# Handwritten Digits Recognition Technology Based on SAE-SVM Classifier

Xiaoting Du[1, a]

[1] Electronics and Information Engineering College, Anhui University, Hefei, 230009,China

[a]email: 1836231234@qq.com

**Keywords:** Handwritten Digits Recognition; Stacked Auto Encoder; BP Algorithm

**Abstract.** For the purpose of this article, there is a huge amount of work in feature extraction, selection and other aspects using traditional supervised machine learning methods, and features for different applications often required different scenarios which need to manually design, but with the final result not ideal. The paper shows a unsupervised feature extraction method-combine Stacked Auto Encoder and Support Vector Machine, experiments had shown that the algorithm's accuracy is 99.31% in MINIST better than other algorithms. This study can help Handwritten Digits Recognition get better development in various fields, such as ZIP code automatic identification, automatic processing of bank checks.

## Introduction

Currently, Optical Character Recognition [1] is the most widely used in character recognition. The statistical pattern recognition method of it focuses on the number of features and is easy of feature extraction, analysis and calculation. However, the stochastic was considered as a random two-dimensional lattice, not considering the structural features and structural information of the character. Therefore, this method is effective for a single character, but poor in distinguishing different font characters. Structural pattern recognition methods consider the be identified pattern as a set including numbers of relatively simple sub-mode configuration, in which any mode can be described by a set of primitives and certain combination relations. Since the characters are rich in structural information, we could try to extract structural features containing such information as the basis for character recognition. However, because the structure character is complex, there is still difficult in practical application. Recent years, there have been some methods combining statistics and Structural pattern recognition, which not only attract the advantages of statistical pattern recognition, but also use the structural information about the character. Handwritten digits recognition is a specific direction in character recognition. Due to the special nature of the problem itself, the traditional OCR methods can't effectively solve this problem. Thus, recognition method of handwritten digits should be a kind of adaptive，immunity method that can solve handwritten digital dividing, combine between statistical pattern recognition and structural pattern recognition effectively.

Since machine learning leader Hinton firstly proposed deep belief network [2] in his article in 2006, the deep learning has been greatly developed, and has become upsurge in Internet big data and artificial intelligence. Deep neural network utilizes multi-layer structure of the human brain, form ideal feature for pattern classification characteristic finally through the feature extraction of input data level by level from the bottom to the top. In this paper, we utilize Stacked Auto Encoder

 

[3] as unsupervised learning algorithm which was commonly used in deep learning to build the depth of the network model. Then we use the preprocessing training sample of handwritten digits recognition so that we can execute fine-tuning in representation and reconstruction of each layer to complete the entire training process system through unsupervised learning. At last, we use Support Vector Machine which is supervised learning methods to fine-tune the entire system.

## Stacked Auto Encoder

### *The Structure of Stacked Auto Encoder*

*The Mathematical Model of Sparse Auto Encoder*

Auto-Encoder Neural Network is an unsupervised learning algorithm, in which optimization goal is that the output value equals to the input values: $y^{(i)} = x^{(i)}$. The traditional auto-encoder neural network has three layers which are input layer, hidden layer and output layer, just like Figure 1. Hidden layer called Features1 must include the inner structure of the input data ,however, the compressed representation will be very difficult to learn if the input of internet is completely random .But if some specific structure of the input data are implied, for example ,certain input characteristics are related to each other ,the algorithm can find the correlation of input data and learn low-dimensional representation of raw data which is extremely similar to the result of Principal Component Analysis .

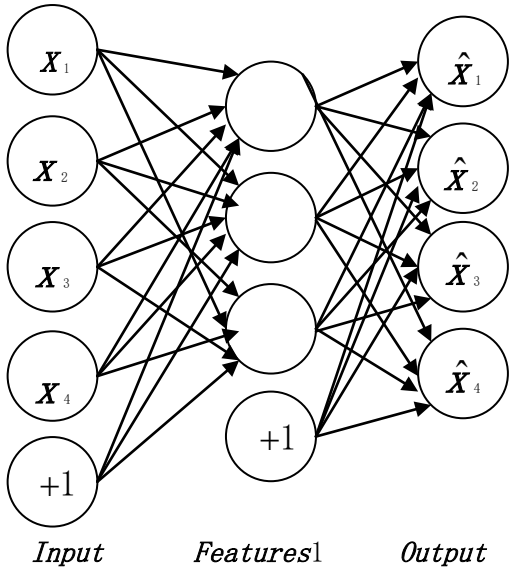


Fig.1. Auto-Encoder Neural Network

For a fixed training set $\{(x^{(1)}, y^{(2)}), \mathrm{K}\ (x^{(m)}, y^{(m)})\}$ its total number of samples is m, then define the overall cost function as Eq.(1).

