WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 2
CII 2023

Unsupervised Image Classifier based on Manifold Learning

Jinghao Situ *

School of Mathematical Sciences, Beihang University, Beijing 100191, China

* Corresponding Author

Abstract. Currently most of image classification tasks are achieved by supervised learning. High-quality datasets naturally bring difficulties in annotation, and the datasets in real-world applications present a nonlinear structure, and the annotation cost grows exponentially with the number of targets and the difficulty of recognisability. In this context, research about unsupervised image classification is the way to go. Traditional unsupervised learning for classification is mostly based on the Euclidean distance and various paradigms, which is unable to extract the nonlinear structure of the dataset. This shortcoming makes the accuracy of traditional unsupervised image classification drop drastically. In this paper, we propose to first extract the nonlinear structure of the original dataset using the manifold learning method, and then produce pseudo-labels through the agglomerative clustering algorithm. The pseudo-labels obtained in this way can effectively retain the special mathematical structure of the original data with high accuracy. The neural network is trained with these pseudo labels to obtain an unsupervised usable image classifier. The classifier can be trained on small-scale data and then applied to large-scale data sets, thus saving the cost of manual labelling. The experiments are carried out by setting up a control group and two manifold learning groups for the extraction of non-linear structures using LLE and Isomap algorithms respectively. After that, the production of pseudo-labels and the training of neural networks are completed, and the accuracy of the three groups is compared. Finally, it is concluded that the correct rate of the two groups that have gone through the manifold learning algorithm to extract the nonlinear structure is much higher than that of the other one, and the image classifier based on the Isomap algorithm achieves an accuracy of 85% in the test set, which is highly practical.

Keywords: Manifold Learning; Agglomerative Clustering; Neural Networks; Unsupervised Learning; Image Classification.

1. Introduction

1.1. Background

The types of machine learning in AI terminology include: supervised learning, semi-supervised learning, reinforcement learning, and unsupervised learning. Supervised learning refers to the training of samples with labels, while unsupervised learning involves the training of samples without labels. However, most of the samples are unlabelled, so unsupervised learning is more widely used than supervised learning.[1]

Solving problems in pattern recognition based on training samples of unknown class (unlabelled) is called unsupervised learning, and the meaning of "supervised" can be intuitively understood as "with or without labelled data". Unsupervised learning is characterised by the fact that the data passed to the algorithm is complex in internal structure, but the goals and rewards for training are very sparse.

Most of image classification task is achieved by supervised learning, i.e., each sample has its corresponding label, and a deep neural network is used to continuously learn the features corresponding to each label and finally achieve classification. At this time, the capacity of the dataset and the quality of the labels often play a decisive role in the performance of the model. A large datasets naturally bring difficulties in labelling, and it takes about 2 seconds to label a single image, but datasets in practical applications often contain thousands of images, which makes the whole labelling process becomes extraordinarily long. Especially when it comes to fine-grained

classification and multi-label classification tasks, the annotation cost will grow exponentially with the number of targets and the difficulty of identifying them.

Unsupervised classification methods at present are mainly divided into two main categories. One is dataset transformation i.e. constructing a new representation of an existing dataset so that the features it contains are more easily understood, also known as dimensionality reduction and noise reduction, and the main methods are PCA and T-SNE. The other is clustering algorithms. Cluster analysis is one of the most commonly used methods in unsupervised learning and has achieved a very wide range of applications in the field of machine learning, and many researchers have tried to combine mature clustering algorithms with deep learning to achieve more efficient learning strategies and design models with higher accuracy.

1.2. Main Research Content

However, traditional unsupervised image classifiers also have certain drawbacks. Traditional unsupervised learning for classification is mostly based on Euclidean distance and various paradigms, and such an approach fails to extract the nonlinear structure of the dataset. The data of Swiss volumes or surfaces presents a non-linear structure. There are cases where Euclidean distances between two data points are small but not familiar with the same class. This can be explained by geodesic distances from the point of view of manifold learning. There is no direct relationship between the Euclidean distance between data points and the size of the geodesic distance. Even if the Euclidean distances between data points are very small, they can still be considered as different classes due to the long geodesic distance.

The study in this paper, is an unsupervised image classifier. Compared to the traditional unsupervised image classifier using Euclidean distance, the model is changed to use geodesic distance in manifold learning, whereby the nonlinear structure of the dataset can be extracted. In addition, a clustering approach with higher applicability and better accuracy is chosen, thus enabling geodesic distance-based clustering classification.

