

Choosing allowability boundaries for describing objects in subject areas

Musulmon Lolaev¹, Shavkat Madrakhimov², Kodirbek Makharov^{2,3}, Doniyor Saidov²

¹Center for Resilient and Evolving Intelligence, Kyungpook National University, Daegu, South Korea

²Department of Algorithms and programming technologies, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan

³Department of Applied Informatics, Kimyo International University in Tashkent, Tashkent, Uzbekistan

Article Info

Article history:

Received Jun 1, 2022

Revised Mar 31, 2023

Accepted Apr 12, 2023

Keywords:

Data cleaning

Dirty data

Invalid objects

Machine learning

Outliers

Preprocessing

Valid intervals

ABSTRACT

Anomaly detection is one of the most promising problems for study and can be used as independent units and preprocessing tools before solving any fundamental data mining problems. This article proposes a method for detecting specific errors with the involvement of experts from subject areas to fill knowledge. The proposed method about outliers hypothesizes that they locate closer to logical boundaries of intervals derived from pair features, and the interval ranges vary in different domains. We construct intervals leveraging pair feature values. While forming knowledge in a specific field, a domain specialist checks the logical allowability of objects based on the range of the intervals. If the objects are logical outliers, the specialist ignores or corrects them. We offer the general algorithm for the formation of the database based on the proposed method in the form of a pseudo-code, and we provide comparison results with existing methods.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Kodirbek Makharov

Department of Algorithms and programming technologies, National University of Uzbekistan named after Mirzo Ulugbek

Niyazova Street, Tashkent, Uzbekistan

Email: maxarov.qodirbek@gmail.com

1. INTRODUCTION

The problem of data preprocessing to detect and remove invalid data values to improve the efficiency of algorithms for solving problems becomes an urgent task due to the growth in the volume of processed data. In the works of [1], various approaches have been proposed for data cleaning based on metric functional dependencies and minimizing statistical distortions measured with the Earth Mover's Distance [2]. It should be noted that searching for logically incompatible data values in describing objects of subject areas for various sets of 2 or more features is a difficult task for implementation in practice. The problem lies in the need for more methods for checking the allowability of feature value relationships on such sets. Hypothetically, the answer to the question of the allowability of relationships can be obtained from competent experts in the subject areas. The process requires: i) development of special methods for analyzing and visualizing data to detect outliers; ii) constructing and filling of knowledge bases to control the belonging of data values to intervals with allowable boundaries.

The logical incompatibility of values for a pair of quantitative attributes within the framework of the subject area is considered a solution to the problem of finding the boundaries of the allowability of relations for this pair. Acceptance limits define intervals to control the correctness of the data used. Among the universal

restrictions on the use of intervals can be attributed invariance to the scale of measurements of features. The following example will show the incompatibility of the concepts of the allowability of values for each feature and a pair of features. A whale is the age of 5 and weighs 45,000 kilograms, two values are acceptable for their ranges separately, but when analyzed by pairs of these feature values, they are unacceptable or doubtful.

Building a regression relationship between features is one of the ways to describe the relationship between them. An example of using such a dependency is filling in missing values in data. The imperfection of this method can be judged by the decrease in the generalizing ability of recognition algorithms on samples, in which the missing values in data are randomly generated. Jouravlev proved the need for introducing additional restrictions on the concept of “allowable object” [3]. The introduced restrictions expand the possibilities for controlling the correctness of the data used. The proposed method for checking the correctness of data does not in any way override the existing ones. It is recommended to use it in cases where the “classical methods” could not detect errors. The use of this method in practice is preceded by the creation of a knowledge base, with the help of which the user is able to analyze the reasons for the incorrectness of the data. The relevance of research on this issue is increasing in connection with the use of cloud storage technologies for the processing of big data.

The efficiency of using the boundaries of permissible values when forming a training data set can be checked through the indicators of the generalizing ability of the algorithms. For comparison “clean” samples can be used, and samples with additional falsifier objects, in the description of which there are violations of the allowability boundaries. For example, in classification problems, noise objects are unallowable relative to their class and negatively affect the accuracy of the solution algorithm. This factor can be used to test the stability of algorithms, in particular, in [4], the “data + noise” technique is proposed, the use of which, on the one hand, contributes to a “smoother” convergence of the procedure for interactive search for logical patterns. On the other hand, “noisy” objects perform an essential function of falsifiers, the “collision” contributing to an increase in the robustness of the solutions obtained.

