МНОГОЗНАЧНАЯ КЛАССИФИКАЦИЯ МЕТОК КЛАССОВ СИСТЕМНЫХ ЖУРНАЛОВ КОМПЬЮТЕРНЫХ СЕТЕЙ. СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ КЛАССИФИКАТОРОВ

Шелухин О. И.[1], Раковский Д.И.[2]

***Цель исследования**: проведение сравнительного анализа бинарного (БК), многоклассового (МклК) и многозначного (МзнК) методов классификации в задачах обеспечения информационной безопасности посредством анализа записей системных журналов, порожденных компьютерной сетью (КС), на примере экспериментальных данных (ЭД) разной атрибутной размерности путем сопоставления результатов классификации по бинарным метрикам оценки качества для каждой размерности.*

***Метод.** Исследовались алгоритмы классификации «Дерево решений», Decision Tree Classifier, (DTC); «Дополнительные деревья решений», Extra Trees Classifier, (ETC); «К ближайших соседей», KNeighbors Classifier, (KNC); «Случайный лес», Random Forest Classifier, (RFC). Исследование проводилось по трем метрикам, основанным на площади под кривой рабочей характеристики приемника (Area Under the Receiver Operating Characteristic Curve: ROC AUC Micro, ROC AUC Macro, ROC AUC Weighted) двумя методами «Один против одного» (One-vs-one, OVO) или «один против всех» (One-vs-everyone, OVE или One-vs-rest - OVR). Эксперимент подразумевал итерационную оценку качества классификации в зависимости от количества атрибутов ЭД. Атрибуты ЭД ранжировались по убыванию их совокупной информативности и статистической значимости.*

***Результаты исследования.** Проведен анализ бинарной, многоклассовой и многозначной реализаций алгоритмов DTC, ETC, RNC, RFC по параметру ROC-AUC (метрики - $ROC\text{-}AUC_{score\ ovo\ macro}$, $ROC\text{-}AUC_{score\ ovo\ weighted}$, $ROC\text{-}AUC_{score\ ovr\ macro}$, $ROC\text{-}AUC_{score\ ovr\ micro}$, $ROC\text{-}AUC_{score\ ovo\ micro}$, $ROC\text{-}AUC_{score\ ovr\ weighted}$). Эксперимент проводился для 28 различных размерностей атрибутного пространства ЭД. Результаты исследования метрики $AUC_{ovo\ micro}$ классификаторов МзнК, МклК и БК от размерности первичных атрибутов показали, что выигрыш МзкК в сравнении с МклК в среднем составляет 15% при ETC и достигает 20% для RFC. Выигрыш по метрике $AUC_{ovo\ micro}$ МклК в сравнении с ВК составляет в среднем 20% при большом числе атрибутов и снижается при уменьшении числа атрибутов в ЭД. Алгоритмы DTC и KNC показывают несколько худшие результаты, хотя общая закономерность сохраняется. Исследование зависимости эффективности МзнК по параметру ROC-AUC от размерности первичных атрибутов в ЭД показало, что метрика $AUC_{ovo\ micro}$ демонстрирует наилучшие результаты для алгоритмов ETC и RFC и составляет в среднем 80% даже при классификации в малом атрибутном пространстве. Исследование показало, что применение многозначной классификации способно увеличить точность классификации до 20% по метрике $AUC_{ovo\ micro}$.*

***Научная новизна** заключается в исследовании эффективности указанных методов классификации применительно к ЭД КС по множеству выходных метрик. Показано, что выигрыш МзнК перед иными методами классификации составляет, суммарно, до 35% (МзнК против БК).*

***Ключевые слова:** интеллектуальный анализ данных; аномальное состояние; multi-label; бинарная классификация; многоклассовая классификация; feature importance; Decision Tree Classifier; Extra Trees Classifier; KNeighbors Classifier; Random Forest Classifier.*

DOI:10.21681/2311-3456-3-62-77

1 Шелухин Олег Иванович, доктор технических наук, профессор Московского технического университет связи и информатики, Москва, Россия. E-mail: sheluhin@mail.ru, ORCID: https://orcid.org/0000-0001-7564-6744

2 Раковский Дмитрий Игоревич, аспирант Московского технического университета связи и информатики, Москва, Россия. E-mail: Prophet_alpha@mail.ru, ORCID: https://orcid.org/0000-0001-7689-4678

Введение и постановка задачи

Современные компьютерные сети (КС) обладают сложной инфраструктурой, требующей постоянного мониторинга с целью выявления аномальных состояний, вызывающих сбои в работе систем (см. [1, 2], а также публикации *Ruan W., Liub Y., Zhaob R.*[1] и *Lima A.C.E.S., de Castro L.N.*[2]). Под состоянием КС будем понимать совокупность значений системных атрибутов, характеризующих основные показатели функционирования КС формируемых в виде категориальных значений с временной меткой.

В качестве системных показателей, характеризующих качество функционирования компьютерной сети, как правило используется уровень обслуживания (*Service Level Objectives, SLO*), и соглашение об уровне предоставляемого сервиса (*Service Level Agreement, SLA)*[3].

Важной проблемой интеллектуальной обработки данных системных журналов является классификация сразу нескольких целевых столбцов, приводящая к решению *задачи многозначной классификации* [3].

Многозначная классификация встречается в ряде практических задач [4]. Например, в рамках информационной безопасности могут решаться задачи одновременного обнаружения множества сетевых атак. В работе [5] с целью повышения точности многозначной классификации исследуется метод, основанный на обнаружении аномалий с помощью нейронной сети с архитектурой типа «автокодировщик». Полученные результаты существенно зависят от типа проводимой атаки (разброс оценок точности по метрике *Accuracy* составляет 0,61 ... 0,99). Подчеркивается возможность работы предложенного метода как с открытым, так и с зашифрованным сетевым трафиком.

В работе [6] рассматривается классификации сетевого трафика методами многозначного анализа. Показано, что бустинговые алгоритмы способны присваивать многозначные метки классов с точностью 0,98 по метрике «площадь под кривой рабочей характеристики приемника» (*Area Under the Receiver Operating Characteristic Curve*, или *ROC-AUC)*.

В работе авторского коллектива *Shalaginov A., Franke K.*[4] исследуется многозначная классификация вредоносного программного обеспечения (ВПО) на основе нечеткой логики и нейронных сетей глубокого обучения и достигается точность многозначной классификации ВПО по параметру Accuracy на уровне 0,69. В работе [7], посвященной автоматизации детектирования вредоносного программного обеспечения и присвоению ему специальных типизирующих тегов, показано, что точность по параметру Accuracy может достигать 0,7. В частных случаях (присвоение отдельных тегов), в оговоренных условиях точность по параметру AUC достигает 0,98.

В работах Д.А. Молодцова[5] вводится в рассмотрение мягкая вероятность, предлагается построение многозначных зависимостей на их основе. Несмотря на экзотический математический аппарат, свободный от необходимости принятия гипотезы о случайной составляющей, идеи, заложенные в указанных работах, нашли применение в задачах регрессионного анализа и прогнозирования [8]. Суть предложенного метода заключалась в том, что закономерность описывалась не однозначной функцией, а многозначным отображением в форме мультимножества.

В работе [9] рассматривается задача прогнозирования состояний КС с помощью использования многозначных отображений, для которых любой набор результатов опытов, представленных в виде таблицы, можно рассматривать как график точечно множественного отображения $D_n = \left\{ (x_1, y_1), \ldots, (x_n, y_n) \middle| (x_i, y_i) \in X \times Y \right\}$. Здесь $X \times Y$ означает декартово произведение двух множеств – X и Y – элементами которого являются все возможные упорядоченные пары «входных» - x_i и «выходных» - y_i элементов исходных множеств.

Работы, в том или ином виде исследующие проблемы многозначности, объединены термином: *многозначное обучение*, *Multi-Label Learning, MLL* [10-11] и иллюстрируют актуальность этой задачи, особенно для обеспечения информационной безопасности КС. Наиболее подробно методы решения задачи MLL рас-

1 Ruan W., Liub Y., Zhaob R. Pattern Discovery in DNS Query Traffic // Procedia Computer Science. 2013. Т. 17. С. 80–87. DOI: 10.1016/j.procs.2013.05.012

2 Lima A.C.E.S., de Castro L.N. A multi-label, semi-supervised classification approach applied to personality prediction in social media // Neural Networks. 2014. Т. 58. С. 122-130

3 Gnanasekar J. Autonomous Intelligent Agent Indemnification in SLA (IAIS) Architecture for Effortless Monitoring of SLA Violations // Ictact journal on soft computing. 2015. № 5. С. 979-984. DOI: 10.21917/ijsc.2015.0137.

4 Shalaginov A., Franke K. A deep neuro-fuzzy method for multi-label malware classification and fuzzy rules extraction // В сборнике: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). 2017. С. 1-8. DOI: 10.1109/SSCI.2017.8280788.

5 Молодцов Д.А. Идеи мягкой вероятности как новый подход к построению теории вероятностей: Гипотезы стохастической устойчивости и вероятность. М.: URSS, 2015. 112 с. ISBN 978-5-9710-1514-7; Молодцов Д. А. Экстраполяция многозначных зависимостей // Нечеткие системы и мягкие вычисления. 2017. Т. 12. № 1. с. 45–63

сматриваются в публикации *Tsoumakas G., Katakis I., Vlahavas I.*[6] Актуальность MLL также может быть подтверждена наличием разнообразного программного обеспечения, работающего с многозначными метками7: WEKA[8]; KEEL[9]; Scikit-learn[10].

