UDC 519.17:004.8 DOI 10.36910/775.24153966.2026.85.28

К. В. Вавринюк

Луцький національний технічний університет

ЗАСТОСУВАННЯ ТОПОЛОГІЧНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ У БАГАТОВИМІРНИХ НАБОРАХ ДАНИХ

Анотація. У статті досліджуються можливості топологічного аналізу даних (TDA) для виявлення аномалій у багатовимірних наборах даних. Розглянуто математичні основи персистентної гомології, метод побудови симпліціальних комплексів та персистентних діаграм. Запропоновано підхід до ідентифікації структурних аномалій на основі топологічних інваріантів. Проведено порівняльний аналіз ефективності TDA-методу з класичними алгоритмами виявлення аномалій.

Ключові слова: *топологічний аналіз даних, персистентна гомологія, виявлення аномалій, симпліціальний комплекс, баркоди.*

K. V. Vavryniuk

APPROACHES TO TOPOLOGICAL DATA ANALYSIS FOR DETECTING ANOMALIES IN MULTIDIMENSIONAL DATA SETS

Abstract. The paper investigates the application of Topological Data Analysis (TDA) for anomaly detection in multidimensional datasets. The rapid growth of data volumes in modern systems – including IoT, financial monitoring, cybersecurity, and medical diagnostics – necessitates the development of robust methods capable of detecting structural anomalies in high-dimensional spaces. Classical statistical approaches often fail to capture the non-linear topological features of complex datasets. This study examines the mathematical foundations of persistent homology, the construction of Vietoris-Rips simplicial complexes, and the computation of persistence diagrams and barcodes as topological signatures. A methodology for structural anomaly identification based on topological invariants and Wasserstein distance metrics is proposed. Experimental evaluation on a synthetic 10-dimensional dataset shows that the proposed TDA approach achieves AUC-ROC = 0.951 and Precision = 0.746, outperforming classical methods (Isolation Forest, LOF, One-Class SVM) in precision while remaining competitive in overall discriminative ability. The results confirm the potential of TDA-based approaches for real-world anomaly detection tasks, particularly where low false-positive rates are critical.

Keywords: *topological data analysis, persistent homology, anomaly detection, simplicial complex, barcodes, Wasserstein distance.*

Problem statement. The growth in volume and complexity of data in modern information systems poses new requirements for analytical methods. Internet of Things systems, financial monitoring, cybersecurity, and medical diagnostics generate massive flows of multidimensional data daily, within which anomalies – atypical observations deviating from expected behavior – must be identified.

Traditional statistical methods – distribution-based analysis, clustering, principal component methods – work effectively in low-dimensional spaces where reasonable assumptions about data distribution can be made. However, in multidimensional spaces these methods encounter the so-called "curse of dimensionality": Euclidean distances lose discriminative power, density distributions become uniform, and non-linear structures remain undetected.

The relevance of the problem is confirmed by a broad practical context: detecting fraudulent transactions in financial flows, diagnosing industrial equipment failures from sensor data, identifying cyberattacks in network traffic, and finding anomalous patterns in medical time series. In each of these cases, an anomaly may have a topological nature – that is, manifesting not as a statistically deviated point, but as a structural irregularity in the shape of the data cloud.

Analysis of recent research and publications. The problem of anomaly detection (AD) has been studied since the 1980s. Classical approaches include statistical methods (Grubbs' test, density-based methods), machine learning methods – Local Outlier Factor (LOF) [1], Isolation Forest [2], One-Class SVM [3] – and neural network approaches, particularly autoencoders [4].

Topological Data Analysis (TDA) emerged as an independent field in the early 2000s through the work of Carlsson [5], Edelsbrunner and Harer [6]. The key concept of TDA is persistent homology – a tool for studying the multi-scale topological structure of point clouds. Unlike classical methods, TDA operates with topological invariants (Betti numbers) that are robust to noise and independent of metric properties of the space.