Detecting anomalies of objects that differ from the main part of the data in the subject areas during data mining [5], there are several approaches to detecting outliers depending on the type of task: parsing, data transformation, methods of enforcing integrity constraints, duplicate detection methods, and others [6]. Almost fully similar works were overviewed by [7] in 2018. According to that paper, there are some categories of outlier methods: global vs. local; labeling vs. scoring; supervised vs. unsupervised; parametric vs. non-parametric methods. While, over the years there are many types of outlier detection techniques have been proposed for various purposes [7], [8]: statistical - checking the input value in the interval, which is formed from the standard deviation based on the Chebyshev inequality [9]; deviation-based; density-based techniques based on a distance between objects; cluster-based such as “density-based spatial clustering of applications with noise” [10], [11]; association templates of rules - validate the input against some template, the representation of which is considered correct.

2. RELATED WORKS

One of the fastest ways to detect abnormal objects is the Isolation Forest method [12]. Most current methods for detecting anomalous objects first create a normal object model, and then it is required to check all objects of the sample for abnormal. The process creates significant computational complexity. The classical method of Isolation Forest is thoroughly analyzed and augmented by bringing an innovative approach [13]. This approach is a k-Means-based Isolation Forest that allows building a search tree based on many branches in contrast to the only two considered in the original method. A set of methods enhancing the Isolation Forest based on Fuzzy C-Means was proposed [14]. The main goal of the study is to analyze the possibilities of using the grouping method using Fuzzy C-Means at the stage of building a search tree [15]. In particular, it is utilized the information on the degree of membership of a given object to the group of similar objects positioned close to a given tree node. The memberships are determined based on distance from the so-called middle of the cluster, i.e., the average value of the feature [16].

The idea behind the algorithm density-based spatial clustering of applications with noise (DBSCAN) is that there is a higher density of objects inside each cluster than the density outside the cluster [10]. The selection of noise objects is made on the assumption that the density in these regions is lower than in any of the clusters. Moreover, for each point of the cluster, its neighborhood of a given radius ϵ (eps) must contain at least a certain number of points, which is set by the threshold value MinPts. A new concept of moveability and

constructed new, comprehensive, hybrid feature-based density measurement method was defined that considers temporal and spatial properties [17]. After, proposed the improved DBSCAN algorithm using the new density measurement method. Another density-based clustering algorithm has been presented based on DBSCAN, and computational geometry [18]. It represents three significant modifications or extensions to DBSCAN: selection of parameter ϵ (eps) using the radii of empty or Voronoi circles; selection of parameter MinPts for the same epsilon; redistribution of noise points to suitable clusters using the concept of centroid hinged clustering.

The main drawbacks of most existing approaches to anomaly detection are summarized [19], [20]. These approaches are not optimized for detecting anomalies, a consequence, these approaches are often not efficient enough, which leads to too many false alarms (when normal instances are identified as anomalies) or too many anomalies; many existing methods are limited to low-dimensional data and small data size due to their legacy algorithms. Based on this overview in the previous paragraphs, we compare our method with these methods regarding our statement of the task, and the main contributions of the paper in continuity [20] are:

- We construct a latent feature by a simple-linear combination of two features in order to conduct logical analysis based on two features;
- We propose an expert-based model which leverages the latent feature to solve the problem of logical inconsistencies;
- We suggest how to utilize this model providing a general example, and we also compare the results with results of existing methods, even though many of them are suitable for another task. For example, the DBSCAN is a clustering algorithm, however, it can separate the noisy clusters too [21].

3. METHOD

A method for detecting logically invalid values by pairs of quantitative features in the description of sample objects in the studied subject area is proposed and is focused on detecting errors in the input data [3]. Dividing the feature values by medians or means of the current features provides a scale-invariant property for the method. Since the median is a random variable, its values for a real sample of data can be considered and interpreted according to the law of large numbers [22]. As the sample size grows, the mathematical expectation of the median tends to a stable value, i.e., is an unbiased estimate.