1.3. Organization

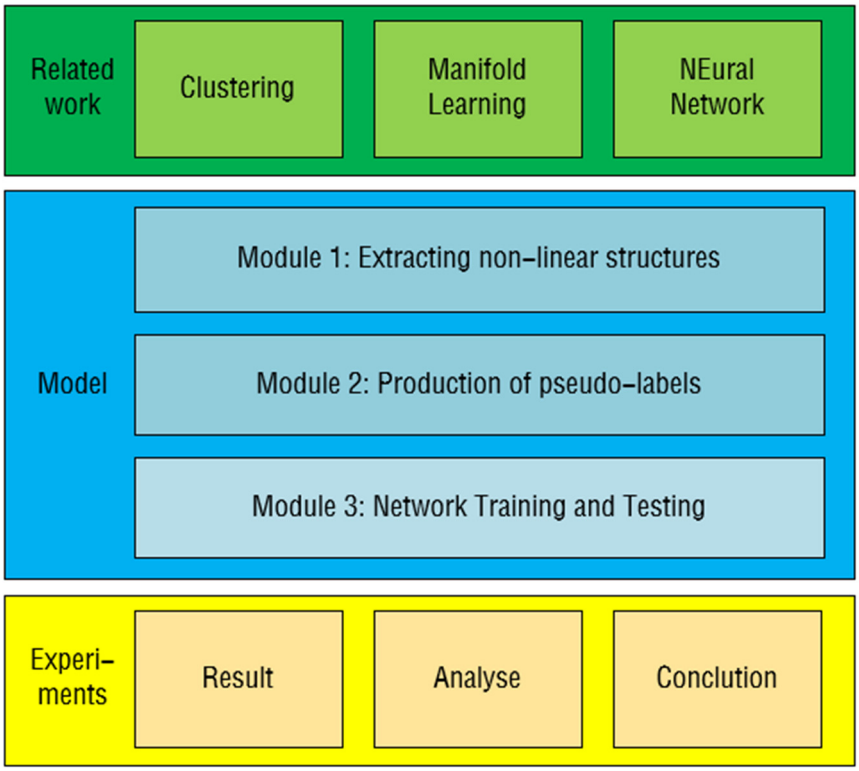

Figure 1. Full Text Architecture Diagram

The organization of the article is as follows: the second part starts with an introduction to related work in the field, including the widely used clustering methods in unsupervised classification; the main ideas and specific methods of manifold learning; and neural networks and their mathematical and theoretical foundations and applied methods. The third part of the paper will provide a detailed description and illustration of the model split into three specific steps: extraction of nonlinear structures, production of pseudo-labels, network training and testing. The main innovation of the model is the combination of existing unsupervised learning algorithms with the manifold learning method. The fourth part is the experimental and comparison part, which derives the effect of unsupervised classification, and compares this effect with other traditional models to reflect the superiority of this model.

2. Related Work

2.1. Clustering

Clustering is one of the main methods of data mining and an important part of unsupervised learning. Objects are divided into subsets or clusters based on their similarity measures, so that objects within clusters are highly similar and objects belonging to different clusters are highly dissimilar. Cluster analysis can mine potential patterns and extract valuable information from complex data, and it is widely used in bio-detection, image processing, intrusion detection, etc [2].

The K-means algorithm, first proposed by Mac in 1967, is a classical algorithm that is highly applicable and widely used. [3] The specific algorithm is that it divides the data into k classes, while satisfying that each class contains at least one sample, and each sample must belong to and only belong to one class [4]. In recent years the K-means algorithm has seen a variety of improvement schemes. Cai Yuhao and other scholars proposed the WLV-K-means algorithm, which heuristically selects the initial centroids of clusters using a weighted local variance measure of the density of the samples [5]. Jia Ruiyu et al. redefined the method of calculating the density of sample objects, and used the method of residual analysis to automatically obtain the initial cluster centre and the number of class clusters from the decision diagram to obtain better clustering results [6]. Based on the nature of exponential function, weight adjustment, paranoid term and the basic idea of elbow method, Wang Jianren et al. proposed an improved k-value selection algorithm ET-SSE algorithm, which effectively improves the accuracy [7]. Comprehensively, K-means cluster analysis has the advantages of fast operation speed and small computational volume, thus it is suitable for analysing and processing large sample data, and can effectively shorten the operation time and improve the operation efficiency [8].

Another widely used clustering method is the agglomerative clustering method. Agglomerative clustering as the name suggests is the operation of clustering a sample set according to some level or some defined distance. [9] Cohesive agglomerative clustering uses a bottom-up approach, where individual objects are clustered, and then the clusters are merged until eventually all objects are merged into one cluster. [10] Early agglomerative clustering methods include AGNES cohesive nested clustering algorithm and DIANA split decomposition clustering algorithm proposed by Kaufman and Rousseeuw. [11] The CURE hierarchical clustering algorithm proposed by Guha et al. employs random sampling and division of partitions, so it can obtain better time efficiency. [12] Chu Kexin and several scholars proposed a agglomerative clustering analysis algorithm for categorical data based on similarity mean and boundary data object allocation strategy, which effectively improves the accuracy of boundary data object allocation and clustering quality. [13] In summary, the agglomerative clustering algorithm does not need to specify the number of class clusters in advance, which is more suitable for the scenarios where the number of class clusters cannot be predicted in advance. [14]

2.2. Manifold Learning

The concept of manifold learning was proposed in 2000, which is a class of feature extraction and dimensionality approximation methods that draw on the concept of topological manifold, and its main

idea is to mine the low-dimensional manifold structures embedded in the high-dimensional space and maintain them in the lowdimensional space, and then extract the salient features of the data. In recent years, manifold learning has received much attention and has been applied to many different fields, such as data visualization, information retrieval, pattern classification, fault diagnosis, etc. [15] The most widely used of mainfold learning, on the other hand, are isomap algorithm and LLE algorithm.

The Isomap algorithm is based on MDS and was first proposed in 2000 [16]. The algorithm suggests that on a curved or manifold surface, the shortest distance is no longer a straight line distance, but rather the length of the path travelled by a geodesic line along that curved or manifold surface, i.e. the geodesic distance [17]. Variations have since been developed. The Landmark Isometric Mapping of Points (L-ISOMAP) has great potential in hyperspectral image visualization. Wang Guoli et al. proposed an improved L-ISOMAP algorithm (KL-ISOMAP) based on the K-medoids clustering algorithm to address the problems of under-representation of marker points and high computational cost of the L-ISOMAP algorithm [18]. Later, the IMM-ISOMAP algorithm appeared, which has good recognition ability for multifluid data with equal dimensions and independent of each other, and can process new data quickly and accurately. [19]

LLE is also a dimensionality reduction method used to process data. In contrast to ISOMAP, which wants to maintain the positional relationship between items, LLE wants to maintain the relative relationship between each item and its neighbours. The LLE algorithm was proposed by Roweis et al [20]. The LLE method models the relationship between an image element and its neighbours, and uses the neighbouring image elements to linearly reconstruct the image element to obtain a local linear function model for a single image element [21]. LLE expresses every high-dimensional data as a linearly weighted combination of its k-nearest neighbours, and solves the corresponding low-dimensional embedding based on the linearly weighted representation [22]. However, the robustness of the algorithm is poor in specific cases. Lv Bingqian et al. proposed the CRLLE algorithm based on commute time distance (CTD) and Rank-Order distance, and achieved better dimensionality reduction in face dataset [23].