$$\begin{aligned} J(W,b) &= \left[\frac{1}{m}\sum_{i=1}^{m} J\left(W,b;x^{(i)},y^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(W_{ji}^{(L)}\right)^2 \\ &= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}(x)-y\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(W_{ji}^{(L)}\right)^2 \end{aligned} \tag{1}$$

The first term J(W,b) of the above formula is mean variance item, in which W is weight vector, b is vector of bias node; while the second is regularization item, used to constrain the solution for achieving the structural risk minimization, which can prevent over-fitting [4].

When the neuron in hidden layer has a larger number, we can find the structure of input data through adding sparsity limitation in hidden neurons. Sparsity limitation is the limit that neuron is suppressed for most of the time, we can form Spare Auto Encoder through adding this constraints the basis of automatic encode, which is more effectively than other expression. We use $a_j^{(2)}(x)$ to represent the flexibility of Auto Encoder's hidden neurons j in a given input x, and use $\hat{P}_j = \frac{1}{m}\sum_{i=1}^{m} a_j^{(2)}\left(x^{(i)}\right)$ to represent the average activity of hidden neurons j on the training set. For sparsity representation of data, we entry restriction which p is the sparsity parameter, usually closed to 0. Relative entropy which is used to measure between two distributions in optimizing the objective function can be used as additional penalty factor $\sum_{j=1}^{S_2} KL\left(P \middle\| \hat{P}_j\right)$ so that we can achieve this limit. Relative entropy may be expressed as Eq.(2).

$$KL\left(p \middle\| \hat{p}_j\right) = pIn\frac{p}{\hat{p}_j} + (1-p)In\left(\frac{1-p}{1-\hat{p}_j}\right) \tag{2}$$

Then the overall cost function can be expressed as Eq.3.

$$J_{spqrse}(W,b) = J(W,b) + \beta\sum_{j=1}^{S_2} KL\left(p \middle\| \hat{p}_j\right) \tag{3}$$

In this formula, β is the weight of controlling sparsity penalty factor. After calculating cost function, we can use the backpropagation to calculate the partial derivative of the cost function

*The Mathematical Model of Stacked Auto Encoder*

Stacked Auto Encoder Neural Network is composed by multilayer Sparse Auto Encoders, which contains the input layer, several hidden layer and output layer, almost like Figure. 2. The multilayer structure can not only improve the representation ability of model, but also avoid the exponential growth of the node number. Preprocessing data accesses to the network via the input layer, then implement the abstraction of data and feature extraction from features 1 to features L.

