Feature Extraction, Essential Dimension Estimation and Dimension Reduction Method of Colony Images

Wei-Zhong Sun, Fu-Jun Guo, Jie-Sheng Wang *, Lin Chen, Dong Wei, Xin-Feng Du

***Abstract*—A colony is a group of microorganisms produced by the growth and reproduction of a single microorganism species. A method of colony image feature extraction, essential dimension estimation and dimension reduction based on digital image processing technology was proposed. Firstly, based on the HSI image color moments (first, second and third order moments), nine color characteristics of the colony images were extracted. Based on gray-level co-occurrence matrix (GLCM), twenty texture features of colony images were obtained. Then, three essential dimension estimation methods, namely, the correlation dimension estimator, the maximum likelihood estimator and the packing numbers estimator, are used to estimate the dimension of the inner low-dimensional structure of the high-dimensional data of the colony images. The results of different dimension estimators were combined with six different dimension reduction techniques (PCA, LDA, MDS, ISOMAP, SNE and NCA) respectively to form 18 different data dimension reduction methods. Based on the distance criterion function, an optimal dimension reduction method for the high-dimensional feature data of the colony images corresponding to the maximum distance criterion function was obtained.**

***Index Terms*—Colony Image, Feature Extraction, Essential Dimension Estimation, Data Dimension Reduction**

I. Introduction

MICROORGANISMS refer to tiny biological individuals that humans cannot see or see clearly with the naked eyes and need to observe through a microscope. With the human exploration of nature, microbial technology has been derived at the microbial level. The study of microbiological technology often relies on the study of microbial images [1]. A colony is a microbial community formed by the growth and reproduction of a single microbial species. This development process is a transformation process from a world invisible to human eyes at the micro level to a world visible to human eyes at the macro level, which enable human eyes to observe microbial populations without the help of any auxiliary instruments [2]. If we want to obtain some pure microorganisms with target characteristics, it will inevitably face the problems of isolation and purification of microbial strains. Compared with the microscopic images, the colony image can provide convenience for the isolation and purification of strain to the greatest extent in both observation and operation. However, due to the characteristics of the process, the task is a significant amount of work, which requires the selection of hundreds or even thousands of colonies. The realization of intelligent classification of colony images will greatly improve the efficiency of related works. Colony image is an important research object in microbial related scientific research and quality detection, which can be divided into bacterial colony, mold colony and actinomycetes colony. In life, bacteria are extremely widespread and can be found almost anywhere, including in water, on biological surfaces, on mucous membranes and so on. Mold is more distributed in wet place, including soil, wet dead wood and so on. Most actinomycetes are found in the soil, where they produce an "earthy smell" after rain. In appearance, bacterial colonies is more moist, transparent, smooth and uniform in texture. Mold colonies are characterized by thick and long hyphae, dry appearance, opacity, spidery, villous or cotton-floss appearance. Actinomycetes colonies are compact, firm, dry and wrinkled. These features can help researchers to provide important reference for isolation, purification and medium selection of strains, and also provide the possibility for feature extraction and intelligent classification of colony images. Through the extraction of image features, many attribute feature variables about each image are obtained. With the increasing number of samples, the computational complexity and time also increase exponentially, which results in the "dimensional disaster" [3]. In order to reduce the amount of computation and improve the efficiency of operation, usually high-dimensional data will be processed by dimension reduction, the inner low-dimensional structure will be extracted, and the low-dimensional space vector will be constructed to describe the data samples in high-dimensional space, so as to transform high-dimension

Manuscript received November 18, 2020; revised February 18, 2021. This work was supported by the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. 2017FWDF10), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 2019-ZD-0031).

Wei-Zhong Sun is a lecturer of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (e-mail: weizhongsun@126.com).

Fu-Jun Guo is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (e-mail:1468239974@qq.com).

Jie-Sheng Wang is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. (Corresponding author, phone: 86-0412-2538246; fax: 86-0412-2538244; e-mail: wang_jiesheng@126.com).

Lin Chen is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: EDMM888888@163.com).

Dong Wei is an associate professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: asweidong@126.com).

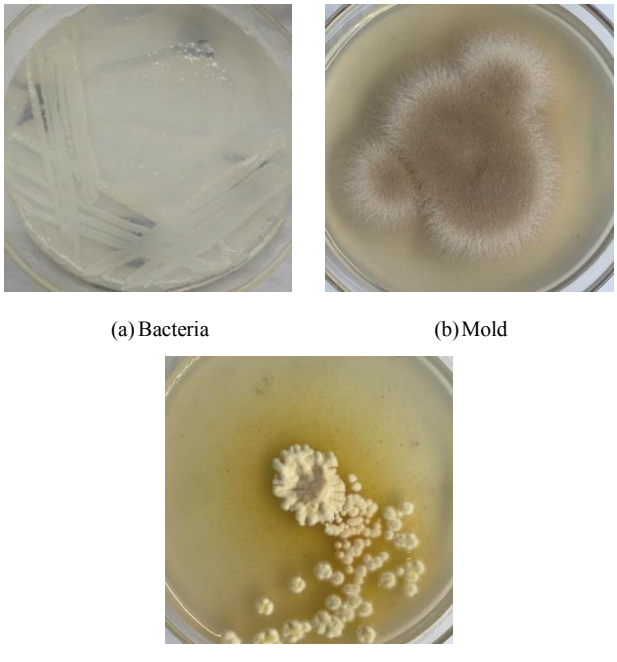
Xin-Feng Du is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China (e-mail: 13569510917@163.com).

data into low-dimensional data for operation. At the same time, the low-dimensional data is required to retain the effective information of the high-dimensional data for the research problem as much as possible. Therefore, this paper proposed a method of colony image feature extraction, essential dimension estimation and dimension reduction based on digital image processing technology, and verified the effectiveness of the proposed strategy through simulation experiments.

II. Colony Images and Feature Extraction

A. Colony Images

Microorganisms refer to the tiny individual organisms that cannot be seen or seen with the naked eye and need to be observed through a microscope, mainly including bacteria, mold and actinomycetes. The study of microbial technology often depends on the study of microbial images, and the bacterial colony is exactly a macroscopic microbe cell group that can be seen by naked eyes after the mass reproduction of a single microbe. Fig. 1 shows some examples of the above strains. Different strains of cells correspond to different growth conditions, which include culture time, temperature, PH, and culture medium. The culture medium is a nutrient substrate prepared by a combination of different nutrients, which is essential for the growth and reproduction of microorganisms. The morphological characteristics of the colony in the colony image are closely related to the separation and purification of different strains. In traditional methods, researchers will select colonies based on their own morphological characteristics of different colonies and their own experience. However, due to the different subjective understandings of researchers, the mastery of morphological characteristics is not comprehensive, which will inevitably cause human errors in the selection, classification and purification of colonies.



(a) Bacteria (b) Mold

(c) Actinomycetes

Fig. 1 Colony images.

Therefore, it is particularly important to classify colony images with the help of modern digital image technology. Modern digital image technology can realize the recognition and classification of different colony images. The important basis of this work is that different images of bacterial colonies have different image characteristics. The commonly used image features include color features, texture features and shape features.

B. Color Feature Extractions

Digital images exist in the form of a matrix in which each element corresponds to a pixel. According to color and grayscale, there are four basic types: binary image, grayscale image, index image and true-color RGB image. Among them, each pixel point of the true color RGB image is composed of R (red), G (green) and B (blue), and the intensity of the primary color is also determined by each primary color component. Their varying degrees of superposition make up RGB color patterns with different colors. Currently, this color pattern is one of the most widely used color systems. Any color can be represented by R, G and B. The three dimensional function can be expressed as follows.

$$f(x,y,z)=\left\{f_{red}(x,y,z),f_{green}(x,y,z),f_{blue}(x,y,z)\right\} \quad (1)$$

where, $f(x,y,z)$ is the color in space coordinate spot (x,y,z); $f_{red}(x,y,z)$, $f_{green}(x,y,z)$ and $f_{blue}(x,y,z)$ are respectively R, G and B component values in the spatial coordinate spot (x,y,z). However, for the colony image in the plane image, the position is determined by the two-dimensional coordinates, so the two-dimensional function of the position coordinates can be expressed as:

$$f(x,y)=\left\{f_{red}(x,y),f_{green}(x,y),f_{blue}(x,y)\right\} \quad (2)$$

where, $f(x,y)$ is the value of each pixel in graphic image.

