Advances in Engineering Research (AER), volume 124
2nd International Symposium on Advances in Electrical, Electronics and Computer Engineering
(ISAEECE 2017)

Unified Entropy in Self-organizing Feature Maps Neural Network

Chunyang Zhu
Shanghai Polytechnic University
zhuzhuMotto@hotmail.com

Keywords: Unified Entropy, Mutual Information, Self-organizing Feature Maps.

Abstract. Pattern Recognition is a very urgent research area in intelligent information processing and computer intelligent perception, such as computer vision, content-based retrieval, image-processing, etc. In general, the research on pattern recognition is carried out partial separately as feature extraction, classification, etc. in which samples of feature extraction could not be reliable and the global optimum could not been achieved. In this paper the unified entropy theory on Pattern Recognition is presented firstly, in which the information procedures in learning and recognition and the determine role of Mutual Information have been discovered. Secondly build SOFM neural network and apply Mutual Information entropy to compute reliability of training samples, through which selecting excellent data samples is presented to get optimum recognition performance, which is crucial for difficult pattern recognition problems. Experiments on device state recognition prove their effective and efficient.

Introduction

Pattern recognition is important and growing fast in intelligent information processing, and efficiently and effectively affect today's progresses on computer vision, biometrics, video surveillance, etc. because Pattern Recognition is the fundamental of intelligent activities.

There are two important problems in pattern recognition: one is classifier design, and the other is feature extraction/selection. More papers have been published for classifier design and for feature extraction independently[1], but few for the relation between them, i.e., few from whole reliability of learning recognition and corresponding samples. More papers published for the research on this area partial separately, such as machine learning, classifier design, and feature selection, etc., but few papers to studied feature extraction/samples selection globally together with classifier design, or with samples learning and recognition. In fact, pattern recognition is close relative with both feature extraction/samples selection and classifier design, and with both reliability of learning recognition procedure and samples. The research on pattern recognition on the global and relative way rather is a very urgent problem.

Designing a good performance recognition system is still an urgent issue in pattern recognition research. The researchers know that the classifier design and the feature extraction are the most important, but they need to know how classifier is reliable. In this chapter, first, we extend the Information Theory into Pattern Recognition area. An unified entropy theory of Pattern Recognition is represented, in which pattern recognition can be described by an information entropy procedure. Also the mutual information generated from learning procedure and selecting excellent data samples is crucial affecting and determining the recognition performance. Second, recognition procedure is crucial for getting optimum recognition performance and is an urgent step for solving difficult pattern recognition problems, here build SOFM neural network, based on experiments on device state recognition, we select excellent data samples through the reliability of mutual information entropy of SOFM corresponding recognition samples, which prove the effective and efficient to achieve the excellent pattern recognition. Even now more statistical learning theory and algorithms have been studied and made more progresses[2,3], which are more direct based on the samples learning for pattern classification, such as SVM, Adaboosting, etc. The statistical learning theory studies the learning methods for classifier design, avoiding the most difficult probability distribution estimation problems in general pattern recognition model. Even though, there still could not regardless the existence of various probability distributions of random variables. Therefore, the unified information entropy theory is useful to the reliability of algorithm classification recognition , Despite of various learning methods, the entropy analysis and unified Entropy frame is useful for deeper analysis and comprehensive understanding of pattern recognition. Because mutual information is the discriminate entropy for recognition, which will reduce the indiscriminate components and noise to achieve the good recognition performance for difficult pattern recognition problems[4,5,6,7]

Unified Entropy Theory[8]

Recognition is a procedure to determine the category of an unknown testing sample based on some known category samples, which can be described by an information entropy procedure. The information entropy system of pattern recognition is composed by feature entropy $H(F)$, system ntropy $H(E)$, conditional entropy H (F/E),a posterior entropy H (E/F) , and mutual information $I(F, E)$, which described in the Appendix.