2.3. Neural Network

An artificial neural network is a structured network consisting of a number of interlinked and parallel distributed work units. The work units are linearly distributed to form the layers of the artificial neural network. Each operator processes the input data accordingly and transfers the output to the next operator for training. [24] Two commonly used neural networks are described below: the BP neural network and the convolutional neural network.

The BP algorithm is a back-propagation algorithm used to train multi-layer feed-forward neural networks. It updates these parameters by calculating the gradient of the loss function corresponding to each weight and bias in the network, allowing the network to progressively approximate the minimum value of the objective function. [25] Recently, a variety of improved methods have appeared, which are applied in several different fields. In order to accurately predict the air temperature at the bottom of the mine intake shaft, the Pearson correlation coefficient analysis combined with genetic algorithm (GA) to optimize the prediction model of BP neural network appeared [26]. Zhang Liang et al. proposed Sine chaotic mapping improved sparrow algorithm optimized BP neural network (Sine-SSA-BP) prediction model for training, which can more accurately carry out the prediction of fan leaf root loads [27]. In conclusion BP network is very applicable.

Convolutional Neural Networks (CNN) is a feed-forward neural network with convolutional layer computation, which is one of the representative algorithms for deep learning, and is essentially a multilayer perceptual machine [28]. LeCun et al. designed the LeNet model to implement handwritten characters and English characters recognition and classification in 1994, which utilises two-dimensional convolution for image processing [29]. Later, Simonyan et al. improved the performance by improving the network structure, and proposed the VGGNet model, which is mainly used for

large-scale image recognition, with certain migration learning ability [30]. In 2014, the GoogLeNet model appeared, which effectively avoids the problems of gradient vanishing and overfitting [31]. With the development of time, neural networks have become more and more mainstream methods for image classification.

3. Model

This model is an image classifier based on manifold learning and neural network.

There are many existing classification models, classification models such as k-nearest neighbour, Bayesian classifier, linear classifier, SVM, neural network, random forest, Adaboost, etc., but with the gradual increase in the classification accuracy requirements of users, the existing models cannot fully meet the needs of use. The high-dimensional vectors of some images after embedding operations have many nonlinear features, how to extract these nonlinear features and train the classifier without supervision has become a big problem.

Table 1. Symbol Table

Notation	Explanation
$\boldsymbol{x_i}$	The ith item
$\boldsymbol{k}$	Number of neighbours
$\boldsymbol{x_{ij}}$	The jth neighbor of the ith item
$\boldsymbol{w_{ij}}$	Linearity coefficient of the jth neighbour of the ith data point to the ith item
$\boldsymbol{w_i}$	The k linear coefficients of the ith data point are arranged in a vector
$\boldsymbol{I_{k\times 1}}$	Unit vectors of shape $k \times 1$
$\boldsymbol{X_i}$	The k nearest neighbours of the ith item arranged as a vector
$\boldsymbol{w_i^*}$	The obtained optimal w_i
$\boldsymbol{y_i}$	The ith item after dimensionality reduction
$\boldsymbol{Y_i}$	The k nearest neighbours of the ith item arranged as a vector after dimensionality reduction
$\boldsymbol{x_{i,t}}$	The tth component of the ith item
Function	**Explanation**
$\boldsymbol{knn(x,k)}$	K nearest neighbour function, x is the item and k is the number of neighbours
$\boldsymbol{argmin(x)}$	Minimising functions
$\boldsymbol{L(w,\lambda)}$	Lagrange function, w is the vector to be optimised, λ is the Lagrange coefficients
$\boldsymbol{tr(X)}$	Function of finding the trace of the $\boldsymbol{X}$ matrix

3.1. Manifold Learning to Extract Nonlinear Structures

The model in this paper uses the idea of manifold learning to make pseudo labels. Compared with ordinary clustering to make pseudo-labels, this model gives priority to local linear embedding algorithm and Isomap algorithm. The advantage is that the non-linear structure of the original data is given priority in making pseudo-labels.

3.1.1. LLE Algorithm to Extract Nonlinear Structures

The idea of LLE algorithm is that each data point can be represented as a linear combination of multiple data points in its domain, so the algorithm is specifically divided into three steps. The first step named K nearest neighbours, first find the nearest K data points of the data point; the second part is to calculate the weight coefficients, here the weight is a linear combination of points in the field to

get the linear coefficients of the point; the third step is to form the weight matrix; the fourth step is to calculate the first d + 1 eigenvalues; finally get from the second eigenvectors to the second eigenvectors of the d +1 eigenvectors of the matrix is the result of the downscaling. The result obtained by LLE is a better report of the nonlinear features in the original dataset, in addition to the effect of dimensionality reduction.

$$x_{ij} = knn(x_i, k) \tag{1}$$

Use K-nearest neighbour algorithm to get "neighbourhood"

$$x_i \approx \sum_{j=1}^{k} w_{ij} x_{ij} \tag{2}$$

Each data point can be obtained by a linear combination of other data points in the neighbourhood.

$$argmin \sum_{i=1}^{N} \left(\left\| x_i - \sum_{j=1}^{k} w_{ij} x_{ij} \right\|_2^2 \right), s.t. \sum_{j=1}^{k} w_{ij} = 1 \tag{3}$$

Get the loss function and find the weights to minimise the loss function.