Although RGB color model is commonly used and relatively simple, it is not the most suitable for the characteristics of the colony images studied in this paper. The colony image studied in this paper involves bacteria, mold and actinomycetes. As each category of bacteria, mold and actinomycetes has a variety of colonies of different colors. So color, as an external representation of the colony, cannot be distinguished from the types of bacteria, mold and actinomycetes simply by R, G and B trichromatic fractions. However, due to obvious differences in the brightness and transparency of the colony images, the conversion of color features from RGB mode to HSI mode can be used as an important basis for discrimination.

HIS color space is a color pattern more consistent with the characteristics of human vision, which reflects the way the human visual system perceives color, with Hue, Saturation and Intensity to describe color. Compared with RGB color space, HSI color space simplifies image analysis and processing. HIS color space model is shown in Fig. 2 [4].

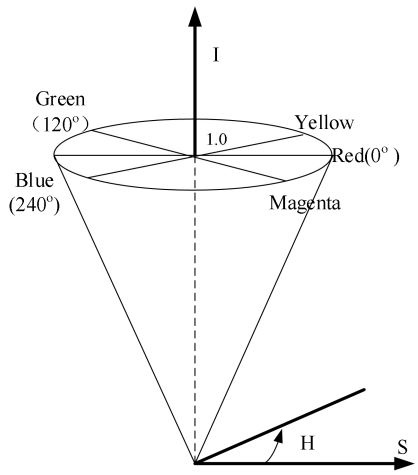


Fig. 2 HIS color space model.

In the HSI model, hue (*H)* is measured from 0 to 360 degrees. It can not only represent people's feelings towards different colors, such as red and green, but also represent a certain range of colors, such as warm colors and cool colors. For example 0° is pure red, 2π/3 is pure green, and 4π/3 is pure blue. Saturation (*S*) is the distance between any point in the color space and axis *I*, indicating the depth or intensity of the color. The brightness (*I*) is mainly affected by the light source and represents the degree of lightness and darkness of the color. In industrial applications, the value range of S [0,1] corresponds to unsaturated to full saturation (no white). The value range of I [0,1] corresponds to the color from dark to light. HSI color space and RGB color space have different representations of the same physical quantity, so they can be converted to each other.

$$I = \frac{R+G+B}{3} \tag{3}$$

$$S = 1 - \frac{3}{R+G+B}\left[\min\left(R,G,B\right)\right] \tag{4}$$

$$H = \cos^{-1}\left[\frac{\left(R-G\right)+\left(R-B\right)}{2\sqrt{\left(R-G\right)^2+\left(R-B\right)\left(G-B\right)}}\right] \tag{5}$$

$R \neq B \; or \; G \neq B$

As a representation method of color characteristics, color moments are proposed by Stricker and Orengo [5]. This method is mainly composed of three parts, namely, first-order moments (Mean), second-order moments (Variance) and third-order moments (Skewness). Since the color information distribution of an image is very important for image recognition, nine characteristic components of the color moment are used to represent the color characteristics of the image, namely the first, second and third moments of the H, S and I components respectively. First moment u_i, second moment σ_i and third moments s_i can be calculated by:

$$u_i = \frac{1}{N}\sum_{j=1}^{N} p_{i,j} \tag{6}$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{i,j}-u_i\right)^2\right)^{1/2} \tag{7}$$

$$s_i = \left(\frac{1}{N}\sum_{j=1}^{N}\left(p_{i,j}-u_i\right)^3\right)^{1/3} \tag{8}$$

where, N is the total pixel number of the image, and $p_{i,j}$ represents the occurrence probability of the pixel whose median value is j in i -th component of the HSI image. The feature data of color feature extraction is listed in Table 1.

C. Texture Feature Extractions

The texture of an image reflects the features of the image itself. Generally speaking, texture is a pattern with small shapes and regular arrangement in a certain range of the image, which is an important feature to describe the image. As can be seen from Fig. 1, the colony texture in the image is quite characteristic, so the statistical analysis method is adopted. Among the statistical analysis methods, the commonly used methods are histogram analysis and GLCM. Although histogram is relatively simple and intuitive, it is a measure of similarity, that is to say that even similar images may have different texture features.

TABLE 1. FEATURE DATA BASED ON COLOR FEATURE EXTRACTION

Samples	First order moment			Second order moment			Third order moment		
	H	S	I	H	S	I	H	S	I
1	0.1 04	0.3 37	0.3 08	0.0 46	0.2 84	0.3 01	0.2 22	0.1 01	0.3 18
2	0.1 12	0.1 02	0.5 07	0.0 21	0.0 22	0.3 50	0.2 41	0.1 08	0.3 42
3	0.1 13	0.1 01	0.5 24	0.0 20	0.0 20	0.3 48	0.2 39	0.1 10	0.3 22
4	0.1 10	0.1 99	0.5 85	0.0 17	0.0 79	0.1 92	0.2 84	0.0 95	0.2 34
5	0.1 09	0.1 94	0.5 92	0.0 14	0.0 77	0.2 16	0.2 68	0.1 04	0.1 84
6	0.1 08	0.1 51	0.4 95	0.0 11	0.0 37	0.2 35	0.2 45	0.0 83	0.2 06
7	0.1 05	0.1 87	0.4 78	0.0 13	0.0 82	0.2 84	0.2 69	0.0 83	0.1 49
8	0.1 17	0.0 99	0.5 65	0.0 24	0.0 20	0.3 81	0.2 36	0.1 15	0.2 60
9	0.0 94	0.1 99	0.3 89	0.0 13	0.0 85	0.3 01	0.2 38	0.1 22	0.2 45
10	0.7 18	0.1 64	0.8 09	0.7 77	0.2 85	0.6 26	0.2 15	0.0 91	0.2 49
11	0.6 32	0.2 03	0.7 51	0.9 23	0.3 34	0.6 19	0.2 33	0.0 91	0.2 86
12	0.7 47	0.1 78	0.7 25	0.7 68	0.2 86	0.6 05	0.2 09	0.0 91	0.2 56
13	0.5 23	0.2 11	0.6 98	0.9 34	0.3 16	0.5 16	0.2 45	0.0 97	0.2 49
14	0.1 53	0.1 39	0.6 07	0.3 64	0.1 98	0.4 06	0.3 51	0.0 85	0.2 69
15	0.1 03	0.1 75	0.5 31	0.0 15	0.0 64	0.1 16	0.2 66	0.0 70	0.1 22
16	0.1 65	0.0 79	0.7 12	0.4 65	0.1 96	0.4 10	0.3 16	0.1 10	0.1 14
17	0.1 61	0.1 23	0.6 12	0.4 01	0.1 95	0.4 42	0.3 39	0.0 92	0.2 28
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	0.1 21	0.0 95	0.7 16	0.3 49	0.2 61	0.2 63	0.3 60	0.1 31	0.0 94

However, in the spatial position, the repeated occurrence of grayscale forms the texture. Although there is a certain distance between two pixels in the image, there will still be grayscale relationship, which is called the grayscale spatial correlation property. GLCM was proposed by Haralick et al. [6] to describe an analysis method of texture, which is based on the second-order combinational conditional probability density function of image estimation. The joint probability density of two pixels in the image space is defined as the symbiotic matrix, which is used as the basis for defining a set of texture features. Fig. 3 is a schematic diagram of grayscale symbiosis matrix, where i and j represent the grayscale values of corresponding pixel points [4].

Set $f(x,y)$ is a two-dimensional digital image, and the GLCM refers to the pixel with the gray level i in the image $f(x,y)$ appears at the same time with the pixel $(x+\Delta x, y+\Delta y)$ having the deflection angle θ, the distance δ, and the value j at the probability $P(i,j,\delta,\theta)$.

$$P(i,j,\delta,\theta)=\left\{\left[(x,y),(x+\Delta x,y+\Delta y)\right]\middle| f(x,y)=i,\ f(x+\Delta x,y+\Delta y)=j; x=0,1,\cdots,N_x-1; y=0,1,\cdots,N_y-1\right\} \tag{9}$$

where, $i,j=0,1,\cdots,L-1$; x and y are the image pixel coordinates; L is the gray series; N_x and N_y are the number of lines and columns in the image.

