Expert Systems with Applications 39 (2012) 11924–11933

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A new approach for data clustering and visualization using self-organizing maps

Shu-Ling Shieh [a,b], I-En Liao [a,*]

[a] Department of Computer Science and Engineering, National Chung-Hsing University, 250, Kuo-Kuang Road, Taichung, Taiwan
[b] Department of Information Technology, Ling-Tung University, 1, Ling-Tung Road, Taichung, Taiwan

ARTICLE INFO

Keywords:
Self-organizing map
Unsupervised learning
Visualization
Clustering method

ABSTRACT

A self-organizing map (SOM) is a nonlinear, unsupervised neural network model that could be used for applications of data clustering and visualization. One of the major shortcomings of the SOM algorithm is the difficulty for non-expert users to interpret the information involved in a trained SOM. In this paper, this problem is tackled by introducing an enhanced version of the proposed visualization method which consists of three major steps: (1) calculating single-linkage inter-neuron distance, (2) calculating the number of data points in each neuron, and (3) finding cluster boundary. The experimental results show that the proposed approach has the strong ability to demonstrate the data distribution, inter-neuron distances, and cluster boundary, effectively. The experimental results indicate that the effects of visualization of the proposed algorithm are better than that of other visualization methods. Furthermore, our proposed visualization scheme is not only intuitively easy understanding of the clustering results, but also having good visualization effects on unlabeled data sets.

1. Introduction

A self-organizing map (SOM), originally suggested by Kohonen (1990), is a very popular unsupervised neural network model for the analysis of high-dimensional patterns in data mining applications. In the area of artificial neural networks, self-organizing feature map is an excellent data-exploring tool as well. It can project high-dimensional patterns onto a low-dimensional topology map. Moreover, the SOMs are able to exhibit the visualized features and to discovery data distribution, i.e., each input data point can be assigned to a cluster according to its nearest output neuron. The neurons in SOM can reveal, in certain topological structures, the output space according to the features of the input vectors, while such types of topological structures would also denote the characteristics of the input samples. SOM has been employed in a wide range of applications in various domains, including speech recognition, image data compression, robot control, pattern recognition, and medical diagnosis (Chen, 2005; Murtagh, 1995; Petrilis & Halatsis, 2008; PÄolzlbauer, Rauber, & Dittenbach, 2005; Ressom, Wang, & Natarajan, 2003; Su, Liu, & Chang, 1999; Tangsripairoj & Samadzadeh, 2006).

There are many information visualization methods can be used to present the information learned in order to elucidate and separate the target data sets, as well as the SOM related approaches (Brugger, Bogdan, & Rosenstiel, 2008; Chakma & Umemura, 2003; Fernandez & Balzarini, 2007; Yuan-chao, Chong, & Ming, 2011; Ritter, Martinetz, & Schulten, 1992; Shieh & Liao, 2009; Su et al., 1999; Vesanto & Alhoniemi, 2000; Ultsch, 2003; Taşdemir & Merényi, 2008; Diri & Albayrak, 2008; Shieh, Liao, Hwang, & Chen, 2009). Recently the Visualization induced SOM (ViSOM) (Yin, 2002) and Probabilistic Regularized SOM (PRSOM) (Wu & Chow, 2005) have been proposed to enhance SOM's visualization and to preserve the inter-neuron distances. In the work of Wu and Chow (2005), ViSOM and PRSOM have already been proven able to provide better effects of data visualization than results made by the Curvilinear Component Analysis (CCA), Multidimensional Scaling (MDS) and Sammon's Mapping approaches. However, clustering is the problem of partitioning a set of unlabeled data into self-similar clusters. These visualization methods do not deal with unlabeled data sets at all. Our proposed visualization method is not only able to cluster unlabeled data set, but also provide the inter-neuron distances directly visible, using a colorful scheme and measurable probability of each neuron, and to find cluster boundaries on the SOM grid.

Computer aided visualization processes offer to understand information (Card & Mackinlay, 1997). The manner of obtaining clustering data from a well-disciplined feature projection map is a big challenge for SOM visualization. SOM visualization usually uses the map grids as a visualization platform. Ultsh and Siemon proposed U-matrix (1990) that shows the local cluster boundaries by depicting pair-wise distances between neighboring neurons. Another form of visualization techniques is hit histograms that

* Corresponding author.
E-mail addresses: ltcc63@teamail.ltu.edu.tw (S.-L. Shieh), ieliao@nchu.edu.tw (I-En Liao).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.181

take the distribution of the data samples into account, and present the amount of data samples which are mapped to a grid. However, due to visual weakness, the U-matrix and hit histograms approaches suffer from several drawbacks that will be discussed later in Section 2.2 in detail. The major problem of these approaches is its poor visualization map, because it does not provide an appropriate scheme for strengthening the effect of visualization and cluster label assignment. Therefore, in this paper we focus on developing a hybrid data clustering and visualization strategy that the inter-neuron distances are clearly visible using a coloring scheme and measurable the probability of each neuron and to find cluster boundary on SOM grid. Experimental results show that our new proposed method can efficiently help discovering the data distribution and improve the visualization effects of the SOM.

The remaining sections of this paper are organized as follows. Section 2 briefly presents Kohonen's SOM algorithm and the related works of data visualization techniques. Section 3 introduces the theory of visualization and clustering by the SOM approach, followed by a detailed description of the proposed algorithm. Section 4 provides the experimental results and reveals that the proposed algorithm is able to visualize the input data and find the decision boundary of clusters. The experimental results indicate that the effects of visualization of the proposed algorithm are better than that of other visualization methods. Finally, Section 5 concludes this paper.

2. Related works

This section provides a brief review on Kohonen's SOM algorithm and the related work on clustering methods and visualization techniques.

2.1. Self-organizing maps

The SOM algorithm is applicable to large data sets. The goal of SOM is to transform the patterns of high dimensionality into a low-dimensional topological map. The training algorithm proposed by Kohonen for forming a feature map is stated as follows (Murtagh, 1995).

Step (1) **Initialization:** Choose random values for the initial weights w_i.

Step (2) **Winner Finding:** Find the winning neuron c at time t, using the minimum Euclidean distance criterion

$$c = \arg\min_i \|x - w_i\|, \quad i = 1, 2, \ldots, M \tag{1}$$

where $x = [x_1, \ldots, x_m] \in R^m$ represents an input vector at time t, M is the total number of neurons, and $\|\cdot\|$ indicates the Euclidean norm.