$$\sum_{j=1}^{k} w_{ij} = 1 \leftrightarrow \boldsymbol{w_i} I_{k\times 1} = 1 \tag{4}$$

$$\left\| x_i - \sum_{j=1}^{k} w_{ij} x_{ij} \right\|_2^2 = \|x_i - \boldsymbol{w_i X}_i\|_2^2 = \boldsymbol{w_i} (I_{k\times 1} x_i - \boldsymbol{X}_i)(I_{k\times 1} x_i - \boldsymbol{X}_i)^T \boldsymbol{w_i}^T \tag{5}$$

Remember that $A_i = (I_{k\times 1} x_i - \boldsymbol{X}_i)(I_{k\times 1} x_i - \boldsymbol{X}_i)^T$ For optimisation problems with constraints we use the Lagrange multiplier method.

$$L(w_i, \lambda) = \frac{1}{2} w_i A_i {w_i}^T + \lambda(w_i I_{k\times 1} - 1) \tag{6}$$

The above equation is derived for w_i and λ, respectively, and the derivatives are made equal to 0, from which the weight coefficients can be obtained.

$$\boldsymbol{w_i^*} = \frac{{I_{k\times 1}}^T {A_i}^{-1}}{{I_{k\times 1}}^T {A_i}^{-1} I_{k\times 1}} \tag{7}$$

The next step is to optimise the data after dimensionality reduction, noting that the data after x_i dimensionality reduction is y_i, and the weight coefficients of its neighbourhood before and after dimensionality reduction remain unchanged. The benefits of doing so are 1) extracting the nonlinear structure of the data; 2) retaining the features of the original dataset to a certain extent; 3) reducing the dimensionality to facilitate subsequent processing.

$$argmin \sum_{i=1}^{N} (\|y_i - w_i^* \boldsymbol{Y}_i\|_2^2) \tag{8}$$

Remember that $E = \frac{1}{n} Y^T Y$. $W_i^*$is that k components in the middle belonging to the neighbourhood of sample x_i we fill with w_i^* and set the rest of the vector to 0. This ensures that the size of the vector is $n \times 1$. Finally, W_i^* is merged by rows to form the $n \times n$ matrix W.

$$\sum_{i=1}^{N} (\|y_i - w_i^* \boldsymbol{Y}_i\|_2^2) = \sum_{i=1}^{N} (\|y_i - W_i^* \boldsymbol{Y}\|_2^2) = tr[\boldsymbol{Y}(E - W)(E - W)^T \boldsymbol{Y}^T] \tag{9}$$

As above, for optimisation problems with constraints, they can be solved using the Lagrange multiplier method.

$$L(Y,\lambda) = tr[\boldsymbol{Y}(E-W)(E-W)^T\boldsymbol{Y}^T] + \lambda tr(Y^TY - nE) \tag{10}$$

So it is sufficient to let $\boldsymbol{Y}$ be spelled by the d eigenvectors of the matrix $(E-W)(E-W)^T$. Substituting the above equation back into the Lagrangian function results in λ which is the corresponding eigenvalue. And since this is the minimisation goal, we want the eigenvalues to be as small as possible, so we finally take the d smallest eigenvalues corresponding to the eigenvectors of this matrix and put them together as $\boldsymbol{Y}$. At this point, the LLE algorithm to extract the nonlinear structure is partially completed, and the obtained $\boldsymbol{Y}$ is the result of the original dataset after extracting the nonlinear features.

LLE is a widely used method for dimensionality reduction of graphical images, which is simple to implement. The advantages are: 1) it can learn low-dimensional manifolds of arbitrary dimensions that are locally linear; 2) the algorithm boils down to sparse matrix feature decomposition, which has relatively small computational complexity and is easy to implement. However, the disadvantages are: 1) the manifold learned by the algorithm can only be unclosed and the sample set is dense and uniform; 2) the algorithm is sensitive to the choice of the number of nearest neighbour samples, and different numbers of nearest neighbours have a great impact on the final dimensionality reduction result.

3.1.2. Isomap Algorithm to Extract Nonlinear Structures

The specific steps of Isomap algorithm are as follows:(1) Set the number of nearest neighbours of each point k, construct the connectivity graph and adjacency matrix. (2) Construct the distance matrix in the original space by the shortest path of the graph. (3) Compute the inner product matrix. (4) Perform eigenvalue decomposition of matrix B to obtain eigenvalue matrix and eigenvector matrix. (5) Take the largest first t terms of the eigenvalue matrix and their corresponding eigenvectors. The result obtained by Isomap algorithm reports the nonlinear features in the original dataset better, in addition to the effect of dimensionality reduction.

$$d(x_i, x_j) = \sqrt[2]{\sum_{t=1}^{n} (x_{i,t} - x_{j,t})^2} \tag{11}$$

$$d(x_i, x_j) = \begin{cases} 0 \quad if\ \boldsymbol{X_j}\ is\ not\ the\ \boldsymbol{K}\ \text{of the nearest neighbors}\ \boldsymbol{of}\ \boldsymbol{X_i} \\ d(x_i, x_j) \quad if \boldsymbol{X_j} is\ the\ \boldsymbol{K} \text{of the nearest neighbors} \boldsymbol{of} \boldsymbol{X_i} \end{cases} \tag{12}$$

Calculate the Euclidean distance between two pieces of data. And redefine the distance according to the principle of K-nearest neighbour, so as to construct the distance matrix B. Using the principle of K-nearest neighbour can make the neighbouring points regarded as connected, and the distant points regarded as disconnected, when calculating the shortest path, it is necessary to transit through other points. Use Floyd's algorithm to find the shortest path to the obtained distance matrix B to get the matrix D. At this point, the transformation from Euclidean distance to geodetic distance is completed, so as to avoid the large differences between the two data points due to the short Euclidean distance between the two points, resulting in the misclassification of the image classifiers. The pseudo-code for the Floyd in Isomap is as follows.

