TR－A－0094

Extraction of the Nonlinear Global Coordinate System of a Manifold by a Five Layered Hour-Glass Network

Bunpei IRIE and Mitsuo KAWATO

1990. 11.21

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷 ☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

Extraction of the Nonlinear Global Coordinate System of a Manifold by a Five Layered Hour-Glass Network

Bunpei IRIE , Mitsuo KAWATO
ATR Auditory and Visual Perception Research Laboratories
Sanpeidani Inuidani Seika-cho Soraku-gun Kyoto 619-02 Japan

Abstract

One of the advantages of the Multi Layered Perceptron (MLP), combined with Back Propagation (BP) algorithm, is its capability of learning from examples. On the other hand, Memory Based Reasoning (MBR) is also known by its learnability from examples, in which method the system memorizes the entire set of the examples of known input-output correspondence and interpolates them in order to calculate outputs for unknown inputs. Naturally, there arises a question whether MLP is a mere variety of MBR where example data are compressed to some extent. In this paper, we will show that MLP has an additional property, i.e. the capability of acquiring internal representation from examples.

To show this, a five layered perceptron is made to learn the identity mapping from the input layer to the output layer. Input vectors are distributed on a manifold whose dimension is identical to the number of units in the compressed representation of the third layer. In this configuration, we show that the network succeeds in acquiring the global nonlinear coordinate system which is evidently most suitable for the distribution of the example data. The way to make use of the result for some applications is also discussed.

Introduction

There have been various attempts of applying Multi Layered Perceptron (MLP) to information processing problems, such as pattern recognition, image compression, speech production, etc. In such applications, MLP can be looked upon as an continuous mapping from the input vector to the output vector. Since the existence of the solution for arbitrary mapping is guaranteed [2], the main problem is how to determine parameters (connection weights and thresholds) for respective problems in order to obtain desired input-output correspondence. By using such algorithm as Back Propagation (BP), appropriate settings of network paremeters can be automatically determined by iterative presentation of input-output correspondence examples. This property of learning -from-examples is thought to be one of the advantages of MLP, since it can lead to programless information processing. However, the same advantage is shared by Memory Based Reasoning (MBR), in which method, the system memorizes the entire set of the known examples of the input-output correspondence and interpolates them to calculate outputs for unknown inputs. Of course, the latter method needs substantial storage for complicated problems, but the problem of the learning time exists in the case of MLP, instead. Stanfill and Waltz [4] have shown that the system named MBRtalk using MBR method proves at least as efficient as NETtalk [3] which uses MLP for the task of telling English pronunciation from spelling. This result implies that MLP might be a mere variety of MBR, that is, the mechanism of MLP is just to interpolate the known example data rather than to calculate outputs using the rule inferred from the examples. Here arises a question, "is MLP a mere variety of MLP where example data are compressed to some extent?" The objective of this paper is to find an answer to this question.

On the other hand, there are claims that MLP has the capability of feature extraction, or the capability of acquiring the internal representation of the inputs. However, in those reports claiming the capability, statistical analysis methods (e.g. principal component analysis = PCA) are usually used for finding the extracted features or acquired internal representation. Considering that the capability of the statistical analysis system to find something out of the

data is very powerful, it is rather difficult to declare that the features, or the representations have been extracted by the MLP, not by the analysis process. Here arises the second question, "is it really possible for MLP to acquire the internal representation (or extract the features) from examples?

What is internal representation

Let us imagine the effect of the coordinate transformation on the interpolation. Since the interpolation process is greatly influenced by the distance measure (definition of norm) of the space where the data are represented, and since the distance measure is not preserved (i.e. the ordering of two distances can be reversed) even by a linear transformation, the coordinate transformation can cause an essential alteration in the interpolation process. In fact, Stanfill and Waltz [4] lay emphasis on the importance of the data representation for getting good performance of the MBR. While the date representation is fixed in the case of MBR, it is changeable in the case of MLP, where the coordinate transformation is performed by the connection between the input layer and the hidden layer whose connection weights are chaged by the BP algorithm automatically. Therefore, one candidate for the advantage of MLP over MBR is the property of coorindate transformation from the input layer to the hidden layer.

Then, what kind of coordinate transformation is the connection between the input layer and the hidden layer supposed to do? To answer this question, it is reasonable to consider what kind of interpolation is required.