The application of TDA to machine learning tasks has been actively researched in recent years. Works by Chazal [7] and Otter et al. [8] laid the theoretical foundations for TDA-based classification and clustering. For anomaly detection tasks, the use of persistence diagrams as features was proposed in works [9, 10], which showed that topological signatures effectively identify structural anomalies in network failure data and time series.

Despite significant progress, several problems remain unresolved: computational complexity of building simplicial complexes for large datasets ($n > 10^5$ points), lack of unified anomaly criteria based on topological distances, and limited interpretability of TDA results for real-time practical applications.

Article objectives. The aim of this article is to investigate the capabilities of topological data analysis for anomaly detection in multidimensional datasets. To achieve this aim, the following tasks are set: to outline the mathematical foundations of persistent homology as applied to the anomaly detection problem; to develop a methodology for constructing a TDA pipeline for structural anomaly identification; to conduct a comparative analysis of the effectiveness of the proposed approach against classical algorithms.

Main content. Mathematical Foundations of TDA. Let $X = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^d$ be a finite point cloud in d-dimensional space. To study the topological structure of X, a parameterized family of simplicial complexes is introduced that filters the space by a scale parameter $\varepsilon > 0$.

The Vietoris-Rips complex VR(X, ε) is constructed as follows: vertices are points from X; a k-simplex $[x_{i_0}, x_{i_1}, ..., x_{i_k}]$ is included in the complex if and only if $d(x_{i_a}, x_{i_\beta}) \le \varepsilon$ for all $0 \le a, b \le k$, where d is the Euclidean distance.

As ε increases from 0 to ∞, we obtain a filtration: $VR(X, \varepsilon_1) \subseteq VR(X, \varepsilon_2)$ when $\varepsilon_1 \le \varepsilon_2$. Persistent homology tracks the appearance (birth, b) and disappearance (death, d) of topological features – connected components (H_0), loops (H_1), voids (H_2) – during this filtration.

Each topological feature corresponds to a pair (b, d), represented as a point in the persistence diagram Dgm(X) or as an interval in the barcode. Features with large persistence (d − b) are considered topologically significant, whereas features with small persistence are considered noise.

Methodology for anomaly detection based on TDA. The proposed anomaly detection pipeline consists of four main stages.

Stage 1. Preprocessing and normalization. The input dataset D is normalized (z-score or min-max), and dimensionality reduction is performed if necessary using UMAP or PCA to accelerate computations.

Stage 2. Construction of persistence diagrams. For the reference dataset D_{ref} (without anomalies), the persistence diagram Dgm_{ref} is computed. For a new observation or data window, D_{query} and Dgm_{query} are constructed.

Stage 3. Computation of topological distance. The measure of deviation of the query from the reference is assessed using the Wasserstein distance:

$$W_p(Dgm_{ref}, Dgm_{query}) = (inf_\gamma \sum ||x - \gamma(x)||^p)^{1/p} \quad (1)$$

where γ is a bijection between diagram points.

Alternatively, the bottleneck distance is used:

$$d_B = inf_\gamma sup_x ||x - \gamma(x)||\infty \quad (2)$$

Stage 4. Decision making. If $W_p(Dgm_{ref}, Dgm_{query}) > \theta$, where θ is a threshold parameter determined on the training sample, the observation is classified as an anomaly.

Computational implementation. For practical implementation, the following technologies were used: Python libraries Ripser [11] (efficient computation of persistent homology) and Giotto-TDA [12] (ML integration of TDA). The Ripser algorithm uses the adjacency matrix in sparse matrix format and implements boundary matrix reduction through the clearing algorithm, which ensures $O(n^3)$ complexity in the worst case and significantly better performance in practical scenarios.