The unified entropy procedure including:

Learning information procedure, $I(F, E) = H(F) - H$ (F/E),

Recognition information procedure, H (E/F) $= H(E) - I(E, F)$,

in which a whole information procedure happens in pattern recognition

Pattern recognition is to identify the category or index of an unknown sample from a category probability space Ω = (ω1, . . . , ωn; P (ωi) , . . . , P (ωn)), which has preliminary category uncertainty described by the system entropy

$$H(E) = -\sum_{i=1}^{n} P(\omega_i)\log P(\omega_i), \sum_{i=1}^{n} p(\omega_i) = 1. \tag{4}$$

For example, for Chinese character recognition, H(E)=12–16 bits.When the feature X is extracted from sample in the feature probability space,F = (X; P (X)) .

Therefore, for the total sample set, the sample feature matrix could be represented as X = [X1, X2, · · · , XL].

The probability distribution of sample feature X could be estimated from more samples, but it is a very difficult problem forever. The mean feature vector and the feature covariance matrix can be estimated by the training samples.The mean feature vector is estimated as.

$$\hat{M} = \frac{1}{L}\sum_{i=1}^{L} X_i \tag{5}$$

Then the feature covariance matrix can be estimated as

$$\hat{\Sigma}_X = \frac{1}{L}XX^T - \hat{M}\hat{M}^T \tag{6}$$

Provided that the feature probability density $p(X)$ is a Gaussian distribution, then $p(X)$ could be estimated by the mean
feature vector and the feature covariance matrix only:

$$P(X) = \frac{1}{(2\pi)^{\frac{N}{2}}\left|\hat{\Sigma}_X\right|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(X-\hat{M})^T\Sigma_X^{-1}(X-\hat{M})\} \tag{7}$$

Based on the feature probability density $p(X)$, the feature differential entropy $h(F)$ could be calculated as

$$\max_{\Sigma_X=E\left\{XX^T\right\}} h(F) = \frac{1}{2}\log(2\pi e)^N\left|\Sigma_X\right| \tag{9}$$

with equality if $X \in N\left(M_X,\right)$ And the feature entropy will be

$$H(F) = \frac{1}{2\ln 2}\left[\ln\left(2e\pi\right)^N + \ln\left|\hat{\Sigma}_X\right|\right] \tag{10}$$

Learning information procedure

In the training procedure, the features probability distribution and feature conditional probability distribution could be estimated from training samples,then both feature entropy H (F) and category conditional entropy H (F |E) are obtained. The leaning entropy reduction H (F)- H (F |E), and the same as the mutual information I(F, E) = H(F) - F(F |E) will be obtained too.

Based on the training samples, the ith category training sample set is represented as a sample matrix $X_{ii} = \{X_{i1}, X_{i2}, \cdots X_{iL}\}, i = 1,2,\cdots,n.$

Then the *ith* category mean vector and its covariance matrix can be estimated as

$$\hat{M}_i = \frac{1}{L_i}\sum_{j=1}^{L_i} X_{ij}, i = 1,2,\cdots,n. \tag{11}$$

$$\hat{\Sigma}_i = \frac{1}{L} X_i X_i^T - M_i M_i^T, i = 1,2,\cdots,n. \tag{12}$$

Provided that the ith class-conditional probability density p (X |ωi) (i = 1,2, · · · , n) is a Gaussian distribution, it will be represented as

$$p(X/\omega_i) = \frac{1}{(2\pi)^{\frac{N}{2}} \left|\hat{\Sigma}_X\right|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(X - \hat{M})^T \Sigma_X^{-1} (X - \hat{M})\} \tag{13}$$

The class-conditional feature entropy H (F |E) will be calculated as

$$H\left(F/E\right) = \sum_{i=1}^{n} P(\omega_i) H\left(F/\omega_i\right) \le \frac{1}{2\ln 2}\left[\ln(2e\pi)^N + \sum_{i=1}^{n} P(\omega_i) \ln\left|\hat{\Sigma}_i\right|\right] \tag{14}$$

with equality if $X \in N(M_X, \Sigma_X)$

Therefore, the learning entropy reduction or the mutual information in feature space of Gaussian feature will be

$$I(F,E) = H(F) - H(F/E) = \frac{1}{2\ln 2}\left[\ln\left|\Sigma_X\right| - \sum_{i=1}^{n} P(\omega_i) \ln\left|\Sigma_i\right|\right] \tag{15}$$