Step (3) **Weights Updating:** Adjust the weights of the winner and its neighbors, using the following rule:

$$w_i(t+1) = w_i(t) + \eta(t) h_{ci}(t)[x_j(t) - w_i(t)], \tag{2}$$

$$h_{ci}(t) = \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right), \tag{3}$$

where $x_j(t)$ represents an input data at time t, $h_{ci}(t)$ is the topological neighborhood function of the winner neuron c at time t, $\eta(t)$ is a positive constant called "learning-rate factor", $r_c \in R^2$ and $r_i \in R^2$ are the location vectors of nodes c and i, respectively. $\sigma(t)$ defines the width of the kernel. Both $\eta(t)$ and $\sigma(t)$ will decrease with time. It should be emphasized that the success of the map formation is critically dependent on the values of the main parameters (i.e., $h_{ci}(t)$ and $\eta(t)$), the initial values of weight vectors, and the pre-specified number of iterations.

In the case of a discrete data set and fixed neighborhood kernel, the sum of the squared-error of SOM can be defined as follows:

$$SSE = \sum_{j=1}^{n} \sum_{i=1}^{M} h_{ci} \|x_j - w_i\|^2, \tag{4}$$

where n is the number of training samples, and M is the number of map units. Neighborhood kernel h_{ci} is centered at unit c, which is the best matching unit of input vector x_j, and evaluated for unit i.

2.2. Clustering and visualization of the SOM

Since traditional SOM methods require the number of clusters for the input data, many researchers have proposed two-level SOMs in order to get the hierarchical structures without knowing the number of clusters. In a general two-level SOM, neurons over the expected number of clusters should be prepared for the first-level SOM. Lampinen and Oja (1992) used a method consisting of a two-level SOM. The first-level SOM is used as a training procedure, while the second-level SOM is used for clustering. Murtagh (1995) applied an agglomerative contiguity constrained clustering method to merge the neurons trained in the first-level SOM, where the merging criterion is to apply the shortest distance between the neurons. Kiang (2001) proposed the first step as recruiting more neurons than required for executing SOM algorithm. Then a refining space topological method is applied by means of a conscience mechanism (DeSieno, 1988). After getting the result of the first-level SOM, a merging criterion of minimum variance is applied to the succeeding steps. Wu and Chow (2004) proposed a clustering validity index based on the inter-cluster, intra-cluster density and inter-cluster distances in the clustering method. The method is able to find a best partition of the input data. Shieh and Liao (2009) proposed a new clustering validity index for two-level SOM algorithms to automatically determine the best number of clusters. This clustering validity index includes the separation rate of inter-cluster, the relative density of inter-cluster and the cohesion rate of intra-cluster. A validity index, $SOM_{clustering\_validity}$, has been defined for evaluating the results of clustering algorithms, and the best number of clusters is obtained by the maximum value of $SOM_{clustreing\_validity}$.

Information visualization techniques are increasingly applied in combination with other data analysis techniques. SOM visualization usually uses a map lattice as a visualization platform (Vesanto, 1999), where quantitative information is most commonly represented as color values or as markers of different sizes. Network topology map (Kohonen, 1990) is a visualization method that is applied after SOM training to a two-dimensional map, using 2D coordinated locations to present neurons and data point locations. In $5 * 5$ SOM shown in Fig. 1,[1] the red points stand for neurons, and other ones are data points (i.e., blue, green, and yellow data points).

The manner of obtaining clustering data from a well-disciplined feature projection map is a big challenge for SOM visualization. Ultsh and Siemon proposed U-matrix (Ultsch, 2003; Ultsch & Siemon, 1990), while PÄolzlbauer et al. (2005) and others pointed out a method which overlays the feature projection map on the two-dimensional or three-dimensional map in order to visualize the cluster structure of the SOM map.

The unified distance matrix (U-matrix) is a visualization technique that shows the local cluster boundaries by depicting pair-wise distances between neighbors, usually through color-coding. It is the most common method associated with SOMs and has been extended in numerous ways. A U-matrix is constructed

[1] For interpretation of colour in Figs. 1, 4 and 5, the reader is referred to the web version of this article.

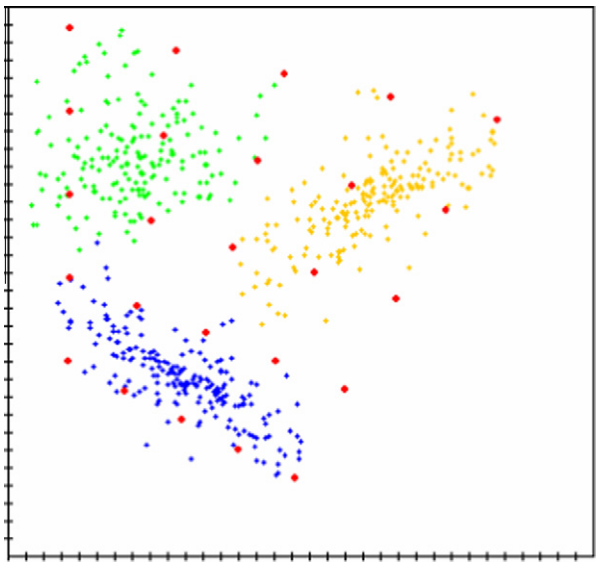

Fig. 1. An example of SOM network topology map.

on the top of a two-dimensional SOM grid. This is achieved by using topological relationships among neurons after completion of the learning process. The main purpose of U-matrix is to calculate the distances between each neuron and all its neighboring neurons revealing the local cluster structure of the map. The further the distance is, the higher the similarity between neurons will be. The U-matrix is to enlarge the output map of SOM from $(n * n)$ to $(2n-1)*(2n-1)$. Fig. 2 presents an example of the $9 * 9$ U-matrix for a $5 * 5$ hexagonal SOM.

A different approach for visualizing clusters is the data histograms, which rely more heavily on distribution of the data on the map than hit histograms (Pampalk, Rauber, & Merkl, 2002). The hit histogram reveals the distribution of the data points on the map; while the size of the marker indicates the frequency of number of times the unit was selected as a best-matching unit by data points. Fig. 2a and b show U-matrix and hit histogram visualizations for a map consisting of $5 * 5$ SOM maps that have been trained on the 2D artificial data set (http://cilab.csie.ncu.edu.tw/course/cluster/FCM'sData.zip).

