Comparison and Evaluation of Classical Dimensionality Reduction Methods

Zilin Shen

College of Science, Minzu University of China, Beijing, China, 100081

szl1428@163.com

Abstract. As one of the tasks of unsupervised learning, data dimensionality reduction is faced with the problem of a lack of evaluation methods. Based on this, three classical dimensionality reduction methods such as PCA, t-SNE and UMAP were selected as the research object in this paper. This article selected 5 three-classification datasets and used the three methods mentioned above to perform dimensionality reduction. This paper plotted 3D scatter graphs after dimensionality reduction to analyze the differentiation effect of the data on different categories of the target variable. Then the data after dimensionality reduction was classified using random forest model and the classification accuracy was obtained. According to the 3D scatter plots and the accuracy of random forest, it is found that PCA has a good dimensionality reduction effect on most of the selected datasets, and t-SNE has a relatively stable dimensionality reduction effect. In contrast, UMAP has good dimensionality reduction performance in some individual datasets but lacks stability. Overall, this paper proposes a dimensionality reduction evaluation method that combines scatter-plot visualization results and classification models, which can effectively predict the performances of the dimensionality reduction methods for a variety of datasets, thereby promoting the comparison and selection of dimensionality reduction methods in the field of unsupervised learning.

Keywords: PCA, t-SNE, UMAP, Dimensionality Reduction.

1. Introduction

As a branch of machine learning, unsupervised learning is usually used to process many unlabeled data to help us explore and analyze the data without a clear goal. Common unsupervised learning tasks include data dimensionality reduction, cluster analysis, association rule mining, etc. Unlike supervised learning, unsupervised learning does not rely on labeled data, but obtains an understanding of the data itself through statistics and analysis of the underlying structure of the input data [1]. Among them, dimensionality reduction of data is an important part of unsupervised learning. Its task is to map high-dimensional data into low-dimensional space and obtain key data information by studying fewer dimensions to improve the efficiency of data analysis.

Similar to other unsupervised learning tasks, due to the lack of clear target values or data labels to guide the learning process, the problem of missing evaluation methods is usually encountered in the actual application of data dimensionality reduction, which leads to the lack of reliable evaluation methods to measure the effect and performance of each dimensionality reduction method, thus limiting the further development and application of dimensionality reduction methods [2-3].

Based on the above problems, this paper selects three commonly used data dimensionality reduction methods, namely PCA, t-SNE and UMAP. It establishes reliable and accurate evaluation methods to analyze the application scenarios and performance of the three-dimensionality reduction methods, thereby helping to solve the problem of lack of evaluation methods for data dimensionality reduction and guiding selecting which dimensionality reduction methods under which scenarios. It will help to improve the performance of dimensionality reduction methods further, provide new research perspectives and ideas for exploring algorithm evaluation of unsupervised learning, and promote its more effective application in various fields.

2. Data Description

Compared with the three-classification dataset, it is difficult to judge the classification effect by using the two-classification dataset to distinguish the results through the scatter plot. The four-classification dataset or above reduces the overall accuracy of dimensionality reduction, which is challenging to achieve the effect of comparison of dimensionality reduction methods. Considering the visualization effect of dimensionality reduction results and the overall accuracy rate of the classifier model, the following five classical three-classification datasets are selected in this paper:

The above data sets are commonly used data sets for machine learning from Kaggle (https://www.kaggle.com/datasets).

(1) Iris dataset: There are four numerical variables on the characteristics of iris. The sample is divided into setosa, versicolor and virginica.

(2) Seeds dataset: There are seven numerical variables of seed characteristics in the dataset of wheat seeds. One categorical variable is divided into three categories.

(3) Palmer Archipelago (Antarctica) penguin dataset: This dataset is about Palmer penguins, with a total of four numerical variables and one categorical variable of species.

(4) Wine Quality dataset: A dataset on wine recognition suitable for testing classifiers. There are thirteen numerical variables related to wine characteristics, and one categorical variable dividing wine into three categories.

(5) Car Information dataset: This dataset is a classified dataset about vehicle feature information. The dataset has seven numerical variables and one categorical variable divides the car into three categories according to the country of origin.

To analyze the internal correlation among independent variables and preliminarily judge the distinguishing effect of a single variable on different categories of data, this paper draws correlation scatter plots for the datasets used. As shown in figure 1, the background color of each plate in the figure is determined by the correlation coefficient between the two variables (the order of lime, light blue, and red indicates an increasing correlation coefficient). Different colors of scatter points represent different categories.