The learning entropy reduction in feature space obtained from machine learning represents the acquired information from training samples, and it presents the relation between feature and category regardless the learning methods, even it is a statistical parameter estimation of probability distributions, or any statistical learning methods. The smaller the Class Feature Entropy H (F |E) is, the bigger the learning entropy reduction I(F, E) will be. The class feature entropy H (F |E) represents the feature variation about the category, or the feature instability of the pattern, which is injuring information for recognition, and will weaken the feature recognition ability.

It should be noted that the mutual information is transferred from mutual information in feature space into mutual information in class space when the cognition procedure happens.

Cluster with Self-Organizing Feature Maps Neural Network

Self-organizing feature maps (SOFM)[9] learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution (as do competitive layers) and topology of the input vectors they are trained on.

The neurons in the layer of an SOFM are arranged originally in physical positions according to a topology function, which can arrange the neurons in a grid, hexagonal, or random topology. Distances between neurons are calculated from their positions with a distance function. There are four distance functions, dist, boxdist, linkdist, and mandist. Link distance is the most common. These topology and distance functions are described in Topologies (gridtop, hextop, randtop) and Distance Functions (dist, linkdist, mandist, boxdist).

Here a self-organizing feature map network identifies a winning neuron i* using the same procedure as employed by a competitive layer. However, instead of updating only the winning neuron, all neurons within a certain neighborhood Ni* (d) of the winning neuron are updated, using the Kohonen rule. Specifically, all such neurons i ϵ Ni* (d) are adjusted as follows:

$$iw(q)=iw(q-1)+\alpha(p(q)-iw(q-1)). \quad (1)$$

or

$$iw(q)=(1-\alpha)iw(q-1)+\alpha p(q). \quad (2)$$

Here the *neighborhood Ni** (*d*) contains the indices for all of the neurons that lie within a radius *d* of the winning neuron *i**.

$$Ni(d)= \{j, d_{ij} \le d \} \quad (3)$$

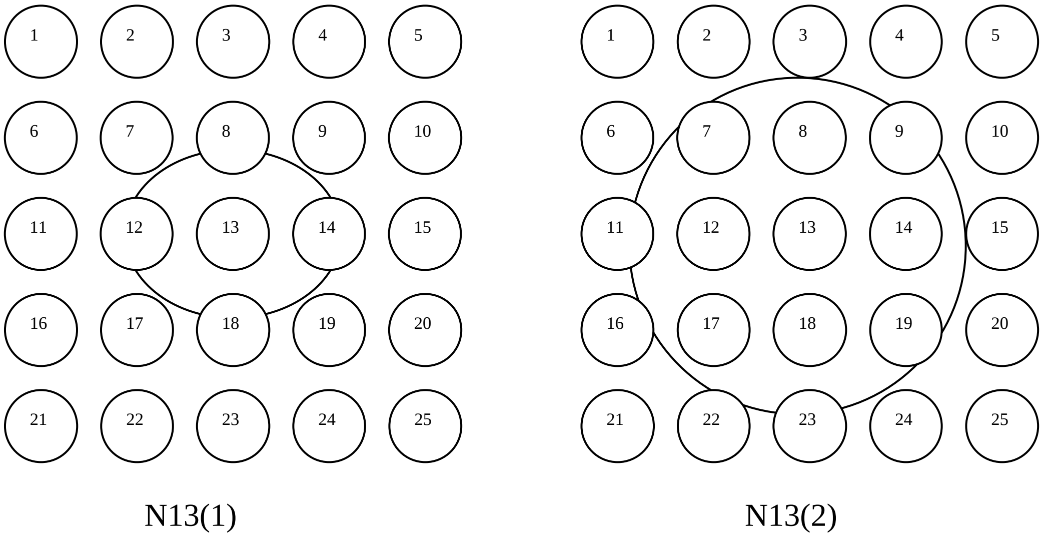
Thus, when a vector p is presented, the weights of the winning neuron and its close neighbors move toward p. Consequently, after many presentations, neighboring neurons have learned vectors similar to each other.