Algorithm 1: Floyd in Isomap

Data: initial distance matrix in Isomap *distance*, number of samples n
Result: distance matrix after Isomap *distance*

```
for k ← 1 to n do
    for i ← 1 to n do
        for j ← 1 to n and i ≠ j do
            if distance(i, j) > distance(i, k) + distance(k, j) then
                distance(i, j) ← distance(i, k) + distance(k, j);
            end
        end
    end
end
```

Finally, we want to downsize the original data to get the result, so we finally take the t smallest eigenvalues of the matrix D corresponding to the eigenvectors spelled $\boldsymbol{Y}$. The Isomap algorithm to extract the nonlinear structure of the partially completed to get the $\boldsymbol{Y}$ is the result of the original data set extracted the nonlinear features.

3.2. Agglomerative Clustering for Pseudo Labelling

The second step is to perform clustering with the data obtained from the above processing. The agglomerative clustering is chosen for this model. It is a bottom-up clustering algorithm that each data point is initially considered as a separate cluster and then the clusters are gradually merged until all the data points are merged into one large cluster. The pseudo-code for the agglomerative clustering algorithm is as follows.

Algorithm 2: Agglomerative Clustering

Data: initial set of X of objects$\{x_1, x_2 ... x_n\}$, a distance function $dict(c_i, c_i)$

Result: a set of labels of each objects $\{l_1, l_2 ... l_n\}$

```
for i ← 1 to n do
    c_1 ← x_1
end
C ← {c_1, c_2 ... c_n};
while C.size > clusters do
    (c_i1, c_i2) = minimum dist(c_i, c_j) for all c_i, c_j in C;
    remove c_i1 and c_i2 from C;
    add {c_i1, c_i2} to C;
end
```

A major focus of using neural networks for unsupervised learning is to produce pseudo-labels. After agglomerative clustering, we obtain a pseudo-label that fully takes into account the nonlinear structure of the original data. It is worth noting that the number of clusters needs to be formulated artificially, and the number of clusters can be roughly formulated based on the specifics of the dataset.

3.3. Neural Network Training

The simplest both neural networks are BP neural network and convolutional neural network. The inputs of them are different, the input segments of BP neural network are vectors while CNN inputs are images. But applicable in this model, there is no essential difference between the two. In this model, a BP neural network is used, which needs to convert the image into a high-dimensional vector first, and then the high-dimensional vector is used as the input to the neural network. Of course, it is possible to use a CNN, which does not require the step of transforming the image into a vector, but only needs to use the image directly as the input to the network.

BP neural network is a kind of error inverse feedback learning and training network with strong nonlinear processing capability, which can adaptively learn from discrete experimental data and store the learning results in weights and thresholds to complete the model building.

CNN is a special multilayer perceptron model designed for recognising 2D images, which adopts a fixed construction mode in its overall architecture, using the convolutional layer and pooling layer as feature extractors to complete the extraction of semantics at different levels of the image, and completing the classification through the fully-connected layer, and then relying on the training of the data samples to continuously fit and adjust the network parameters.

This model, which uses a BP neural network, requires the image to be converted into a high-dimensional vector first, and then the high-dimensional vector is used as the input to the neural network. When the image into the convolutional neural network, or the vector into the fully connected neural network, the resultant output is a vector, each component of the vector represents the probability that the data belongs to that category, the higher the probability the more likely to belong to the category, and ultimately determine that the image belongs to the category with the largest

component. In order to train the neural network, we need to compare the output vector with the pseudo-labels to construct the loss function, and then the loss is propagated backward in the neural network, and then the neural network is trained by gradient descent method. The neural network uses cross entropy loss function and Adam adaptive gradient descent method to finalise the construction and training of the network.

Cross Entropy Loss Function

$$H_p(q) = \sum_x q(x) log_2 \left(\frac{1}{p(x)}\right) \tag{13}$$

3.4. Applications and Follow-up

The trained neural network is a usable image classifier. For this obtained image classifier, a new image is used as input to the neural network to get the classification result. This image classifier is generalisable and can be applied to process large number of images after going through the training set. Secondly, this classifier considers the nonlinear features of the dataset, unlike traditional image classifiers, which consider and extract the nonlinear structure of the dataset when making the pseudo-labels, and unlike the traditional models that use Euclidean distance, this model uses the geodesic distance in the manifold, which will effectively improve the model accuracy.

4. Experiment

4.1. Description of the Data Set

The dataset for the model experiments is the handwritten digit body dataset. The dataset that comes with the sklearn library of python is used. The handwritten dataset image is 8*8 in size, and the imported data is directly transformed into the form of vectors of length 64. It should be emphasised that when the input data are embedded vectors, BP neural networks should be used; when the input data are pictures, convolutional neural networks should be used. In this experiment, since the input data is in vector format, a BP neural network is used. The dataset has a total of 1797 data, of which 1700 are used as training data and 97 are set as test data.