It is not intuitive to use symbiotic matrix to describe texture features. Therefore, some parameters reflecting the condition of the matrix are derived by using symbiotic matrix, which can describe texture features from different angles. Because of the abundant characteristic parameters, gray co-occurrence matrix can describe the texture from different angles. At first, Haralick[7] gave 14 characteristic parameters, such as contrast, uniformity, entropy, and variance, and average, energy, correlation, and entropy based on GLCM. In this paper, 20 texture feature parameters, such as angular second moment, entropy and non-similarity were used to describe the texture features of bacterial colony images. The calculation formula is shown in Table 2. Table 3 and Table 4 shows the texture feature parameters based on GLCM.

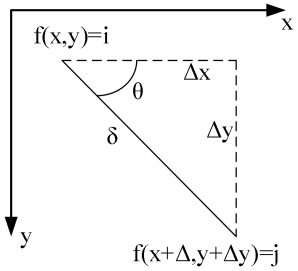


Fig. 3 Gray-level co-occurrence matrix.

TABLE 2. TEXTURE FEATURE CALCULATION METHODS BASED ON GLCM

Texture feature	Calculation method
Angular second moment	$f_1=\sum_i\sum_j\{p(i,j)\}^2$
Entropy	$f_2=-\sum_i\sum_j p(i,j)\log(p(i,j))$
Dissimilarity	$f_3=\sum_{n=0}^{N_g-1}\lvert i-j\rvert\cdot\left\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}p(i,j)\right\}$
Contrast	$f_4=\sum_{n=0}^{N_g-1}\lvert i-j\rvert^2\left\{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}p(i,j)\right\}$
Uniformity	$f_5=\sum_i\sum_j\frac{1}{1+(i-j)}p(i,j)$
Correlation	$f_6=\frac{\sum_i\sum_j(ij)p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$
Moment of deficit	$f_7=\sum_i\sum_j\frac{1}{1+(i-j)^2}p(i,j)$
Auto-correlation	$f_8=\sum_i\sum_j i\cdot j\cdot p(i,j)$
Cluster shadow	$f_9=\sum_i\sum_j(i-\mu+j-\mu)^3p(i,j)$
Cluster protrusion	$f_{10}=\sum_i\sum_j(i-\mu+j-\mu)^4p(i,j)$
Maximum probability	$f_{11}=\max_i\left\{\max_j p(i,j)\right\}$
Sum variance	$f_{12}=\sum_i\sum_j(i-\mu)^2p(i,j)$
Sum average	$f_{13}=\sum_{i=2}^{2N_g}ip_{x+y}(i)$
Sum variance	$f_{14}=\sum_{i=2}^{2N_g}(i-f_{15})^2p_{x+y}(i)$
Sum entropy	$f_{15}=-\sum_{i=2}^{2N_g}p_{x+y}(i)\log\{p_{x+y}(i)\}$
Variance	$f_{16}=variance\ of\ p_{x-y}$
Differential entropy	$f_{17}=-\sum_{i=0}^{N_g-1}p_{x-y}(i)\log\{p_{x-y}(i)\}$
Related information measures	$f_{18}=\frac{HXY-HXY1}{\max\{HX,HY\}}$ $f_{19}=\left(1-\exp\left[-2.0(HXY2-HXY)\right]\right)^{\frac{1}{2}}$ $HXY=-\sum_i\sum_j p(i,j)\log(p(i,j))$ $HXY1=-\sum_i\sum_j p(i,j)\log\{p_x(i)p_y(j)\}$ $HXY2=-\sum_i\sum_j p_x(i)p_y(i)\log\{p_x(i)p_y(j)\}$
Maximum correlation Coefficient	$f_{20}=(Second\ largest\ eigenvalue\ of\ Q)^{\frac{1}{2}}$ $Q(i,j)=\sum_k\frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$

TABLE 3. TEXTURE FEATURE DATA BASED ON GLCM (1)

Sample	Angular second moment	Entropy	Dissimilarity	Contrast	Uniformity	Correlation	Moment of deficit	Auto-correlation	Cluster shadow	Cluster protrusion
1	0.012	0.908	0.563	0.460	0.407	0.544	0.118	0.614	0.478	0.613

2	0.040	0.689	0.166	0.064	0.800	0.936	0.263	0.811	0.640	0.850
3	0.052	0.652	0.149	0.053	0.820	0.947	0.291	0.807	0.609	0.835
4	0.131	0.514	0.202	0.161	0.786	0.840	0.528	0.765	0.651	0.884
5	0.233	0.418	0.138	0.116	0.856	0.884	0.692	0.820	0.612	0.908
6	0.046	0.658	0.282	0.143	0.674	0.856	0.205	0.764	0.540	0.664
7	0.069	0.584	0.315	0.169	0.640	0.829	0.228	0.721	0.514	0.703
8	0.056	0.613	0.161	0.058	0.805	0.942	0.285	0.792	0.569	0.788
9	0.021	0.801	0.303	0.163	0.653	0.835	0.162	0.713	0.588	0.662
10	0.028	0.737	0.312	0.143	0.635	0.858	0.155	0.868	0.537	0.729
11	0.020	0.789	0.315	0.159	0.636	0.844	0.141	0.889	0.480	0.763
12	0.020	0.772	0.331	0.174	0.620	0.828	0.137	0.859	0.518	0.712
13	0.017	0.804	0.348	0.177	0.599	0.823	0.119	0.846	0.383	0.689
14	0.045	0.729	0.244	0.141	0.725	0.861	0.343	0.904	0.390	0.798
15	0.056	0.634	0.273	0.155	0.689	0.844	0.232	0.768	0.557	0.767
16	0.068	0.599	0.163	0.085	0.813	0.915	0.432	0.893	0.485	0.902
17	0.052	0.687	0.221	0.117	0.746	0.884	0.324	0.877	0.437	0.839
18	0.058	0.659	0.275	0.151	0.687	0.850	0.340	0.838	0.506	0.855
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	0.066	0.644	0.326	0.207	0.638	0.794	0.316	0.796	0.477	0.821

TABLE 4. TEXTURE FEATURE DATA BASED ON GLCM (2)

Sample	Maximum probability	Sum variance	Sum average	Sum variance	Sum entropy	Variance	Differential entropy	Related information measures 1	Related information measures 2	Maximum correlation
1	0.062	0.675	0.875	0.108	0.125	0.463	0.220	0.833	0.559	0.586
2	0.051	0.456	0.833	0.361	0.400	0.067	0.593	0.471	0.923	0.973
3	0.082	0.422	0.824	0.398	0.438	0.056	0.619	0.441	0.931	0.976
4	0.159	0.627	0.835	0.377	0.417	0.163	0.621	0.386	0.932	0.861
5	0.302	0.517	0.871	0.457	0.499	0.117	0.736	0.253	0.964	0.912
6	0.038	0.422	0.840	0.280	0.314	0.148	0.444	0.590	0.836	0.940
7	0.062	0.329	0.821	0.419	0.460	0.174	0.423	0.588	0.818	0.930
8	0.051	0.371	0.816	0.453	0.494	0.061	0.600	0.447	0.921	0.978
9	0.037	0.333	0.808	0.149	0.171	0.169	0.423	0.666	0.797	0.809
10	0.055	0.612	0.930	0.240	0.273	0.149	0.395	0.633	0.816	0.925
11	0.038	0.759	0.949	0.167	0.193	0.165	0.391	0.657	0.804	0.891
12	0.039	0.776	0.927	0.216	0.246	0.180	0.387	0.657	0.800	0.924
13	0.038	0.577	0.933	0.271	0.306	0.184	0.360	0.666	0.798	0.955
14	0.079	0.688	0.959	0.192	0.221	0.144	0.498	0.516	0.905	0.914
15	0.050	0.452	0.850	0.221	0.251	0.159	0.461	0.588	0.831	0.803
16	0.106	0.580	0.921	0.379	0.420	0.088	0.620	0.393	0.943	0.987
17	0.077	0.676	0.916	0.309	0.346	0.121	0.521	0.503	0.905	0.967
18	0.106	0.581	0.908	0.387	0.427	0.156	0.477	0.507	0.897	0.977
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	0.155	0.598	0.900	0.335	0.374	0.212	0.433	0.584	0.837	0.954

III. Essential Dimension Estimation of High Dimensional Features

By extracting color features and texture features from all samples, image feature data containing a lot of feature variable information are obtained, and these data have high dimensional characteristics. It will cause huge computing complexity and storage space overhead. High-dimensional data generally contain a lot of redundant information relative to the target problem. Such high-dimensional characteristics often hide the internal relations and laws behind the data, which is not conducive to discovering the data relations and laws, and it becomes very difficult to process high-dimensional data. Therefore, it is necessary to find a suitable way to effectively process the high-dimensional data and turn it into the low-dimensional data closely related to the high-dimensional data to be processed. Such low-dimensional data can better highlight the relationships and laws among variables hidden in the high-dimensional data. But the determination of the "essential" dimension of the low-dimensional structure hidden in the high-dimensional data is the key to the dimension reduction. The ability to accurately find out the intrinsic "essential" dimension of high dimensional data determines the ability to better map high dimensional data to low dimensional data.