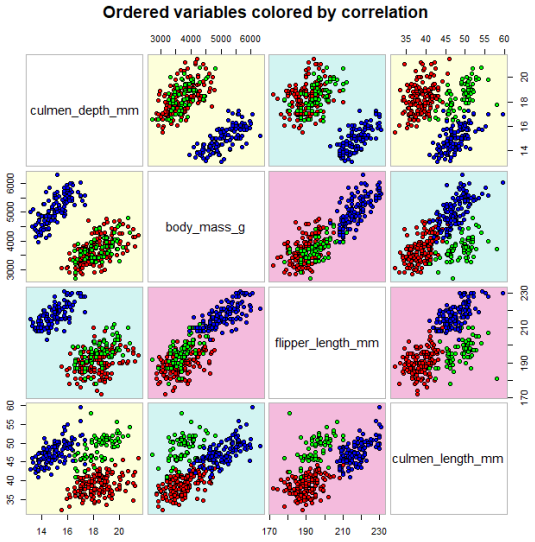

Figure 1. Correlation scatter plot of the Palmer Penguin dataset

3. Methods

3.1. Principle of dimensionality reduction method

(1) PCA

Principal component analysis, first proposed by Pearson (1901) and further developed by Hotelling (1933), is still widely used today as a classical dimensionality reduction method [4].

Let $X=\left(x_1,x_2,\cdots,x_p\right)'\cdot\mu=E(X)$, $\sigma^2=\mathrm{var}(X)$, consider the following linear transformation:

$$\begin{cases} y_1=a_{11}x_1+a_{21}x_2+\cdots+a_{p1}x_p=\boldsymbol{a}_1'X \\ \quad\vdots \\ y_p=a_{1p}x_1+a_{2p}x_2+\cdots+a_{pp}x_p=\boldsymbol{a}_p'X \end{cases} \tag{1}$$

The vector $\boldsymbol{a}_1$ is found under the constraint $\|\boldsymbol{a}_1\|=1$ so that var (y_1) is maximized, and y_1 is called the first principal component.

If we consider using y_2 as the other principal component, we also need to find a vector $\boldsymbol{a}_2$ under the constraint $\|\boldsymbol{a}_2=1\|$ so that $\mathrm{var}(y_2)$ reaches its maximum. In order to ensure that the information contained in y_2 is different from that in y_1, we need to enable:

$$\mathrm{Cov}(y_1,y_2)=0 \tag{2}$$

The $y_2=\boldsymbol{a}_2'X$ thus found is the second principal component. Similarly, the *i* principal component of X can be defined.

(2) t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction method proposed by Laurens van der Maaten and Geoffrey Hinton in 2008 [5]. Based on non-convex optimization, t-SNE is a method widely used in data visualization [6].

Let $X=\{x_1,\cdots,x_N\}$ be the high-dimensional dataset, t-SNE transfers the similarity of datapoints to the conditional probability:

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2N} \tag{3}$$

Where

$$p_{j|i}=\frac{\exp(-\|x_i-x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i-x_k\|^2/2\sigma_i^2)}. \tag{4}$$

σ_i is the variance of the Gaussian that is centered on x_i. It is impossible that there is a single value of σ_i which is optimal for all points. Considering that different values of σ_i can be used to induce different probability distributions for the dataset, and the entropy of this distribution increases with the increase of σ_i, the perplexity is defined:

$$Perp(p_i)=2^{H(p_i)} \tag{5}$$

And $H(p_i)$ is the Shannon entropy of p_i measured in bits:

$$H(p_i)=-\sum_j p_{j|i}\log_2 p_{j|i} \tag{6}$$

To solve the crowding problem that points with large distances in high dimensional space will become smaller in low dimensional space during the transformation process, the t-SNE method employs the Student-t distribution with one degree of freedom as the distribution of low-dimensional data. For the low-dimensional dataset $\{Y=y_1,\cdots,y_N\}$, the distance between data points is converted into conditional probability:

$$q_{ij} = \frac{\left(1+\left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|y_k - y_l\right\|^2\right)^{-1}} \tag{7}$$

In order to minimize the mismatch between low-dimensional data and high-dimensional data, t-SNE selects Kullback-Leibler divergence on all data points and constructs the following loss function:

$$C = KL\left(P\,\|\,Q\right) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{8}$$

(3) UMAP

UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction method proposed by Leland McInnes, John Healy, and James Melville in 2018 [7]. As a recently proposed technique for dimensionality reduction, UMAP can reveal the global structure of the data while preserving the local structure, thus effectively improving the performance [8].