In general, the data we want to process have some probabilistic distribution in the coordinate system (of the input layer units) by which they are originally represented. Here, let us assume that the data are distributed on some lower dimensional manifold in the original space. This assumption may seem to be too restrictive. But when we want to do some information processing, the input data are usually generated through some physical processes. It follows that the data source has some physical mechanism, in which case, the input data are distributed around a quite low dimensional manifold because any physical process can be expressed in simple equations

of real parameters. Now, if the data are tranformed into the the global coordinate system which is naturally defined in the manifold on which the data are distributed, then we can get the best interpolation for the data because the interpolated points naturally fall into the manifold.

This is illustrated in Fig. 1. In the figure, the data are originally represented in the two dimensional space S using the coordinate system (x, y). However, by the physical characteristics of the generating system of data, the data points distribute on the one dimensional manifold (curved line) M. When we want to interpolate between point a and b to find the center point, if we do so in the space S using (x, y) coordinate system, what we get will be c', which doesn't fall into the original distribution. On the other hand, if we interpolate a and b in the manifold M using the single global coordinate axis p of M, the resultant point c will be inside the manifold. If the MLP automatically transforms the (x, y) representation into p representation, it is reasonable to

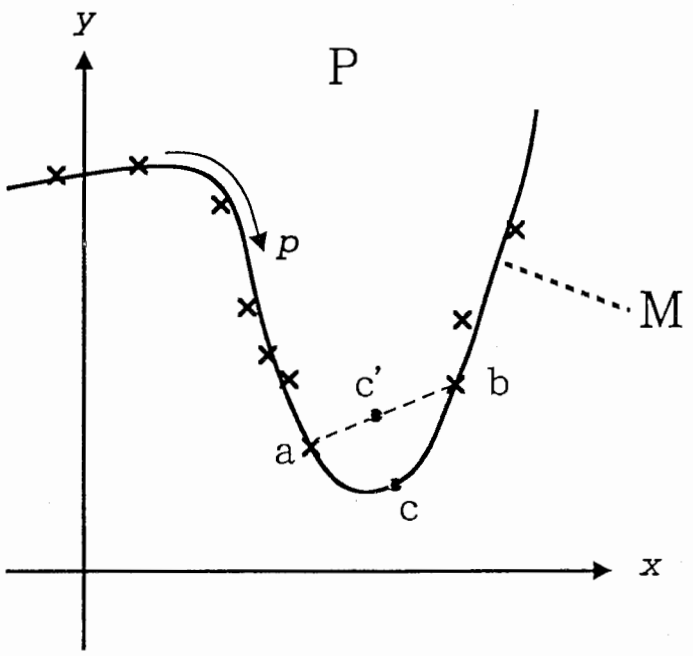

Fig.1 interpolation in the original space and in the manifold on which the data distribute

say that the system has found the internal representation (= manifold M).

By using the conventional Principal Component Analysis (PCA), a linear subspace of the original space on which the data distribute can be found. However, it is not likely that the manifold on which the data are distributed happens to be the linear subspace of the original space. In the following sections, we will show that the MLP has the capability of finding the global coordinate system of the manifold which is not necessarily the linear subspace of the original spac. On analyzing the transformation, we just observe the firing level of the units, which will prevent the "feature extraction by analyzing the network, not by the network itself".

Network topology

Bourlard and Kamp [1] have shown that PCA can be executed by the three layered hour-glass model. A three layered hour-glass model is a kind of MLP of which the input and output layer has the same number of units and the hidden layer has fewer. They have shown that if the same data is used for input and output in each step for training, the firing level of the hidden layer units converge to the principal components or their linear combination of the distribution. In this case, a linear projective transformation is executed in the connection between the input layer and the hidden layer, and its inverse transformation is performed in the connection between the hidden layer and the output layer. (See Fig. 2) In order to enable nonlinear coordinate transformation, we have added extra hidden layers before and after the single hidden layer of the three layered hour-glass model. The resultant five layered hour-glass model is shown in Fig.3. The same network model has been used for image compression by Katayama. It is guaranteed by Irie and Miyake [2] that an arbitrary continuous nonlinear coordinate transformation from the first (input) layer to the third layer, and its inverse transformation from the third to the fifth (output) layer can be realized by increasing the units of the second and the forth layer. Each of the units of the third layer is supposed to correspond to the coordinate axis of the manifold on which the input data points distribute.