Постановка задачи

Задача выявления нарушений нормального функционирования КС за счет классификации соответствующих состояний может быть решена одним из трех методов: бинарной (БК), многоклассовой (МклК) и многозначной (МзнК) классификации.

Целью работы является сравнительный анализ этих трех методов классификации на экспериментальных данных (ЭД) разной атрибутной размерности путем сопоставления результатов классификации по бинарным метрикам оценки качества для каждой размерности.

Сформируем общие рекомендации по использованию методов бинарной, многоклассовой и многозначной классификации.

Для БК необходимо зафиксировать факт возникновения аномалии хотя бы по одному вторичному атрибуту исследуемой КС.

Оценка эффективности многоклассовых и многозначных алгоритмов классификации может быть осуществлена по шести метрикам, основанным на *Area under curve (AUC)*, площадью под *receiver operating characteristic* (*ROC*)[11].

В зависимости от методов вычисления *AUC* метрики подразделялись на – «Один против одного» (*One-vs-one, OVO*) или «один против всех» (*One-vs-everyone, OVE* или *One-vs-rest - OVR*).

В каждом методе метрики могут быть вычислены тремя разными способами:

Micro – Микро-подход заключается в агрегации результатов классификации по каждому из *M* состояний отдельно по каждой метрике, после чего происходит вычисление итоговой метрики:

$$B_{micro} = B\left(\sum_{m=1}^{M} TP_m, \sum_{m=1}^{M} TN_m, \sum_{m=1}^{M} FP_m, \sum_{m=1}^{M} FN_m\right). \quad (1)$$

Macro – Макро-подход заключается в вычислении метрик для каждого из *M* состояний КС и взятия их среднего арифметического:

$$B_{macro} = \frac{1}{M}\sum_{m=1}^{M} B(TP_m, FP_m, TN_m, FN_m) \quad (2)$$

Weighted – Взвешенный подход заключается в агрегации результатов классификации по каждому из *M* состояний отдельно по каждой метрике. После агрегации вычисляется *Accuracy* для каждого состояния КС. Каждая метрика – *TP, FP, FN, TN* - нормируется на *Accuracy* и вычисляется итоговая метрика:

$$B_{micro} = B\left(\sum_{m=1}^{M} TP_m / A_m, \sum_{m=1}^{M} TN_m / A_m, \sum_{m=1}^{M} FP_m / A_m, \sum_{m=1}^{M} FN_m / A_m\right), \quad (3)$$

$$A_m = \frac{TP_m + TN_m}{TP_m + TN_m + FP_m + FN_m}$$

где A_m – *Accuracy*.

Используя рассмотренные метрики, необходимо не только установить факт возникновения аномалии, но и конкретизировать текущее состояние КС: нормальное или аномальное. Если принимается решение о том, что состояние КС аномальное, необходимо дополнительно оценить, какая именно аномалия реализуется в текущий момент.

Для этого требуется выполнить сравнительный анализ многозначных и многоклассовых алгоритмов классификации между собой по совокупности выходных результатов эксперимента. Необходимо исследовать влияние разнообразия первичных атрибутов на итоговый результат классификации.

Процесс проведения исследования может быть разделен два этапа. На этапе №1 выполняется предобработки исходных ЭД. На этапе №2 ЭД разделяются на первичные и вторичные атрибуты. Вторичные атрибуты кодируются состояниями КС, после чего осуществляется классификация данных после предобработки.

В работе рассматриваются результаты исследования БК, МклК и МзнК классификаторов с помощью

6 Gibaja E., Ventura S. A Tutorial on Multi-Label Learning // ACM Computing Surveys. 2015. №47. С. 1-40. DOI: 10.1145/2716262

7 Tsoumakas G., Katakis I., Vlahavas I. Mining Multi-label Data. Data Mining and Knowledge Discovery Handbook. 2 изд. Stanford, California: Springer Series in Statistics (SSS), 2010. 1383 с. С. 667 – 685. DOI: 10.1007/978-0-387-09823-4

8 Hall M., Frank E., Holmes G., Pfahringer B., Peter R., Witten I. The WEKA data mining software: An update // SIGKDD Explorations, 2009, Т. 11, № 1.

9 Triguero I., González S., Moyano J. M., García S., Alcalá-Fdez J., Luengo J., Fernández A., Jesus M. J., Sánchez L., Herrera F. KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining // International Journal of Computational Intelligence Systems. 2017. № 10. С. 1238-1249

10 Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas I., Passos A., Cournapeau D., Brucher M.,,Perrot M., Duchesnay E. Scikit-learn: Machine Learning in Python // JMLR 2011, Т. 95, №12, С. 2825-2830

11 Hand D.J., Till R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems // Machine Learning, 2001, Т. 45 № 2, С. 171-186

разработанного фреймворка, реализованного на ПО *Python* версии 3.8.

С этой целью исследовались ЭД полученные в результате предварительной очистки экспериментальные данные представленные в [12 - 15] применительно к задаче выявления нарушений нормального функционирования КС.

В каждом из рассмотренных случаев первичные атрибуты КС ранжировались по убыванию их совокупной информативности и статистической значимости, после чего подавались на вход набору алгоритмов классификации в цикле.

Разделение ЭД на первичные и вторичные атрибуты

КС можно представить в виде множества из *M* наборов значений дискретно изменяющихся атрибутов («исторических данных») КС:

$$\begin{aligned} \mathbf{A} \subseteq A_{перв} \cup A_{втор} = \\ = \{A_{перв\,1}, A_{перв\,2}, ..., A_{перв\,2\,len_1}\} \cup \\ \cup \{A_{втор\,1}, A_{втор\,2}, ..., A_{втор\,len_2}\}; \\ A_m = \{a_{mn}; m = \overline{1, M}, n = \overline{1, N}\}, \end{aligned} \quad (4)$$

$$A_m \subset \mathbf{A}, M = len_1 + len_2.$$

Атрибуты КС в (4), могут подразделяться на два типа: первичные $\{A_{перв\,k_1}; k_1 = 1, len_1\}$ и вторичные $\{A_{втор\,k_2}; k_2 = 1, len_2\}$.

Заметим, что определение аномальных состояний КС согласно правилам *SLO* чаще всего выполняется на основании именно вторичных атрибутов [16, 17].

В дальнейшем вторичными считаются атрибуты, на основании которых выносится решение о соответствии КС уровню обслуживания *SLO*. Остальные атрибуты считаются первичными.

Конкретизируем показатели уровней обслуживания *SLO* и будем считать, что КС функционирует в штатном режиме, если ни один порог уровня обслуживания *SLO* не превышен. В противном случае будем считать, что КС нарушила уровень обслуживания. Руководствуясь результатами статистического анализа, проведенного в [14], сформируем требования к *SLO* и связанные с ним состояния КС в виде порогов, определяющих категориальные маркеры. Для исследуемых ЭД КС эти уровни представлены в табл. 1.

На этапе классификации вторичные атрибуты исключаются, поскольку рассматривается ситуация наличия скрытой переменной, отображающейся в соответствующие категориальные понятия.

В качестве входных данных при проведении вычислительного эксперимента использовались следующие параметры:

— Логическая переменная, отвечающая за тип классификации: L_{value1} = **{бинарная, многоклассовая, многозначная};**

— Логическая переменная, отвечающая за необходимость предварительного перемешивания данных: L_{value2} = **{без перемешивания};**

Таблица 1

Условия возникновения состояний КС в зависимости от нарушаемых порогов SLO

Условие	Атрибут КС, связанный с условием	Соответствующее состояние КС
время задержки сигнала к тестовому серверу > 5 мс.	*ping_avg*	*signal_delay*
время ответа тестового сервера > 1.5 с.	*server_response_timetotal*	*server_response_delay*
количество пакетов, потерянных при передаче к тестовому серверу > 0 шт.	*network_outdropped*	*packets_dropped*
время обработки запроса диском хостовой машины > 2 с.	*disk_ioreadmergespersec*	*disk_iowriteawait*
Иначе	-	*normal*

Таблица 2

Таблица соответствия наименований атрибутов КС и кодовых значений

Наименование атрибута	Соответствующий код	Наименование атрибута	Соответствующий код
cpu_iowait	A1	**load_fifteenminutes**	A18
cpu_nice	A2	**load_fiveminutes**	A19
cpu_softirq	A3	**load_oneminute**	A20
cpu_system	A4	**network_inbytes**	A21
cpu_user	A5	**network_inpackets**	A22
memory_actualfree	A6	**network_outbytes**	A23
memory_free	A7	**network_outdropped**	A24
memory_swapusedpct	A8	**dns_answerscount**	A25
disk_await	A9	**dns_networkbytes**	A26
disk_busy	A10	**http_requestbytes**	A27
disk_ioreadmergespersec	A11	**http_responsebytes**	A28
disk_ioreadrequestspersec	A12	**ping_avg**	A29
disk_iostatrequestavgsize	A13	**ping_max**	A30
disk_iowriteawait	A14	**ping_min**	A31
disk_iowritemergespersec	A15	**server_response_ timenamelookup**	A32
disk_iowriterequestspersec	A16	**server_response_ timestarttransfer**	A33
disk_writebytes	A17	**server_response_timetotal**	A34

— Логическая переменная, отвечающая за необходимость трансформации атрибутов ЭД: L_{value3} = **{трансформация необходима}**;

— Количество блоков разделения ЭД в режиме перекрестной проверки (кросс-валидации) по нотации *K-Fold*: L_{value4} = **{разделение на 2 блока}**;

— Массив, содержащий в себе наименование всех вторичных атрибутов, исследуемых в ЭД: L_{value5} = **{'ping_avg', 'server_response_ timetotal', 'network_outdropped', 'disk_ ioreadmergespersec'}**.