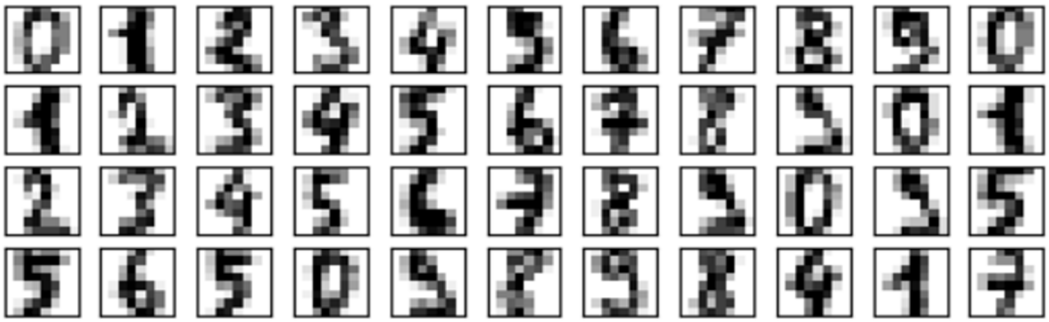

Figure 2. Handwritten Numerals Dataset

4.2. Comparison on Extracting Nonlinear Structures

Module 1: Extracting the nonlinear structure of the dataset using manifold learning algorithm, LLE algorithm and Isomap algorithm are used in this model. In order to make a comparison, three groups of experiments are conducted, which are control group (without any processing), LLE group (extracting nonlinear structure using LLE algorithm), and Isomap group (extracting nonlinear structure using Isomap algorithm). The LLE algorithm has the effect of dimensionality reduction on the original data while extracting features, here the other LLE algorithm reduces the 64 dimensions of the original data to 5. The Isomap algorithm is designed based on the geodesic distance, and the dimensionality of the original data will be less than or equal to 64 dimensions after the processing, so here the choice is to let the processed data still maintain the 64 dimensionality, i.e., only extracting

the nonlinear structures but not reducing the dimensionality. Both flow learning algorithms use the K-nearest neighbour principle, and the 15 nearest neighbours of the selected points are taken into account.

Table 2. Group Setting

	CONTROL GROUP	LLE GROUP	ISOMAP GROUP
NEIGHBORHOOD PROCESSING	—	K nearest neighbours (K=15)	K nearest neighbours (K=15)
(REDUCED DIMENSIONAL) RESULTS	64	5	64
WHETHER TO EXTRACT NON-LINEAR STRUCTURES	NO	YES	NO

4.3. Pseudo-labelling

Module 2: Pseudo-labelling using agglomerative clustering algorithm. For the above three groups: control group, LLE group, Isomap group respectively use agglomerative clustering to produce pseudo-labels, and compare the pseudo-labels with the real labels of the images. It is easy to find that the labels generated by the control group deviate from the real labels, the LLE group is slightly better than the control group, but still has a certain error, and the smallest error is in the Isomap group. However, since it is an unsupervised learning algorithm, its accuracy is generally much lower than that of supervised learning, so it is inevitable that each group has an error with the real label group.

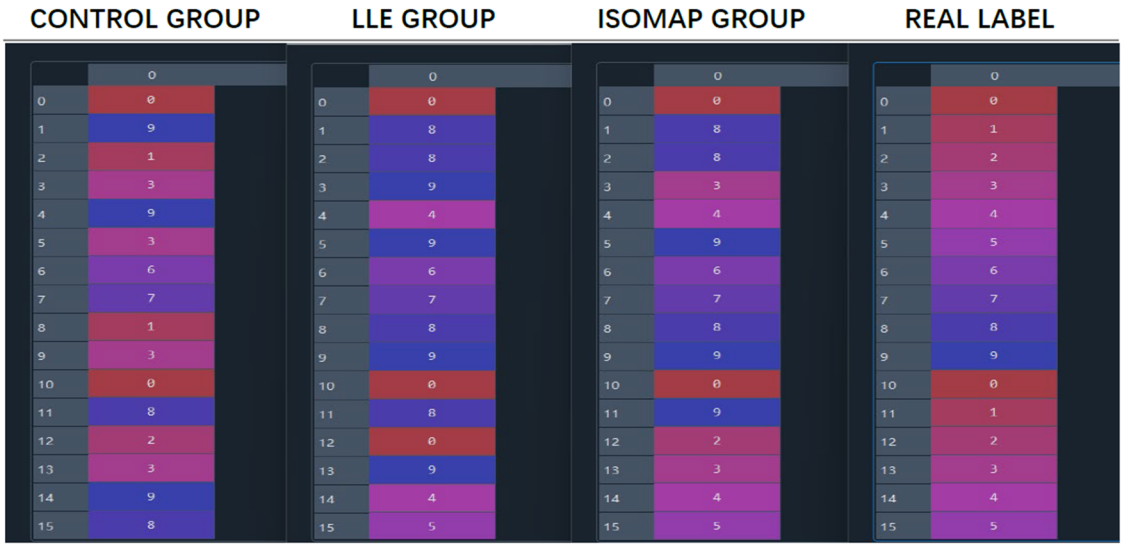

Figure 3. Partial Pseudo-labelling

After counting, the difference between the pseudo-labels and the real labels in each group is as follows.

Table 3. Pseudo-labelling Accuracy

	CONTROL GROUP	LLE GROUP	ISOMAP GROUP
PSEUDO-LABELLING ACCURACY	48.71%	63.53%	86.18%

4.4. Neural Network Test Results

This experiment uses a BP neural network with 6 layers of neural network with 64, 12, 32, 64, 70, and 10 neurons at a time, respectively, and the activation function dimensional ReLU function used. The training of the neural network is set to a group of 32 data, with a learning rate of 0.1, for a total of 100 times, resulting in a usable image classifier generated unsupervised. The results of testing three groups of trained networks on the test set of data are as follow.