$X = \{x_1, x_2, \cdots, x_N\}$ is the input dataset. The hyper-parameter k indicates the set $\{x_{i_1}, \cdots, x_{i_k}\}$ with the nearest neighbors of x_i. For each x_i, define the parameters:

$$\rho_i = \min\left\{d(x_i, x_{i_j}) \,|\, 1 \le j \le k, d(x_i, x_{i_j}) > 0\right\} \tag{9}$$

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right) = \log_2(k) \tag{10}$$

In high-dimensional space, the conditional probability is set as

$$p_{ij} = \exp\left(-\frac{d(x_i, x_{i_j}) - \rho_i}{\sigma_i}\right) \tag{11}$$

Similar to t-SNE, the meaning of formula (11) is to convert the nearest neighbors of points in high-dimensional space into conditional probability.

In the low-dimensional space, we use the joint probability to approximate the similarity between data points. The joint probability is given by:

$$q_{ij} = \left(1 + a\left(\left\|\mathrm{x} - \mathrm{y}\right\|_2\right)^{2b}\right)^{-1} \tag{12}$$

The hyper-parameters *a* and *b* are chosen by non-linear squares fitting to make equation (12) approximate the following piecewise function:

$$\psi(\mathrm{x}, \mathrm{y}) = \begin{cases} 1 & if\ \left\|\mathrm{x} - \mathrm{y}\right\|_2 \le min\ dist \\ \exp\left(-\left(\left\|\mathrm{x} - \mathrm{y}\right\|_2 - min\ dist\right)\right) & otherwisse \end{cases} \tag{13}$$

$min\ dist$ is introduced artificially. When the distance between two points in low-dimensional space is less than $min\ dist$, the joint probability between two points is 1.

$\mu(a)$ and $\nu(a)$ represent joint probability p_{ij} and q_{ij} respectively. Fuzzy set cross entropy is used as the loss function, the minimum value is sought by Gradient Descent method, and the loss function is given by:

$$C\left((A,\mu),(A,\nu)\right) = \sum_{a\in A} \mu(a) \log\left(\frac{\mu(a)}{\nu(a)}\right) + \left(1-\mu(a)\right) \log\left(\frac{1-\mu(a)}{1-\nu(a)}\right) \tag{14}$$

3.2. Random Forest classification

Based on the above three methods, this paper considers establishing a classification model to calculate the accuracy of data classification after dimensionality reduction. As a classification method that can obtain high-precision results faster and repeatedly, random forest is insensitive to noise, and the overfitting problem of the decision tree is overcome. Therefore, this paper uses random forest model to quantify and compare the classification performance of three dimensionality reduction methods [9-10].

Random forest is a machine learning algorithm proposed by Leo Breiman in 2001. First, Bootstrap sampling is used to extract training sets $\{T_k, k=1,\cdots,K\}$, from the original sample set $\boldsymbol{X}$. Construct a decision tree for each training set. Based on the idea of feature subspace, each node in the decision tree is split. In this process, not all attributes are used for node splitting, but a subset of attributes is randomly selected, and node splitting is carried out in the best splitting way of this subset. The final output result is achieved by most voting methods.

Given the classifiers $h(\boldsymbol{X})$, the original sample set is $\{(x_i, y_i), x_i \in \boldsymbol{X}, y_i \in \boldsymbol{Y}\}$, where $\boldsymbol{X}$ is the sample set composed of M-dimensional vectors, $\boldsymbol{Y}$ contains j categories. Define the margin function as

$$mg(\boldsymbol{X},\boldsymbol{Y}) = av_k I(h_k(\boldsymbol{X}) = \boldsymbol{Y}) - \max_{j \neq Y} av_k I(h_k(\boldsymbol{X}) = j) \tag{15}$$

Where I is the indicator function. The generalization error is:

$$PE^* = P_{\boldsymbol{X},\boldsymbol{Y}}(mg(\boldsymbol{X},\boldsymbol{Y}) < 0) \tag{16}$$

4. Results

Since 3D results retain more original data information than 2D results, this paper uses the data scatter plot after 3D dimensionality reduction to analyze the results. As shown in Figure 2, the random forest classifier is used to classify the data after dimensionality reduction, and the effects of PCA, t-SNE and UMAP under each data set are calculated according to the classification accuracy of the test set. As shown in Figure 3, if there is little or no overlap of different colors in the scatter plot, this dimensionality reduction method can effectively distinguish different categories of data and has a strong ability to extract key information.