Исходя из указанных входных параметров ЭД исследовались три типа классификаторов: БК, МклК, МзнК.

Использовались алгоритмы классификации со следующими гиперпараметрами:

— «**Дерево решений**», **Decision Tree Classifier, DTC**; в качестве гиперпараметров выбраны стандартные рекомендации библиотеки *scikit-learn* с фиксированным начальным значением *random_state=0*;

— «**Дополнительные деревья решений**», **Extra Trees Classifier, ETC**; в качестве гиперпараметров выбраны стандартные рекомендации библиотеки scikit-learn [16] с фиксированным начальным значением *random_state=0*;

— «**К ближайших соседей**», **KNeighbors Classifier, KNC**; в качестве гиперпараметров выбраны: стандартные рекомендации библиотеки scikit-learn [16] с фиксированным начальным значением *random_state=0,* в дополнение метрическая величина, описывающая количество соседей, используемых по умолчанию для запросов *kneighbors, n_neighbors=3*;

— «**Случайный лес**», **Random Forest Classifier, RFC**; в качестве гиперпараметров выбраны: стандартные рекомендации библиотеки *scikit-learn* с фиксированным начальным значением *random_ state=0,* в дополнение метрическая величина, описывающая глубину дерева, *max_depth = 3*;

Согласно заданным параметрам перекрестной проверки для каждой итерации цикла набор данных разделялся на обучающую и тестовую выборку, после чего происходило поочередное обучение и тестирование каждого из указанных алгоритмов классификации.

Таблица 3

Описательная статистика исследуемого набора ЭД

Атр. КС	mean	std	min	25%	50%	75%	max	Ун. знач. атриб.
A1	0,04	0,04	0,00	0,01	0,04	0,04	0,67	1810
A2	0,01	0,06	0,00	0,00	0,00	0,00	1,16	530
A3	0,02	0,01	0,01	0,02	0,02	0,02	0,12	831
A4	0,38	0,05	0,19	0,36	0,38	0,39	0,96	3061
A5	1,34	0,71	0,42	0,71	1,40	1,45	7,62	3759
A6	3,5E+10	5,3E+09	2,7E+10	3,0E+10	3,2E+10	4,2E+10	4,5E+10	4177
A7	1,6E+09	2,3E+09	3,6E+08	4,6E+08	6,2E+08	1,4E+09	1,9E+10	4171
A8	0,07	0,06	0,00	0,01	0,12	0,14	0,18	248
A9	1,3E+12	8,6E+13	0,00	0,37	0,43	0,57	5,4E+15	316
A10	2,26	3,48	0,33	0,80	1,97	2,20	48,80	411
A11	0,22	1,93	0,00	0,00	0,00	0,00	67,27	112
A12	3,0E+11	1,4E+14	0	0	0	0	7,2E+16	212
A13	8049	16283	2088	3112	4450	6158	235426	4150
A14	1,75	12,01	0,00	0,37	0,40	0,53	313	320
A15	2,29	24,06	0,00	0,00	0,07	0,13	1478,72	204
A16	29,23	23,75	10,00	13,40	31,63	33,17	797,4	981
A17	9,2E+06	2,2E+06	5,7E+06	7,1E+06	9,0E+06	1,1E+07	1,2E+07	4210
A18	1,97	0,98	0,93	1,54	1,86	2,11	8,50	1816
A19	1,97	1,14	0,74	1,49	1,82	2,13	11,35	1969
A20	1,97	1,31	0,38	1,35	1,76	2,17	15,95	2217
A21	3,0E+10	4,4E+10	0	1,2E+08	3,2E+09	4,8E+10	2,3E+11	3397
A22	7,2E+07	1,2E+08	0	257826,3	2005107	1,0E+08	7,0E+08	3397
A23	4,3E+10	7,9E+10	0	2,7E+08	3,8E+09	6,0E+10	6,4E+11	3401
A24	0,57	1,67	0	0	0	0	10	22
A25	0,05	0,17	0	0	0	0	6	113
A26	82	67	24	58	58	58	662	1077
A27	120	11	93	120	120	120	443	15
A28	171	2066	137	137	137	137	143262	18
A29	2,20	0,61	1,83	2,01	2,05	2,10	12,08	749
A30	2,48	1,41	1,89	2,08	2,15	2,23	27,67	957
A31	1,99	0,25	1,77	1,92	1,96	2,00	7,00	471
A32	0,02	0,09	0,00	0,01	0,01	0,01	5,51	27
A33	1,68	6,71	0,00	0,96	1,04	1,13	84,85	635
A34	523,54	2156,69	0,01	0,99	1,07	1,18	15067,35	982

Результаты работы каждого из алгоритмов классификации **DTC, ETC, KNC, RFC** оценивался по трем метрикам (см. формулы (1) – (3)) двумя методами «один против одного» (*One-vs-one, OVO*) или «один против всех» (*One-vs-everyone, OVE* или *One-vs-rest - OVR*).

В конце итерации из исходного множества исключался первичный атрибут с наивысшей важностью.

После окончания эксперимента все значения эффективности классификации на разных блоках перекрестной проверки усреднялись.

Перед проведением вычислительных экспериментов необходимо проведение разведочного анализа, предобработки ЭД и выполнить оценку их совокупной информативности и статистической значимости атрибутов ЭД.

Разведочный анализ ЭД

Рассмотрим результаты разведочного анализа ЭД, позволяющие получить описательную статистику исследуемого набора. Для упрощения записи закодируем названия атрибутов следующими порядковыми номерами, приведёнными в табл. 2.

Результаты обработки ЭД, полученные при помощи функции *describe*[12], сведены в табл. 3. Результатом работы функции *describe* является формирование описательной статистики по каждому атрибуту[13], включающей вычисление: среднего *(mean);* среднеквадратического отклонения *(standard deviation, STD, STDev);* минимального и максимального значения набора; перцентилей (по умолчанию: 25%, 50% и 75%); количества отсутствующих значений атрибутов; количества некорректных значений атрибутов (NaN).

Дополнительно формировался столбец с количеством уникальных значений атрибутов КС.

Из табл. 3 видна значительная флуктуация абсолютных величин атрибутов, что актуализирует необходимость их нормировки.

В ЭД не наблюдалось отсутствующих значений (все столбцы ЭД одинаковы по количеству элементов) и некорректных *(NaN)* значений метрического типа. Атрибуты категориального типа были исключены из исследуемых ЭД.

Предобработка входных ЭД

Процесс предобработки данных осуществлялся с помощью стандартных библиотек Python и в соответствии с логическими переменными L_{value2}, L_{value3}, L_{value4} включал перемешивание, трансформацию и удаление статичных значений.

Метки классов (состояния КС) кодировались под стандарты библиотеки *scikit-learn* в зависимости от поставленной задачи: БК, МклК или МзнК.

Результаты кодирования приведены в табл. 4. Отметим, что метки классов в многоклассовой и многозначной задаче были объединены методом трансформации задачи *Label Powerset* [18].

Графическое представление данных, приведенных в табл. 4, дано на рис. 1. Как видно из диаграммы на рис. 1.а, число состояний КС, ассоциированных с наличием аномальных состояний КС, составляет ~28% от общего числа записей. Аномальными считаются состояния, ассоциированные с нарушением как по одному, так и по нескольким вторичным атрибутам КС.

Объединив все аномальные состояния КС в один класс и сведя значения целевого столбца к двоичному множеству, получим диаграмму распределения экспериментальных данных по наличию/отсутствию аномалии (рис. 1.б).

По диаграмме видно, что большая часть аномальных состояний КС ассоциирована с состояниями «*packets_dropped*» и «*server_response_delay*». На эти состояния совокупно приходится 23% от всех записей в ЭД. Остальные аномальные состояния составляют в совокупности 5% от всех записей в ЭД, что иллюстрирует значительный дисбаланс классов [19, 20] (также см. работу авторского коллектива Haixiang G.14), что необходимо учитывать при обработке данных при классификации.

Оценка важности атрибутов ЭД

Под важностью атрибутов (*feature importance*) будем понимать совокупную информативность и статистическую значимость атрибутов ЭД [21 - 23].

Сортировка атрибутов ЭД по убыванию важности позволяет поочередно исключать атрибуты, наиболее сильно связанные с целевым столбцом и оказывающие значительное влияние на качество последующей классификации ЭД по данному целевому столбцу. Итерационное исключение наиболее важных атрибутов КС позволяет оценить поведение алгоритмов классификации в условиях возрастающей неопределенности.

Перед вычислительным экспериментом с классификацией, была оценена важность исходных ЭД для трех случаев предобработки данных:

- целевой столбец с бинарными состояниями КС (см. табл. 4, рис. 1, справа);
- целевой столбец с множеством состояний КС (см. табл. 4, рис. 1, слева);
- множество целевых столбцов, соответствующих многозначному случаю.

Оценка важности атрибутов проводилась по нескольким группам критериев: *f*-меры, вычисленная между метками класса и значениями атрибутов с помощью дисперсионного анализа (*ANalysis Of VAriance, ANOVA*) [24], взаимной информации [25], и критерия важности путем вычисления индекса Джини [26].

Для многозначной классификации существенным аспектом оценки важности атрибутов является множество целевых столбцов. Поскольку таких столбцов

12 Pandas.DataFrame.describe // Pandas URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html (дата обращения: 24.02.2023).