Table 4. Image Classifier Test Accuracy

	CONTROL GROUP	LLE GROUP	ISOMAP GROUP
ACCURACY	44.33%	65.98%	85.56%

Obviously the accuracy of the image classifier is strongly correlated with the accuracy of the pseudo-labels. The more accurate the pseudo-labels are, the better the network classifier is naturally. The control group does not extract the nonlinear structure of the original data, resulting in the pseudo-labels deviating too much from the real labels, and the image classifier obtained from the training is no longer practical. And the accuracy of both groups through the manifold learning algorithm is above 60%, much higher than the control group, which argues the necessity of manifold learning in Module 1. Comparing the latter two groups, the accuracy rate of the LLE group is again much lower than that of the Isomap group, mainly due to two reasons: the first one is that the Isomap algorithm happens to be able to extract the information of handwritten digitized body images better, and for different datasets it is necessary to use different manifold learning algorithms in extracting their nonlinear structures; secondly the data processed by the LLE algorithm is inevitably the dimensionality reduced data, while the Isomap algorithm can choose not to reduce the dimensionality of the processed data, so that the nonlinear structure can be extracted at the same time to retain more information about the original data, in the case of both, certainly the image classifier is more effective.

Since the Isomap group image classifiers are far more effective than the other two groups and the results are more satisfactory, the following will focus on analysing the training process and results of the Isomap group image classifiers.

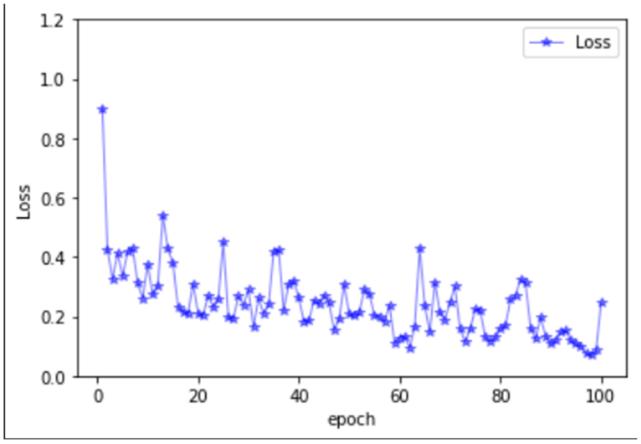

Figure 4. Plot of Change in Loss Function

From the point of view of the Loss function of the neural network, the classifier has a rapid decline in the Loss function in the initial stage of training, the subsequent decline is slow and fluctuations occur, and basically reaches the level where the loss value is less than 0.2 after 30 times of training.

Looking at the accuracy of the training set, it has reached over 90% at 20 epochs. It fluctuates up in 20-100 epochs and finally stabilises at around 97%, indicating good results on the training set.

The accuracy rate of the test set also shows a trend of rapid increase and then levelling off, and finally reaches about 85%, which is the best effect among the three groups. Unsupervised learning can reduce

the requirement of the original data and save the cost of manual classification, and the 85% accuracy rate under this condition indicates that the present image classifier can classify pictures well and achieve the expected effect.

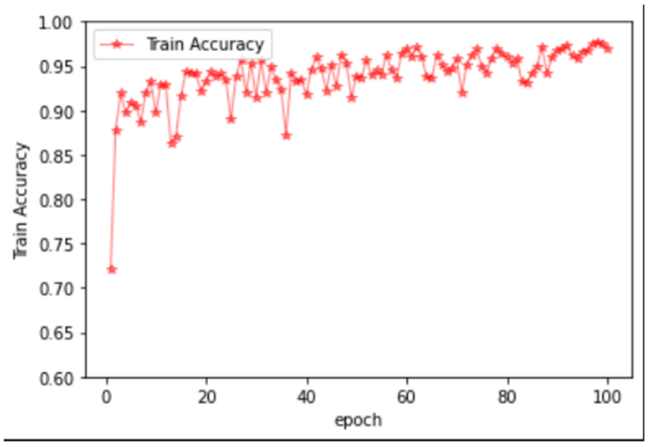

Figure 5. Plot of Change in Training Set Accuracy

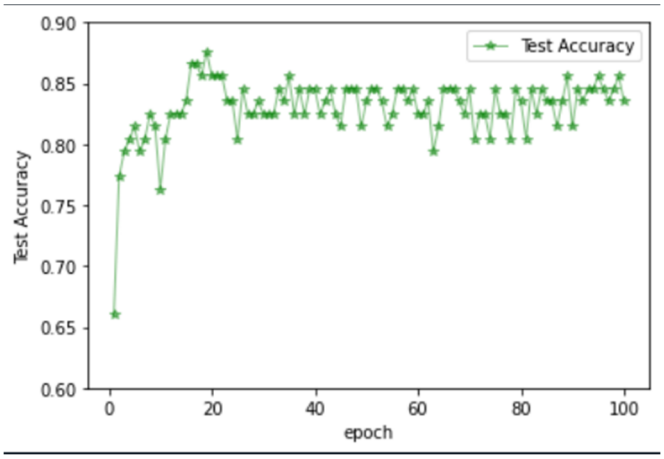

Figure 6. Plot of Test Set Accuracy Changes

5. Summary

Starting from previous research results, this paper briefly introduces clustering algorithm, manifold learning and neural network algorithm. And then the three modules of the algorithms are described in detail: manifold methods for extracting nonlinear structures, the agglomerative clustering for making pseudo-labels, and the neural network for training and testing. Finally, the experimental results are analysed in detail. An image classifier with higher accuracy that fully considers the nonlinear structure of the original dataset is obtained. It is concluded that this model fully considers the nonlinear structure of the dataset and achieves high accuracy and good generalisability.

Looking ahead, image classifiers will become increasingly accurate. By appropriately reducing the input, fully extracting information from the dataset to enhance its real-time processing capability, and in conjunction with the popularity of cloud computing, it is expected to be applied in the field of human interaction such as automated driving and intelligent identification, to improve the timeliness and safety at the application level. Furthermore, the data of some personalised programmes and objects have their uniqueness and their datasets are bound to have more complex structures. By extracting the non-linear structure of these personalised objects, the user's behaviour and preferences can be better analysed, thus providing a usage experience that is more in line with the user's preferences. In addition, cross-domain applications, collaboration and interaction are also an

important direction for machine learning, and all of these developments will greatly facilitate the application and promotion of AI technology, bringing more convenience and progress to society.