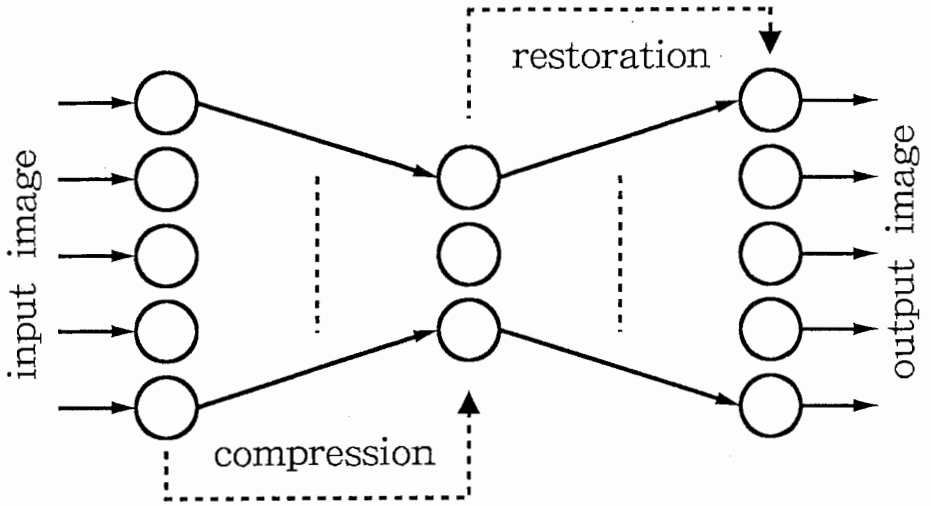

Fig.2 Image compression by a three layered hour-glass model

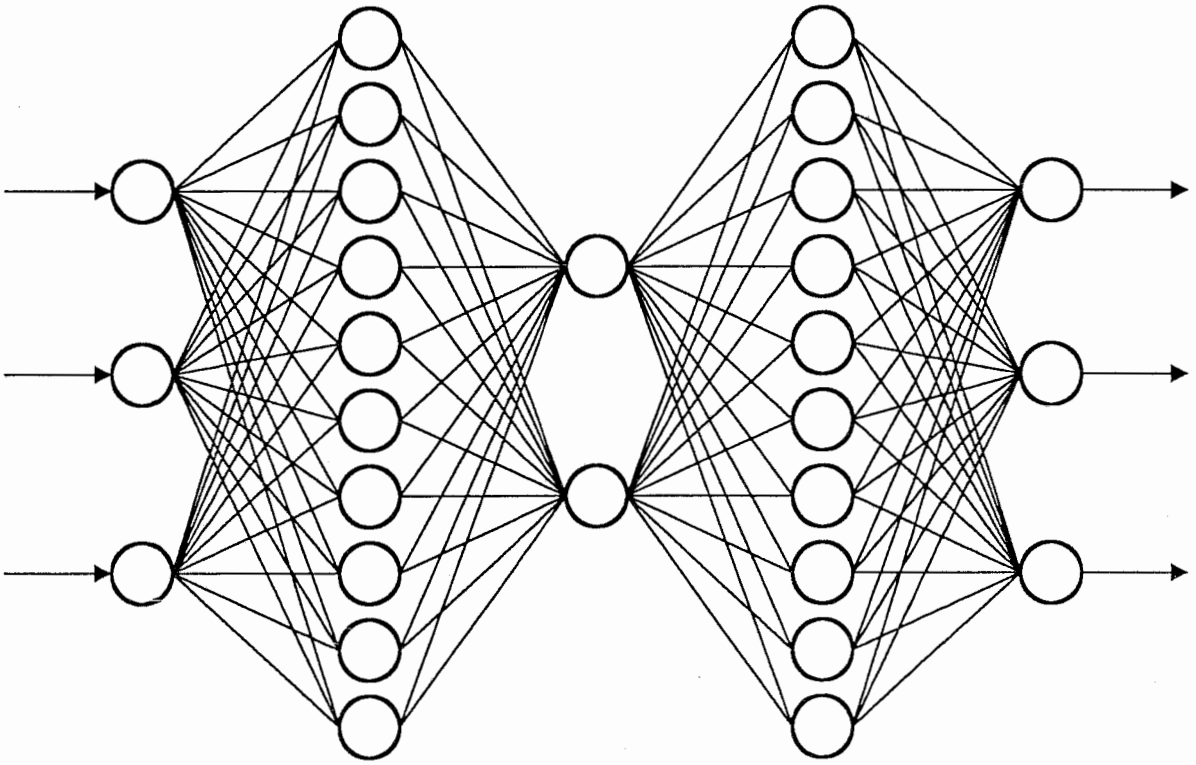

Fig.3 Five layered hour-glass model

Simulation 1

We fixed the manifold, chose random points from it and used them for the learning examples of the hour-glass model. If the firing level of the third layer units correspond to the nonlinear coordinates of the manifold after learning, we can conclude that the network has acquired the internal representation. First, we conducted a simulation for the case of one dimensional manifold (a semicircle) in the two dimensional Euclidean space with Cartesian coordinate system (x,y) (See Fig. 4). Both the number of the units of the input and output layer was set to two, while that of the third layer was set to one. The number of units of the second and the forth layer correspond to the degree of nonlinearity of the coordinate transformation from input layer to the third layer, and the inverse transformation from the third layer to the output layer, respectively. As these numbers were not so essential