13 Bandaru S., Ng A.H.C., Deb K.Expert Data mining methods for knowledge discovery in multi-objective optimization: part A – Survey // Systems with Applications. 2017. Т. 70. С. 139-159. DOI: 10.1016/j.eswa.2016.10.015

14 Haixiang G., Yijing L., Mingyun G., Yuanyue H., Shang J., Bing G. Learning from class-imbalanced data: review of methods and applications // Expert Systems with Applications. 2017. Т. 73. С. 220-239. DOI: 10.1016/j.eswa.2016.12.035

Таблица 4

Распределение состояний КС в ЭД

Состояние КС, многоклассовое представление	Кол-во	Количество, относительное значение	Соответствие бинарной метке класса	Количество	Количество, относительное значение
normal	170931	0,718696	Аномалии нет	170931	0,718696
packets_dropped	29294	0,123169	Аномалия есть	66904	0,281304
server_response_delay	25705	0,108079			
server_response_delay packets_dropped	4674	0,019652			
disk_iowriteawait	4209	0,017697			
signal_delay	1239	0,005209			
packets_dropped disk_iowriteawait	727	0,003057			
signal_delay server_ response_delay	473	0,001989			
signal_delay packets_dropped	234	0,000984			
server_response_delay disk_iowriteawait	174	0,000732			
signal_delay server_ response_delay packets_dropped	121	0,000509			
server_response_delay packets_dropped disk_iowriteawait	54	0,000227			
Сумма	237835	1	**Сумма**	237835	1

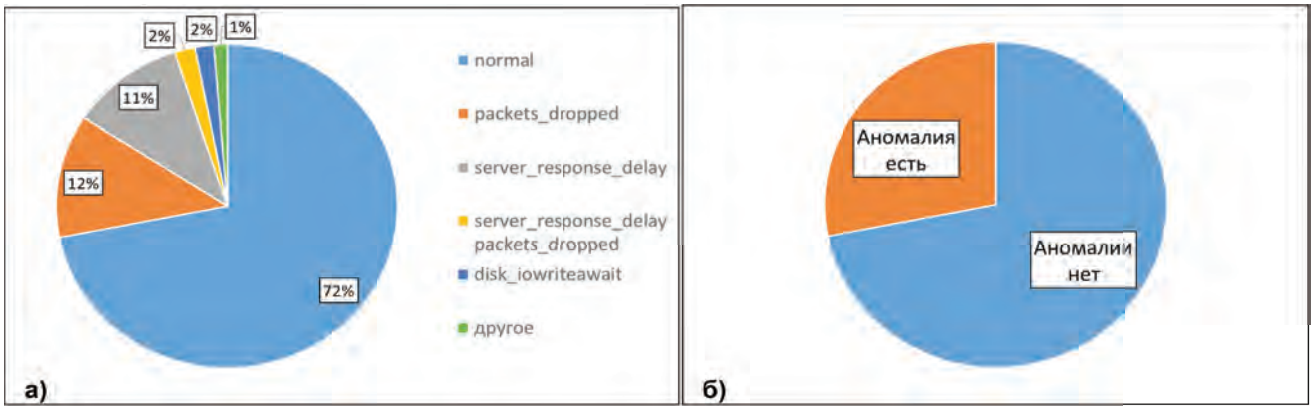

Рис.1. Распределение экспериментальных данных по количеству одновременно нарушаемых показателей уровня обслуживания: а) - по состояниям КС; б) - по наличию аномалии

в случае многозначной классификации несколько, то в каждом отдельном случае атрибуты оценивались по важности и ранжировались «по-своему».

Исследования показали, что в случае оценки атрибутной размерности по критерию определенного состояния КС (целевой столбец – наличие/отсутствие состояния КС), присвоенного по логическим правилам *SLO*, атрибуты, однозначно ассоциированные с присваиваемым меткам, маркировались как самые важные. При сравнении перечней ранжированных по важности атрибутов КС наблюдалось почти полное несовпадение их рангов.

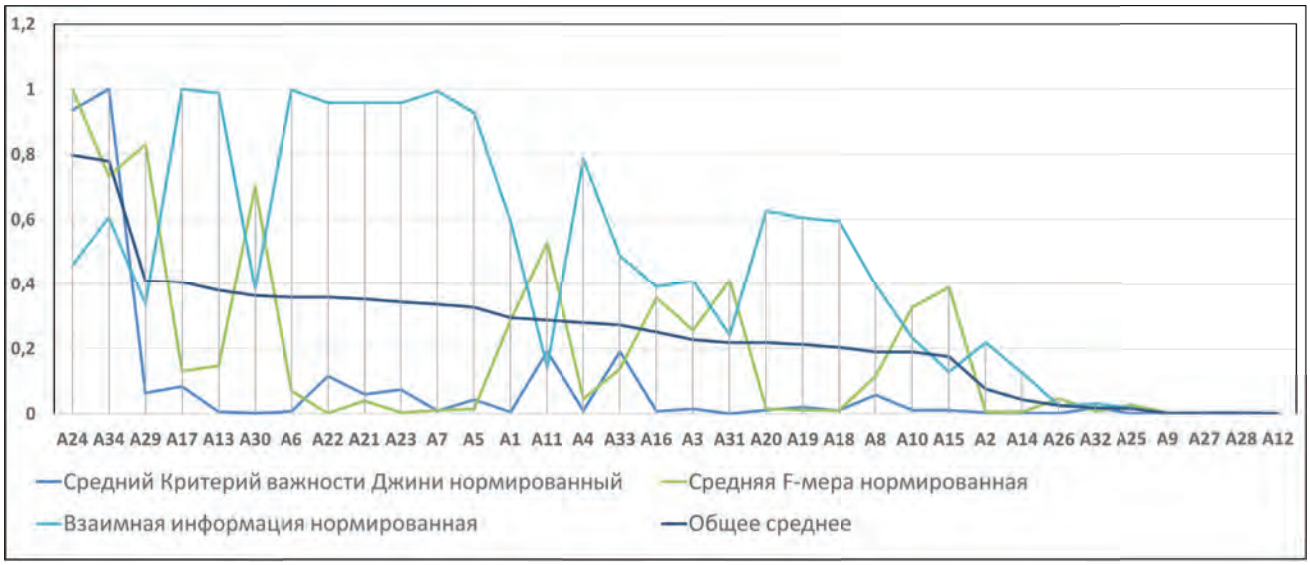

Рис.2. График распределения атрибутов КС по важности (многозначная классификации)

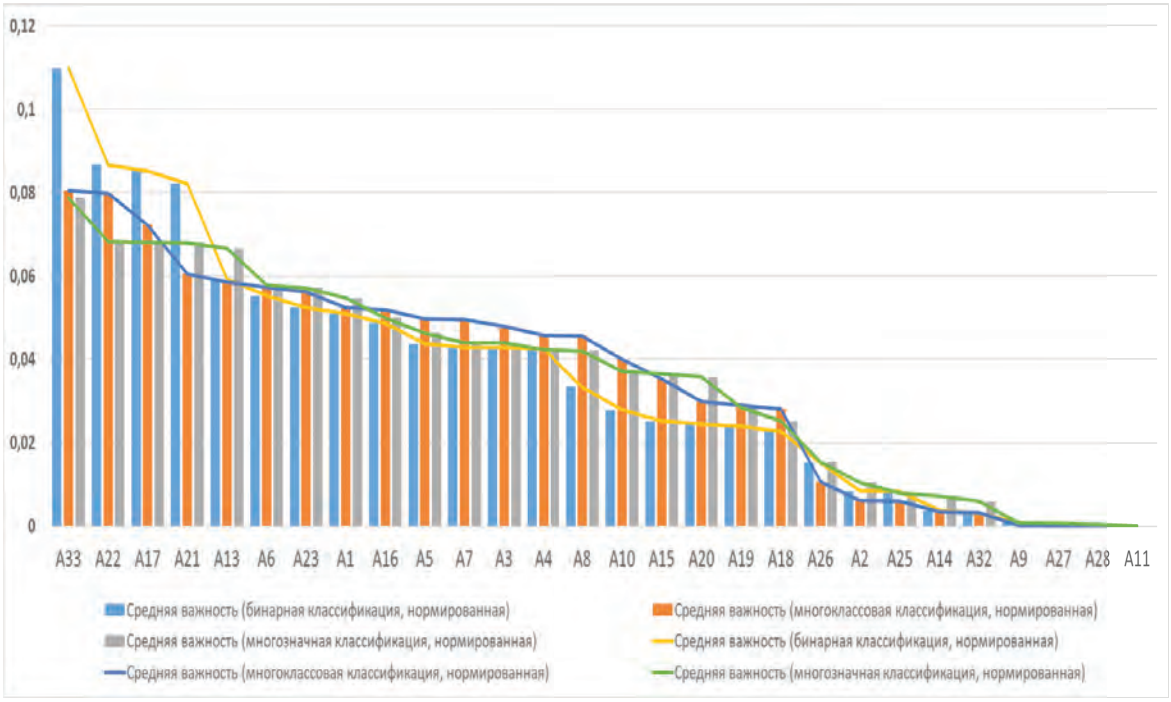

Рис.3. Нормированное распределение первичных атрибутов КС по критериям средней важности

Распределение атрибутов КС по важности приведено на рис. 2.

Из представленных зависимостей видно значительное влияние средней взаимной информации (информационной ценности) атрибутов на вид усредненной кривой. Средняя важность атрибутов имеет плавный убывающий, почти монотонный, характер по сравнению с каждой из групп метрик по отдельности.