4.1. Dimensionality reduction results for each dataset

In the Car Information dataset, the scatter plots of the three-dimensionality reduction methods have a relatively insignificant effect on the classification of the three categories of data, but the three methods can well distinguish the vehicles of the category USA. From the perspective of the random forest classification accuracy rate, the accuracy of the three methods is lower than that of other datasets. t-SNE has the highest accuracy.

In the Iris dataset, scatter plots of UMAP and PCA can effectively distinguish points of three colors. UMAP has the highest classification accuracy while t-SNE is inferior to the other two methods regarding scatter plot distinguishing effect and accuracy.

In the Palmer penguin dataset, scatter plots of PCA and t-SNE can significantly distinguish three categories of penguin. In terms of classification accuracy, PCA and t-SNE also have higher accuracy. PCA has the highest accuracy. Therefore, PCA and t-SNE have better performance than UMAP.

In the Seeds dataset, the distinguishing effect of the scatter plot corresponding to UMAP is better than that of PCA and t-SNE. The classification accuracy of UMAP is also higher than that of the other two methods. In this dataset, UMAP has the best effect.

In the Wine Quality dataset, PCA has the best distinction effect and classification accuracy. In contrast, t-SNE has no obvious distinction effect and is slightly lower than PCA in terms of accuracy, while UMAP has weak performance.

Overall, the average classification accuracy of PCA, t-SNE and UMAP are 0.91, 0.86 and 0.79. PCA has the highest accuracy but has weak effect on Car Information dataset. The accuracy of t-SNE in each dataset is higher than 0.75. Although the overall effect is lower than PCA, t-SNE shows good stability. UMAP has the best effect in Iris and Seeds datasets, with the accuracy of 0.97 and 0.96. But the effect of UMAP in other datasets is weak.