Assuming R^p have n samples in high dimensional space represented by $X_1, X_2, \cdots, X_n$, while in low dimensional space $R^m (m << p)$ represented by $Y_1, Y_2, \cdots, Y_n$. The mapping from high-dimensional space to low-dimensional space can be expressed as $X_i = g(Y_i)$, where $i = 1, 2, \cdots, n$ and m is the essential dimension.

Because the research of high-dimensional data processing involves the development of many industries and fields, it has been widely concerned by scholars at home and abroad. The problem of essential dimension estimation is one of the important research contents in the field of high-dimensional data processing. As early as 1969, the definition of "essential" dimension was proposed by Bennett [8], who applied principal component analysis (PCA) to the estimation of essential dimension of linear data. Later, scholars made innovations and improvements on this basis, and produced a series of PCA based methods applied to global or local applications, which are called eigenvalue methods. After entering the 21st century, with the continuous development of manifold learning, people have a deeper understanding of essential dimension, and put forward many essential dimension estimation methods, such as correlation dimension estimator [9], maximum likelihood estimator [10], and cluster number estimator [11].

A. Estimation of Correlation Dimension

The correlation dimension estimator (CDE) calculates the relative number of data points within the hypersphere with radius r and use the number of data points within the hypersphere in the radius r proportional to the r^d. The relative number of data points in the hypersphere with radius r is calculated by:

$$C(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} c$$
$$where \quad c = \begin{cases} 1, if \left\| x_i - x_j \right\| \le r \\ 0, if \left\| x_i - x_j \right\| > r \end{cases} \tag{10}$$

Due to the value $C(r)$ is proportional to the r^d, $C(r)$ is can be used to estimate the intrinsic dimension d. The proper dimension d is given by the limit.

$$d = \lim_{r \to 0} \frac{\log C(r)}{\log r} \tag{11}$$

Since the solution in Eq. (11) cannot be directly obtained, $C(r)$ can be calculated by two r values. The essential dimension d of data is given by the ratio shown in Eq. (12).

$$\hat{d} = \frac{\log\left(C(r_2) - C(r_1)\right)}{\log(r_2 - r_1)} \tag{12}$$

B. Maximum Likelihood Estimate

The maximum likelihood estimation (MLE) can realize the essential dimension estimation of high-dimensional data. Suppose that given a point x, within a sufficiently small range of R and a sphere $S_x(R)$ with its radius, there is f(x) ≈ C(C is the fixed value).

$$\left\{N(t,x), 0 \le t \le R\right\}, N(t,x) = \sum_{i=1}^{n} I\left\{X_i \in S_x(t)\right\} \tag{13}$$

It can be seen that N(t,x) is the number of points $X_1, X_2, \cdots, X_n$ falling into $S_x(t)$. $N(t,x) = \sum_{i=1}^{n} I\left\{X_i \in S_x(t)\right\}$ is similar to Poisson process. For $X_1, X_2, \cdots, X_n$, the distance from the k -th nearest neighbor in $X_1, X_2, \cdots, X_n$ to x is denoted as $T_k(x)$. Then obtain:

$$\frac{k}{n} \approx f(x) V(d) \left[T_k(x)\right]^d \tag{14}$$

where, $V(d) = \pi^{d/2} [\Gamma(d/2) + 1]^{-1}$ is the d -dimensional unit sphere volume. Assuming that t is fixed, $\lambda(t) = f(x) V(d) dt^{d-1}$, where $\lambda(t)$ is the relative rate of change of $N(t)$ to t; $\theta = \ln f(x)$. Then Eq. (15) can be established [12].

$$\ln L(d, \theta) = \ln\left(\int_0^R \ln \lambda(t) dN(t) - \int_0^R \lambda(t) dt\right) \tag{15}$$

Eq. (15) satisfies the following likelihood equation.

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{\int_0^R \ln \lambda(t) dN(t) - \int_0^R \lambda(t) dt} \cdot$$
$$\left(\int_0^R dN(t) - \int_0^R \lambda(t) dt\right) = 0 \tag{16}$$

It can be reduced to:

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{\int_0^R \ln \lambda(t) dN(t) - \int_0^R \lambda(t) dt} \cdot$$
$$\left(N(R) - e^{\theta} V(d) R^d\right) = 0 \tag{17}$$

By $\int_0^R \ln\lambda(t)dN(t)-\int_0^R \lambda(t)dt \neq 0$, obtain:

$$N(R)-e^{\theta}V(d)R^{d}=0 \tag{18}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial d}=&\frac{1}{\int_0^R \ln\lambda(t)dN(t)-\int_0^R \lambda(t)dt}\\ &\cdot\left\{\left(\frac{1}{d}+\frac{V(d)}{V(d)}\right)N(R)+\int_0^R \ln t dN(t)\right.\\ &\left.-e^{\theta}V(d)R^{d}\left(\ln R+\frac{V(d)}{V(d)}\right)\right\}=0\end{aligned} \tag{19}$$

By $\int_0^R \ln\lambda(t)dN(t)-\int_0^R \lambda(t)dt \neq 0$, obtain:

$$\begin{aligned}&\left(\frac{1}{d}+\frac{V(d)}{V(d)}\right)N(R)+\\ &\int_0^R \ln t dN(t)-e^{\theta}V(d)R^{d}\left(\ln R+\frac{V(d)}{V(d)}\right)=0\end{aligned} \tag{20}$$

By combining Eq. (18) and Eq. (20), obtain:

$$\hat{d}_R(x)=\left[\frac{1}{N(R,x)}\sum_{j=1}^{N(R,x)}\ln\frac{R}{T_j(x)}\right]^{-1} \tag{21}$$

In order to facilitate the calculation, the spherical neighborhood is discarded and the neighborhood k is taken for calculation, so Eq. (21) is transformed into:

$$\hat{d}_k(x)=\left[\frac{1}{k-1}\sum_{j=1}^{k-1}\ln\frac{T_k(x)}{T_j(x)}\right]^{-1} \tag{22}$$

where, $T_k(x_i)$ represents the radius of a hyperplane centered on x_i encompassing k neighboring points.

C. Packing Numbers Estimation

The packing numbers estimator (PNE) is based on the number $N(r)$ covered by the radius r of the hypersphere covering all data points in the data set X , which is proportional to r^{-d} . Among them, this hypersphere is a collection covering all data points in the data set X . Just because $N(r)$ is proportional to r^{-d} , the essential dimension of the data set X can be obtained by:

$$d=-\lim_{r\to 0}\frac{\log N(r)}{\log r} \tag{23}$$

Generally, it is difficult to obtain the number $N(r)$ covered by the radius r in the data set X . Therefore, the number of clusters $M(r)$ under the radius r is used instead of $N(r)$ to solve this difficulty. This cluster is defined as the largest cluster of independent subsets in the data set X . In other words, $M(r)$ is the maximum number of data points in X that can be covered by a single hypersphere with radius r . In a data set of reasonable size, it is feasible to use $M(r)$ to complete the calculation. The essential dimension of the data set X can be obtained by the following limit.

$$d=-\lim_{r\to 0}\frac{\log M(r)}{\log r} \tag{24}$$

Since the essential dimension cannot be calculated by the limit in Eq. (24), Eq. (25) can be used to estimate the essential dimension.

$$\hat{d}=-\frac{\log\left(M(r_2)-M(r_1)\right)}{\log(r_2-r_1)} \tag{25}$$

IV. Dimension Reduction Method for High Dimensional Characteristic Data

The essential dimension estimation methods of high-dimensional data introduced in the previous section can be used to obtain the essential dimension of high-dimensional data set. Next, the application and discussion of the dimension reduction methods for high-dimensional data will be carried out. Common dimension reduction methods include Principal Component Analysis (PCA) [13], Linear Discriminant Analysis (LDA) [14], Multi-dimensional Scaling (MDS) [15], Isomapetric Mapping (ISOMAP) [16], Stochastic Neighbor Embedding (SNE) [17], Neighborhood Component Analysis (NCA) [18], etc.