Разделение атрибутов ЭД

Разделим атрибуты КС, в соответствии с *SLO*, на первичные и вторичные. Первичными атрибутами будем считать все атрибуты КС (см. табл. 2), за исключением множества L_{value5}. Атрибуты множества L_{value5} {*'ping_avg', 'server_response_timetotal', 'network_outdropped', 'disk_ioreadmergespersec'*} считаются вторичными и, поскольку они кодируются состояниями КС, в дальнейшем не рассматриваются в качестве атрибутов (параметров) классификации.

Таблица 5

Ранжирование первичных атрибутов в соответствии с исходными параметрами эксперимента

БК	МклК	МзнК	Номер итерации, после которого данный атрибут исключается
server_response_ timestarttransfer - 0,79	network_inpackets - 0,62	disk_writebytes - 0,56	1
disk_writebytes - 0,62	server_response_ timestarttransfer - 0,62	server_response_ timestarttransfer - 0,48	2
network_inbytes - 0,61	disk_writebytes - 0,56	network_inpackets - 0,48	3
network_inpackets - 0,59	disk_iostatrequestavgsize - 0,47	memory_free - 0,48	4
memory_actualfree - 0,43	cpu_iowait - 0,45	network_inbytes - 0,47	5
network_outbytes - 0,4	disk_iowriterequestspersec - 0,44	cpu_user - 0,41	6
disk_iostatrequestavgsize - 0,38	network_inbytes - 0,43	memory_actualfree - 0,41	7
cpu_user - 0,37	memory_actualfree - 0,4	disk_iostatrequestavgsize - 0,39	8
memory_free - 0,35	network_outbytes - 0,4	network_outbytes - 0,35	9
disk_iowriterequestspersec - 0,31	cpu_softirq - 0,38	cpu_system - 0,33	10
cpu_iowait - 0,31	disk_iowritemergespersec - 0,38	cpu_softirq - 0,31	11
cpu_system - 0,31	disk_busy - 0,37	memory_swapusedpct - 0,31	12
memory_swapusedpct - 0,3	memory_free - 0,35	cpu_iowait - 0,3	13
cpu_softirq - 0,24	cpu_user - 0,35	load_fifteenminutes - 0,3	14
disk_busy - 0,2	cpu_system - 0,31	load_fiveminutes - 0,26	15
load_oneminute - 0,18	memory_swapusedpct - 0,27	disk_iowriterequestspersec - 0,26	16
load_fiveminutes - 0,18	load_oneminute - 0,23	load_oneminute - 0,25	17
load_fifteenminutes - 0,17	load_fiveminutes - 0,22	disk_busy - 0,2	18
disk_iowritemergespersec - 0,16	load_fifteenminutes - 0,22	disk_iowritemergespersec - 0,18	19
dns_networkbytes - 0,11	cpu_nice - 0,08	cpu_nice - 0,11	20
cpu_nice - 0,06	dns_networkbytes - 0,05	dns_networkbytes - 0,07	21
dns_answerscount - 0,06	disk_iowriteawait - 0,05	disk_iowriteawait - 0,06	22
disk_iowriteawait - 0,03	dns_answerscount - 0,03	dns_answerscount - 0,05	23
server_response_ timenamelookup - 0,02	server_response_ timenamelookup - 0,02	server_response_ timenamelookup - 0,04	24
disk_await - 0,01	disk_await - 0	http_requestbytes - 0,01	25
http_responsebytes - 0	http_requestbytes - 0	disk_await - 0	26
http_requestbytes - 0	http_responsebytes - 0	http_responsebytes - 0	27
disk_ioreadrequestspersec - 0	disk_ioreadrequestspersec - 0	disk_ioreadrequestspersec - 0	28

Кроме того, из первичных атрибутов были исключены атрибуты, имеющие корреляцию с *ping_avg* - 'ping_max', 'ping_min', большую, чем 0,8.

Сравним среднюю важность оставшихся первичных атрибутов КС для анализируемых трех способов классификации. С этой целью рассмотрим гистограммы распределения каждого из атрибутов по трем критериям: средняя важность БК; средняя важность МклК, средняя важность МзнК. Гистограммы представлены на рис. 3. Для удобства визуализации дан-

ные каждого столбца нормировались и к каждому распределению добавлены аппроксимирующие линии.

Видно, что в случае БК важность первичных атрибутов имеет наиболее неравномерное распределение среди всех рассматриваемых алгоритмов классификации. Причиной неравномерности является концентрация основной массы весов в первых четырех, наиболее важных, атрибутах КС.

Видно, что средняя важность БК имеет наиболее высокие темпы убывания, в сравнении с двумя остальными классификаторами.

Сравнение МзнК с МклК демонстрирует значительную неопределенность в распределении важности вплоть до атрибута A18. Начиная с A18 более плавное снижение важности атрибутов демонстрирует МзнК.

Многозначная классификация выявляет больше информации (корреляционные связи более высокого порядка) по всему набору, включая наименее значимые атрибуты КС. Это, потенциально, приводит к более эффективной классификации по параметру *AUC* методом *One-vs-One* на *micro*-уровне.

Сравнительный анализ методов классификации

В результате вычислительного эксперимента в соответствии с логической переменной L_{value1} получено 4480 различных метрик. С целью обобщения выводов и возможности графического отображения, метрики усреднялись по каждому блоку перекрёстной проверки.

В табл. 5 отражена последовательность исключения атрибутов в эксперименте по мере получения оценочных метрик. В каждой строке по столбцам БК, МклК, МзнК представлены атрибуты КС, ранжированные по средней важности. В каждой ячейке записано метрическое значением данной величины.

На рисунках 4а...4г приведены зависимости величины $AUC_{ovo\ micro}$ от номера итерации, после которого данный атрибут исключается из рассмотрения для многозначной, многоклассовой и бинарной классификации ЭД.

Метрика $AUC_{ovo\ micro}$ (вычисленная методом *OVO* в соответствии с микро-подходом к нахождению бинарных метрик), выбрана для сравнительной оценки эффективности рассматриваемых алгоритмов классификации при сопоставлении всех возможных значений целевых столбцов.

Выбранная метрика отражает ошибки первого и второго родов, допущенные для каждого состояния КС. Метрика $AUC_{ovo\ micro}$ оказалась наиболее чувствительной в ситуации, когда некоторые состояния КС невозможно классифицировать.

На рисунках представлены зависимости AUCovo micro каждого из анализируемых классификаторов – МзнК, МклК и БК – от количества атрибутов для пяти алгоритмов классификации.

Из представленных зависимостей видно преимущество МзнК перед МклК и БК для всех алгоритмов классификации.

Наилучшие результаты наблюдаются для алгоритмов классификации ETC и RFC, у которых выигрыш МзкК, в сравнении с МклК, в среднем, составляет 15% при ETC и достигает 20% для *RFC*.

Выигрыш по метрике $AUC_{ovo\ micro}$ для МклК, в сравнении с ВК, составляет, в среднем, 20% при большом числе атрибутов (от 28 до 10) и снижается при малом числе атрибутов.

Алгоритмы DTC и KNC показывают несколько худшие результаты, хотя общая закономерность преимущества МзнК над МклК и БК сохраняется.

На рис. 5 представлены зависимости эффективности МзнК по параметру *ROC-AUC* для различных алг ритмов классификации и метрик: $ROC\text{-}AUC_{score\ ovo\ macro}$, $ROC\text{-}AUC_{score\ ovo\ weighted}$, $ROC\text{-}AUC_{score\ ovr\ macro}$, $ROC\text{-}AUC_{score\ ovr\ micro}$, $ROC\text{-}AUC_{score\ ovo\ micro}$, $ROC\text{-}AUC_{score\ ovr\ weighted}$.

Из представленных зависимостей видно, что метрика $AUC_{ovo\ micro}$ (*AUC*, вычисленная методом *OVO* в соответствии с микро-подходом к нахождению бинарных метрик) - показывает наилучшие результаты для алгоритмов *ETC* и *RFC*. Величина $AUC_{ovo\ micro}$ составляет в среднем 80%. Ансамблевый алгоритм *RFC* демонстрирует наилучшие результаты во всем диапазоне изменения исследуемой атрибутной размерности. Эффективность классификации по метрике $AUC_{ovo\ micro}$ алгоритмом ETC выше RFC на интервале 28 – 25 атрибутов.

По иным метрикам наибольшую эффективность демонстрирует алгоритм KNC.

Замечания

В рамках проведенного исследования гиперпараметры фиксировались, считались константами и не оптимизировались. Реализация этих позиций может привести к изменениям относительно ранговых позиций многоклассовых и многозначных классификаторов.

