Comparison of Data Visualization, Outlier Detection and Data Dimensionality Reduction Methods

Xingyu Zhao *

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S10 2TN, England

* Corresponding author: xzhao82sheffield@ldy.edu.rs

Abstract. With the deepening of the digital age of information, people's daily data is getting larger and larger, and it is more and more difficult to quantify and process. At this time, the data processing means becomes particularly important. This paper compares and analyzes some methods from data visualization to data dimensionality reduction to outlier detection. In this paper, two different types of datasets, ModelNet40, and red wine quality, are used to introduce the visualization method of the Farthest Point Sampling (FPS). This method can have a clear visual effect on the data dimension and scale, and allow users to observe the structure, type, and scale of the data. In data dimensionality reduction, the study uses Principal Component Analysis (PCA), T-Distributed Stochastic Neighbor Embedding (t-SNE), Triplets Manifold Approximation and Projection (TriMAP), Uniform Manifold Approximation and Projection (UMAP), Pairwise Controlled Manifold Approximation Projection (PaCMAP), and Autoencoder to compare their dimensionality reduction effects. Through these methods, this paper finds that different methods have different effects on different datasets. Therefore, in data dimensionality reduction, it can get twice the result with half the effort by choosing the appropriate method. Finally, this paper also detects outliers. Outliers in datasets will make it difficult for people to process data and make subsequent results inaccurate, so it is necessary to identify outliers. This paper involves methods such as isolation forest and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Through this paper, the methods of different datasets are analyzed and summarized.

Keywords: Data Visualization, Data Dimensionality Reduction, FPS, PCA, t-SNE.

1. Introduction

In the digital information age, with the development of artificial intelligence technology, data is constantly being generated and growing at an amazing speed. How to extract useful knowledge from this massive data has become a big challenge in the field of modern information technology. Traditional data processing methods, such as statistical analysis and prediction models, rely on artificially designed characteristics and assumptions, and their processing methods are often single and limited by data types and quality. However, the types of modern data are complex and diverse, and the quality of data is uneven the existing tools and methods cannot meet this demand in a reasonable time. Facing the challenge of big data, innovative and robust data processing methods are needed.

In the past few years, the field of machine learning has made great progress, especially in preprocessing multi-dimensional data. In the face of huge and complex datasets, the dimensionality reduction method has become an important technical means, which aims to reduce the dimensions of data while preserving the internal structure and characteristics of data to the greatest extent. In the growing data, the importance of dimensionality reduction methods and anomaly detection technology is becoming more and more prominent. They provide effective means for processing complex data and also provide strong support for the further development of data analysis and pattern recognition.

2. Method

2.1. Dataset

2.1.1. ModelNet40 - Princeton 3D Object Dataset

ModelNet40 is a set of Computer Aided Design (CAD) models with clear 3D objects. It consists of 40 categories, with a total of 12,311 models pre-aligned. As seen in Fig. 1, the CAD model adopts the object file format (closed). Princeton Vision Toolkit (PVT) provides Matlab functions for reading and visualizing files.