A. PCA Dimension Reduction

Principle component analysis (PCA) reduces the dimension of high-dimensional data to low-dimensional data. It is to extract a set of basis vectors that express the variable information of the high-dimensional original data to the maximum extent, and then select the size of each sample point in the high-level feature space. Several feature directions with the largest changes are used as projection directions to achieve feature extraction of data. Suppose sample data $X=(x_1,x_2,\cdots,x_n)$ in d -dimensional space, where n is the number of samples. The mean value of the samples can be obtained by:

$$\bar{x}=E(X)=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{26}$$

PCA adopts the orthogonal transformation on variable X to obtain a set of linearly independent variables $Y=(y_1,y_2,\cdots,y_m)$.

$$\begin{cases}y_1=\theta_{11}x_1+\theta_{12}x_2+\cdots+\theta_{1n}x_n\\ y_2=\theta_{21}x_1+\theta_{22}x_2+\cdots+\theta_{2n}x_n\\ \vdots\\ y_m=\theta_{m1}x_1+\theta_{m2}x_2+\cdots+\theta_{mn}x_n\end{cases} \tag{27}$$

where, $\theta_i=(\theta_{i1},\theta_{i2},\cdots,\theta_{im})^T$ is the unit projection vector on subspace, where $i=1,2,\cdots,m$. Arrange Eq. (27) to obtain:

$$\begin{aligned}&y=\sum_{i=1}^{n}\theta_i x_i=\theta^T x\\ &s.t.\quad \|\theta\|=\sqrt{\theta_i\theta_j^T}=1,\quad i=j\\ &\qquad\ \theta_i\theta_j^T=0,\qquad\ \ i=j\end{aligned} \tag{28}$$

The core idea of this method is to find the direction of the largest variance, that is to say $E\left[y^2\right]$ is maximized as much as possible.

$$E\left[y^2\right]=E\left[\left(\theta^T X\right)^2\right]=\theta^T E\left[xx^T\right]\theta=\theta^T C\theta \tag{29}$$

where, C represents the covariance matrix corresponding to the sample data, which is defined as:

$$C=\begin{bmatrix}\operatorname{cov}\left(x_1,x_1\right) & \operatorname{cov}\left(x_1,x_2\right) & \cdots & \operatorname{cov}\left(x_1,x_n\right)\\ \operatorname{cov}\left(x_2,x_1\right) & \operatorname{cov}\left(x_2,x_2\right) & \cdots & \operatorname{cov}\left(x_2,x_n\right)\\ \vdots & \vdots & & \vdots\\ \operatorname{cov}\left(x_n,x_1\right) & \operatorname{cov}\left(x_n,x_2\right) & \cdots & \operatorname{cov}\left(x_n,x_n\right)\end{bmatrix} \tag{30}$$

where, $\operatorname{cov}\left(x_i,x_j\right)=E\left[\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)^{'}\right]$. In order to make $E\left[y^2\right]$ reach the maximum value, θ also needs to satisfy the following conditions:

$$C\theta_i=\lambda_i\theta_i,\quad i=1,2,\cdots,n \tag{31}$$

where, θ_i is uniquely determined by matrix C, which is corresponding to the eigenvector obtained by matrix C decomposition. Generally, when eigenvectors are used to solve practical problems, not all eigenvectors are involved in the problem, and most of the sample information can be expressed by using only some principal components with a large contribution rate. The contribution rate e of the principal component determines the number of retained feature vectors m, and the calculation formula is described as follows.

$$e=\frac{\lambda_1+\lambda_2+\cdots+\lambda_m}{\sum_{i=1}^{n}\lambda_i}\times 100\% \tag{32}$$

The eigenvectors corresponding to the first m largest eigenvalues are thus obtained, and the new eigenspace completes the mapping from the original high-dimensional data to the low-dimensional space. This mapping can be expressed as:

$$Y=PX \tag{33}$$

where, X is in the high-dimensional space, Y is in the low-dimensional space, and P is a matrix composed of m eigenvectors.

B. LDA Dimension Reduction

Linear discriminant analysis (LDA) reduces data from high-dimensional space to low-dimensional space by looking for low-dimensional features with the most categorical information. These characteristics need to meet certain conditions. While taking into account that the sample data of the same type should be as compact as possible, the sample data of different types should be dispersed as much as possible. In order to meet the above conditions, it is necessary to find a transformation matrix W_0 to satisfy the Fisher criterion.

$$R\left(W_0\right)=\arg\max\frac{\left|W^T S_m W\right|}{\left|W^T S_n W\right|} \tag{34}$$

In Eq. (34), the inter-class dispersion S_m and the intra-class dispersion S_n are given by:

$$S_m=\sum_{i=1}^{c}\frac{k_i}{k}\left(\overline{x_i}-\bar{x}\right)\left(\overline{x_i}-\bar{x}\right)^T \tag{35}$$

$$S_n=\sum_{i=1}^{c}\sum_{j=1}^{k_i}\frac{1}{k}\left(x_{ij}-\overline{x_i}\right)\left(\overline{x_{ij}}-\overline{x_i}\right)^T \tag{36}$$

where, c represents the total number of sample categories, and the mean of all samples $\bar{x}$ is defined as:

$$\bar{x}=\frac{1}{k}\sum_{i=1}^{k}x_i \tag{37}$$

Eq. (37) is the generalized Rayleigh entropy of S_m relative to S_n. When S_m is positive, the extreme value of $R(W)$ is obtained on the ellipsoid $W^T S_n W=1$. At this time, there is the characteristic equation $S_m W=\lambda S_n W$, where the characteristic vector corresponding to the characteristic value λ is W, and there is $R(W)=\lambda$. Remember that the optimal solution of $R(W)$ is $W_0=\left(w_1,w_2,\cdots,w_s\right)$, the corresponding eigenvectors of the first s non-zero eigenvalues $\lambda_1,\lambda_2,\cdots,\lambda_s$ of S_m/S_n are $w_1,w_2,\cdots,w_s$.

C. NCA Dimension Reduction

Nearest neighbor analysis (NCA) is to measure the distance on high-dimensional data sets, and use this distance measure $d\left(x_i,x_j\right)$ to characterize a certain similarity between x_i and x_j, so as to achieve the purpose of dimension reduction. The algorithm randomly selects nearest neighbors, and calculates the transformation matrix in the Mahalanobis distance by optimizing the left-one-out (LOO) cross-check result. Suppose there are n input samples $x_1,x_2,\cdots,x_n$ in the R^D space, and the corresponding class labels $c_1,c_2,\cdots,c_n$ are attached, and a distance measure is found to make the nearest neighbor classification effect as optimal as possible. The Mahalanobis distance transformation matrix Q is a symmetric positive semi-definite matrix, that is $Q=A^T A$. The Mahalanobis distance between two sample points can be calculated by:

$$\begin{aligned}d\left(x_i,x_j\right)&=\sqrt{\left(x_i,x_j\right)^T Q\left(x_i,x_j\right)}\\&=\sqrt{\left(Ax_i-xA_j\right)^T Q\left(Ax_i-Ax_j\right)}\end{aligned} \tag{38}$$

where, $i,j=1,2,\cdots,n$. When using LOO to calculate the classification error, a differentiable Softmax function needs to be introduced:

$$P_{ij}=\frac{\exp\left(-\left\|Ax_i-Ax_j\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|Ax_i-Ax_k\right\|^2\right)},\quad p_{ij}=0 \tag{39}$$

The relevant definition of P_{ij} is described as follows. The sample point x_i selects a neighbor in a random manner, and

it finally selects another sample point x_j as its neighbor and inherits the probability of its class label c_j. The probability that the sample point x_i is correctly classified can be calculated by:

$$p_i = \sum_{j \in C_i} p_{ij} \tag{40}$$

where, $C_i = \{ j | c_i = c_j \}$.

Since the objective function needs to maximize the number of correctly classified points, it is defined as:

$$f(A) = \sum_i \sum_{j \in C_i} P_{ij} = \sum_i p_i \tag{41}$$

Eq. (41) is a continuously differentiable matrix function, and A can be obtained by stochastic gradient method or conjugate gradient method. The gradient method can be expressed as:

$$\frac{\partial f}{\partial A} = -2A \sum_i \sum_{j \in C_i} p_{ij} \left(x_{ij} x_{ij}^T - \sum_k p_{ik} x_{ik} x_{ik}^T \right) \tag{42}$$

where, $x_{ij} = x_i - x_j$. When matrix A is a square matrix, the dimension after learning from the NCA distance measure will remain unchanged.