4.2. Consistency of visualization results and classification accuracy

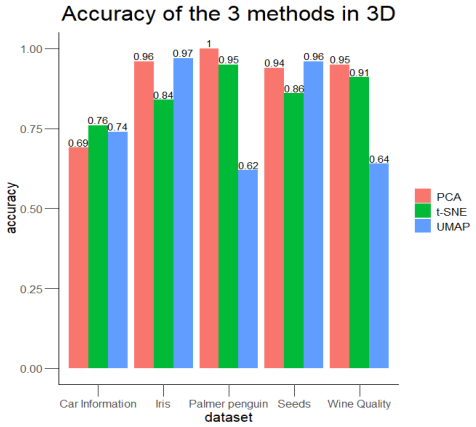

Figure 2. Random forest classification accuracy rate

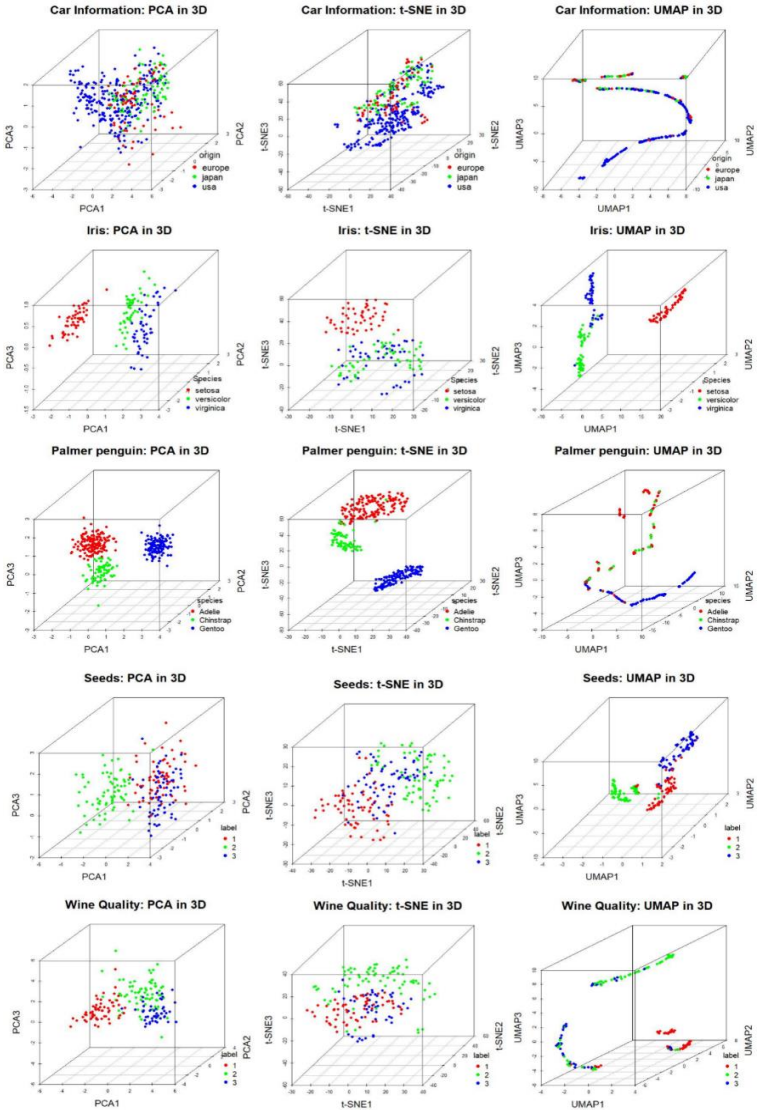

Figure 3. 3D dimensionality reduction scatter plot

This paper found that the visualization results of scatter plots after dimensionality reduction are highly consistent with the accuracy of random forest classification. For example, for the Car Information dataset, the scatter plots of three methods cannot effectively distinguish the three categories, and the classification accuracy of the three methods is significantly lower than that of other datasets. For the Iris dataset, the scatter plot of UMAP and PCA can significantly distinguish data points of three colors, and the corresponding classification accuracy is 0.97 and 0.96, higher than that of t-SNE. This consistency is not obvious in the PCA of Seeds and t-SNE of Wine Quality. In addition to these two cases, the distinguishing effect of the scatter plot and the corresponding classification accuracy show high consistency.

5. Conclusion

PCA, t-SNE and UMAP have their advantages and disadvantages. PCA shows the best classification accuracy, while t-SNE is the most stable method. UMAP has the best performance in Iris and Seeds datasets, but not effective enough in other datasets.

This paper finds and proves that in machine learning data mining, the effect of dimensionality reduction can be initially judged by the data scatter plot after dimensionality reduction. Thus, we can avoid establishing an evaluation model for each dimensionality reduction method, which reduces the calculation cost.

This paper evaluates the effects of PCA, t-SNE and UMAP from multiple perspectives such as dimensionality reduction visualization and random forest classification accuracy. It is concluded that PCA has a good dimensionality reduction effect on classical datasets while t-SNE is the most stable method. UMAP has good performance in individual datasets but lacks stability. At the same time, this paper proposes an evaluation method combining scatter graph visualization and random forest classification, which can judge the effectiveness of dimensionality reduction methods on the adopted dataset. The result in this paper provides perspective for solving the problem of data dimensionality reduction, which lacks evaluation methods.

References

[1] Mahesh B. Machine learning algorithms-a review [J]. International Journal of Science and Research (IJSR). [Internet], 2020, 9 (1): 381-386.

[2] Zebari R, Abdulazeez A, Zeebaree D, et al. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction [J]. Journal of Applied Science and Technology Trends, 2020, 1 (2): 56-70.

[3] Anowar F, Sadaoui S, Selim B. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne) [J]. Computer Science Review, 2021, 40: 100378.

[4] Shlens J. A tutorial on principal component analysis [J]. arXiv preprint arXiv: 1404.1100, 2014.

[5] Van der Maaten L, Hinton G. Visualizing data using t-SNE [J]. Journal of machine learning research, 2008, 9 (11).

[6] Arora S, Hu W, Kothari P K. An analysis of the t-sne algorithm for data visualization [C] // Conference on learning theory. PMLR, 2018: 1455-1462.

[7] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction [J]. arXiv preprint arXiv: 1802.03426, 2018.

[8] Allaoui M, Kherfi M L, Cheriet A. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study [C] // International conference on image and signal processing. Cham: Springer International Publishing, 2020: 317-325.

[9] Boateng E Y, Otoo J, Abaye D A. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review [J]. Journal of Data Analysis and Information Processing, 2020, 8 (4): 341-357.

[10] Wang Y S, Xia S T. A Survey of Random Forest Algorithms [J]. Information and communication technologies, 2018, 12 (01): 49-55.