3. Proposed method

SOMs are the kind of widely used unsupervised dimensional reduction method. There are two advantages of visualization when applying SOMs. The first one is its exploration of the clustering structure, and the second one is that it is much easier to visualize the distribution of the data set. In this section, a new scheme that integrates clustering and visualization strategy in order to represent data topology projected on the SOM grid is proposed. The proposed scheme could help in the discovery of data distribution and improve the effects of visualization of the SOM. Fig. 3 presents the proposed method's flowchart.

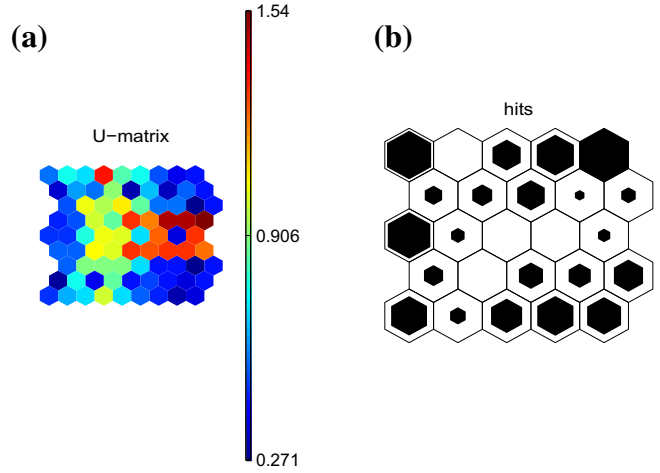

Fig. 2. 2D artificial data set in $5 * 5$ SOM: (a) $9 * 9$ U-matrix and (b) hit histogram.

3.1. Clustering using SOM

First, we have to choose the experimental data set in order to determine whether there has been marked as classified. If the data set of this experiment is unlabeled, we indicate that it finds the best number of clusters and proposed by several researchers, such as (PÄolzlbauer et al., 2005; Murtagh, 1995; Shieh & Liao, 2009; Wu & Chow, 2004), and then assign the cluster labels to the trained SOM. We apply our proposed method to enhance the effects of visualization. In contrast, when the chosen data set is labeled, our visualization method will clearly represent cluster structures in an enhanced visual form, which reveals better understanding of the structures and the geographic processes.

In our previous experimental results (Shieh & Liao, 2009), we proposed a clustering validity index for determining the most proper number of clusters in SOM. In other words, the clustering algorithm (Shieh & Liao, 2009) determined the best number of clusters by maximizing the validity index and achieved the best clustering result with the highest clustering accuracy. The clustering accuracy in Shieh and Liao (2009) is better than those of using the CDbw (Wu & Chow, 2004), the extended SOM (Kiang, 2001), and the five traditional inter-cluster distances approaches (Han & Kamber, 2006; Jain, Murt, & Flynn, 1999). Accordingly, in this paper, we shall apply this clustering validity index (Shieh & Liao, 2009) to find the best number of clusters as well. If the data set is unlabeled, after training of SOM, we must find the best number of clusters, and then assign the cluster labels to the trained SOM. After these procedures are completed, we use the trained SOM data set with cluster labels as the input of the proposed visualization method. On the other hand, if the data set is labeled, then the proposed visualization method is directly applied to enhance visualization effects.

3.2. Visualization of the SOM

SOMs are a kind of well-known tools for exploratory data analysis and visualization of high-dimensional data, and several visualization tools have been applied to represent the trained SOMs. In this sub-section, we evaluate the visualization effects by using the Wine data set (Blake & Merz, 1998), which has 178 13D data with three known clusters. The numbers of data samples in the three clusters are 59, 71 and 48, respectively. For better visual perception, we use a map of $5 * 5$ neurons in our experiments because it is small enough to visualize.

A U-matrix is a map of distances between each of neurons and all its neighbors and slices are the two-dimensional combinations of the SOM weights. With the U-matrix map, the easiest way to examine whether there are distinct clusters present in a given data set is to utilize the distance matrix. This matrix represents the distances between each neuron and all its neighboring ones, and is able to reveal the local cluster structure of the map. The further the distance is, the higher the difference between them will be, which is resulted in a higher similarity. Fig. 4a and b show visualizations results of the Wine data set for the U-matrix and hit histogram approaches for a map consisting of $5 * 5$ trained SOM