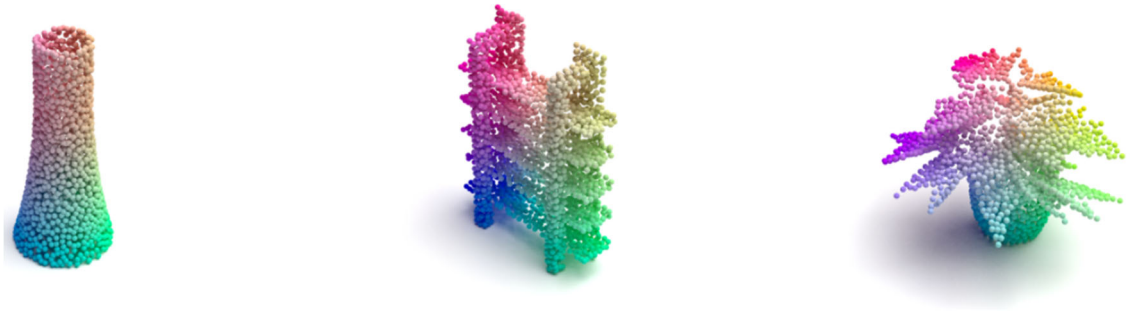

Figure 1. ModelNet40 Dataset

2.1.2. Red Wine Quality

This dataset contains the quality of Portuguese wine, which is composed of density, alcohol content, residual sugar and other factors, which affect the quality of wine. These datasets can be regarded as classification or regression tasks. Grades are orderly, not balanced (for example, ordinary wines are much more than good or bad wines).

2.2. Dimensionality Reduction Methods

2.2.1. PCA

PCA is a common method for decreasing the dimensionality of a dataset while aiming to maintain the underlying information. The primary objective of PCA is to recognize the inherent patterns within the data and represent it in a manner best explained by its principal components. These principal components are new irrelevant variables, which are sorted according to their variance explained in the original dataset. The first principal component represents the maximum variance possible, and each successive component also captures the maximum variance possible while being perpendicular to the preceding component [1, 2].

2.2.2. t-SNE

t-SNE, widely utilized in machine learning, data analysis, and visualization, is a potent method for reducing dimensions and visually representing high-dimensional data. This technique aims to maintain the local structure of the data while transforming it from high-dimensional space to a lower-dimensional space (often 2D or 3D). Unlike some linear techniques, t-SNE is especially effective at capturing non-linear structures, making it suitable for visualizing complex relationships among data points [3].

2.2.3. UMAP

UMAP stands out as a potent method for reducing dimensionality. Its design revolves around the preservation of both local and global structures present in high-dimensional datasets when projected into a lower-dimensional space. This characteristic makes it especially adept at visualizing and exploring intricate datasets, particularly in cases where the local data point structures exist in nonlinear or manifold spaces [4].

2.2.4. TriMap

TriMap shares similarities with t-SNE and UMAP in its objective to capture intricate nonlinear relationships within high-dimensional data and transform it into low-dimensional space. An interesting characteristic of TriMap is its capacity to simultaneously learn three distinct low-dimensional embeddings of the identical dataset, setting it apart from conventional techniques. This approach enables TriMap to seek a representation that effectively preserves both the local and global structure of the data, distinguishing it from traditional methods [5].

2.2.5. PaCMAP

PaCMAP is a UMAP class method, which focuses on maintaining global and local structures. Its main innovation is to advocate the inclusion of "middle-near" pairs in optimization. These points are not the nearest neighbors, but on average they are closer than the random points in the dataset. The cost function in PaCMAP is also quite simple while maintaining certain important attributes to avoid distorting the local manifold structure and maintain the good clustering behavior of UMAP [6].

2.2.6. Isolation Forest

The Isolation Forest algorithm operates through the creation of an ensemble of isolation trees. These trees are formed by iteratively dividing the input space into finer subspaces, effectively isolating the outliers. This process involves the random selection of a feature and a split value for the feature at each iteration, contributing to the construction of the isolation trees. An isolation tree is constructed in such a way that anomalies are expected to be isolated into smaller, shorter paths within the tree as compared to normal data points, making them easier to identify. By building many such trees and combining their results, Isolation Forest is able to effectively isolate anomalies and identify outliers by measuring how many splits are required to separate a data point, with the assumption that anomalies will require fewer splits on average to be isolated [7].

2.2.7. DBSCAN

DBSCAN, a widely adopted clustering algorithm in machine learning and data mining, is recognized for its capacity to detect clusters of diverse structures and accurately differentiate noise from legitimate clusters. Unlike some clustering algorithms, DBSCAN eliminates the need for users to predefine the number of clusters, rendering it particularly suitable for datasets with irregular cluster shapes and varying densities [8]. Renowned for its robust performance in handling clusters of different configurations and dimensions, DBSCAN exhibits resilience to noise and excels in managing extensive spatial databases [9]. Its applications span various domains including pattern recognition, image processing, and anomaly detection, where it is widely implemented.