Исходные данные не перемешивались. Однако, как было показано в [27] перемешивание исходных данных может приводить к увеличению точности классификации более 30% на типе классификатора *Support Vector Data Description (SVDD).*

В версии разработанного фреймворка была реализована перекрестная проверка по нотации *K-Fold*

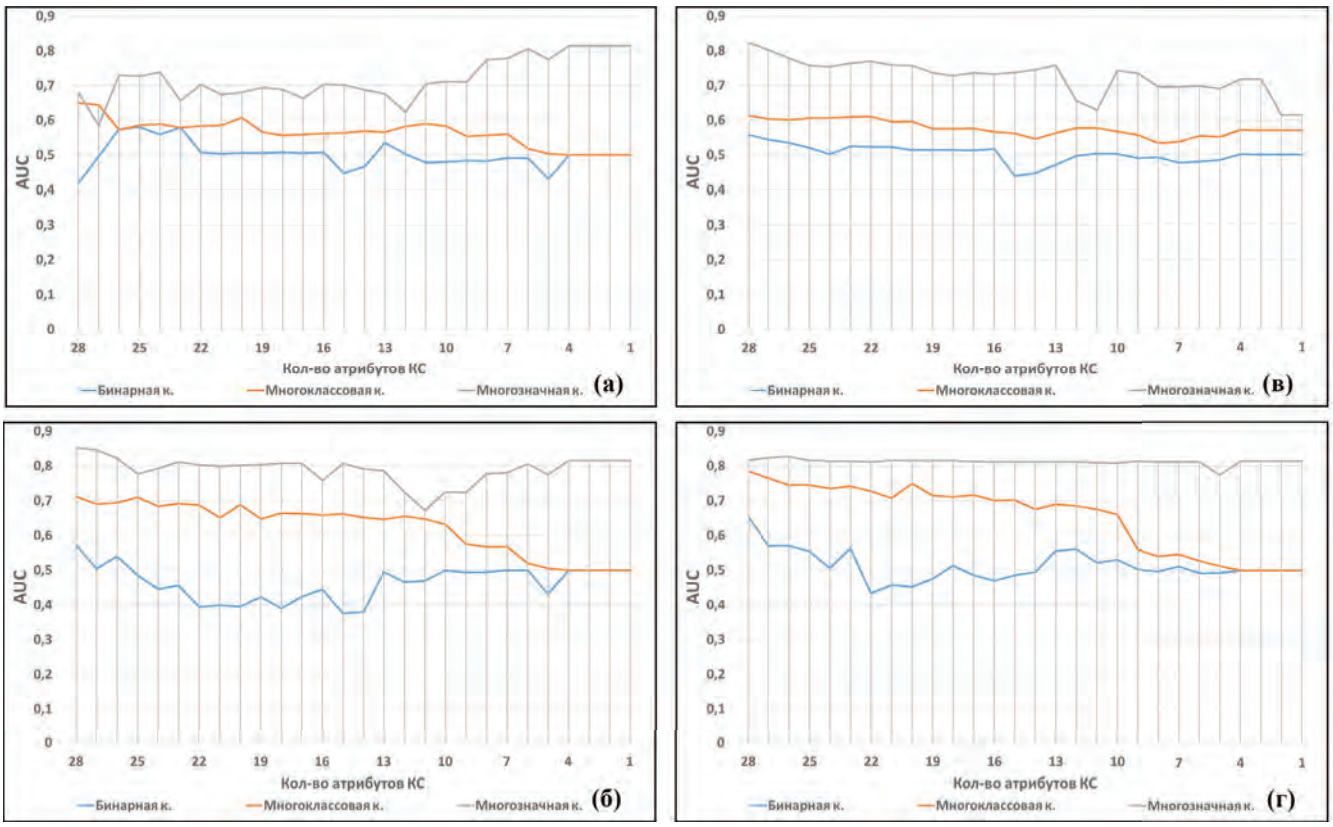

Рис.4. Зависимость характеристики $AUC_{ovo\ micro}$ классификаторов МзнК, МклК и БК от количества атрибутов для алгоритмов классификации: а) - DTC; б) - ETC; в) - KNC; г) – RFC.

с разделением только на обучающую и тестовую выборки (без валидационной выборки), что может снизить эффективность рассмотренных алгоритмов классификации (см. *Hastie T., Tibshirani R., Friedman J.* [15] и [28]).

Выводы

Проведенный сравнительный анализ эффективности классификаторов МзнК, МклК и БК по метрике $AUC_{ovo\ micro}$ от количества атрибутов показал, что выигрыш МзкК в сравнении с МклК в среднем составляет 15% при алгоритме *ETC* и достигает 20% для алгоритма RFC. Алгоритмы классификации *DTC* и *KNC* показывают несколько худшие результаты, хотя общая закономерность сохраняется.

Выигрыш по метрике $AUC_{ovo\ micro}$ классификатора МклК в сравнении с БК составляет в среднем 20% при большом числе атрибутов и снижается при уменьшении числа атрибутов в ЭД.

Исследование зависимости эффективности МзнК по параметру *ROC-AUC* от количества первичных атрибутов в ЭД показало, что метрика $AUC_{ovo\ micro}$ демонстрирует наилучшие результаты для алгоритмов *ETC* и *RFC* и составляет в среднем 80% даже при классификации в малом атрибутном пространстве,

Ансамблевый алгоритм *RFC* по качеству доминирует над иными классификаторами, в среднем, по всему диапазону исследуемой атрибутной размерности за исключением зоны высокой размерности атрибутного пространства (28 – 25 атрибутов). В указанной зоне наиболее эффективным по метрике $AUC_{ovo\ micro}$ является алгоритм *ETC*.

Наблюдаемые преимущества многозначного метода классификации перед многоклассовым и бинарным могут иметь комплексную природу, обусловленную рядом факторов таких как наличие или отсутствие оптимизации гиперпараметров, наличие или отсутствие перемешивания исходных ЭД, а также выбранным методом перекрестной проверки.

15 Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition. 2 изд. Stanford, California: Springer Series in Statistics (SSS), 2009. 764 с.

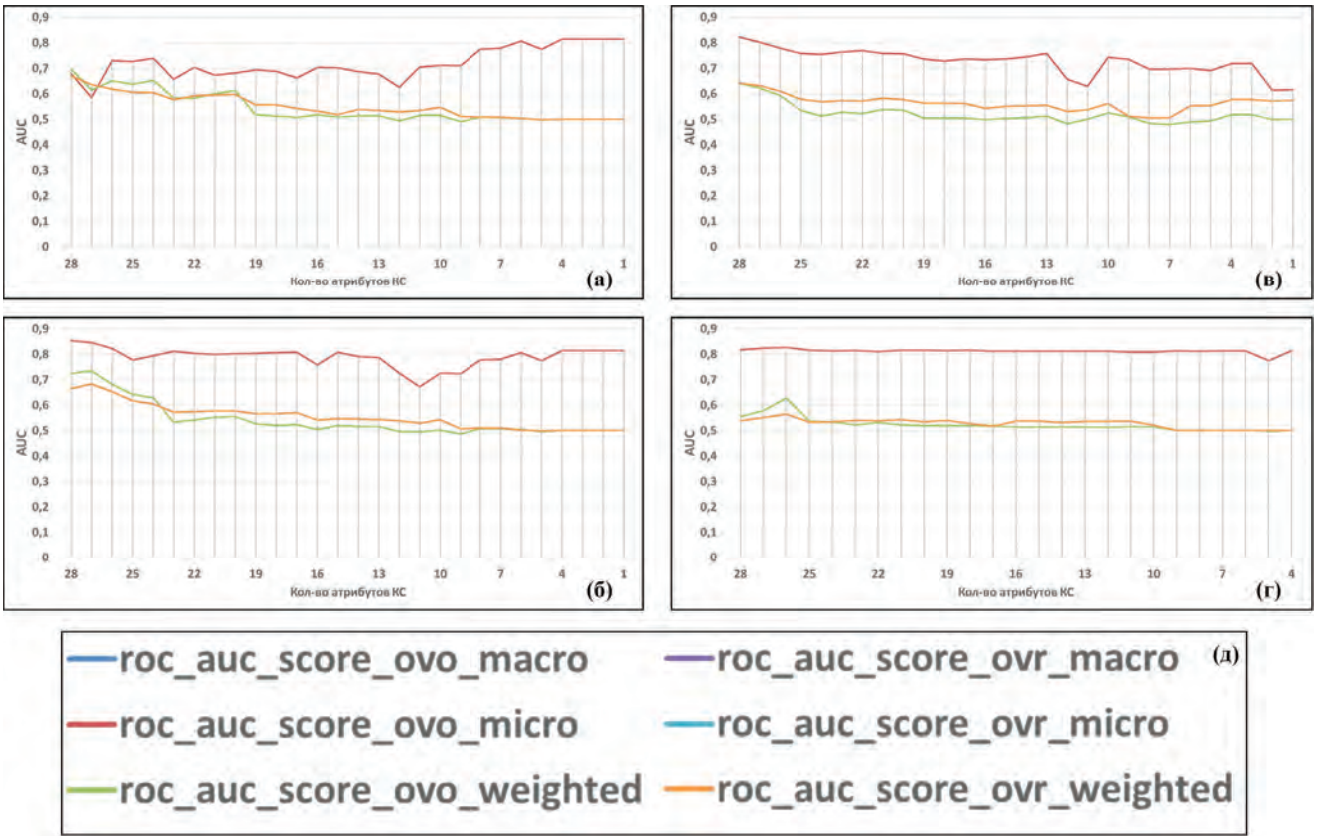

Рис.5. Зависимость эффективности МзнК от количества первичных атрибутов в ЭД для различных алгоритмов классификации по параметру ROC-AUC: а) – DTC; б) - ETC; в) - KNC; г) - RFC; д) – легенда с соответствиями метрик и соответствующего цвета.