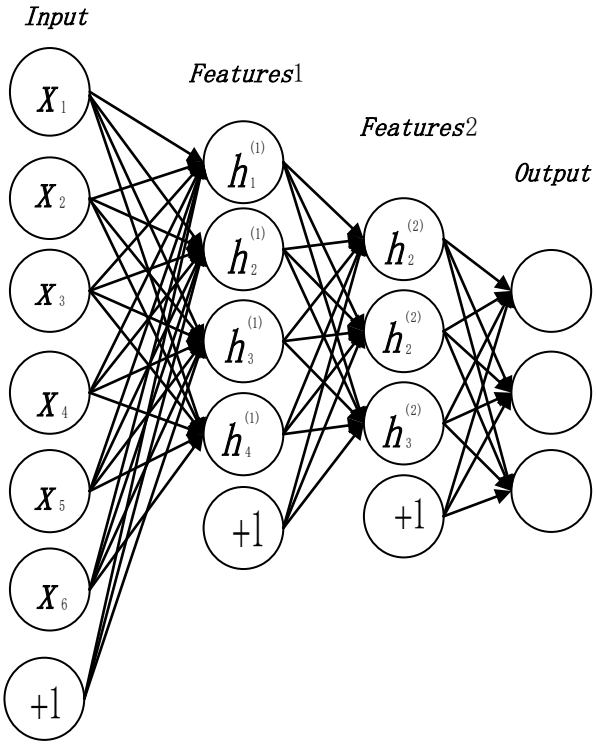


Fig.2. the Example of SAE Neural Network

*Training Methods of Stacked Auto Encoder*

As deep neural network, the most important problem of Stacked Auto Encoders is how to train parameter optimization. Greedy layer-wise approach is an effective method of training depth network. First of all, it overcomes the drawback that the depth of the network get data costly, it can access to abundant unlabeled data to learn the best initial weights W' of all layers inexpensively and effectively so that the network parameters of lower layer have been fully trained. Secondly, it overcomes the disadvantages of gradient descent method's local extremum and gradient instability in depth network training. After training depth network with unlabeled data and greedy layer-wise approach, initialization weights W' will be located in a range with better parameter space compared with the weight of the random initialization. We can do further fine-tuning of weight with the initial weights of weight. Greedy layer-wise approach based on Auto-Encoder trains, as follow.

(1) construct structure of Stacked Auto Encoder network and input training results without a label $\left\{x^{(1)}, x^{(2)}, \ldots x^{(m)}\right\}$;

(2) obtain initialized weight parameter $\left[w^2, w^3, \ldots w^L\right]$ through greedy layer-wise approach in spare auto encoder;

(3) add support vector machine, then construct classification structure of Stacked Auto Encoder neural networks and data pair for input labels $\left\{x^{(i)}, y^{(j)}\right\}$;

(4) use deviation's backpropagation algorithm for global tuning of weight $\left[w^2, w^3, \ldots w^L\right]$;

(5) export weighting factor $\left[w^2, w^3, \ldots w^L\right]$;

## Experiment and Analysis

### *Experimental Data Presentation*

During digital stage of training process, the sample gallery in this article is handwritten numeral database (MNIST) which is established by Corinna Cortes in Google laboratory and Yann LeCun in New York University's Courant Institute. MNIST has a training library whose sample data is about 60000 and a test data set whose sample data is almost 10000, which contain corresponding number, just like Figure.3. MNIST data set is a subset of the NIST, which add some amendments to make the data more standardized, digital size of all graphics is in the center of the image through standardized treatment.

Fig.3. some Samples of MINIST Data Set

### *Experimental results of SAE in MINIST*

*SAE's edge features of learning*

As seen from Figure.4, we can know that the SAE weight learned is sparse. The learned characteristic is with the direction of edge features from the standpoint of computer vision. Nobel Laureate David Hubel and Torsten Wiesel found a kind of neurons called Orientation Selective Cell in the cerebral cortex [5]. When the pupil found the edge of the object in front of the eyes which is in a certain direction, this kind of neuron will be active. That is to say, some "specific direction neurons" are only incentive or excited to particular direction of image edge. So the feature detection principle of SAE is to use Maximizing the action, what has a certain similarity to the brain's visual system.

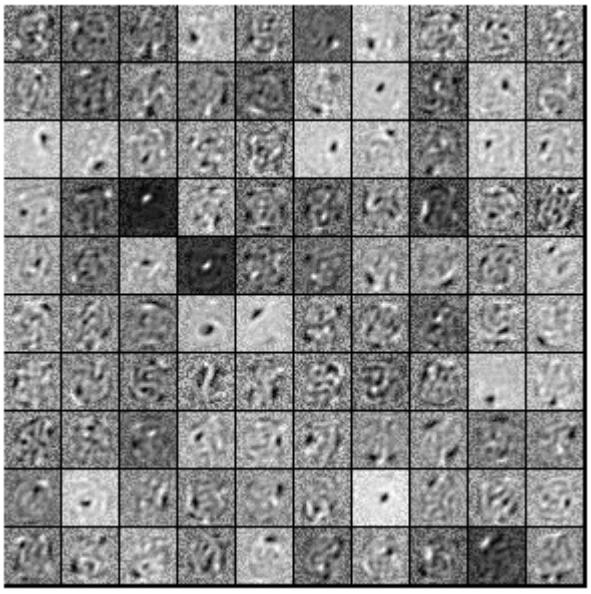

Fig.4.Visualization Feature of SAE

*Hidden layer effect on the accuracy*

Determining the depth of the network is completely important to the final classification results in depth model. In the SAE network, when the number of hidden layer nodes is small, we mayn't be able to fully learn the characteristics of the data; when it's too small, feature vector is too sparse to describe the features of the data effectively and increase the burden of learning, which will result in too long training time. After several experiments, we selected the best hidden layer nodes is 400. Moreover, we change the SAE' number of hidden layers on the basis of best nodes so that the experimental results are shown in Figure.5, we can obtain that handwritten digits recognition accuracy is the best when the SAE' number of hidden layers is 3.