2.2.8. Autoencoder

An Autoencoder is used to take a dataset as input and usually takes it as the target to predict. The Autoencoder therefore learns its input. The input is compressed and reduced to a low-dimensional space, which is done by the encoder. Then, the original dataset is recovered from the low-dimensional vector by a decoder [10].

2.3. FPS

FPS is an effective sampling technique for point clouds, which is used in Pointnet ++ algorithms. Compared with other sampling techniques, FPS can cover the whole point set better, because it can find a subset of points that are farthest from each other [10].

3. Results and Discussion

3.1. Red Wine Quality

Based on variance optimization (PCA), distance (PCA, t-SNE, UMAP, TriMap, PaCMAP, DBSCAN), and neural network using weights, this paper judges the quality of wine. According to the benchmark the higher the alcohol content, the better the wine.

3.1.1. Dimensionality Reduction Methods

The following Fig. 2 shows the effects of six data dimensionality reduction methods. It can be seen that PCA and Autoencoder have good dimensionality reduction effects in the global scope of data, and the effect of Autoencoder is more obvious. UMAP and T-SNE have similar effects. TriMAP and PaCMAP are also good dimensionality reduction methods, but compared with PCA, they have a more obvious local dimensionality reduction effect and ignore the global data [9].

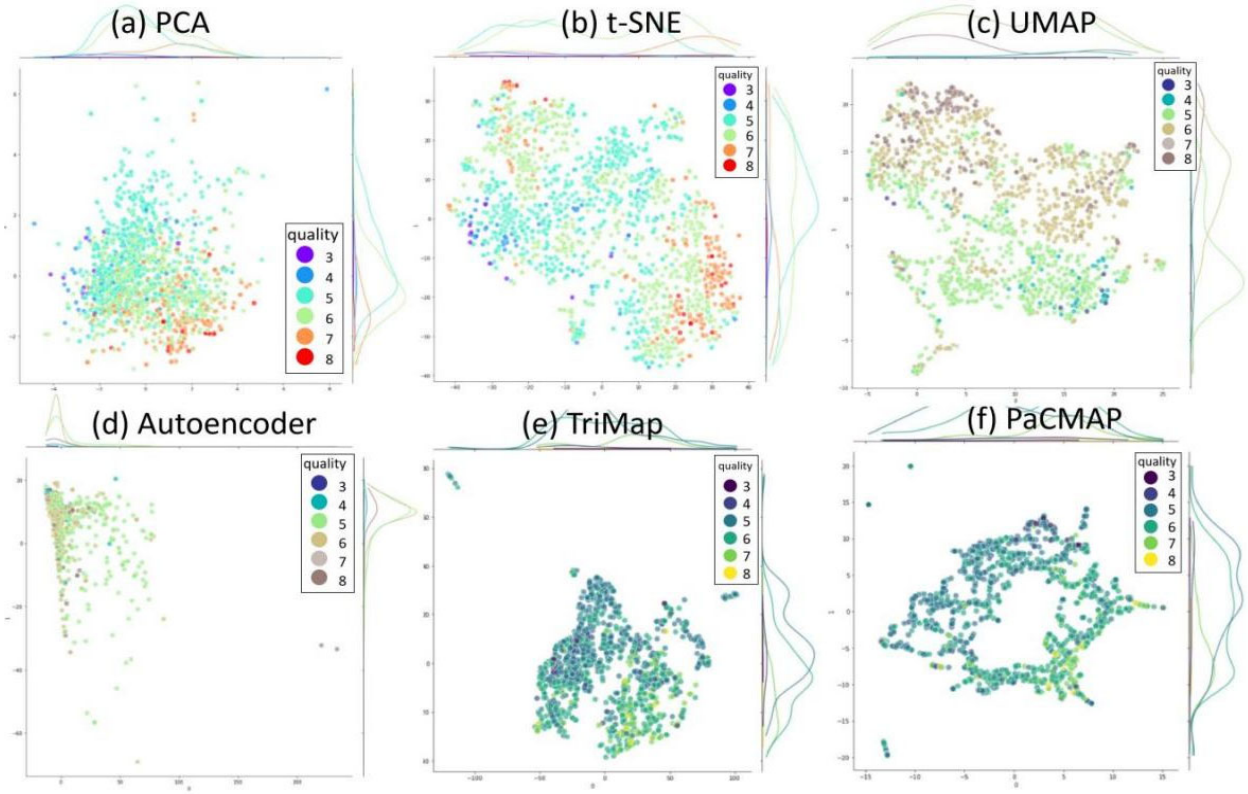

Figure 2. Various Dimensionality Reduction Methods

3.1.2. The Comparison of Various Methods

This article explains each variable in the dataset in two dimensions. Fig.3 Shows the R-Square distribution. According to the maximum error, As can be seen from the following two infundibuliform comparison diagrams, it can be seen that PaCMAP and TriMap show the most robust results, followed by Autoencoder and UMAP. T-SNE and PCA are in the last position.

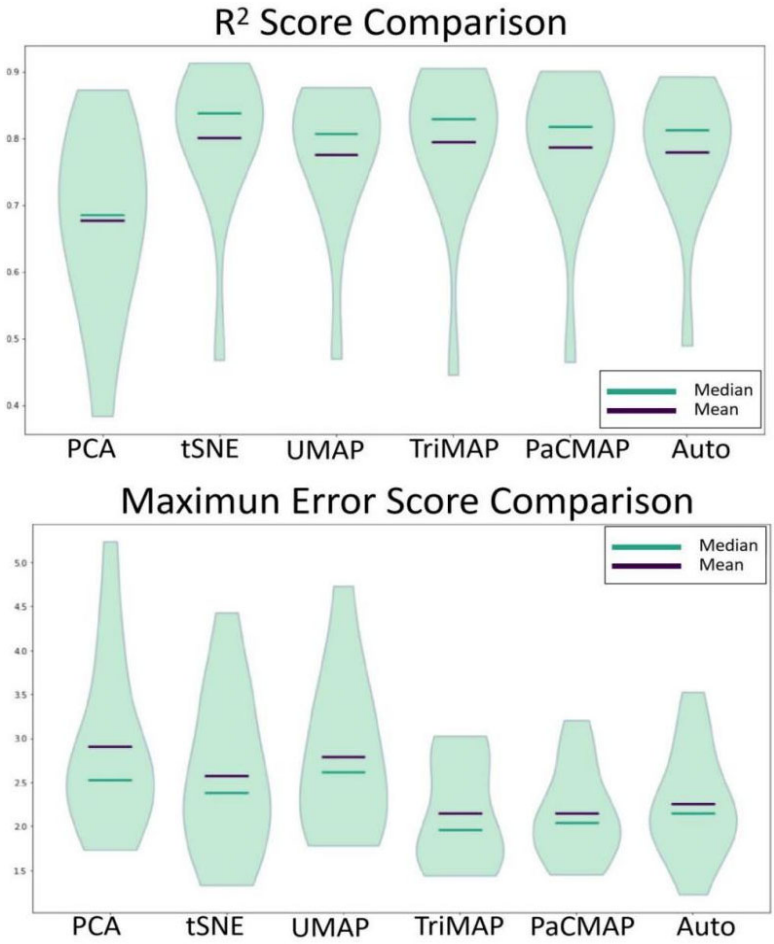

Figure 3. Comparison R^2 Score and Maximum Error Score

3.2. ModelNet40 Dataset

3.2.1. Farthest Point Sampling

As can be seen from Fig. 4, by using the FPS algorithm, two planes in the CAD dataset are well displayed, and the accuracy of visualization can be changed by changing the values of parameters. This method can be used to deal with data visualization.