Литература

1. Mirsky Y., Doitshman T., Elovici Y., Shabtai A. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection // В сборнике: Network and Distributed System Security Symposium. 2018. С. 1 – 16. DOI: 10.14722/ndss.2018.23211.
2. Hu J., Li Y., Xu G., Gao W. Dynamic subspace dual-graph regularized multi-label feature selection // Neurocomputing. 2022. Т. 467. С. 184-196. DOI: 10.1016/j.neucom.2021.10.022
3. Dong Q., Gong S., Zhu X. Imbalanced deep learning by minority class incremental rectification // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. Т. 41. № 6. С. 1367-1381. DOI: 10.1109/TPAMI.2018.2832629
4. Шелухин, О. И., Рыбаков С.Ю., Ванюшина А.В. Модификация алгоритма обнаружения сетевых атак методом фиксации скачков фрактальной размерности в режиме online // Труды учебных заведений связи. 2022. Т. 8. № 3. С. 117-126. DOI 10.31854/1813-324X-2022-8-3-117-126
5. Gurina A., Eliseev V. Anomaly-Based Method for Detecting Multiple Classes of Network Attacks // Information. 2019. Т. 84. №10. С. 1-24 DOI: 10.3390/info10030084.
6. Machoke M., Mbelwa J., Agbinya J., Sam A. Performance Comparison of Ensemble Learning and Supervised Algorithms in Classifying Multi-label Network Traffic Flow // Engineering, Technology & Applied Science Research. 2022. №. 12. С. 8667-8674. DOI: 10.48084/etasr.4852
7. Ducau F. N., Rudd E. M., Heppner T. M., Long A., Berlin K. Automatic Malware Description via Attribute Tagging and Similarity Embedding // arXiv preprint arXiv:1905.06262. 2019. С. 1 – 17. DOI: 10.48550/arXiv.1905.06262
8. Молодцов Д. А., Осин А. В. Новый метод применения многозначных закономерностей // Нечеткие системы и мягкие вычисления. №2. 2020. с. 83-95 DOI: 10.26456/fssc72
9. Sheluhin O. I., Kostin D. V., Polkovnikov M. V. Forecasting of Computer Network Anomalous States Based on Sequential Pattern Analysis of “Historical Data” // Automatic Control and Computer Sciences. 2021. № 6. С. 522–533. DOI: 10.3103/S0146411621060067
10. Шелухин О.И., Осин А.В., Костин Д.В. Мониторинг и диагностика аномальных состояний компьютерной сети на основе изучения “исторических данных” // T-Comm: Телекоммуникации и транспорт. 2020. №4. С. 23-30. DOI: 10.36724/2072-8735-2020-14-4-23-30
11. Шелухин О.И., Осин А.В., Костин Д.В. Диагностика “здоровья” компьютерной сети на основе секвенциального анализа последовательностных паттернов // T-Comm: Телекоммуникации и транспорт.
12. Шелухин О.И., Раковский Д.И. Выбор категориальных атрибутов редких аномальных событий компьютерной системы методами символьного анализа // В сборнике: Технологии Информационного Общества. Сборник трудов XV Международной отраслевой научно-технической конференции «Технологии информационного общества». 2021. С. 179-181
13. Шелухин О.И., Раковский Д.И. Прогнозирование профиля функционирования компьютерной системы на основе многозначных закономерностей // Вопросы кибербезопасности. 2022. № 6. С. 28-45. DOI:10.21681/2311-3456-2022-6-53-70

14. Шелухин О.И., Раковский Д.И. Выбор метрических атрибутов редких аномальных событий компьютерной системы методами интеллектуального анализа данных // T-Comm: Телекоммуникации и транспорт. 2021. Т. 15. № 6. С. 40-47. DOI: 10.36724/2072-8735-2021-15-6-40-47
15. Awad, W., El-Attar N. Adaptive SLA mechanism based on fuzzy system for dynamic cloud environment // International Journal of Computers and Applications. 2019. Т. 44. С. 1-11. DOI: 10.1080/1206212X.2019.1683956.
16. Kapassa, E., Touloupou, M., Kyriazis, D. SLAs in 5G: A complete framework facilitating VNF-and NS-tailored SLAs management // 5GTANGO - 5G Development and Validation Platform for global Industry-specific Network Services and Apps. AINA 2018. Krakow, Poland: 2018. С. 1-7. DOI:10.1109/WAINA.2018.00130.
17. Freeborn L., Andringa S., Lunansky G., Rispens J. Network analysis for modeling complex systems in SLA research // Studies in Second Language Acquisition. 2022. С. 1 – 33. DOI: 10.1017/S0272263122000407
18. Maltoudoglou L., Paisios A., Papadopoulos H., Lenc L., Martínek J., Král P. Well-calibrated confidence measures for multi-label text classification with a large number of labels // Pattern Recognition. 2022. Т. 122. С. 108271. DOI: 10.1016/j.patcog.2021.108271
19. Шелухин О. И., Раковский Д.И. Визуализация аномальных событий при прогнозировании состояний компьютерных систем на основе “исторических данных” // REDS: Телекоммуникационные устройства и системы. 2022. Т. 12. № 2. С. 53-58.
20. Раковский Д.И. Прогнозирование профиля функционирования компьютерной системы с применением аппарата точечно-множественных отображений // Сборник трудов II Всероссийской научно-практической конференции «Теория и Практика Обеспечения Информационной Безопасности», Москва, Россия. 2022. С. 222 – 231.
21. Zhong S., Zhang K., Yu X., Zhang H., Bagheri M., Burken J.G., Gu A., Li B., Wang T., Ma X., Marrone B.L., Ren Z.J., Zhu J.-J., Schrier J., Shi W., Tan H., Wang X., Wong B.M., Xiao X. Machine learning: new ideas and tools in environmental science and engineering // Environmental Science & Technology. 2021 Т. 55 № 19, С. 12741-12754 DOI: 10.1021/acs.est.1c01339
22. Saarela M., Kärkkäinen T. Can we automate expert-based journal rankings? analysis of the finnish publication indicator // Journal of Informetrics. 2020. Т. 14. № 2. С. 101008 DOI: 10.1016/j.joi.2020.101008
23. Wang K., Zhou L., Zhang D., Lim J., Liu Z.What is more important for touch dynamics based mobile user authentication? // В сборнике: Proceedings of the 24th Pacific Asia Conference on Information Systems: Information Systems (IS) for the Future, PACIS 2020. 24, Information Systems (IS) for the Future. Dubai, UAE. 2020.
24. Mohan V.M., Satyanarayana K.V.V. Multi-objective optimization of composing tasks from distributed workflows in cloud computing networks // Advances in Intelligent Systems and Computing. 2020. Т. 1090. С. 467-480. DOI: 10.1007/978-981-15-1480-7_39
25. Hasanin T., Khoshgoftaar T.M., Leevy J.L., Seliya N. Examining characteristics of predictive models with imbalanced big data // Journal of Big Data. 2019. Т. 6. № 1. С. 1 - 21. DOI: 10.1186/s40537-019-0231-2
26. Jung H., Jeon J., Choi D., Park A.J.-Y. Application of machine learning techniques in injection molding quality prediction: implications on sustainable manufacturing industry // Sustainability. 2021. Т. 13. № 8. С. 1 – 16. DOI: 10.3390/su13084120
27. Шелухин О.И., Раковский Д.И. Бинарная классификация многоатрибутных размеченных аномальных событий компьютерных систем с помощью алгоритма SVDD // Наукоемкие технологии в космических исследованиях Земли. 2021. Т. 13. № 2. С. 74-84. DOI: 10.36724/2409-5419-2021-13-2-74-84
28. Mohammadreza Q., Rohit B. Adversarial examples for extreme multilabel text classification // Machine Learning. 2022. Т. 111. № 1. С. 4539-4564. DOI:111. 10.1007/s10994-022-06263-z

MULTI-LABEL CLASSIFICATION OF SYSTEM LOGS OF COMPUTER NETWORKS. COMPARATIVE ANALYSIS OF CLASSIFIER EFFICIENCY

Sheluhin O.I.[16], Rakovskiy D.I.[17]

The aim of the study. *The aim of the study is to conduct a comparative analysis of binary (BC), multiclass (MCC) and multivalued (MLC) classification methods in information security problems. The boundaries of the study are the system logs formed by the computer network (CN).*

Method. *Classification algorithms were analysed: Decision Tree Classifier (DTC); Extra Trees Classifier (ETC); KNeighbors Classifier (KNC); Random Forest Classifier (RFC). The study was conducted on three metrics based on the Area Under the Receiver Operating Characteristic Curve (ROC-AUC): ROC-AUC$_{Micro}$, ROC-AUC$_{Macro}$, ROC-AUC$_{Weighted}$ using two methods One-vs-one, OVO or One-vs-everyone, OVE (in some sources - One-vs-rest - OVR). The experiment*

16 Oleg I. Sheluhin., Dr.Sc., Full Professor, Moscow Technical University of Communications and Informatics, Moscow, Russia. E-mail: sheluhin@mail.ru; ORCID: https://orcid.org/0000-0001-7564-6744

17 Dmitry I. Rakovskiy, Postgraduate student, Moscow Technical University of Communication and Informatics, Moscow, Russia. E-mail: Prophet_alpha@mail.ru. ORCID: https://orcid.org/0000-0001-7689-4678

implied an iterative assessment of the classification quality depending on the number of ED attributes. The ED attributes were ranked in descending order of their total informative value and statistical significance.