References

[1] Yin Ruigang, Wei Shuai, Li Han et al. A review of unsupervised learning methods in deep learning. Computer System Applications,2016,25(08):1-7. DOI:10.15888/j.cnki.csa.005283.

[2] Zhou Y, Xia H, Liu HY et al. DPC clustering algorithm based on K mutual nearest neighbours and kernel density estimation[J/OL]. Journal of Beijing University of Aeronautics and Astronautics:1-16[2023-10-21]. https: //doi.org/ 10. 13700/j.bh.1001-5965.2023.0342.

[3] MAC Q J. Some methods for classification and analysis of multivariate observations. Berkeley: Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[4] Chuang, Chuan-Chi. Research on multivariate discriminative clustering algorithm. Nanjing University of Aeronautics and Astronautics,2011.

[5] Yuhao Cai, Yongquan Liang, Jiancong Fan et al. K-means algorithm for weighted local variance optimization of initial cluster centres. Computer Science and Exploration,2016,10(05):732-741.

[6] Jia Ruiyu, Li Yugong. K-means algorithm for self-determination of the number of class clusters and initial centroids [J]. Computer Engineering and Applications,2018,54(07):152-158.

[7] Jianren Wang, Xin Ma, Ganglong Duan. Improved k-means clustering k-value selection algorithm. Computer Engineering and Applications,2019,55(08):27-33.

[8] Liu Quanhong, Tang Fuxing. Optimisation of site selection for faulty shared bicycle recycling centre based on K-means clustering algorithm and centre of gravity method. Operations Research and Management,2023,32(07):85-91.

[9] Johnson SC. Hierarchical clustering schemes. Psychometrika, 1967, 32(2): 241-254.

[10] Lv L, YU YQ, REN M et al. Cohesive hierarchical clustering based on ant colony optimization algorithm. Computer Application Research,2017,34(01):114-117.

[11] Kaufman L, Rousseeuw PJ. Finding Groups in Data: an Introduction to Cluster Analysis. New York: John Wiley & Sons, 1990.

[12] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for clustering large databases // ACM SIGMOD International Conference on Management of Data, 1998: 73-84

[13] Chu Kexin, Xun Yaling. A hierarchical cluster analysis algorithm for categorical data based on similarity mean. Computer Technology and Development,2022,32(11):154-163.

[14] Liu Yi-Wei. Application research on person name disambiguation based on improved cohesive hierarchical clustering. Shandong University of Science and Technology, 2021. DOI: 10.27275/d.cnki.gsdku.2020.001929.

[15] Zhang R. Research on bearing fault diagnosis method based on improved LLE algorithm. Northeast Petroleum University, 2023. DOI: 10.26995/d.cnki.gdqsc.2023.000376.

[16] TENENBAUM J B, DESILVA V, LANGFORD J C. A global geometric framework for nonlinear dimension reduction. Sci- ence, 2000, 290(5500):2319-2323.

[17] Li Gapeng. Research on point cloud denoising algorithm based on non-local self-similarity and stream learning. Southwest University, 2023. DOI: 10.27684/d.cnki.gxndx.2023.001242.

[18] L.G. Wang, F. Wu. Colour visualization of hyperspectral images based on KL-ISOMAP. Journal of Nanjing University of Information Engineering (Natural Science Edition),2018,10(01): 63-71. DOI:10. 13878/j. cnki. jnuist. 2018. 01.006.

[19] Jiefei Liu. Research on multi-stream shape learning algorithm for high-dimensional data. Shanxi University,2018.

[20] Roweis S T, Saul L K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 2000, 290 (5500): 232

[21] Yu WB. Research on Feature Extraction from Hyperspectral Data Based on Stream Embedding and Deep Feature Analysis. Harbin Institute of Technology, 2021. DOI: 10.27061/d.cnki.ghgdu.2021.005341.

[22] Qiuying Yang, Xiaoqing Weng. Time series clustering based on LLE and Gaussian mixture model. Computer Technology and Development,2022,32(08):33-41.

[23] Lv Bingqian, Fan Linyuan. Improvement of LLE algorithm based on commuting time distance and Rank-Order distance. Information Systems Engineering,2021(07):141-144.

[24] Jiang Sanshan. Research on gas outflow prediction based on LLE-WPA-BP algorithm. Liaoning University of Engineering and Technology, 2023. DOI: 10.27210/d.cnki.glnju.2022.000747.

[25] LIU Tao, WANG Shen, WU Jiajing et al. Research and application of intelligent power technology based on BP neural network metro shield. China Water Transport (the second half of the month),2023,23(11):37-39.

[26] Gao Jia-Nan, Ma Le-Tian, Bai Jin-Yang et al. A wind temperature prediction model for drenching wellbore based on GA-BP neural network. China Mining:1-6[2023-10-22].http://kns.cnki. net/kcms/ detail/11. 3033.TD. 2023 1017. 1044. 004.html.

[27] Zhang Liang, He Shan, Ai Chunyu. Load prediction of wind turbine leaf root based on Sine-SSA-BP neural network model. Renewable Energy,2023,41(10): 1322-1328. DOI: 10.13941/j.cnki.21-1469/tk.2023.10.006.

[28] J. Chen. Research on image style migration based on convolutional neural network. Journal of Hubei Institute of Technology,2023,39(05):35-38.

[29] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient, based learning applied to document recognition. Pro- ceedings of the IEEE, 1998, 86(11): 2278-2324.

[30] SIMONYAN K, ZISSERMAN A. Very deep convolu tional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[31] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions, Proceedings of the 2015 IEEE Confer- ence on Computer Vision and Pattern Recognition, 2015:1-9.