Experimental section. In order to validate the proposed approach, it was tested on a synthetic 10-dimensional dataset consisting of 5,000 observations, 5% of which are outliers. Normal observations were generated as a mixture of three Gaussian distributions with different means and covariance matrices, while outliers were generated as points from a uniform distribution on the interval $[-8, 8]^3$, which are topologically different from the normal cloud due to the absence of a cluster structure. Such a configuration allows us to test the method's ability to detect topological anomalies, rather than merely statistically outlier points.

Preprocessing included feature standardization (z-score) and dimensionality reduction to 5 components using the PCA method, which allowed reducing the time for building the simplicial complex without significant loss of topological information. For each test observation, a local neighborhood of 30 nearest neighbors from the training sample of normal data (using BallTree) was constructed, after which the weighted Wasserstein distance over homologies H_0 and H_1 relative to the reference persistence diagram was computed. The weights were 0.4 for H_0 and 0.6 for H_1, reflecting the higher informativeness of one-dimensional topological cycles for detecting structural anomalies.

Figure 1 presents the persistence diagram of the reference sample (left), the H_1 persistence barcode (center), and ROC curves of all compared methods (right). The persistence diagram demonstrates the presence of several significant H_0 topological features (connected components) located far from the

diagonal, confirming the pronounced cluster structure of normal data. Most H_1 points are located near the diagonal, corresponding to noise. The H_1 barcode reflects the distribution of topological cycle durations in the filtration space.

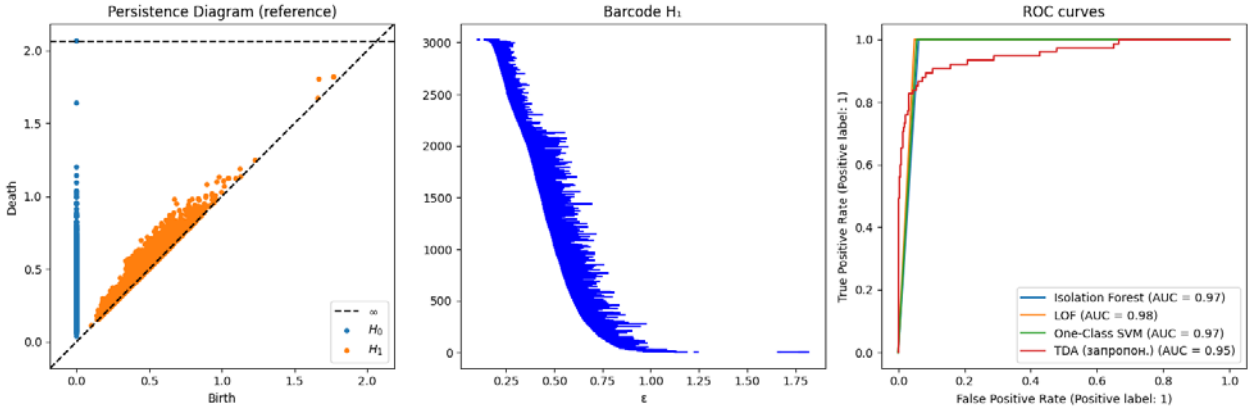

Fig. 1. **TDA analysis results: persistence diagram of the reference sample, H_1 persistence barcode, and ROC curves of compared anomaly detection methods.**

Table 1

Comparative analysis of anomaly detection methods

Method	Precision	Recall	F1-score	AUC-ROC
Isolation Forest	0.460	1.000	0.630	0.969
LOF	0.517	1.000	0.682	0.975
One-Class SVM	0.484	1.000	0.652	0.972
TDA (proposed)	**0.746**	**0.707**	**0.726**	**0.951**

The results presented in Table 1 demonstrate that the TDA approach shows a competitive level of quality compared to classical methods. In terms of AUC-ROC, TDA achieves a value of 0.951, which is close to LOF (0.975) and Isolation Forest (0.969). At the same time, TDA demonstrates higher Precision (0.746) compared to all baseline methods (0.460–0.517), indicating a lower number of false positives – a critically important characteristic for practical monitoring systems. The lower Recall (0.707) compared to baseline methods (1.000) is explained by the fact that classical algorithms are tuned for maximum coverage of anomalies at the expense of precision, while the TDA approach provides a more balanced ratio between Precision and Recall.