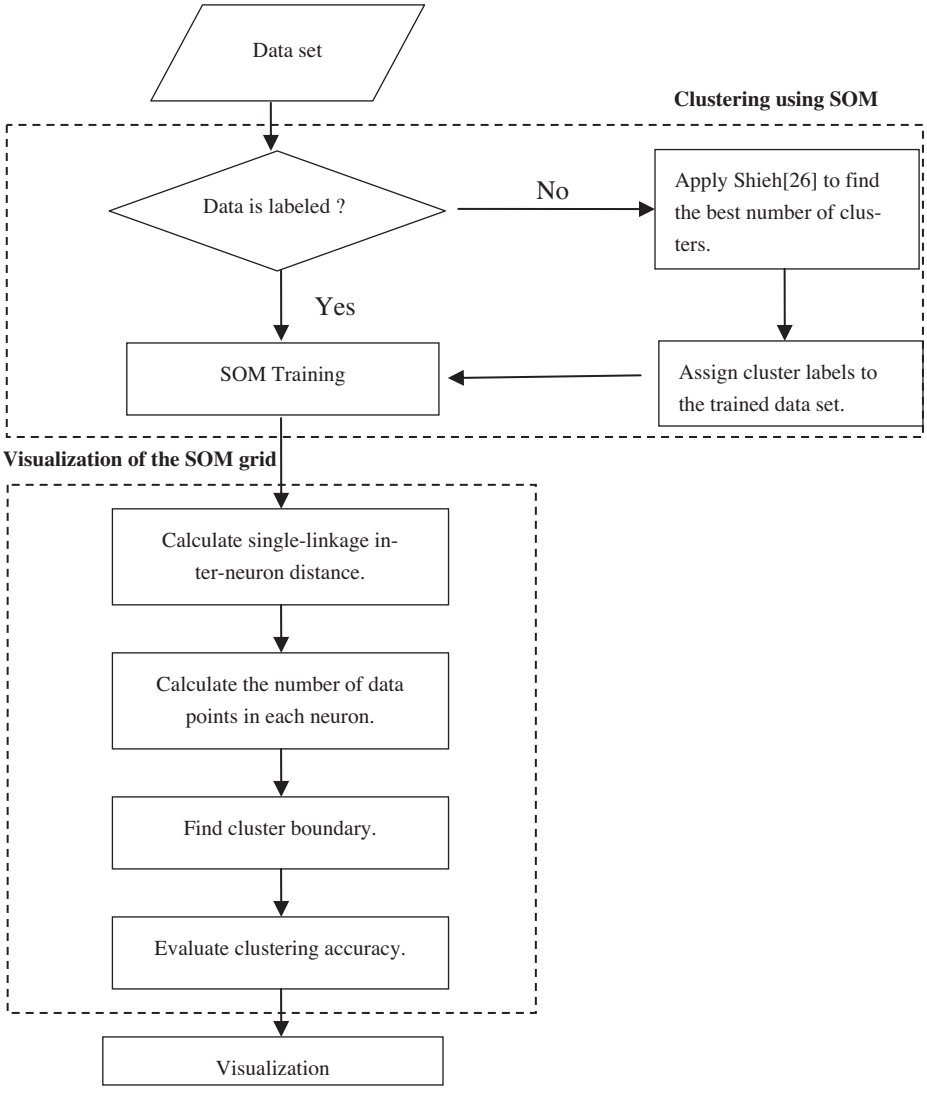

Fig. 3. The flowchart of the proposed method.

maps on, where small and large distances represent in red and blue, respectively.

One disadvantage of the U-matrix is to enlarge the output map of trained SOM from $n * n$ to $(2n-1) * (2n-1)$. Another disadvantage for hit histogram is that it only elucidates the number of each neuron's data points, but it cannot distinguish the exact data points and the numbers of each class in the same neuron. In the following paragraphs, a novel visualization method is proposed in order to tackle the mentioned disadvantages for both the U-matrix and hit histogram approaches, and to enhance the effects of data visualization. The proposed method extends the usage of SOMs and makes it able to provide the data distribution in each neuron. Our proposed method also applies the Bayes theorem to find the clustering boundary, which is a kind of new idea compared with the U-matrix and hit histogram, while our SOM visualization methods can be described as follows.

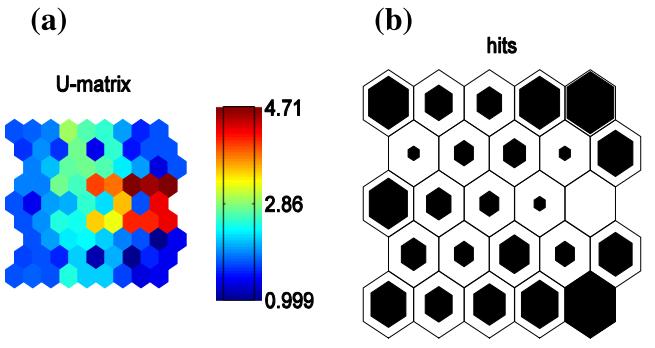

Fig. 4. The Wine data set in $5 * 5$ SOM: (a) $9 * 9$ U-matrix and (b) hit histogram.

First of all, the distance between two neurons is represented by a color-coding scheme. We utilize the single-linkage inter-cluster distances to express the distances between two neighboring neurons as formulated in Eq. (5).

$$d_{\min}(q_i, q_j) = \min_{p \in q_i, p' \in q_j} \|p - p'\|, \tag{5}$$

where q_i and q_j are two neurons, $\|p - p'\|$ represents the Euclidean distance between input data points p and p'. After all the distances between neurons are calculated, the largest one and smallest one are able to be found. Then our scheme calculates the difference of these two values, and divides the difference amount into ten equal parts. Each part will be assigned a different color to depict the specific distance between neurons. Therefore, the relation of distances between any neuron and its neighboring neurons are granted.

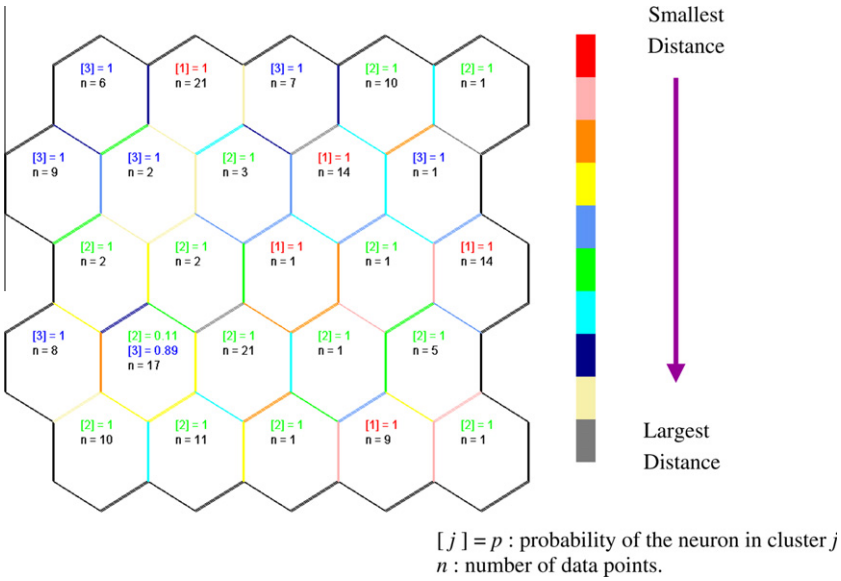

Fig. 5. Visualizing the labeled Wine data set in $5 * 5$ SOM using the proposed method.

Next, we present the novel method to visualize the probability of the various clusters and number of data points within each neuron. For example, the neuron located at the second column of the fourth row in Fig. 5 has 6 neighboring neurons, and the neuron with the shortest distance is at its left, which is with orange border. In addition, the farthest neighboring neuron is in its left-upper corner, which is with blue border color. This neuron has 17 data points in total, while two points (i.e. $17 * 0.11 \approx 2$) of them belong to the second cluster and other 15 points are of the third cluster. In Fig. 5, the probability of every cluster in each neuron and the distance between two neighboring neurons can be observed clearly. Different colors are used to represent different distances, such as red stands for the nearest neighborhood distance, and grey stands for the farthest.

