chief:
Ravshanov N.

Deputy Editors:
Aripov M.M., Shadimetov Kh.M., Akhmedov D.D.

Executive Secretary:
Ubaydullaev M.Sh.

Editorial Council:
Azamov A.A., Aloev R.D., Amirgaliev E.N. (Kazakhstan), Arushanov M.L., Abdullozoda S.F. (Tajikistan), Boborakhimov B.I., Burnashev V.F., Djumayozov U.Z., Zagrebina S.A. (Russia), Zadorin A.I. (Russia), Ignatiev N.A., Ilyin V.P. (Russia), Imankulov T.S. (Kazakhstan), Ismagilov I.I. (Russia), Kabanikhin S.I. (Russia), Kurbonov N.M., Mamatov N.S., Mirzaev N.M., Muradov F.A., Nazirova E.Sh., Normurodov Ch.B., Nuraliev F.A., Nuraliev F.M., Opanasenko V.N. (Ukraine), Sadullaeva Sh.A., Starovoitov V.V. (Belarus), Khayotov A.R., Khaldjigitov A., Khamdamov R.Kh., Khujaev I.K., Khujayorov B.Kh., Eshmamatova D.B., Dustmurodova Sh.J., Chye En Un (Russia), Shabozov M.Sh. (Tajikistan), Dimov I. (Bulgaria), Li Y. (USA), Mascagni M. (USA), Min A. (Germany), Singh M. (South Korea).

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
Certificate of Registration No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

At a reprint of materials the reference to the journal is obligatory.
Authors are responsible for the accuracy of the facts and reliability of the information.

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 71 263-41-98.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIRI printing office.
Signed for print 28.10.2026
Format 60x84 1/8. Order No. 6. Print run of 100 copies.

Содержание

Расулов А.С., Раимова Г.М., Рахматов М.Й.
Метод Монте-Карло для решения нелинейного кинетического уравнения Больцмана 8

Эшкулов М.У., Хамдамов Р.Х.
Многокритериальная оптимизация системы подъёма воды с использованием солнечных фотоэлектрических панелей 22

Абдуллаев К.Д.
Принцип разделения классов и зависимость результатов кластеризации от меры расстояния 33

Рамазонов Ш.Ш.
Селекция объектов по отношениям с ближайшими объектами противоположного класса и сохранение локальной структуры обучающей выборки . . 46

Исомиддинов А.И., Акбарова А.А., Акбаров Б.Х.
Марковская переключающаяся векторная авторегрессия для скрытых гликемических режимов в реальных данных самоконтроля диабета: оценка на отложенной выборке 60

Ким К.К.
Сравнительный анализ XAI-подходов для обнаружения синтетической речи 72

Эшмаматова Д.Б., Солижанова К.А.
Пфаффианова структура и спектральная динамика дискретного оператора Лотки–Вольтерры: подход к интеграции геологических данных 95

Джумаёзов У.З., Абдирахмонова М.Н., Усмонов Л.С.
Численное решение трёхмерной связанной динамической задачи термоупругости в деформациях 119

Бакаев И.И., Ахмеджанова Д.А.
α-справедливая оптимизационная модель и численные алгоритмы выбора общей лексической нормы тюркских языков 137

Равшанов Н., Камолиддинова Ф., Lu Y.
Моделирование переноса и диффузии загрязняющих веществ в пограничном слое атмосферы с учётом орографии местности и метеорологических факторов 148

Contents

Rasulov A.S., Raimova G.M., Rakhmatov M.Y.
Monte Carlo solution of the nonlinear kinetic Boltzmann equation 8
Eshkulov M.U., Khamdamov R.Kh.
Multi-criteria optimization of solar photovoltaic-based water lifting systems . . . 22
Abdullaev K.D.
The principle of class separation and the dependence of clustering results on the distance measure 33
Ramazonov Sh.Sh.
Instance selection based on nearest-enemy relations and preservation of the local structure of a training sample 46
Isomiddinov A.I., Akbarova A.A., Akbarov B.X.
Markov-switching vector autoregression for latent glycaemic regimes in real-world diabetes self-monitoring data: an out-of-sample evaluation 60
Kim K.K.
Comparative analysis of XAI approaches for synthetic speech detection 72
Eshmamatova D.B., Solijanova K.A.
Pfaffian structure and spectral dynamics of a discrete Lotka–Volterra operator: a framework for geological evidence integration 95
Djumayozov U.Z., Abdirakhmonova M.N., Usmonov L.S.
Numerical solution of a three-dimensional coupled dynamic thermoelasticity problem in terms of strains 119
Bakayev I.I., Akhmedjanova D.A.
An α-fair optimization model and numerical algorithms for selecting a common lexical norm of Turkic languages 137
Ravshanov N., Kamoliddinova F., Lu Y.
Modeling pollutant transport and diffusion in the atmospheric boundary layer considering terrain orography and meteorological factors 148