Let's denote object $E_0 = \{S_1, \dots, S_m\}$ is given, described by a set of features $X = (x_1, \dots, x_n)$ with different types, and I, J a set of indices of features, respectively, quantitative and nominal features. For each pair of features $(x_i, x_j) \subset X(n), i \neq j, i, j \in I$, we calculate the latent feature for feature pairs in (1):

$$R(i, j) = \left\{ \frac{a_{ki}}{P_i} - \frac{a_{kj}}{P_j} \right\}_{k \in \{1, \dots, m\}}, \quad (1)$$

where $S_k = (a_{k1}, \dots, a_{kn})$, $S_k \in E_0$, P_i, P_j - are the values of the medians of the features x_i, x_j on the set E_0 , and the boundaries of the intervals z_1, z_2 calculated (2)

$$z_1 = \min_{E_0} R(i, j), z_2 = \max_{E_0} R(i, j). \quad (2)$$

Checking the allowability of nominal features is determined through the membership of their gradations in a finite set of values. The use of dimensionless quantities for calculating the boundaries of feature ratios according to (2) and data visualization allows an expert to interactively determine the correctness of object descriptions and enter information into the knowledge base of the subject area. To inform the user about the interval $[z_1; z_2]$ recommended to split it into a given number (for example, 10 or 100) parts. The choice of the number of chunks depends on the size of the original dataset or the recommendations for conducting exploratory data analysis. The suspicion of an anomalous object is determined when its values belong to the extreme (right or left) parts of the interval $[z_1; z_2]$. The method is implemented in 3 stages:

- Forming the intervals $\{[z_1; z_2]\}$ set

$$\Pi = \bigcup_{i, j \in I, i \neq j} \left\{ S_u, S_v \mid \frac{a_{ui}}{P_i} \frac{a_{uj}}{P_j} = z_1 \text{ and } \frac{a_{vi}}{P_i} \frac{a_{vj}}{P_j} = z_2 \right\}; \quad (3)$$

- $\forall S_u \in \Pi$ and $\frac{a_{ui}}{P_i} \frac{a_{uj}}{P_j} \in \{[z_1, z_2]\}$ making an expert decision on its allowability (inallowability);

- Removing invalid objects from E_0 and forming a set of intervals $Z = \bigcup_{i,j \in I, i \neq j} [z_1; z_2]$ according to (2) and (3).

Pseudocode for this algorithm is given as a following:

```

while True do
    generate the set of objects that are on the boundaries of the intervals
    for all object in set do
        make an expert decision on its allowability or inallowability
        if object's inallowability is true then
            remove this object from dataset  $E_0$ 
        end if
    end for
    if any object is deleted then
        forming a set of intervals from  $E_0$ 
    else
        break while
    end if
end while

```

DBSCAN algorithm scans whole dataset only one time and needs to calculate the distance of any pair of objects in the dataset. Time complexity of original DBSCAN algorithm is high - $O(n^2)$. Using efficient indexing structures complexity can be reduced to $O(n \log n)$ [23]. iForest has time complexities of $O(t\psi \log \psi)$ in the training stage and $O(nt \log \psi)$ in the evaluating stage, where t - the number of trees, ψ - the number of samples, n - testing data size [12]. Proposed method has time complexity of $O(n \cdot \frac{m(m-1)}{2})$ in the evaluating stage, where m - number of features, n - testing data size. The training stage complexity is independent of specialist decisions. In one iteration of the training stage, it will be equal to $O(\frac{m(m-1)}{2})$.

4. RESULTS AND DISCUSSION

We conduct our experiments on a dataset titled “Kalahari Kung San people” to explore Kalahari Kung San people collected by *Nancy Howell*, and publicly available [24], [25]. It contains 545 people's information, and each person is described by four features: height, weight, age, and sex, but we only consume numerical features. Before exploring, we omit logically incorrect and missing rows, therefore the modified version of datasets may also contain some anomaly objects too, but we assume it does not have any logically incorrect instances during our simulations. Table 1 depicts the results of calculating the values of the boundaries of the intervals for each pair and objects that were on the boundaries of the interval for pairs of features.