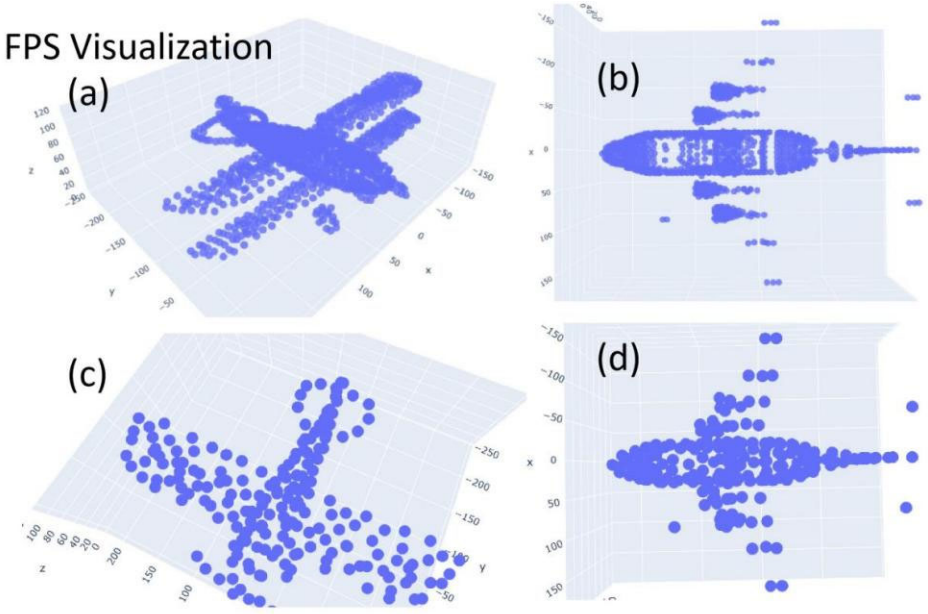

Figure 4. Visualization Method of ModelNet40

3.2.2. Dimensionality Reduction Methods

Similarly, the effect of ModelNet40 in the dataset is shown in the Fig 5. PCA and Autoencoder show a strong concern for global structure. It can be seen that PCA and Autoencoder show obvious aircraft model structures. UMAP and t-SNE, on the other hand, show concern for local structure. TriMap and PaCMAP show that these methods try to represent local and global structures [9].