for the purpose of the simulation, they were both set to ten. As to the activation function, we used sigmoid function only for the units of the second and the forth layer, in which layer the nonlinearity is essential for generating arbitrary

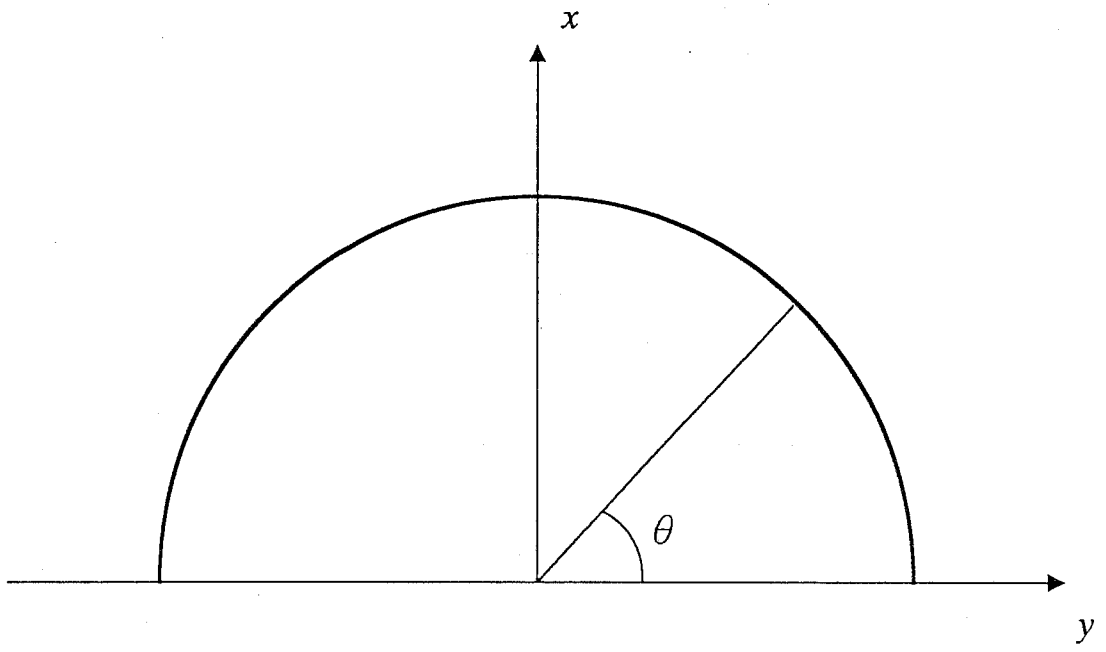

Fig.4 A semicircle used for the simulation

transformations. In the units of other layers, the weighted sum of the inputs were directly used for outputs. This was effective also for speed up of the simulation. Fifty training examples were randomly chosen from the semicircle ($x=\cos\theta$, $y=\sin\theta$, $0\leqq\theta\leqq\pi$). The data were iteratively presented to the input layer and the output layer simultaneously. Fig. 5 is the graph of the relationship between θ and the firing level of the single unit of the third layer after convergence. θ is on the horizontal axis and the firing level of the third layer units is on the vertical axis. A monotone continuous relationship can be seen in the graph. This shows that the third layer unit has acquired the internal representation of the data.