Another version of SOFM training, called the *batch algorithm*, presents the whole data set to the network before any weights are updated. The algorithm then determines a winning neuron for each input vector. Each weight vector then moves to the average position of all of the input vectors for which it is a winner, or for which it is in the neighborhood of a winner.

To illustrate the concept of neighborhoods, consider the figure 1, the batch algorithm of SOFM . The left diagram shows a two-dimensional neighborhood of radius *d* = 1 around neuron 13. The right diagram shows a neighborhood of radius *d* = 2.

These neighborhoods could be written as N13(1) = {8, 12, 13, 14, 18} and N13(2) = {3, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 23}.

The neurons in an SOFM do not have to be arranged in a two-dimensional pattern. You can use a one-dimensional arrangement, or three or more dimensions. For a one-dimensional SOFM, a neuron has only two neighbors within a radius of 1 (or a single neighbor if the neuron is at the end of the line). You can also define distance in different ways, for instance, by using rectangular and hexagonal arrangements of neurons and neighborhoods. The performance of the network is not sensitive to the exact shape of the neighborhoods. Figure.2 show neurons arrangement in different topology.



N13(1) N13(2)

Fgure.1. Batch Algorithm of SOFM

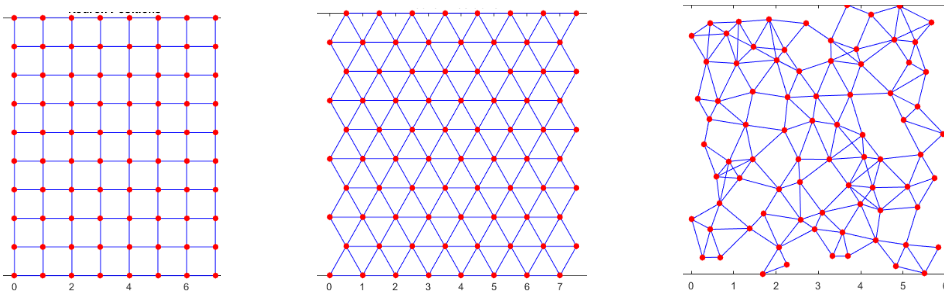


Figure. 2. Neuron Positions in Topology

The graph below shows a home neuron in a two-dimensional (gridtop) layer of neurons. The home neuron has neighborhoods of increasing diameter surrounding it. A neighborhood of diameter 1 includes the home neuron and its immediate neighbors. The neighborhood of diameter 2 includes the diameter 1 neurons and their immediate neighbors.