Table 1. The boundaries of a latent feature formed from pair features

#	Pair-feature names	Interval ranges	The boundary objects	
			left	right
1	(Height, Weight)	[-0.58, 0.46]	5 (163.83, 62.99)	359 (109.22, 11.71)
2	(Height, Age)	[-1.90, 0.72]	164 (140.97, 85.60)	359 (109.22, 2.00)
3	(Weight, Age)	[-1.79, 0.99]	222 (35.81, 82.00)	182 (61.80, 22.00)

Table 2 represents examples of issuing the results of checking for the presence of invalid objects with an indication of the error status: “Critical error” - the value of the attribute according to (1) lies outside the interval; “Possibly a mistake” - the values of the feature according to (1) lie in the “dangerous” proximity to the boundaries of the interval, determined by the entry into $[z_1; z_1 + h]$ or $[z_2 - h; z_2]$, where h - is the step of dividing the interval into a given number of parts. In this experiment, $h = 1\%$ for testing purposes. Moreover, we omit one object in each iteration and test the skipped object to be an anomaly during this experiment. Because forgetting the one-test object, the intervals vary in each row in Table 2.

Figure 1 illustrates an example of objects that are close to the critical ranges regarding their location. The two lines indicate the upper and lower borders of the interval based on pair feature values, and we calculate them using $3h$, $h = (z_2 - z_1)/100$, where $3h$ - the gap between two lines and the ranges. The objects outside of the two lines are considered outliers. To compare methods for finding anomalous objects, we artificially create 20 anomalous objects to the original data [24]. We experiment with two forms of data: we use all data in the

first form, we fit the method using original data, and we use the trained model to identify anomalous objects in the second form. Table 3 illustrates the test results. Since the proposed method works with a pair of features, in the testing process, if an object is anomalous regarding any pair of features, then the object is considered anomalous. Experimenting with separated data on the DBSCAN method is not supported using Scikit-learn library. The proposed and Isolation Forest methods get 100% in accuracy. We leverage precision and F_1 score to compare the results to make the comparison more precise.

Table 2. Examples of defining objects with potentially invalid values

Objects no.	Feature's name	Feature's value	Interval	Value of features by pairs	Error status
6	Height	163.83	-0.41, 0.44	-0.47	Critical error
	Weight	62.99			
20	Height	105.41	-0.47, 0.44	0.36	Possibly a mistake
	Weight	13.95			
57	Height	165.74	-0.47, 0.44	-0.35	Possibly a mistake
	Weight	58.60			
90	Height	136.53	-2.23, 0.66	-2.01	Possibly a mistake
	Age	79.00			
114	Height	124.46	-0.47, 0.44	0.38	Possibly a mistake
	Weight	18.26			
165	Weight	40.94	-2.15, 0.73	-2.15	Critical error
	Age	85.60			