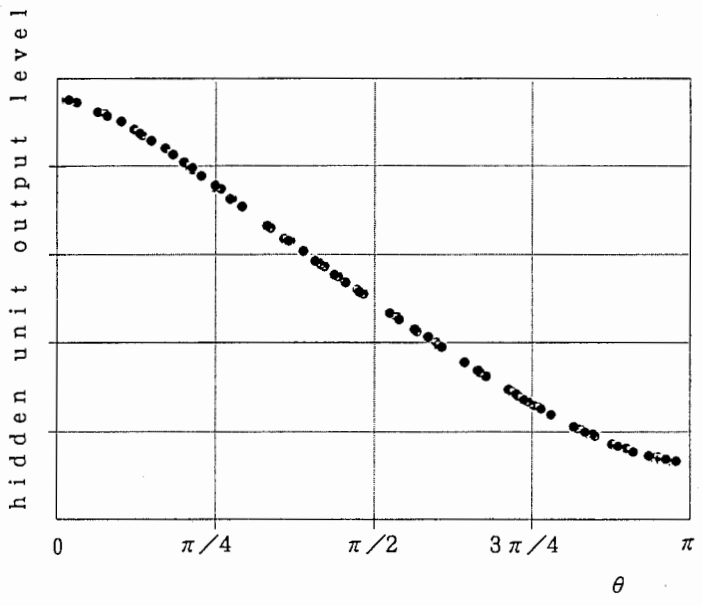

Fig.5 Relationship between θ and the output of 3rd layer unit

Simulation 2

We conducted another simulation increasesing the dimension by 1, i.e. a hemisphere in the three dimensional Euclidean space with Cartesian coordinate system. In this case, the number of units are 3, 10, 2, 10, 3. The exact topology of this network is illustrated in Fig. 3. Again, fifty random points were choson from the hemisphere ($x=\cos\theta\sin\phi$, $y=\sin\theta\sin\phi$, $z=\cos\phi$, $0\leqq\theta\leqq2\pi$, $0<\phi\leqq\pi/2$). This time, we had to devise a scheme to illustrate the resultant relationship between the four parameters : θ, ϕ, and the firing level of the two units of the third layer. For this purpose, we ignored the input and the second layer of the network. We manually set the firing level of the two units of the third layer to grid points ($x_0+m\times\Delta x, y_0+n\times\Delta y$), and calculated the output level of the units of the network, which correspond to points in the three dimensional space. The result is shown in Fig. 6. The adjacent points are connected by line segments. As is evident from the figure, the firing level of the third layer units correspond to a curved global coordinate system on the hemisphere, which is apparently the desired internal representation. Note that if PCA, which is a linear method, is used, this problem of finding the representation for a hemisphere is a very hard one. The linear method can only approximate the hemisphere by a plane intersecting it.

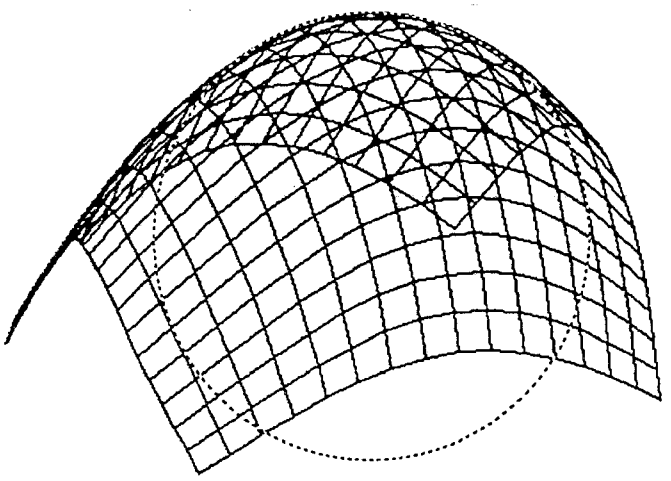

Fig.6 Self organized curved coordinate axes

Conclusion and discussion

The MLP's property of acquiring internal representation from examples has been shown for some examples by simulation. Since the extraction of the internal representation is equivalent to nonlinear coordinate transformation, MLP has the potential to execute interpolation, i.e. generalization, in a different way from MBR. In the simulation, five layered hour-glass model has been employed for nonlinear coordinate transformation. The model is the improved version of the three layered hour-glass model, which can only do linear coordinate transformation. The network model employed here can be used as a feature extractor by eliminating the latter half, and can be used as a preprocessor for pattern recognition, image processing and other applications. (See Fig. 7)