Finally, we propose a decision boundary formula to classify all the neurons. The decision boundary formula will classify that all the neurons on one side of the decision boundary belong to one class and all those on the other side as belonging to other classes. A decision boundary formula is proposed and stated as follows.

Let $x_k^{L_j} = \left[x_{k1}^{L_j}, \ldots, x_{km}^{L_j}\right]^T \in R^m, k \in \{1, \ldots, N_j\}$ and $j \in \{1, \ldots, c\}$, be the m-dimensional training data sample of cluster L_j, where N_j is the number of samples in the cluster L_j, c is the number of clusters, and $N = \sum_{j=1}^{c} N_j$ is the total number of training data samples. In order to calculate the expectation risk $R_i(q)$ for each neuron q belonging to cluster L_i, the risk is estimated by using the conditional posterior probability and the Bayesian rule. By assuming the neuron W is the neighboring neuron of neuron q, the expectation risk $R_i(q)$ is

$$R_i(q) = \sum_{i=1}^{c} \left[\sum_{\substack{j=1 \\ j \neq i}}^{c} p(L_j|q) \right] P(L_i|W), \quad 1 \leqslant i \leqslant c, \quad 1 \leqslant j \leqslant c. \tag{6}$$

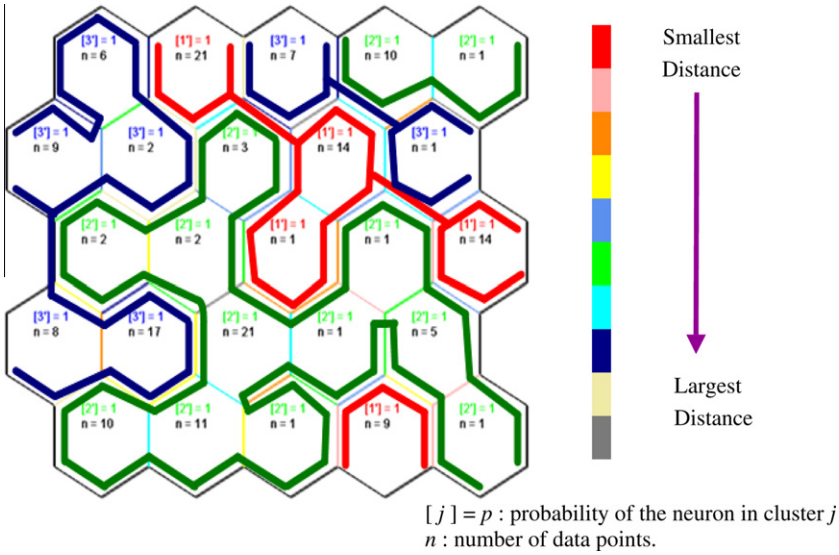

Fig. 6. Results of visualizing the unlabeled Wine data set by the proposed method.

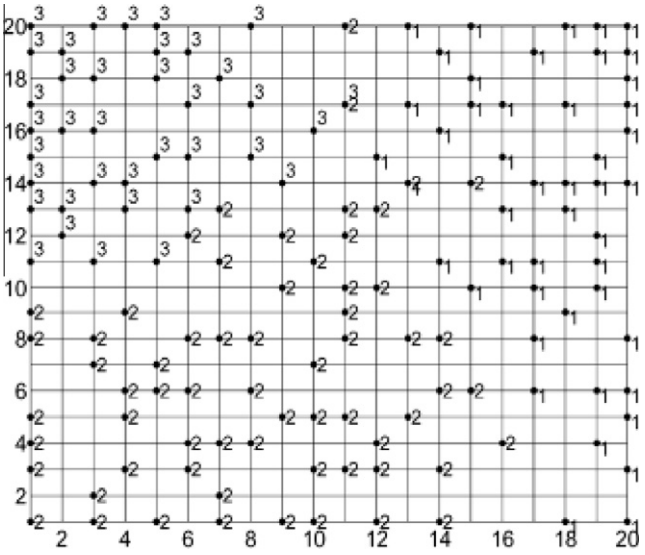

Fig. 7a. Visualization by SOM (size $20 * 20$, 1000 epochs) for Wine data set (adapted from Wu and Chow (2005)).

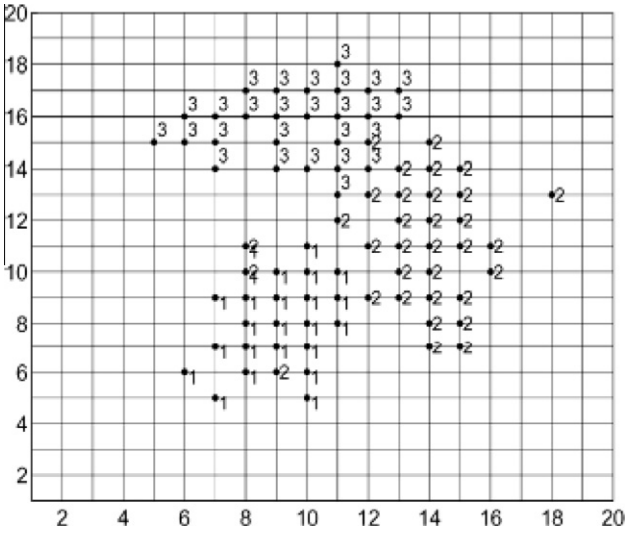

Fig. 7b. Visualization by ViSOM (size $20 * 20$, 1000 epochs) for Wine data set (adapted from Wu and Chow (2005)).

The estimation of the posterior probability of neuron q is defined as:

$$p(L_j|q) = \frac{p(q|L_j)p(L_j)}{\sum_{j=1}^{c} p(q|L_j)p(L_j)}, \quad 1 \leqslant j \leqslant c, \tag{7}$$

where $p(L_j) = N_j/N$ is the *a priori* probability of the cluster L_j and $P(L_i|W)$ denotes that for a given neighboring neuron W, the probability that it belongs to cluster L_i. Use Eq. (6) to calculate the expectation risk $R_i(q)$ for each cluster and then classify the neuron q to the cluster with the minimum expectation risk. If the expectation risks are the same, then assign neuron q the cluster label of the neighboring neuron having shortest distance with q.