Results. *The analysis of binary, multiclass and multivalued implementations of the DTC, ETC, RNC, RFC algorithms in terms of the ROC-AUC parameter (metrics - ROC-AUC$_{score\ ovo\ macro}$, ROC-AUC$_{score\ ovo\ weighted}$, ROC-AUC$_{score\ ovr\ macro}$, ROC-AUC$_{score\ ovr\ micro}$, ROC-AUC$_{score\ ovo\ micro}$, ROC-AUC$_{score\ ovr\ weighted}$). The experiment was carried out for 28 different dimensions of the ED attribute space. The results of the study of the MLC, MCC and BC classifiers according to the AUC$_{ovo\ micro}$ showed that the gain of MLC in comparison with MCC is on average 15% for ETC and reaches 20% for RFC. The gain in the AUC$_{ovo\ micro}$ MCC metric compared to BC averages 20% with a large number of attributes and decreases with a decrease in the number of attributes in ED. Algorithms DTC and KNC show slightly worse results, although the general pattern remains. A study was made of the dependence of the MLC on the ROC-AUC parameter on the dimension of the primary attributes in the ED. It showed that the AUC$_{ovo\ micro}$ metric shows the best results for the ETC and RFC algorithms and averages 80% even when classifying in a small attribute space. The study showed that the use of multivalued classification can increase the classification Accuracy by up to 20% according to the AUC$_{ovo\ micro}$.*

Scientific novelty *lies in the study of the effectiveness of these classification methods in relation to ED KN by a set of output metrics. It is shown that the gain of MLC over other classification methods is up to 35% in total (MLC versus BC).*

Keywords: *data mining; abnormal condition; multi-label; binary classification; multiclass classification; feature importance; Decision Tree Classifier; Extra Trees Classifier; KNeighbors Classifier; Random Forest Classifier.*

References

1. Mirsky Y., Doitshman T., Elovici Y., Shabtai A. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection // V sbornike: Network and Distributed System Security Symposium. 2018. S. 1 – 16. DOI: 10.14722/ndss.2018.23211.
2. Hu J., Li Y., Xu G., Gao W. Dynamic subspace dual-graph regularized multi-label feature selection // Neurocomputing. 2022. T. 467. S. 184-196. DOI: 10.1016/j.neucom.2021.10.022
3. Dong Q., Gong S., Zhu X. Imbalanced deep learning by minority class incremental rectification // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019. T. 41. № 6. S. 1367-1381. DOI: 10.1109/TPAMI.2018.2832629
4. Sheluhin, O. I., Rybakov S.Ju., Vanjushina A.V. Modifikacija algoritma obnaruzhenija setevyh atak metodom fiksacii skachkov fraktal'noj razmernosti v rezhime online // Trudy uchebnyh zavedenij svjazi. 2022. T. 8. № 3. S. 117-126. DOI 10.31854/1813-324X-2022-8-3-117-126
5. Gurina A., Eliseev V. Anomaly-Based Method for Detecting Multiple Classes of Network Attacks // Information. 2019. T. 84. №10. C. 1-24 DOI: 10.3390/info10030084.
6. Machoke M., Mbelwa J., Agbinya J., Sam A. Performance Comparison of Ensemble Learning and Supervised Algorithms in Classifying Multi-label Network Traffic Flow // Engineering, Technology & Applied Science Research. 2022. №. 12. S. 8667-8674. DOI: 10.48084/etasr.4852
7. Ducau F. N., Rudd E. M., Heppner T. M., Long A., Berlin K. Automatic Malware Description via Attribute Tagging and Similarity Embedding // arXiv preprint arXiv:1905.06262. 2019. C. 1 – 17. DOI: 10.48550/arXiv.1905.06262
8. Molodcov D. A., Osin A. V. Novyj metod primenenija mnogoznachnyh zakonomernostej // Nechetkie sistemy i mjagkie vychislenija. №2. 2020. s. 83-95 DOI: 10.26456/fssc72
9. Sheluhin O. I., Kostin D. V., Polkovnikov M. V. Forecasting of Computer Network Anomalous States Based on Sequential Pattern Analysis of "Historical Data" // Automatic Control and Computer Sciences. 2021. № 6. C. 522–533. DOI: 10.3103/S0146411621060067
10. Sheluhin O.I., Osin A.V., Kostin D.V. Monitoring i diagnostika anomal'nyh sostojanij komp'juternoj seti na osnove izuchenija "istoricheskih dannyh" // T-Comm: Telekommunikacii i transport. 2020. №4. S. 23-30. DOI: 10.36724/2072-8735-2020-14-4-23-30
11. Sheluhin O.I., Osin A.V., Kostin D.V. Diagnostika "zdorov'ja" komp'juternoj seti na osnove sekvencial'nogo analiza posledovatel'nostnyh patternov // T-Comm: Telekommunikacii i transport.
12. Sheluhin O.I., Rakovskij D.I. Vybor kategorial'nyh atributov redkih anomal'nyh sobytij komp'juternoj sistemy metodami simvol'nogo analiza // V sbornike: Tehnologii Informacionnogo Obshhestva. Sbornik trudov XV Mezhdunarodnoj otraslevoj nauchno-tehnicheskoj konferencii «Tehnologii informacionnogo obshhestva». 2021. S. 179-181
13. Sheluhin O.I., Rakovskij D.I. Prognozirovanie profilja funkcionirovanija komp'juternoj sistemy na osnove mnogoznachnyh zakonomernostej // Voprosy kiberbezopasnosti. 2022. № 6. S. 28-45. DOI:10.21681/2311-3456-2022-6-53-70
14. Sheluhin O.I., Rakovskij D.I. Vybor metricheskih atributov redkih anomal'nyh sobytij komp'juternoj sistemy metodami intellektual'nogo analiza dannyh // T-Comm: Telekommunikacii i transport. 2021. T. 15. № 6. S. 40-47. DOI: 10.36724/2072-8735-2021-15-6-40-47
15. Awad, W., El-Attar N. Adaptive SLA mechanism based on fuzzy system for dynamic cloud environment // International Journal of Computers and Applications. 2019. T. 44. S. 1-11. DOI: 10.1080/1206212X.2019.1683956.
16. Kapassa, E., Touloupou, M., Kyriazis, D. SLAs in 5G: A complete framework facilitating VNF-and NS-tailored SLAs management // 5GTANGO - 5G Development and Validation Platform for global Industry-specific Network Services and Apps. AINA 2018. Krakow, Poland: 2018. S. 1-7. DOI:10.1109/WAINA.2018.00130.
17. Freeborn L., Andringa S., Lunansky G., Rispens J. Network analysis for modeling complex systems in SLA research // Studies in Second Language Acquisition. 2022. S. 1 – 33. DOI: 10.1017/S0272263122000407

18. Maltoudoglou L., Paisios A., Papadopoulos H., Lenc L., Martínek J., Král P. Well-calibrated confidence measures for multi-label text classification with a large number of labels // Pattern Recognition. 2022. T. 122. S. 108271. DOI: 10.1016/j.patcog.2021.108271
19. Sheluhin O. I., Rakovskij D.I. Vizualizacija anomal'nyh sobytij pri prognozirovanii sostojanij komp'juternyh sistem na osnove "istoricheskih dannyh" // REDS: Telekommunikacionnye ustrojstva i sistemy. 2022. T. 12. № 2. S. 53-58.
20. Rakovskij D.I. Prognozirovanie profilja funkcionirovanija komp'juternoj sistemy s primeneniem apparata tochechno-mnozhestvennyh otobrazhenij // Sbornik trudov II Vserossijskoj nauchno-prakticheskoj konferencii «Teorija i Praktika Obespechenija Informacionnoj Bezopasnosti», Moskva, Rossija. 2022. C. 222 – 231.
21. Zhong S., Zhang K., Yu X., Zhang H., Bagheri M., Burken J.G., Gu A., Li B., Wang T., Ma X., Marrone B.L., Ren Z.J., Zhu J.-J., Schrier J., Shi W., Tan H., Wang X., Wong B.M., Xiao X. Machine learning: new ideas and tools in environmental science and engineering // Environmental Science & Technology. 2021 T. 55 № 19, S. 12741-12754 DOI: 10.1021/acs.est.1c01339
22. Saarela M., Kärkkäinen T. Can we automate expert-based journal rankings? analysis of the finnish publication indicator // Journal of Informetrics. 2020. T. 14. № 2. S. 101008 DOI: 10.1016/j.joi.2020.101008
23. Wang K., Zhou L., Zhang D., Lim J., Liu Z.What is more important for touch dynamics based mobile user authentication? // V sbornike: Proceedings of the 24th Pacific Asia Conference on Information Systems: Information Systems (IS) for the Future, PACIS 2020. 24, Information Systems (IS) for the Future. Dubai, UAE. 2020.
24. Mohan V.M., Satyanarayana K.V.V. Multi-objective optimization of composing tasks from distributed workflows in cloud computing networks // Advances in Intelligent Systems and Computing. 2020. T. 1090. S. 467-480. DOI: 10.1007/978-981-15-1480-7_39
25. Hasanin T., Khoshgoftaar T.M., Leevy J.L., Seliya N. Examining characteristics of predictive models with imbalanced big data // Journal of Big Data. 2019. T. 6. № 1. S. 1 - 21. DOI: 10.1186/s40537-019-0231-2
26. Jung H., Jeon J., Choi D., Park A.J.-Y. Application of machine learning techniques in injection molding quality prediction: implications on sustainable manufacturing industry // Sustainability. 2021. T. 13. № 8. S. 1 – 16. DOI: 10.3390/su13084120
27. Sheluhin O.I., Rakovskij D.I. Binarnaja klassifikacija mnogoatributnyh razmechennyh anomal'nyh sobytij komp'juternyh sistem s pomoshh'ju algoritma SVDD // Naukoemkie tehnologii v kosmicheskih issledovanijah Zemli. 2021. T. 13. № 2. S. 74-84. DOI: 10.36724/2409-5419-2021-13-2-74-84
28. Mohammadreza Q., Rohit B. Adversarial examples for extreme multilabel text classification // Machine Learning. 2022. T. 111. № 1. S. 4539-4564. DOI:111. 10.1007/s10994-022-06263-z