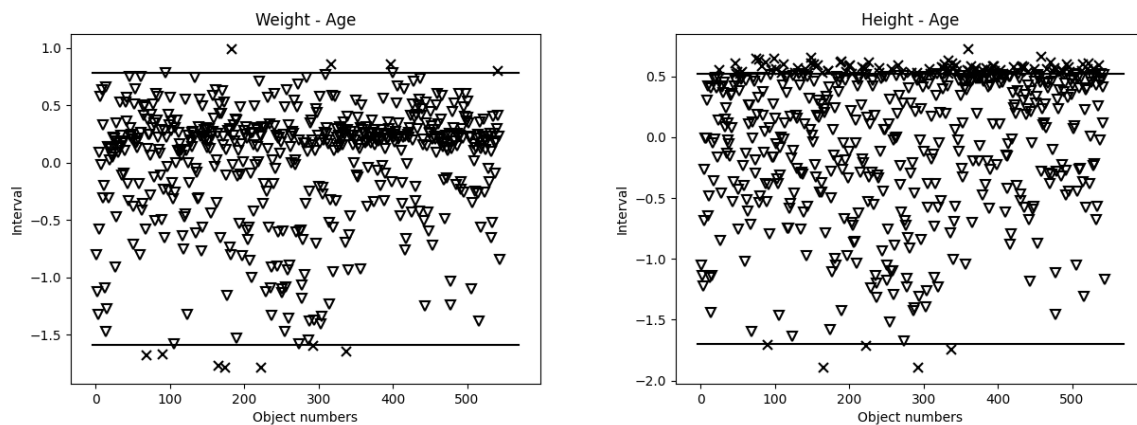


Figure 1. Visual representation of the process of identifying invalid objects

Table 3. The comparison results of identifying anomalous objects by methods

Test number	Isolation Forest		DBSCAN		Proposed	
	F_1 score	Precision	F_1 score	Precision	F_1 score	Precision
Test 1	0.70	0.70	0.92	0.90	0.85	1.00
Test 2	0.80	0.80	0.90	0.86	0.91	1.00
Test 3	0.75	0.75	0.90	0.86	0.88	1.00
Test 4	0.70	0.70	0.90	0.86	0.91	1.00
Test 5	0.70	0.70	0.92	0.92	0.72	0.62
Test 6	0.65	0.65	0.90	0.90	0.88	1.00
Test 7	0.60	0.60	0.90	0.90	0.91	1.00
Test 8	0.70	0.70	0.88	0.82	0.68	0.59
Test 9	0.75	0.75	0.88	0.82	0.91	1.00
Test 10	0.70	0.70	0.85	0.81	0.68	0.59
Expected	0.70	0.70	0.89	0.86	0.83	0.88

According to Table 3, the DBSCAN's results overperform the proposed method in most tests concerning F_1 scores, but the proposed method overperforms in many test cases concerning precision scores. Actually, our task is to find anomaly objects in the dataset, and therefore by the definition of the precision score (also called true positive value), we can conclude that the proposed method can find anomaly objects more than DBSCAN. Moreover, we provide another example, like this experiment, in contrast to the previous one. We

only use two features to find anomaly objects in DBSCAN and the proposed methods. The results of this experiment ended up with the same average of accuracies 0.9390, 1.00, and 0.8864 of scores F_1 , precision, and recall, respectively, in each test, but we removed one feature “age” and some rows from anomaly datasets to be correct with our setup. During this experiment, we made a trade-off with the values of h individually for each test in order to get high performance in the proposed method. That is the reason why the results are the same. In contrast, we did not make a like trade-off technique for the previous experiment due to our main reason being to illustrate a simple method.

4.1. Discussion

The construction of logical intervals for a pair of features in problems related to the solution of practical problems is considered separately for each subject area. According to the example in the Introduction, it is impossible to be a person with a 1-meter height of 250-kilogram weight. However, in other subject areas, for example, animals, a 1-meter tall animal can weigh 250 kilograms, even more. For this reason, we added the phrase “subject area” to the title. Also, we included the term “describing objects” in the title because the same objects can be used to build different boundaries, for example, when objects are grouped.

The proposed model allows the construction intervals for a pair of features with the help of experts. This allows determining an allowable range of values by pair of features. Therefore, the proposed method can be used in the process of filling the knowledge bases of subject areas. The method is nonparametric and has linear computational complexity, which makes it possible to apply it to problems with big data. The main limitation of the proposed work is the need for an expert to build knowledge, and this is a common problem that is considered a limitation of artificial intelligence too in other types of machine learning, such as labeling data in the classification task.

In contrast, other approaches mainly focused on finding anomaly objects in data. For example, DBSCAN is a clustering algorithm based on density level estimation with many modifications. One of the most significant drawbacks is using distance which may cause the problem of the curse of dimensionality. However, this limitation doesn’t affect the results in this paper because we used only three features.

During experiments, we have done several trade-offs on parameters of both DBSCAN and proposed methods to get acceptable accuracies. However, we only cared a little about the Isolation forest method because the nature of this algorithm requires more objects and features in datasets to get high outcomes, and this method may not match our setups. In practice, the field of experts is required to build the ranges which identify anomaly objects using the proposed model, so if the ranges on latent features are built properly, the results of the proposed method can be more accurate than the results. For the reason that the accuracies of the proposed method in experiments did not overperform in all cases. Moreover, we leveraged DBSCAN method in scikit-learn library [26] in our experiments with parameters $\epsilon = 0.1$ and $min_samples = 5$.