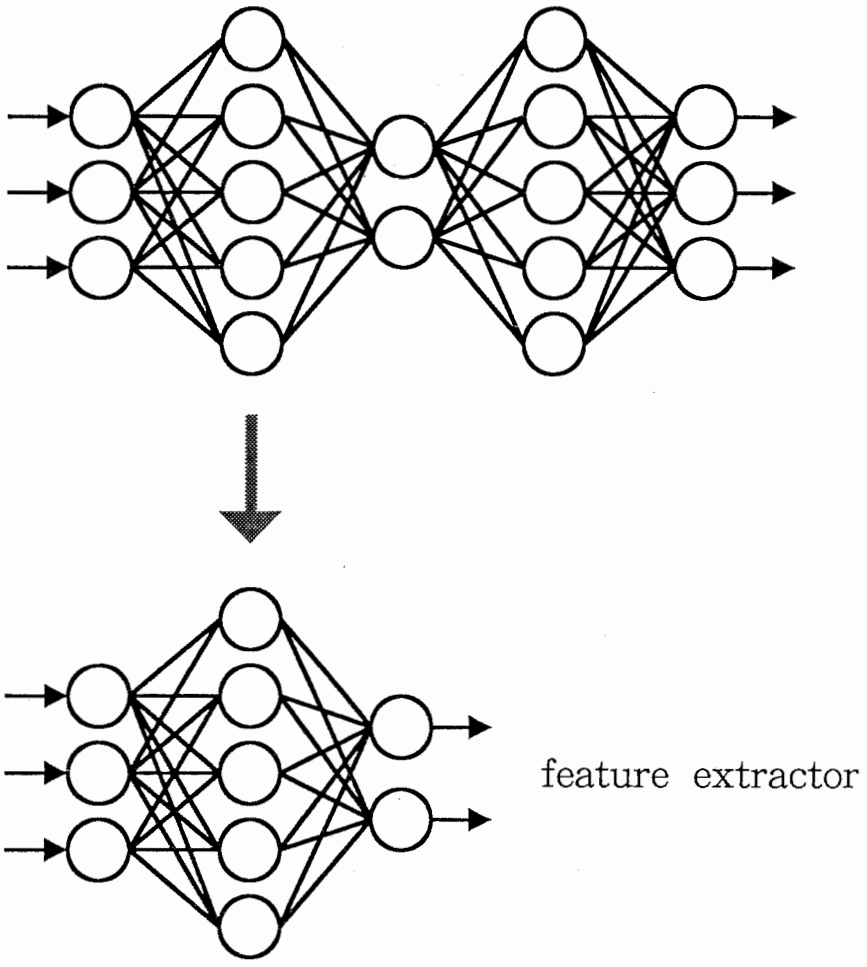

Fig.7 Feature extractor using part of the five layered hour-glass model

In the simulation, the dimension of the manifold, which is of course lower than the dimension of the original space, was known in advance, and we set the number of the third layer unit to this number. Considering that the BP algorithm just executes the gradient descent method, the only constraint for this problem is this setting of the third layer. Therefore, the successful acquisition of internal representation (= nonlinear global coordinate axes of the manifold) is attributed to this dimensionality reduction.

Then, naturally there arises a question, what if the dimension of the manifold is unknown. In connection with this question, we propose a conjecture that if the dimension of the third layer is too small for the manifold, the connection weights will diverge on the way of learning. Using this fact, the appropriate number of unit can be found by increasing the number gradually while divergence is observed.

The simulations for more complicated data, natural data and the theoretical analysis are our future problem.

References

[1] Bourlard, H. and Kamp, Y. : Auto-association by multilayer perceptrons and singular value decomposition,Biological Cybernetics, 59, pp.291-294, 1988.

[2] Irie, B. and Miyake, S. : Capabilities of three-layered perceptrons, Proc. ICNN 88, vol. I , pp.641-648,1988.

[3] Sejnowsky, T. J. and Rosenberg, C. R., NETtalk : a parallel network that learns to read aloud, Johns Hopkins Univ. Tech. Rep. JHU/EECS-86/01, 1986.

[4] Stanfill, C. and Waltz, D. : Toward memory-based reasoning. Communications of the ACM, vol.29, pp.1213-1228, 1986.