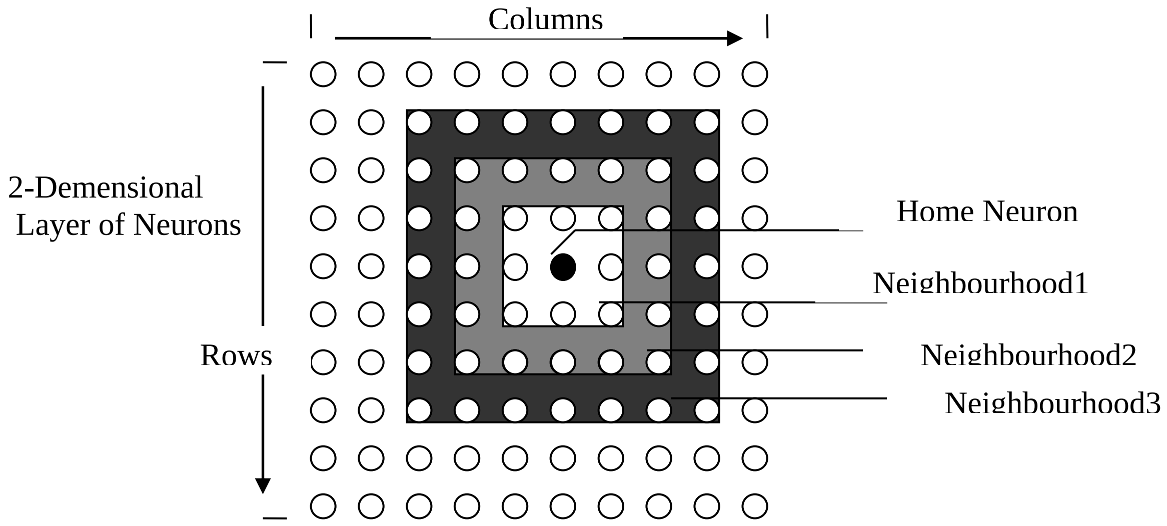


Figure.3. 2-Dimensional layer of Neurons

As for the dist function, all the neighborhoods for an S-neuron layer map are represented by an S-by-S matrix of distances. The particular distances shown above (1 in the immediate neighborhood, 2 in neighborhood 2, etc.), are generated by the function boxdist,The Manhattan distance between two vectors x and y is calculated as D = sum(abs(x-y)).

Now, however, as described above, neurons close to the winning neuron are updated along with the winning neuron. You can choose from various topologies of neurons. Similarly, you can choose from various distance expressions to calculate neurons that are close to the winning neuron.

Unified Entropy Theory in SOFM Neural Network

Algorithm Procedure of Unified Entropy Theory in SOFM Neural Network. SOFM is a kind of nearest neighbor classifier which can get optimum reference input vectors through unsupervised and competitive learning. provided that there are parameter vectors which has differnt states, there are k training input vectors : $X_k=\{x_{k1},x_{k2},\cdots x_{kn}\}$,which are classified into p classes, so, there are corresponding p weights, they are: $W_i=\{w_{i1},w_{i2},\cdots w_{in}\}\{(i=1,2,\cdots P)(n=1,2,\cdots m)\}$,at the same time, W_i is the center of the ith class, algorithm procedure of unified entropy theory in SOFM is showed as follows:

(1)The winning neuro C_j.

$$D_{\min}=\left\|X_k-W_{C_j}\right\|=\min_{i=1}^{p}\left\|X_k-W_{C_j}\right\| \tag{17}$$

equation(17) can reflect corresponding input vector through the nearest distant parameter, which represents similarity between C_j and corresponding input sample.

Provided that there is p classes, which has p corresponding weights $W_i = \{w_{i1}, w_{i2}, \cdots w_{in}\}$.

(2)Self-organizing of weight.

Equation (18) is the nearest field function which modifies the weight between neurons of the nearest field and input neurons, $N_{iC}(t)$ is the distance between someone neuron i of output layer and the winning neuron C, the distance is larger, modified weight is smaller.

$$\Delta W_{ij}(t) = \eta(t) N_{ic}(t)(X_k - W_{ij}(t)) \qquad (18)$$

Equation (19) is a function of modifying the weight of every neuron.

$$W_{ij}(t+1) = W_{ij}(t) + \Delta W_{ij}(t) \qquad (19)$$

Equation (20) supposed that $N_{iC}(t)$ is normal distribution, $\eta(t)$ is the learning rate. $|r_i - r_C|$ is the distance between the i_{th} and the winning C, T is the total learning times, $\sigma(t)$ is the nearest neighbor radius.

$$N_{ic}(t) = \exp(-\frac{|r_i - r_C|^2}{2\sigma^2(t)}) \qquad (20)$$

$$\eta(t) = \eta_0 \times \exp(-t/T) \qquad (21)$$

$$|r_i - r_C|^2 = ((i_x - C_x)^2 + (i_y - C_Y)^2)^{1/2} \qquad (22)$$

$$\sigma(t) = \sigma_0 \times \exp(-t/T) \qquad (23)$$

(3)Computing the mutual information I(F, E)of every winning neuron w_i

Apply the function Equation (15) to compute the mutual information I(F, E)of every winning neuron w_i, w_i is the classification neuron , X are some samples, n numbers is the sum of samples which belong to the classification w_i of output layer. here $P(w_i)$ is the probability of corresponding winning neuron classification w_i and $P(w_i)$ is normal distribution function. So equation(15) is the value of mutual information ,that is entropy value of all samples belong to the same class w_i, the bigger I(F,E)is, the better recognition will be achieved, and it is more reliable.

$$I(F,E) = H(F) - H(F/E) = \frac{1}{2\ln 2}\left[\ln|\Sigma_X| - \sum_{i=1}^{n} P(w_i)\ln|\Sigma_i|\right] \qquad (15)$$

The Application Analysis of Algorithm Procedure. Build 2 dimensions SOFM neural network and there are 9 neurons in output layer.