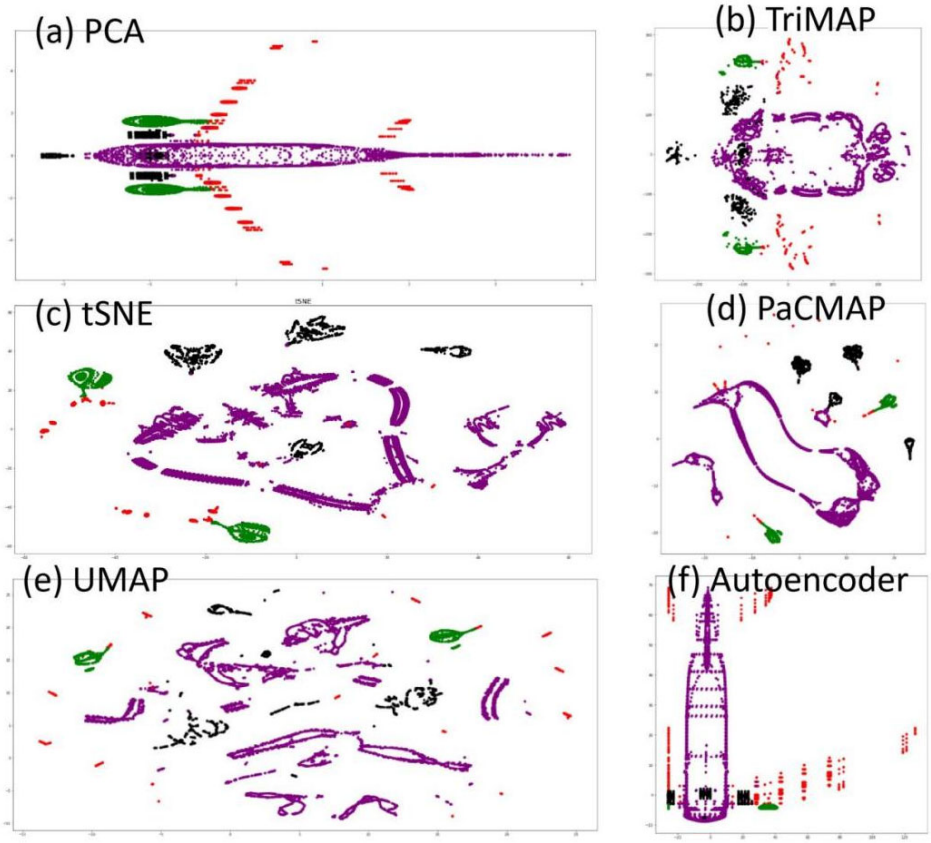

Figure 5. Reduction Method of ModelNet40

3.2.3. The Comparison of Various Methods

From the following wedge diagram, it can still be seen that according to the maximum error, Autoencoder and t-SNE rank the highest, while PaCMAP and TriMap are more stable, although PCA results are also excellent. As seen in Fig. 6.

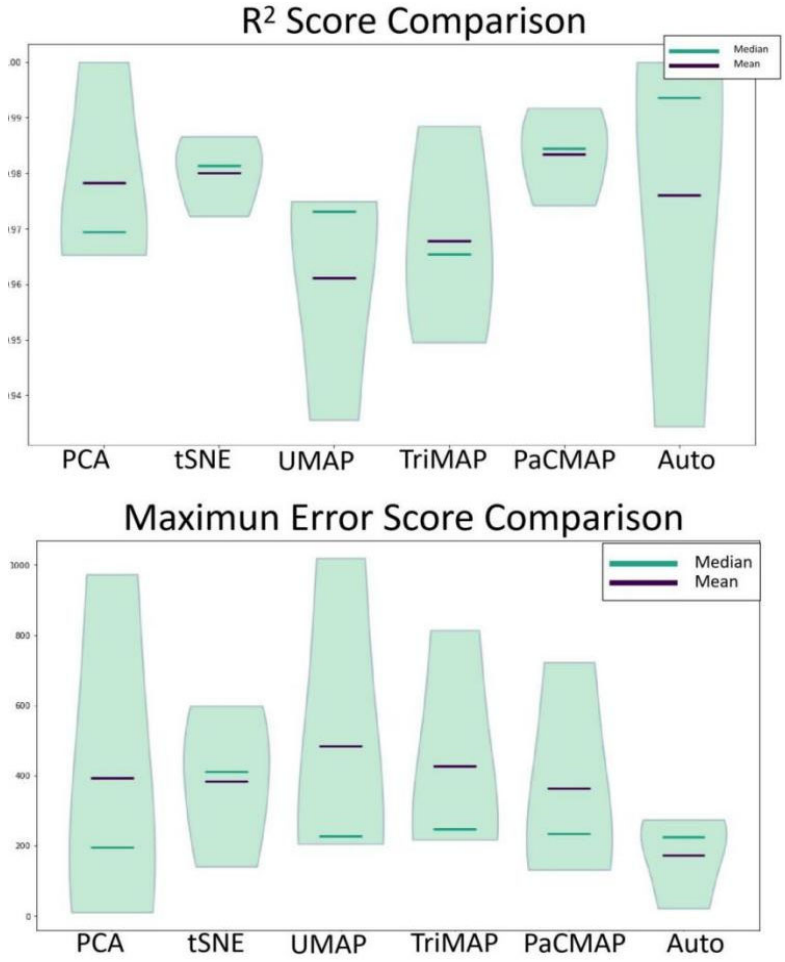

Figure 6. Comparison of R2 Score and Maximum Error Score

3.3. Outlier Detection

As shown in Fig. 7, UMAP brings a large amount of data into a visible range, which means that it tends to bring outliers into the clustering range. Autoencoder and PCA also fail to identify outliers well, which means that they cannot be transferred to invisible data.