3.3. Validation of visualization

As mentioned earlier, one major limitation of many proposed SOM-based clustering visualization algorithms is that they assume that the number of clusters is known. However, in fact, this assumption is not always true. Our proposed visualization method does not assume the prior knowledge on the number of clusters and we present experimental results to validate the visualization effects of the proposed scheme. First, the experiment on visualizing the unlabeled Wine data set is performed. Then the visualization of labeled Wine data set is used to verify the results of the unlabeled case. If the data set is unlabeled, we apply Shieh's methods (Shieh & Liao, 2009) to find that best number of clusters, which is known to be two for the Wine data set. After cluster labels are assigned to neurons, we calculate inter-neuron distances, data distribution in neurons, and finally determine the cluster boundary. The result of applying the proposed algorithm with a map size of $5 * 5$ to visualize the Wine data set is shown in Fig. 6.

Visualization by the SOM, ViSOM and PRSOM approaches for the Wine data set were adapted from [37] and presented in Figs. 7a, 7b and 7c, respectively. ViSOM and PRSOM have been proposed to enrich SOM's visualization and to preserve the inter-neuron distances from high-dimensional input space to low-dimensional output space when there are marked classified label on topological map. Unfortunately, ViSOM and PRSOM have the difficulty of visualizing the data cluster and data distribution if we remove the classified labels from the topological maps, i.e., class labels 1–3 in 7(a)–(c). For example, if class labels 1, 2 and 3 are removed from Fig. 7c, the following visualization problems will encountered. First, we can only see the neurons containing data points on SOM map, but cannot identify the distribution of data points within the neurons. Then the situation of whether the data points in a neuron belong to the same cluster is questioned. Secondly, it is hard to determine the number of clusters. From the grey levels of the figure, we can only distinguish two clusters, i.e., class 1 and class 3.

To validate the visualization effects of the proposed scheme, we run the proposed method again using the labeled Wine data set. The results of visualizing this labeled Wine data set by the proposed method are shown in Fig. 8. The experimental results indicate that the proposed visualization method for unlabeled data finds a very good approximation in the decision boundary for the Wine data set. From our experimental results, the proposed visualization method achieves the clustering accuracy of 98.9%, i.e., $1 - 0.11 * 17/178 \approx 98.9\%$, where the clustering accuracy is calculated by the fraction of data points that are correctly clustered.

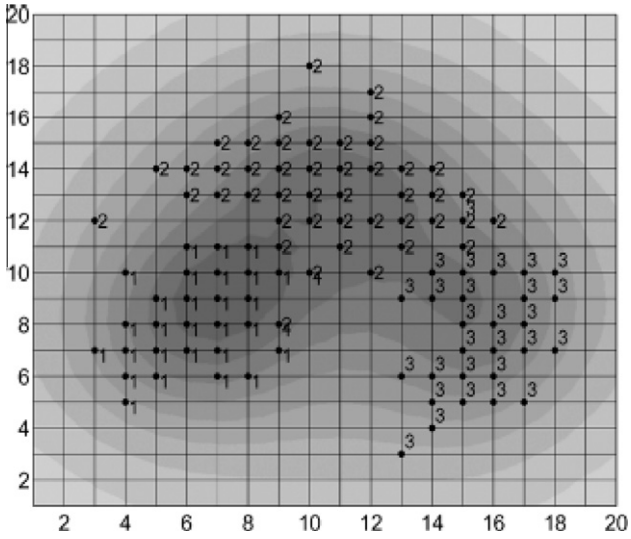

Fig. 7c. Visualization by PRSOM (size $20 * 20$, 1000 epochs) for Wine data set (adapted from Wu and Chow (2005)).

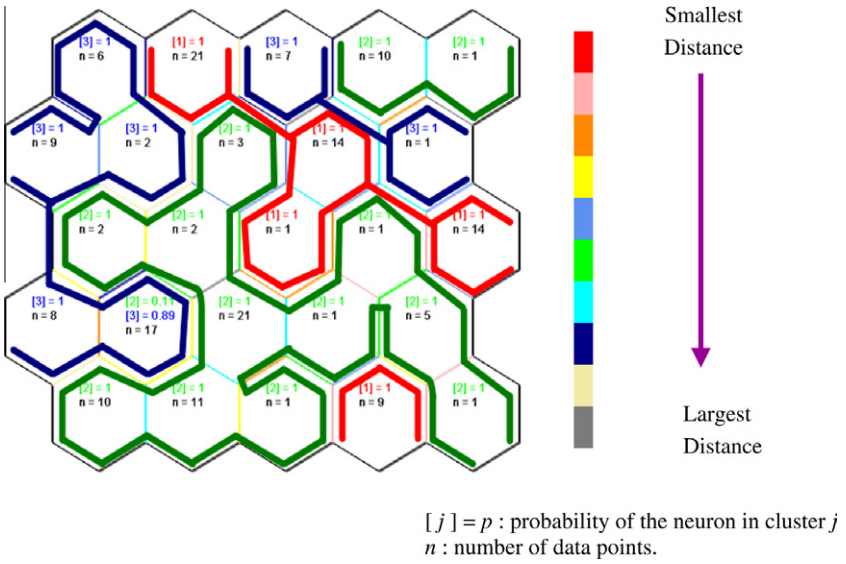

Fig. 8. Final results of visualizing labeled Wine data set using the proposed method.

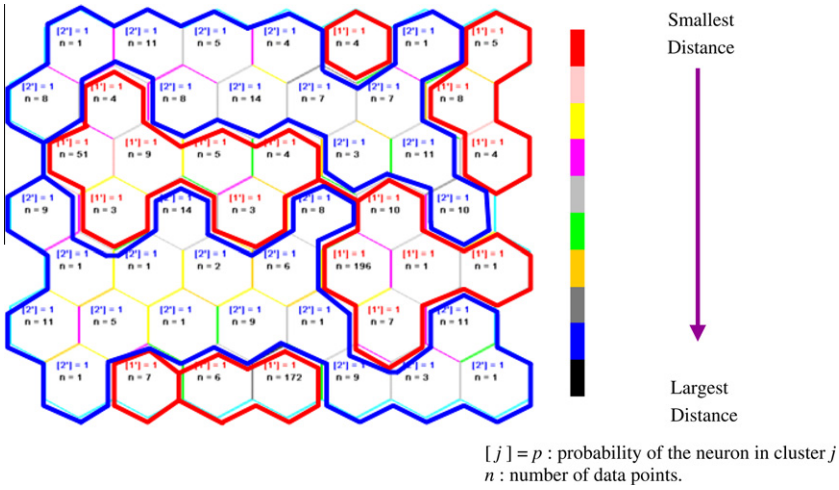

Fig. 9. Results of visualizing unlabeled WBC data set using the proposed method.