(1)Training the SOFM network

There are some1200 samples we know their classification, these samples can be classified into 2 classifications, which are normal and abnormal state. We applied these samples to train this SOFM neural network, initial learning rate $\eta(0) = 0.01$, the training maximum times $T = 500$,table1. shows the winning times of the winning neurons and the mutual information value I(F,E)of every neuron which is corresponding classification. the mutual information value I(F,E) is approximately equal to 1, which shows 9 neurons of output layer are reliably approaching to input samples. Neuron 3 and Neuron 4 represent abnormal state, other neurons represent normal state. If the mutual information value I(F,E) is lower than 0.5. we can delete corresponding data samples.

Table 1. the Last Weight and Mutual Information of the Output Layer Neurons (Training Data Samples)

Neuron weight	w_{i1}	w_{i2}	w_{i3}	w_{i4}
w_1	0.2726	0.2751	0.2784	0.2826
w_2	0.2797	0.2818	0.2836	0.2851
w_3	0.2828	0.2841	0.2852	0.2865
w_4	0.2759	0.2786	0.2813	0.2840
w_5	0.3042	0.3007	0.2971	0.2935
w_6	0.2861	0.2866	0.2872	0.2877
w_7	0.2897	0.2896	0.2895	0.2893
w_8	0.2982	0.2962	0.2938	0.2914
w_9	0.2938	0.2927	0.2916	0.2903
Neuron Weight	w_{i5}	w_{i6}	w_{i7}	w_{i8}
w_1	0.2863	0.2902	0.2936	0.2975
w_2	0.2872	0.2891	0.2914	0.2934
w_3	0.2878	0.2892	0.2905	0.2919
w_4	0.2869	0.2900	0.2931	0.2957
w_5	0.2899	0.2862	0.2825	0.2790
w_6	0.2883	0.2889	0.2895	0.2902
w_7	0.2890	0.2886	0.2883	0.2880
w_8	0.2895	0.2874	0.2851	0.2830
w_9	0.2891	0.2879	0.2868	0.2858
NeuronWeight	w_{i9}	w_{i10}	w_{i11}	w_{i12}
w_1	0.3011	0.3046	0.2723	0.3046
w_2	0.2958	0.2979	0.2794	0.2979
w_3	0.2933	0.2948	0.2826	0.2948
w_4	0.2983	0.3008	0.2758	0.3008
w_5	0.2760	0.2732	0.2726	0.3042
w_6	0.2907	0.2913	0.2859	0.2914
w_7	0.2876	0.2873	0.2870	0.2903
w_8	0.2813	0.2793	0.2789	0.2982
w_9	0.2848	0.2837	0.2833	0.2939

Neuron Weight	w_{i13}	w_{i14}	Output state	Winning Times	$I(F,E)$
w_1	0.0365	－0.0021	Normal	39	1
w_2	0.0202	－0.0005	Normal	79	0.962
w_3	0.0133	0.0004	Abnormal	140	0.952
w_4	0.0281	－0.0012	Abnormal	59	0.987
w_5	－0.0351	－0.0020	Normal	60	0.973
w_6	0.0058	0.0008	Normal	181	0.928
w_7	0.0028	0.0011	Normal	211	0.916
w_8	-0.0212	-0.0008	Normal	134	0.949
w_9	0.0113	0.0006	Normal	177	0.931

(2)Recognizing unknown data samples

There are 100 unknown samples which need to be classified, we make use of the trained SOFM neural network which has gotten the last weight to predict these unknown samples, that is, the minimum distance between the weight of output layer neuron and unknown sample, corresponding winning neuron represents the state of corresponding unknown sample. Only 4 samples are predicted wrongly, So its recognition rate is 96%. Table2 gives the prediction result of 10 samples (14 dimensions vectors),for example, inputting the 1st sample into the trained SOFM is recognized to be normal state , its winning neuron is the 4th inquiry Table1, neuron No.4 represents abnormal state, so the 1st sample should be abnormal state, this sample is wrongly recognized. But other samples are rightly classified.