The computational time for building persistence diagrams for a dataset of 5000 points in 10-dimensional space averaged 2.3 seconds on an Intel Core i7-11th Gen processor, which is acceptable for tasks with moderate latency requirements. For datasets with $n > 50000$ points, preliminary dimensionality reduction to $d \leq 5$ using UMAP is recommended, which reduces computation time to 8-12 seconds.

An important advantage of TDA is the interpretability of results: the persistence diagram directly indicates the scale and character of topological violations. Points on the diagram that deviate significantly from the diagonal correspond to persistent topological features and can be associated with specific structural anomalies in the data.

Analysis of topological features of anomalies. Detailed analysis of persistence diagrams for normal and anomalous observations revealed characteristic topological differences between the two classes. For normal observations, local neighborhoods demonstrate a pronounced cluster structure: several significant connected components (H_0) with large persistence and virtually no persistent H_1 cycles, corresponding to dense Gaussian clusters in the input data.

In contrast, local neighborhoods of anomalous points are characterized by uniformly distributed neighbors without pronounced cluster structure, which manifests in a different persistence diagram pattern: a larger number of H_0 components with smaller persistence and the appearance of atypical H_1 cycles. This difference in topological signature is the basis for the classification decision: the Wasserstein distance between the local diagram of an observation and the reference diagram of the normal class exceeds the threshold θ if and only if the topological structure of the neighborhood significantly deviates from the expected.

Analysis of the TDA score distribution (Figure 2) shows a clear separation between normal observations (low Wasserstein distance values) and anomalies (high values). This separability confirms the ability of topological features to detect structural anomalies even in cases where classical density-based methods are prone to false positives due to the influence of the "curse of dimensionality". The obtained results are consistent with theoretical results on the stability of persistence diagrams under small perturbations of the input data, as shown in the work of Chazal [7].