5. CONCLUSION AND FUTURE WORK

A method of searching for logically incompatible quantitative data in describing objects of subject areas for various sets of two features has been proposed. Using the values of the medians of the analyzed features gives the method the property of invariance to the scale of measurements of features, which expands the range of application of the method. The general use of the method has been illustrated to form a knowledge base about the incompatibility of values based on intervals of paired features from the data of the subject area. The quantitative features space has been considered in this work. Future research will be aimed to adapt the method for different types of feature space, reduce the specialist’s decision-making role, and correct the detected logical errors.

ACKNOWLEDGEMENTS

This work is supported by BK21 Four Project, AI-Driven Convergence Software Education Research Program 4199990214394 2, and also supported by the National Research Foundation of Korea 2020R1A2C1012196.

REFERENCES

- [1] N. Prokoshyna, J. Szlichta, F. Chiang, R. J. Miller, and D. Srivastava, “Combining quantitative and logical data cleaning,” *Proceedings of the VLDB Endowment*, vol. 9, no. 4, pp. 300–311, 2015, doi: 10.14778/2856318.2856325.




- [2] E. Levina, and P. Bickel, "The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics," *Proceedings of ICCV 2001*, vol. 2, pp. 251–256, 2001, doi: 10.1109/ICCV.2001.937632.
- [3] Yu.I. Zhuravlev, "On the algebraic approach to solving recognition and classification problems," *Problems of Cybernetics*, vol. 33, pp. 5–68, 1978.
- [4] V. A. Duke, "Methodology of searching for logical patterns in a subject area with fuzzy systemology (on the example of clinical and experimental research)," D.Sc. dissertation, St Petersburg University, Saint Petersburg, Russia, 2005.
- [5] A. Zimek and E. Schubert, "Outlier Detection," in *Encyclopedia of Database Systems*, New York, NY, USA: Springer, 2017.
- [6] D. Vora and S. Porwal, "A comparative analysis of data cleaning approaches to dirty data," *International Journal of Computer Applications*, vol. 62, no. 17, pp. 30–34, 2013, doi: 10.5120/10175-5041.
- [7] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 2:e1280, 2018, doi: 10.1002/widm.1280.
- [8] J. Maletic and A. Marcus, "Data Cleansing," in *Data mining and knowledge discovery handbook*, New York, NY, USA: Springer Science+Business Media LLC, 2005, pp. 19–33.
- [9] B. G. Amidan, T. A. Ferryman, and S. K. Cooley, "Data Outlier Detection using the Chebyshev Theorem," *Aerospace Conference IEEE*, United States, pp. 3814–3819, 2005, doi: 10.1109/AERO.2005.1559688.
- [10] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowledge Discovery and Data Mining*, pp. 226–231, 1996.
- [11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, doi: 10.48550/arXiv.1201.0490.
- [12] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation Forest," *Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008, doi: 10.1109/ICDM.2008.17.
- [13] P. Karczmarek, A. Kiersztyn, W. Pedrycz, and E. Al, "K-Means-based isolation forest," *Knowledge-Based Systems*, vol. 195, p. 105659, 2020, doi: 10.1016/j.knosys.2020.105659.
- [14] P. Karczmarek, A. Kiersztyn, W. Pedrycz, and D. Czerwinski, "Fuzzy C-Means-based Isolation Forest," *Applied Soft Computing Journal*, vol. 106, p. 107354, 2021, doi: 10.1016/j.asoc.2021.107354.
- [15] C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, pp. 191–203, 1984, doi: 10.1016/0098-3004(84)90020-7.
- [16] P. Karczmarek, A. Kiersztyn, and W. Pedrycz, "Fuzzy set-based isolation forest," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, 2020, doi: 10.1109/FUZZ48607.2020.9177718.
- [17] T. Luo, X. Zheng, G. Xu, K. Fu, and W. Ren, "An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories," *International Journal of Geo-Information*, vol. 6, no. 3, 2017, doi: 10.3390/ijgi6030063.
- [18] K. Giri, T. Biswas, and P. Sarkar, "ECR-DBSCAN: An improved DBSCAN based on computational geometry," *Machine Learning with Applications*, vol. 6, p. 100148, 2021, doi: 10.1016/j.mlwa.2021.100148.
- [19] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-Based Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, pp. 1–39, 2012, doi: 10.1145/2133360.2133363.
- [20] S. F. Madrahimov, K. T. Makharov, and M. Y. Lolaev, "Data preprocessing on input," *AIP Conference Proceedings* 2365, doi: 10.1063/5.0058132.
- [21] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1–21, 2017, doi: 10.1145/3068335.
- [22] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics*, London, UK: Springer, 2005.
- [23] Nidhi and K. A. Patel, "An Efficient and Scalable Density-based Clustering Algorithm for Normalize Data," *Procedia Computer Science*, vol. 92, pp. 136–141, 2016, doi: 10.1016/j.procs.2016.07.336.
- [24] N. Howell, "Demographic behavior of hunter-gatherers: evidence for density-dependent population control," *Demographic behavior: interdisciplinary perspectives on decision-making*, pp. 185–200, 1980.
- [25] N. Howell, "Body Size and Growth," in *Life histories of the Dobe !Kung: Food, fatness, and well-being over the life-span*, California, USA: University of California Press, pp. 49–82, 2010, [Online]. Available: <https://tspace.library.utoronto.ca/handle/1807/10395>.
- [26] L. Buitinck et al., "API design for machine learning software: experiences from the scikit-learn project," *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013, doi: 10.48550/arXiv.1309.0238.