Table.2. Testing Unknown Samples

Samples	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	Classification	Winning Neuron
1	42.94	43.06	43.33	43.62	43.69	43.72	44.07	44.07	44.39	44.5	42.94	44.5	1.723	-0.125	Normal	4
2	37.25	37.65	37.85	38.14	38.48	39.21	39.9	40.01	40.34	41.02	37.25	41.02	4.2164	-0.220	Normal	6
3	43.19	43.11	43.04	43.01	42.93	42.89	42.86	42.81	42.74	42.74	42.74	43.19	-0.501	-0.141	Normal	2
4	35.55	35.78	35.89	36.01	36.08	36.18	36.42	36.55	36.57	36.82	35.55	36.82	1.3085	-0.276	Normal	6
5	51.83	51.9	52.36	52.76	52.77	52.96	53.42	53.88	53.95	54.28	51.83	54.28	2.7982	0.060	Normal	2
6	50.59	50.62	50.67	50.7	50.75	50.78	50.83	50.84	50.59	50.86	50.55	50.86	0.3618	0.0144	Normal	8
7	51.85	51.85	52.26	52.81	52.85	52.97	53.56	54.18	54.18	54.8	51.85	54.8	3.323	0.0626	Normal	1
8	41.95	42.76	43.11	43.67	43.76	44.34	45.16	46.14	46.43	47.17	41.95	47.17	5.6285	-0.111	Abnormal	3
9	36.52	36.95	37.82	38.7	39.38	40.21	40.55	40.66	40.99	41.51	36.52	41.51	5.683	-0.213	Abnormal	4
10	54.94	55.05	55.49	55.92	56.03	56.5	56.95	57.19	57.37	57.83	54.94	57.83	3.2915	0.1265	Abnormal	4

Conclusion

SOFM is a classical clustering algorithm in neural network, which approaches internal character of samples through the minimum distance and apply samples information into the whole neural network, neurons weight of output laye represent information. The unified entropy theory in pattern recognition SOFM neural network improve the reliability of SOFM algorithm,the mutual information is the finally determination for the recognition performance, which has been shown is the best merit for selecting excellent data samples.

References

[1] Biem A, Katagiri S, Juang B H (1997) Pattern Recognition Using Discriminative.

[2] Watanabe H, Yamaguchi T, Katagiri S (1997) Discriminative Metric Design for Robust Pattern Recognition, Transaction on signal Processing, 45(11): 2655 – 2662.

[3]Ding X Q, Chen L, Wu T (2007) Character Independent Font Recognition on a Single Chinese Character, IEEE Transaction on Pattern Recognition and Machine Intelligence, 29(2): 195 – 204.

[4] Maes F, vandermeulen D, Suetens P (2003) Medical Image Registration Using Mutual Feature Extraction, IEEE Tranc on Signal Processing, 45(2): 500 – 504.

[5] Maes F, vandermeulen D, Suetens P (2003) Medical Image Registration Using Mutual Information. Proceeding of IEEE, 91(10): 1699 – 1722.

[6] Ding S, Zhang Y, et al (2009) Research on a Principal Components Decision Algorithm Based on Information Entropy, Journal of Information Science, 35(1): 120 – 127.

[7] Escolano F, Suau P, Bonev B (2009), Information Theory in Computer Vision and Pattern Recognition, Springer, New York.

[8]Xiaoqing, Ding. Pattern Recognition and Machine Learning.

[9]http://cn.mathworks.com/help/nnet/ug/cluster-with-self-organizing-map-neural-network.html?requestedDomain=www.mathworks.com&requestedDomain=cn.mathworks.com#bss4b_l-12