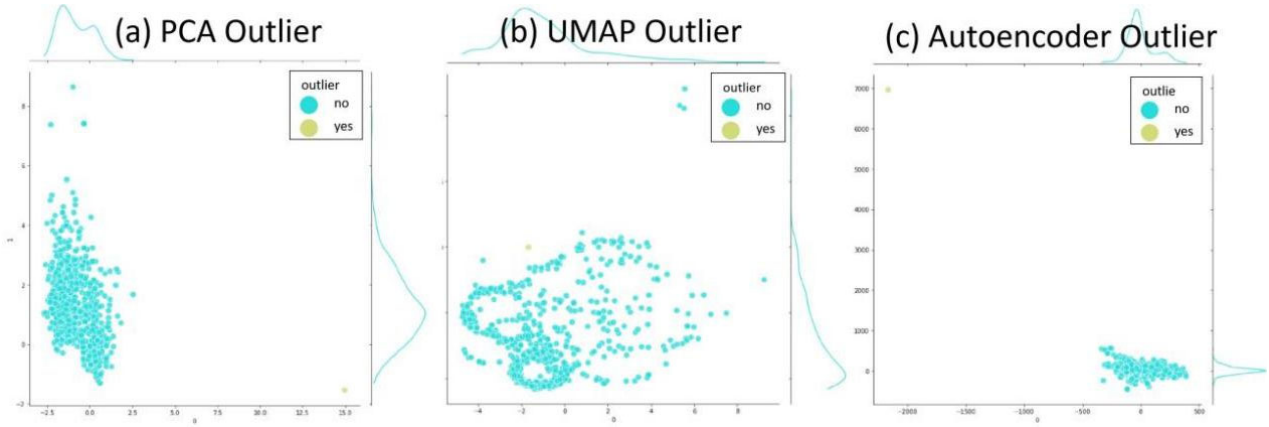

Figure 7. Outlier Distribution of Each Method

As can be seen from Fig. 8, Isolation Forest is powerful in finding outliers, but DBSCAN is not as accurate as it is. Therefore, the introduction of isolated forests is an effective method. The isolation forest isolates all the observations of the dataset from each other. Therefore, it is assumed that observations that are particularly easy to isolate may be outliers [7, 8].