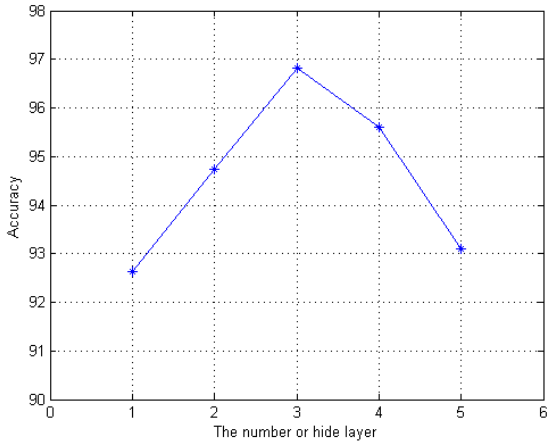


Fig.5. Hidden Layer Effect on the Accuracy

*Comparison with the Traditional Feature Extraction Methods*

The below table.1 shows that accuracy of SAE-SVM in handwritten digits recognition is up to 99.31%, which is much higher than SVM of Gaussian kernel for feature extraction. SAE is deep neural network, which has a strong characteristic expression and modeling capabilities for complex tasks because of multi-layer linear structure. Handwritten digits recognition model based on SAE provides a more appropriate initial value for the network through greedy layer-wise approach. It can

learn the structure of the original input data and find more useful features through abstract representation of original input step by step. What's more, it further determines the optimum network model after fine-turning supervised system (SVM), which improves the accuracy of handwritten digits recognition

Tab.1 Test data accuracy of different Classifier and feature extraction algorithm

| Classifier | Pretreatment | Misclassification rate (%) |
|---|---|---|
| Single-layer perceptron's linear classifier | No | 12.0[6] |
| K-nearest neighbor | No | 5.0 |
| SVM of Gaussian kernel | Anti-distortion | 1.4 |
| Three-layer feedforward network containing 500+150 hidden units | No | 2.95 |
| LeNet-5 | No | 0.95 |
| SAE-SVM | No | 0.69 |

## Conclusion

In this article, we use SAE-SVM algorithm to construct handwritten digits recognition model. Multilayer structure of network improve the data representation ability and effectively portray the nonlinear and stochastic volatility of data. Greedy layer-wise approach overcomes the defects that model is easy to fall into local minimum and improve the scalability of the model to some extent so that it can get better accuracy. Compared with the method of traditional feature extraction, its accuracy improves 99.31% without supervision and manual design, which provides a broader platform for the better development of handwritten digits recognition.

## References

[1] S.v.Rice,G.Nay and T.A.Narker.Optical Character Recognition:An illustrated guide to the frontier. Kluwer Academic Publishers.1999

[2] Hinton Geoffrey E.Osindero Simon,The,Yee-Whye.A Fast Learning Algorithm for Deep Belief Nets[J].NEURAL COMPUTATION,2006, 18(7):1527-1554

[3] Bengio.Y,Lambin.P,Popvici.D,et al. Greedy layerwise training of deep networks[J].Advances in Neural Information Processing Systems,2007(19):153-160

[4] TetKo,I.V;Livingstone,D.J.;Luilk,A.I.Neural network studies 1.Comparison of Overfitting and Overtraining,J.Chem.Inf.Comput.Sci.,1995,35,826-833

[5] Bell A J.Sejnowski T.J. The "independent components" of natural science are edge fiters[J].Vision rescarch.1997.37(23):3327-3338

[6] Y.Lecun,L.Bottou,Y.Bengio and P.Haffer,"Gradient-Based Learning Applied To Document Recognition,"Proceedings of the IEEE,pp.278-2324,10 1998