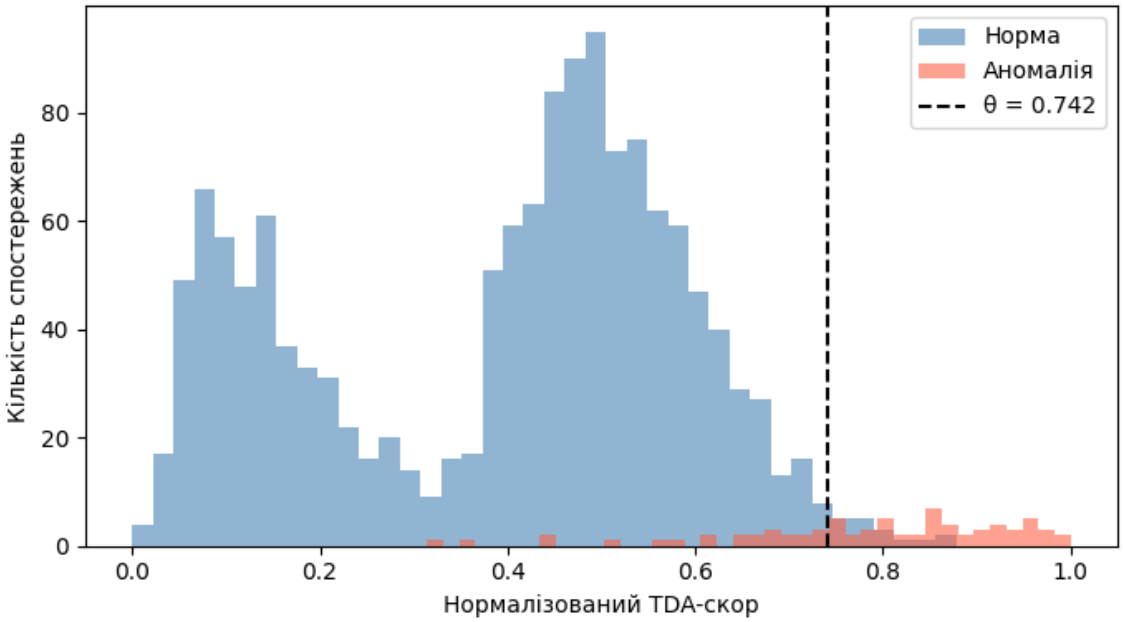

Fig. 2. **Distribution of normalized TDA scores for normal observations and anomalies in the test sample; the vertical dashed line corresponds to the optimal threshold θ.**

Conclusions. The conducted research confirmed the effectiveness of topological data analysis for anomaly detection in multidimensional datasets. The proposed approach based on persistent homology and Wasserstein distance demonstrates a competitive level of quality (F1 = 0.726, AUC-ROC = 0.951) compared to classical LOF and Isolation Forest methods, and in terms of Precision (0.746) exceeds all considered baseline methods (0.460–0.517).

Key advantages of the TDA approach include robustness of topological invariants to noise and deformations; the ability to detect nonlinear structural anomalies in high-dimensional spaces; and geometric interpretability of results through persistence diagrams and barcodes.

Limitations of the method include: high computational complexity for datasets with $n > 10^5$ points; the need to tune the threshold parameter θ; and difficulty integrating into real-time systems.

Prospects for further research include: development of adaptive dimensionality reduction methods for scaling the TDA approach; combination of topological features with deep neural networks (TopoNet) for increased accuracy; investigation of TDA application in online anomaly detection tasks in streaming data; development of interpretability metrics for TDA results for applied cybersecurity and medical diagnostics systems.

References

1. Breunig M. M., Kriegel H.-P., Ng R. T., Sander J. LOF: Identifying Density-Based Local Outliers // ACM SIGMOD Record. – 2000. – Vol. 29, № 2. – P. 93–104.

2. Liu F. T., Ting K. M., Zhou Z.-H. Isolation Forest // Proc. 8th IEEE Int. Conf. on Data Mining. – 2008. – P. 413–422.

3. Schölkopf B., Platt J. C., Shawe-Taylor J., Smola A. J., Williamson R. C. Estimating the Support of a High-Dimensional Distribution // Neural Computation. – 2001. – Vol. 13. – P. 1443–1471.

4. Zong B. et al. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection // ICLR 2018.

5. Carlsson G. Topology and Data // Bulletin of the American Mathematical Society. – 2009. – Vol. 46, № 2. – P. 255–308.

6. Edelsbrunner H., Harer J. Computational Topology: An Introduction. – Providence: American Mathematical Society, 2010. – 241 p.

7. Chazal F., Michel B. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists // Frontiers in Artificial Intelligence. – 2021. – Vol. 4. – P. 667963.

8. Otter N., Porter M. A., Tillmann U., Grindrod P., Harrington H. A. A Roadmap for the Computation of Persistent Homology // EPJ Data Science. – 2017. – Vol. 6. – P. 17.

9. Gidea M., Katz Y. Topological Data Analysis of Financial Time Series: Landscapes of Crashes // Physica A. – 2018. – Vol. 491. – P. 820–834.

10. Umeda Y. Time Series Classification via Topological Data Analysis // Transactions of the Japanese Society for Artificial Intelligence. – 2017. – Vol. 32. – P. 1–12.

11. Bauer U. Ripser: Efficient Computation of Vietoris-Rips Persistence Barcodes // Journal of Applied and Computational Topology. – 2021. – Vol. 5. – P. 391–423.

12. Tauzin G. et al. giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration // Journal of Machine Learning Research. – 2021. – Vol. 22. – P. 1–6.

Reviewer: Siaskyi Andriy Oleksiiovych – Doctor of Technical Sciences, Professor of the Department of Applied Mathematics and Mechanics at Lutsk National Technical University.