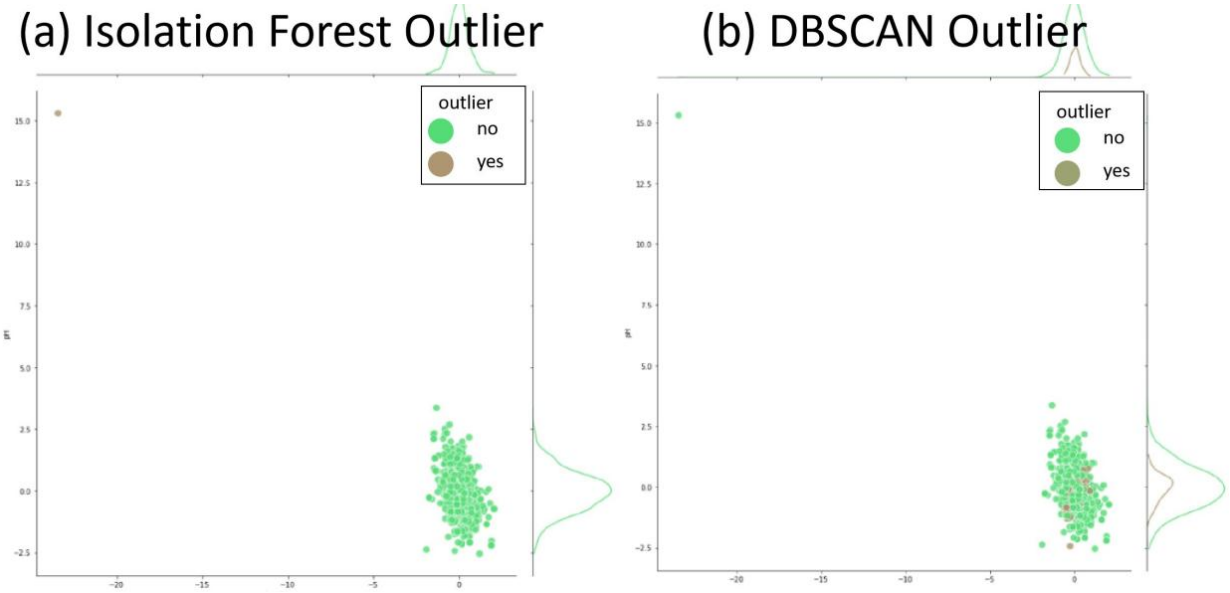

Figure 8. Outlier Distribution of Isolation Forest and DBSCAN

4. Conclusion

In this article, based on two different types of datasets, six dimensionality reduction methods, one visualization method, and five outlier detection are used to compare and analyze their similarities and differences and their effects on the data. The results show that t-SNE, TriMap, and UMAP have the best robustness in the two datasets. Because R squared, average absolute error and maximum error are well balanced. PaCMAP and Autoencoder also do well, and PCA gets the least information in two dimensions. In addition, the concepts of global and local data dimensionality reduction are also presented in this project. PCA and Autoencoder have the best dimensionality reduction in the global dataset, while other data dimensionality reduction effects pay more attention to the local part of the dataset, but the effect is also excellent. In the ModelNet40 dataset, the global and local structures can be seen by the data dimension reduction method, in which FPS can display the distribution of data points in a three-dimensional macro way.

As a method of outlier analysis, UMAP is not suitable for the prediction of these datasets, and it will forcibly bring Outliers into the sphere of influence of clusters. As the only clustering algorithm used here, the results of DBSCAN in cross-sectional data are not satisfactory. The isolated forest shows good results. In addition, the automatic encoder can easily separate outliers in the cross-section.

References

[1] Roweis S. EM algorithms for PCA and SPCA Advances in neural information processing systems1997, 10.

[2] Daffertshofer A., Lamoth C. J., Meijer O. G. & Beek P. J. PCA in studying coordination and variability: a tutorial Clinical biomechanics, 2004 19 (4) 415 - 428.

[3] Wattenberg M., Viégas F. & Johnson I. How to use t-SNE effectively Distill, 2016, 1 (10) e2.

[4] McInnes L., Healy J., & Melville J. Umap: Uniform manifold approximation and projection for dimension reduction arXiv preprint arXiv: 1802. 03426, 2018.

[5] Amid E. & Warmuth M. K. TriMap: Large-scale dimensionality reduction using triplets arXiv preprint arXiv, 1019, 1910. 00204.

[6] Tuncer O., Leung V. J. & Coskun A. K. Pacmap: Topology mapping of unstructured communication patterns onto non-contiguous allocations In Proceedings of the 29th ACM on International Conference on Supercomputing2015, 37 - 46.

[7] Hariri S., Kind M. C. & Brunner R. J. Extended isolation forest IEEE transactions on knowledge and data engineering, 2015, 33 (4) 1479 - 1489.

[8] Wang Y., Huang H., Rudin C. & Shaposhnikov Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE UMAP TriMAP and PaCMAP for data visualization The Journal of Machine Learning Research, 2021, 22 (1) 9129 - 9201.

[9] Schubert. Sander, J., Ester M., Kriegel H.P. & Xu, X. DBSCAN revisited revisited: why and how you should (still) use DBSCAN ACM Transactions on Database Systems (TODS), 2017, 42 (3) 1 - 21.

[10] Eldar Y., Lindenbaum M., Porat M. & Zeevi Y. Y. The farthest point strategy for progressive image sampling IEEE Transactions on Image Processing1997, 6 (9) 1305 - 1315.