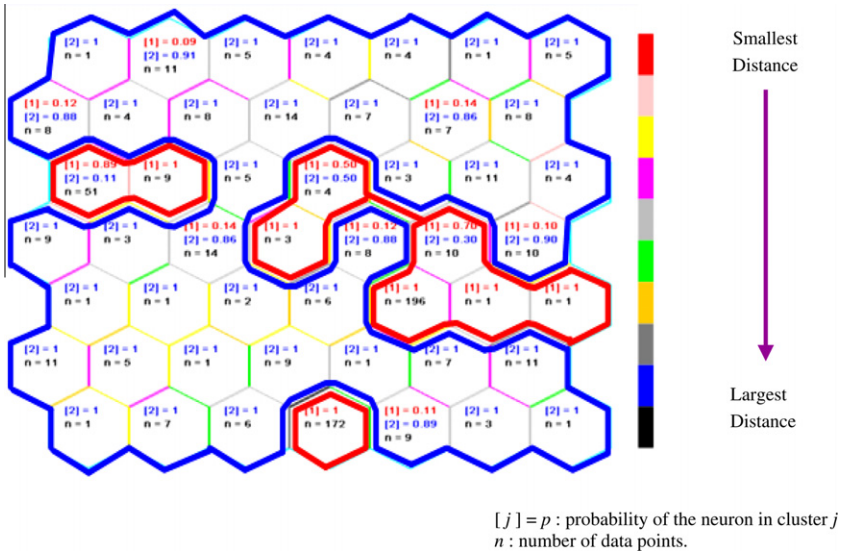

Fig. 10. Results of visualizing labeled WBC data set using the proposed method.

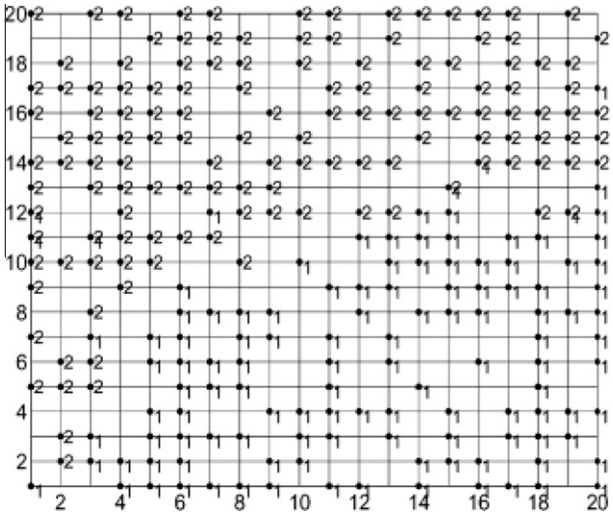

Fig. 11a. Visualization by SOM (size $20 * 20$, 1000 epochs) for WBC data set (adapted from Wu and Chow (2005)).

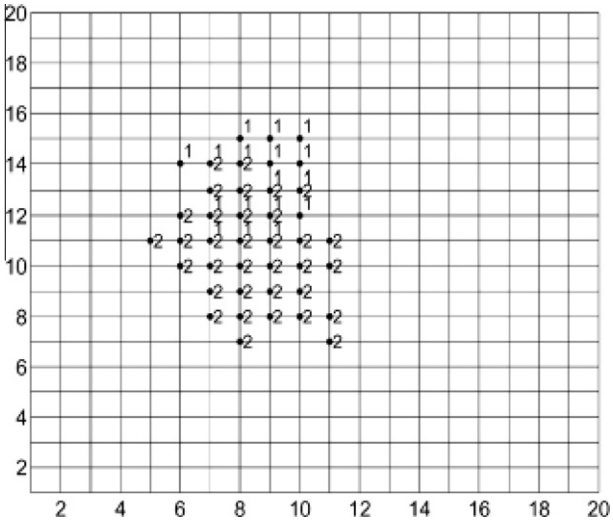

Fig. 11b. Visualization by ViSOM (size $20 * 20$, 1000 epochs) for WBC data set (adapted from Wu and Chow (2005)).

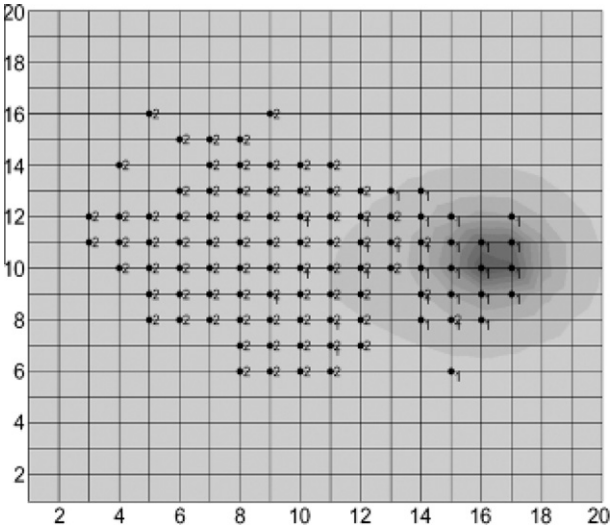

Fig. 11c. Visualization by PRSOM (size $20 * 20$, 1000 epochs) for WBC data set (adapted from Wu and Chow (2005)).

4. Experimental results

In these experiments, the program is written in Java using Borland JBuilder 9 Enterprise Edition. The platform used is an Intel Pentium 4 3.2 GHz with 512 MB DRAM and 80 GB hard disk running Windows 2000 Server Pack3. Two data sets were used to demonstrate the effectiveness of the proposed algorithm in the experiments. To achieve a better clustering result and to avoid the negative effects produced by noises and outliers, all data sets were pre-processed using data cleaning normalization schemes.