D. MDS Dimension Reduction

Multidimensional scaling (MDS) is a more classic distance-preserving dimension reduction method. Its core is to minimize the relative position and difference between sample points after projecting high-dimensional variables. In other words, the distance or similarity between the sample points in the low-dimensional space and the high-order space should be as consistent as possible. For the data set $X = \{x_1, x_2, \cdots, x_n\} \in R^D$, MDS usually uses distance information to represent the difference between samples, that is, the smaller the distance, the similarity between samples will be higher. Its distance expression in high-level space is described as:

$$d_{ij} = \left[\sum_{m=1}^{D} \left| X_{im} - X_{jm} \right|^r \right]^{1/r} \tag{43}$$

where, m is the feature dimension of the samples in the high-level space, and when r is 2, it represents the Euclidean distance.

According to Eq. (43), the distance matrix $D = (d_{ij})_{n \times n}$ of n samples in the data set X in the high-dimensional space is calculated, and the characteristic of the samples in the low-dimensional space is represented as $Y = \{y_1, y_2, \cdots, y_n\}$. Because it is necessary to maintain the consistency of the distance between the samples before and after the transformation, that is to say that $d_{ij} = \| y_i - y_j \|$ must be satisfied. So Eq. (43) can be transformed into:

$$\begin{aligned} d_{ij} &= (y_i - y_j)^T (y_i - y_j) \\ &= y_i^T y_i - y_i^T y_j - y_j^T y_i - y_j^T y_j \\ &= y_i^T y_i - 2 y_i^T y_j + y_j^T y_j \end{aligned} \tag{44}$$

It can be seen that the distance in the high-dimensional space can be represented by the inner product of the samples in the low-dimensional space. Assuming that the sample data after dimension reduction is the data after the centering processing and the inner product matrix is A, where $a_{ij} = y_i^T y_j$, then

$$\sum_{i=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ij} = 0 \tag{45}$$

The inner product matrix A in the low-dimensional space can be expressed as:

$$a_{ij} = -\frac{1}{2} \left(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2 \right) \tag{46}$$

Among them:

$$d_{i\cdot}^2 = \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 \tag{47}$$

$$d_{\cdot j}^2 = \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 \tag{48}$$

$$d_{\cdot\cdot}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 \tag{49}$$

Inner product matrix $A = Y^T Y$. Perform eigen decomposition of matrix A to solve sample data Y in low-dimensional space.

$$A = U \Lambda U^T = Y^T Y \tag{50}$$

$$Y = \Lambda^{1/2} U^T \tag{51}$$

$$\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_d) \tag{52}$$

where, Λ is a diagonal matrix composed of eigenvalues, and its eigenvalues are arranged from large to small, namely $\lambda_1, \lambda_2, \cdots, \lambda_d$; U is a matrix composed of eigenvectors corresponding to the eigenvalues in Λ, and all feature vectors are pairwise orthogonal.

E. ISOMAP Dimension Reduction

Isometric mapping (ISOMAP) is an improvement based on MDS. Its purpose is to keep the distance between two data sample points as much as possible. This distance uses the Geodesic distance $d_G(x_i, x_j)$, not Euclidean the distance $d(x_i, x_j)$. After the low-dimensional embedding mapping is obtained by the MDS algorithm, the error value between Geodesic distance $d_G(x_i, x_j)$, and Euclidean distance $d(x_i, x_j)$ is minimized, so that the low-dimensional embedding coordinates of the high-dimensional space can be obtained, and the purpose of dimension reduction is realized.

Calculate the Euclidean distance $d(x_i, x_j)$ between all sample points in the data sample set $X = [x_1, x_2, \cdots, x_n]^T \in R^{n \times m}$, and determine the distance between the sample point's proximity. A nearby graph G with a sampling point as a node is used to represent the neighborhood relationship of the sampling point. Each node

in G corresponds to each point in the high-dimensional data set, and the connected edges represent its neighbor relationship. By using Dijkstra's algorithm on the nearest neighbor graph G, we can get the shortest distance $d_G(x_i, x_j)$ between each node in G, and the geodesic distance matrix $D_G(i,j) = \{d_G(x_i, x_j), i, j = 1, 2, \cdots, n\}$. The MDS algorithm is used to solve the low-dimensional embedding in high-level sampling data.

F. SNE Dimension Reduction

Stochastic neighbor embedding (SNE) adopts the affine transformation to map data points to a probability distribution. First convert the Euclidean distance to conditional probability to express the similarity between two points. The number of samples in the high-dimensional sample space is N, and the high-dimensional data is $x_1, x_2, \cdots, x_N$. The probability of x_i and x_j is calculated for all data point combinations.

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / \left(2\sigma_i^2\right)\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / \left(2\sigma_i^2\right)\right)} \tag{53}$$

The parameter σ_i here will vary with different data points x_i, and the parameter can be searched for an optimal σ_i by means of binary search. For low-dimensional y_i, you can specify that the variance of the Gaussian distribution is $1/\sqrt{2}$, and their similarity can be defined as:

$$q_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2\right)} \tag{54}$$

If the dimension reduction effect is better, then $p_{j|i} = q_{j|i}$. Therefore, the objective function is:

$$\varphi(Y) = \sum_{ij} p_{ij} \log \frac{p_{j|i}}{q_{j|i}} \tag{55}$$

Make the acquired value of the objective function $\varphi(Y)$ as small as possible.

V. Dimension Reduction Results and Analysis

A. Dimension Reduction Performance Evaluation Index

The performance evaluation index of the dimension reduction methods can use the distance criterion function J between the classes and the samples within the class.

$$J = tr\left[S_W^{-1} S_B\right] \tag{56}$$

where, S_W is the within-class dispersion matrix, S_B is the between-class dispersion matrix, and $tr[\cdot]$ represents the trace of the solution matrix. The J value is used to reflect the intra-class distance and the inter-class distance of the data. When the intra-class distance of the data is large and the intra-class distance is small, the current data has better separability. The J value is larger, and the dimension reduction effect is better. Through the study of 3 kinds of common high-dimensional data essential dimension estimation methods and 6 common dimension reduction techniques, a more suitable dimension reduction method for high-dimensional feature data processing of colony images is found. The selected sample set is randomly divided into training set and validation set at a ratio of 2:1. Then the standard support vector machine (SVM) is adopted to classify the sample set to obtain the classification accuracy of different dimensions, and verify whether the dimension reduction method found is the most suitable method for the data processing for colony image feature data set.

B. Essential Dimension Estimation Result

The correlation dimension estimation, the maximum likelihood estimation and the cluster number estimation are adopted to estimate the essential dimension of the sample set. The dimension estimation results are listed in Table 5. According to the simulation algorithm, the result of the essential dimension estimation listed in Table 5 is obtained. Next, these three different dimensions are taken as the target dimension of the low-dimensional space, and different dimension reduction methods are used to reduce the high-dimensional feature data of the colony images.

C. Comparison of Data Dimension Reduction Results

In order to find a more suitable combination of dimension reduction method, three different dimension estimation results are combined with six different dimension reduction techniques (PCA, LDA, MDS, ISOMAP, SNE, and NCA) to form 18 different methods as listed in Table 6. Compare the value of the distance criterion function J to find the most suitable dimension reduction method.

The CDE-based essential dimension estimation results and six different dimension reduction techniques are combined and tested, and the value of the distance criterion function J is listed in Table 7, and the corresponding J value line graph is shown in Fig. 4.