BIOGRAPHIES OF AUTHORS






Musulmon Lolaev holds a master of science degree from National University of Uzbekistan in 2019 with dissertation "Analytical description of the own features space of objects for modeling decision-making processes" and he finished bachelor degree in "Applied math and informatics" in 2017. He has published about 15 scientific papers in international and national journals and conferences. He is currently a Computer Science Ph.D. student at Center for Resilient and Evolving Intelligence of Kyungpook National University, Daegu, South Korea, and his research interests include artificial intelligence, data mining, computer vision, pattern recognition, numerical modeling and software engineering. He can be contacted at email: musulmon.lolayev.94@gmail.com or musulmon@knu.ac.kr.






Shavkat Madрахimov    received the D.Sc. degree in computer science from the Tashkent University of Information Technologies, Tashkent, Uzbekistan, with the dissertation “Detection systems of hidden regularity on the basis of the methods of calculation of generalized estimates”. He is a Professor of the faculty Applied mathematics and intellectual technologies, National University of Uzbekistan, Tashkent, Uzbekistan, since 1989. In addition, he is Head of the department “Artificial intelligence”. His research interests are in data mining, knowledge discovery, pattern recognition, and software engineering. He has published more than 70 scientific papers in journals and conferences. He can be contacted at email: sh.madрахimov@nuu.uz.



Kodirbek Makharov    holds a master of science degree from National University of Uzbekistan in 2013 with dissertation “Determining the regularities by the dominance intervals of the features of objects” and he finished bachelor’s degree in “Informatics and information technologies” in 2011. He has published about 25 scientific papers in international and national journals and conferences. He is currently a computer science teacher and researcher, and his research interests include artificial intelligence, data mining, pattern recognition, and software engineering. He can be contacted at email: maxarov.qodirbek@gmail.com.



Doniyor Saidov    holds a Ph.D. degree in computer science field from the National University of Uzbekistan (NUUz), Uzbekistan in 2017. He also received his B.Sc. (Computer science) and M.Sc. (Computational mathematics) from NUUz in 2008 and 2011, respectively. He is currently an associated professor at the Department of Algorithms and Programming Technologies in NUUz. His research includes machine learning, data mining, graph theory, and data structures. He has published over 10 papers in international journals and conferences. From February 2018 to July 2018, he was a post Ph.D. research fellow in Keele University, United Kingdom. He can be contacted at email: doniyorsaidov86@gmail.com.