TABLE 5. ESSENTIAL DIMENSION ESTIMATION RESULTS OBTAINED BY THREE DIMENSIONAL ESTIMATION METHODS

Estimation method	Correlation dimension	Maximum likelihood	Number of clusters
Dimension	2	6	7

TABLE 6. 18 DIFFERENT COMBINATIONS OF DIMENSION REDUCTION METHODS

	CDE	MLE	PNE
PCA	CDE-PCA	MLE-PCA	PNE-PCA
LDA	CDE-LDA	MLE-LDA	PNE-LDA
MDS	CDE-MDS	MLE-MDS	PNE-MDS
ISOMAP	CDE-ISOMAP	MLE-ISOMAP	PNE-ISOMAP
SNE	CDE-SNE	MLE-SNE	PNE-SNE
NCA	CDE-NCA	MLE-NCA	PNE-NCA

TABLE 7. TEST RESULTS BASED ON THE COMBINATION OF CDE AND SIX DIMENSION REDUCTION METHODS

Dimension estimator	Dimension reduction technology	Combination method	J value
CDE	PCA	CDE-PCA	2.49813
	LDA	CDE-LDA	0.43819
	MDS	CDE-MDS	2.49813
	ISOMAP	CDE-ISOMAP	2.67416
	SNE	CDE-SNE	2.45279
	NCA	CDE-NCA	0.84417

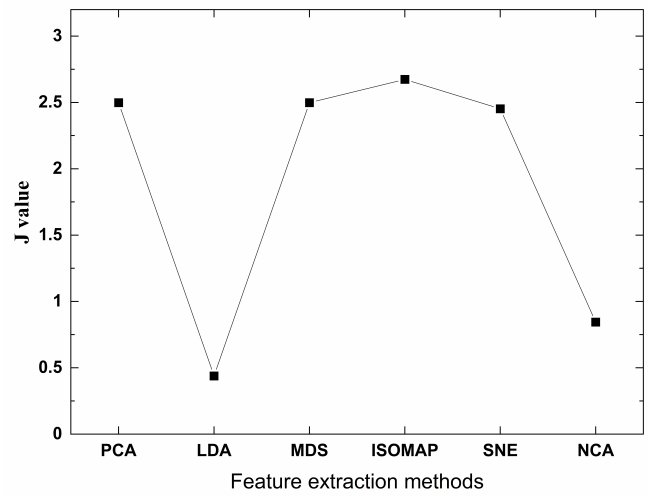


Fig. 4 Six dimension reduction results based on CDE estimation.

The MLE-based essential dimension estimation results and six different dimension reduction techniques are combined and tested, and the value of the distance criterion function J is listed in Table 8, and the corresponding J value line graph is shown in Fig. 5. The PNE-based intrinsic dimension estimation results and six different dimension reduction techniques are combined and tested, and the value of the distance criterion function J is listed in Table 9, and the corresponding J value line graph is shown in Fig. 6.

TABLE 8. TEST RESULTS BASED ON MLE COMBINED WITH SIX DIMENSIONAL REDUCTION METHODS

Dimension estimator	Dimension reduction technology	Combination method	J value
MLE	PCA	MLE-PCA	2.90922
	LDA	MLE-LDA	2.82590
	MDS	MLE-MDS	2.90922
	ISOMAP	MLE-ISOMAP	2.94842
	SNE	MLE-SNE	2.83531
	NCA	MLE-NCA	2.80593

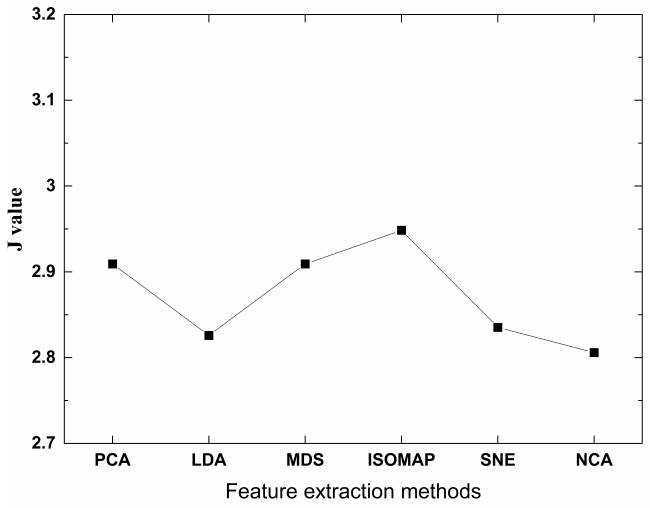


Fig. 5 Six dimension reduction results based on MLE estimation.

TABLE 9. TEST RESULTS BASED ON PNE COMBINED WITH SIX DIMENSIONAL REDUCTION METHODS

Dimension estimator	Dimension reduction technology	Combination method	J value
PNE	PCA	PNE-PCA	2.93246
	LDA	PNE-LDA	2.93273
	MDS	PNE-MDS	2.93246
	ISOMAP	PNE-ISOMAP	2.94951
	SNE	PNE-SNE	2.84864
	NCA	PNE-NCA	2.89570

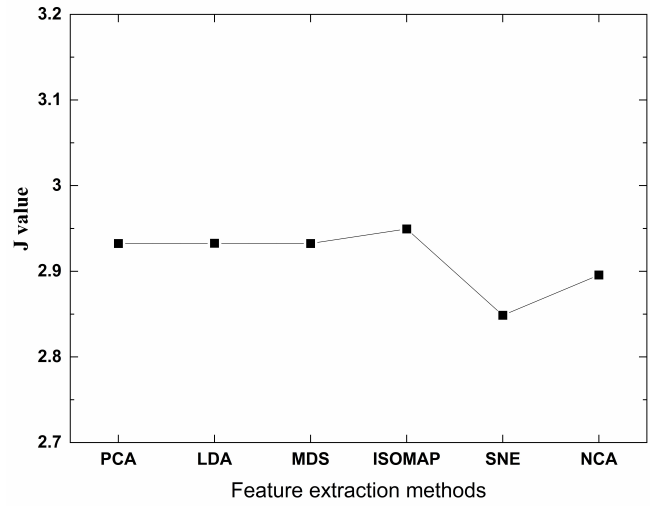


Fig. 6 Six dimension reduction results based on PNE estimation.

Seen from the J value shown in Table 7 and Fig. 4, it can be concluded that the dimension reduction effect of CDE-ISOMAP is the best, and the dimension reduction effect of CDE-PCA, CDE-MDS and CDE-SNE is not much different. But the dimension reduction effect of CDE-LDA and CDE-NCA is far inferior to other four combinations. Seen from the J value data in Table 8 and Fig. 5, it can be concluded that the effect achieved between the six combination methods based on MLE dimension estimation

is not as great as the difference between six combination methods based on CDE, and there is the difference between advantages and disadvantages in the dimension reduction effect, but the difference in J value is small. Among them, the J value of MLE-ISOMAP is the largest, and the dimension reduction effect is the best. From the J value data in Table 9 and Fig. 6, it can be concluded that the J value of the six combinations estimated based on the essential dimension of PNE are basically consistent with the results based on MLE in terms of overall differences, and the J value of LDA has a small margin improvement, but the best combination is PNE-ISOMAP.

In summary, by comparing the optimal combination of each dimension reduction techniques, it can be found that the optimal dimension reduction method on this experimental data set is ISOMAP. But based on the ISOMAP dimension reduction processing under three estimators, the J value is shown in Table 10 and Fig. 7. It is found that the J value difference between MLE-ISOMAP and PNE-ISOMAP is very small, but there is a difference in the data space by one dimension. In large-scale data calculations, each additional dimension of data will cause considerable computational burden. Therefore, due to the consideration of time complexity and other calculation costs, it is considered that MLE-ISOMAP is more suitable for the dimension reduction processing of this data set than PNE-ISOMAP.

D. MLE-ISOMAP Dimension Test

In order to further determine whether ISOMAP is suitable for dimension reduction of the data set according to the results of the essential dimension estimation of MLE, ISOMAP is used to reduce the dimension of the data set to a different target dimension, that is from the lowest dimension of 1 to the highest dimension of the data set, 29 dimensions. The standard SVM optimized by the classic wolf pack algorithm is used to classify it, and the accuracy corresponding to each dimension are compared to verify the appropriate dimension. The ISOMAP is used to reduce the impact of different dimensions on the accuracy of SVM classification is shown in Fig. 8.

It can be seen from Fig. 8 that by comparing the accuracy rates corresponding to different dimensions, the overall higher accuracy rates are concentrated between 84% and 85%, and the feature dimension that first reaches this accuracy range is 6, It is the essential dimension estimation result given by MLE. In the process of data processing, each additional one-dimensional feature data will greatly increase the computational cost. For this reason, it is considered necessary to use the least feature dimension to ensure better accuracy. Therefore, it is considered that the dimension reduction method of MLE-ISOMAP is very suitable when performing dimension reduction processing on the data set in this paper.

TABLE 10. ISOMAP DIMENSION REDUCTION TEST RESULTS UNDER THREE ESTIMATORS

Dimension estimator	Dimension reduction technology	Combination method	J value
CDE	ISOMAP	CDE-ISOMAP	2.67416
MLE		MLE-ISOMAP	2.94842
PNE		PNE-ISOMAP	2.94951

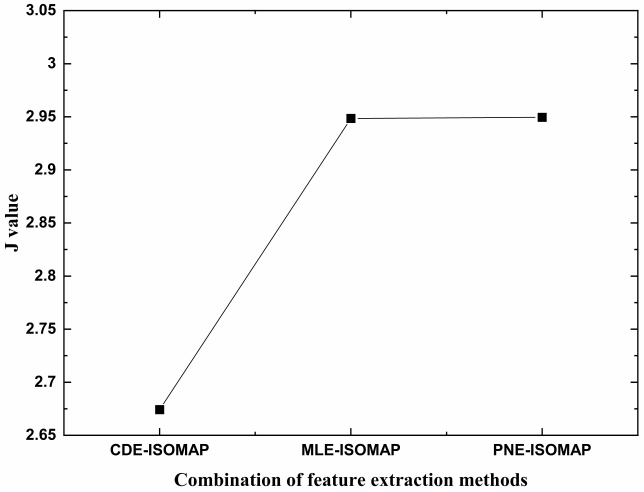


Fig. 7 Dimension reduction results of three combined methods based on ISOMAP.

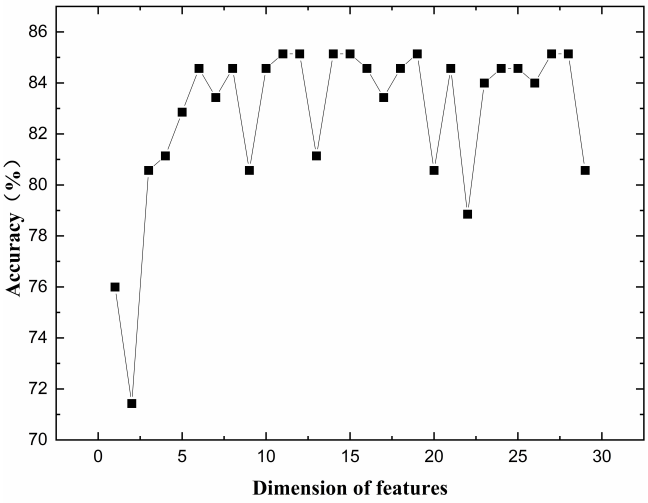


Fig. 8 Effects of ISOMAP with different reduced dimensions on SVM classification accuracy.

VI. CONCLUSIONS

This article is dedicated to finding a more suitable method for colony image feature extraction and sample data processing, including the extraction of color feature information and texture feature information of colony images, and dimension reduction processing of high-dimensional feature data. In terms of color feature information extraction, the first-order moment, the second-order moment and the third-order moment are respectively used to calculate the H , S , and I components in the HSI model to obtain the colony image sample set. The texture feature information of each colony image is obtained through the GLCM. The color features and texture features

together form the feature variables with up to 29 dimensions to form the high-dimensional information data to be processed. Before the dimension reduction processing, the high-dimensional data is estimated by three essential dimension estimators for the inherent low-dimensional structure. Then different dimension estimation results are compared with PCA, LDA, MDS, ISOMAP, SNE and NCA. These six different dimension reduction techniques are combined to find the optimal data dimension reduction strategy based on the distance criterion function. Finally, the support vector machine (SVM) optimized by the classic wolf pack algorithm is used to carry out the classification simulation experiments. The results show that the classification accuracy is concentrated between 84% and 85%; the comprehensive data dimension is as small as possible. If the accuracy rate is as large as possible, it is determined that the MLE-ISOMAP method performs better dimension reduction processing on the high-dimensional feature data of the colony images.

REFERENCES

[1] S. D. Brugger, B. Christian, J. Marcel, J. Werner, B. Urs, and Kathrin Mühlemann, "Automated Counting of Bacterial Colony Forming Units on Agar Plates," *Plos One*, vol. 7, no. 3, pp. e33695, 2020.

[2] Y. S. Huang, Y. C. Chen, M. L. Chen, A. Cheng, I. C. Hung, J. T. Wang, W. H. Sheng, and S. C. Chang, "Comparing Visual Inspection, Aerobic Colony Counts, and Adenosine Triphosphate Bioluminescence Assay for Evaluating Surface Cleanliness at a Medical Center," *American Journal of Infection Control*, vol. 43, no. 8, pp. 882-886, 2015.

[3] E. Kushilevitz, R. Ostrovsky, and Y. Rabani, "Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces," *Siam Journal on Computing*, vol. 30, no. 2, pp. 457-474, 2006.

[4] J. S. Wang, and X. D. Ren, "GLCM Based Extraction of Flame Image Texture Features and KPCA-GLVQ Recognition Method for Rotary Kiln Combustion Working Conditions," *International Journal of Automation & Computing*, vol. 11, no. 1, pp. 72-77, 2014.

[5] T. Gevers, "Image Segmentation and Similarity of Color-texture Objects," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 509-516, 2003.

[6] G. Himabindu, C. Anuradha, and P. Murty, "Assessment of Combined Shape, Color and Textural Features for Video Duplication," *Traitement Du Signal*, vol. 36, no. 2, pp. 193-199, 2019.

[7] R. M. Haralick, "Statistical and Structural Approaches to Texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786-804, 1979.

[8] R. Bennett, "The Intrinsic Dimensionality of Signal Collections," *IEEE Transactions on Information Theory*, vol. 15, no. 5, pp. 517-525, 2003.

[9] A. Guerrero, L. A. Smith, "Towards Coherent Estimation of Correlation Dimension," *Physics Letters A*, vol. 318, no. 4-5, pp. 373-379, 2003.

[10] C. A. Taconeli, and S. R. Giolo, "Maximum Likelihood Estimation Based on Ranked Set Sampling Designs for Two Extensions of the Lindley Distribution with Uncensored and Right-censored Data," *Computational Statistics*, vol. 35, no. 4, pp. 1-25, 2020.

[11] F. Camastra, and A. Staiano, "Intrinsic Dimension Estimation: Advances and Open Problems," *Information Sciences*, vol. 328, no. 4, pp. 26-41, 2016.

[12] E. Kokiopoulou, and Y. Saad, "Orthogonal Neighborhood Preserving Projections: a Projection-based Dimensionality Reduction Technique," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 29, no. 12, pp. 2143-2156, 2007.

[13] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, vol. 24, no. 7, pp. 498-520, 1993.

[14] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Human Genetics*, vol. 7, no. 7, pp. 179-188, 2012.

[15] T. Li, Q. Yin, R. Song, M. Gao, and Y. Chen, "Multidimensional Scaling Method for Prediction of Lysine Glycation Sites," *Computing*, vol. 101, no. 6, pp. 705-724, 2019.

[16] L. Chen, D. Wei, and J. S. Wang, "Research on Magnetic Resonance Imaging Segmentation Algorithm," *Engineering Letters*, vol. 27, no. 3, pp. 559-567, 2019.

[17] K. Bunte, S. Haase, M. Biehl, and T. Villmann, " Stochastic Neighbor Embedding (SNE) for Dimension Reduction and Visualization Using Arbitrary Divergences," *Neurocomputing*, vol. 90, no. 8, pp. 23-45, 2012.

[18] W. Yang, K. Wang, and W. Zuo, "Neighborhood Component Feature Selection for High-Dimensional Data," *Journal of Computers*, vol. 7, no. 1, pp. 161-168, 2012.

Wei-Zhong Sun received his B. Sc. degree in Electronic and Information Engineering and his M. Sc. degree in Control Science from University of Science and Technology Liaoning, China in 2004 and 2008, respectively. He is a lecturer of School of Computer Science and Software Engineering, University of Science and Technology Liaoning. His main research interest is image processing, pattern recognition and embedded system.

Fu-Jun Guo is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. His main research interest is image processing and pattern recognition.

Jie-Sheng Wang received his B. Sc. and M. Sc. degrees in Control Science from University of Science and Technology Liaoning, China in 1999 and 2002, respectively, and his Ph. D. degree in Control Science from Dalian University of Technology, China in 2006. He is a professor and Doctor's Supervisor of School of Electronic and Information Engineering, University of Science and Technology Liaoning. His main research interest is intelligent control and computer integrated manufacturing.

Lin Chen is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, China. His main research interest is image processing and pattern recognition.

Dong Wei is an associate professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. His main research interest is modeling of complex industry process, intelligent control and computer integrated manufacturing.

Xin-Feng Du is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, P. R. China. His main research interest is image processing and pattern recognition.