4.1. The Wisconsin breast cancer data set

In this experiment, the Wisconsin Breast Cancer (WBC) data set (Blake & Merz, 1998) was used and analyzed, which was collected by Dr. William H. Wolberg (1989–1991) at the University of Wisconsin–Madison Hospitals. There are 699 data samples in this dataset, while 16 samples contain a single missing (i.e., unavailable) attribute value. We do firstly pre-processing to remove the 16 samples. Therefore, the WBC dataset has 683 9D data samples with two known clusters. The data set is then divided into two classes: benign and malignant samples, whilethe sample numbers of these two classes are 444 and 239, respectively.

The Shieh's methods (Shieh & Liao, 2009) was applied and found that, for the unlabeled WBC data set, the best number of clusters is two as well. After completed the cluster label assignment to neurons, the inter-neuron distances and data distribution in neurons are calculated, and finally the cluster boundary is determined. The proposed algorithm is used with a map size of $7 * 7$ to visualize unlabeled WBC dataset, as shown in Fig. 9.

To validate the visualization effects of the proposed scheme, we also run the proposed method again using the labeled WBC data set, where the results are shown in Fig. 10. Visualization by the SOM, ViSOM and PRSOM for the WBC data set were adapted from (Wu & Chow, 2005) and presented in Figs. 11(a)–(c), respectively, for making a clear comparison. Experimental detail for these SOM, ViSOM and PRSOM to construct their visualizations for the WBC dataset can be found in Wu and Chow (2005).

In this experiment, there are 444 data points in the benign class, and 239 ones in the malignant class. In Fig. 10, it can be found that the two classes of data points are denoted as class 1 and class 2 for benign and malignant, respectively. There are seven data points belonging to class 1, but are misjudged to be in class 2, i.e., the error rate is $7/683 \approx 1.02\%$. On the other hand, the error ratio of class 2 is $64/683 \approx 9.37\%$, since there are 64 data points belonging to class 2, but are misjudged to be assigned to class 1. As a result, the clustering accuracy of the proposed visualization method is 89.61, i.e., $1 - 71/683 \approx 89.61\%$.

4.2. The IMOX data set

The IMOX data set (Chen, 2005) consists of 192 8D data and with four known clusters. The dataset is the IEEE data file of letters I, M, O, and X. We apply Shieh's methods (Shieh & Liao, 2009) and find that the best number of clusters is four for the unlabeled IMOX one. After finished the cluster label assignment to neurons, the inter-neuron distances and data distribution in neurons are calculated, and the cluster boundaries are determined as well. The proposed algorithm is used with a map size of $6 * 6$ to visualize unlabeled IMOX dataset, as shown in Fig. 12.

To validate the visualization effects of the proposed scheme, we also run the proposed method again by applying the labeled IMOX data-set, while the results are shown in Fig. 13. In this simulation,

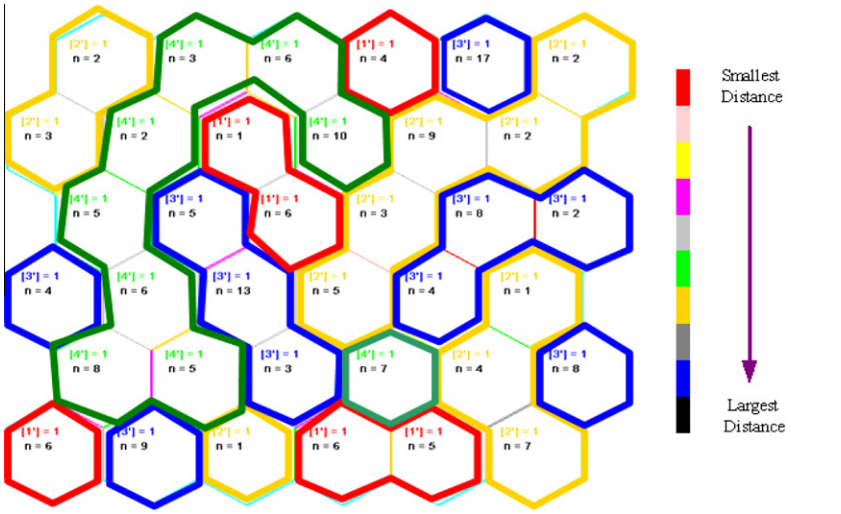

Fig. 12. Final results of visualizing unlabeled IMOX data set using the proposed method.

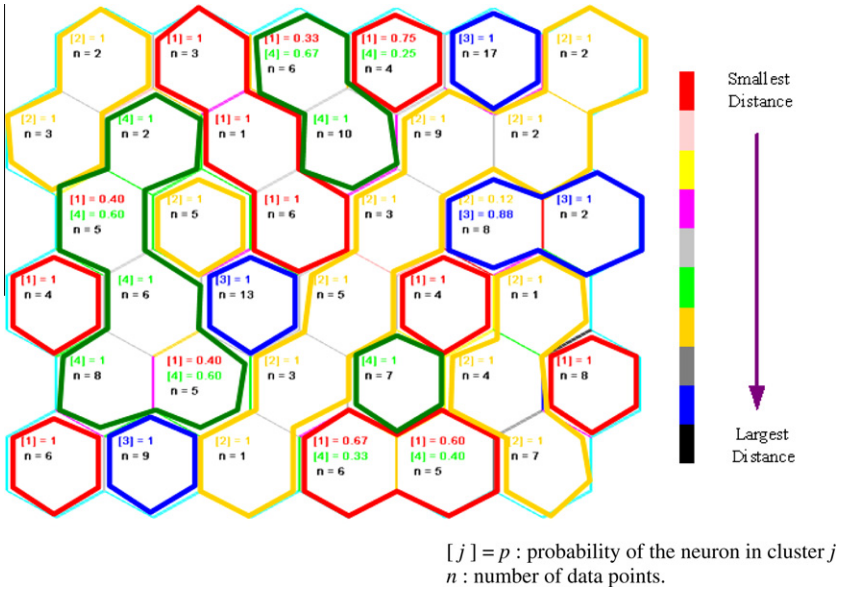

Fig. 13. Final results of visualizing labeled IMOX data set using the proposed method.

the numbers of data points in the four clusters are equally set to 48. Since there are 16 data points belong to class 1 but are misjudged to be in class 3, the error ratio is $16/192 \approx 8.33\%$. As there are nine data points should be in class 1, misjudged to be in class 4, the error rate of this part is $9/192 \approx 4.69\%$. Besides, because there are nine data points belong to class 2, but misjudged to be in class 3, the error ratio is $9/192 \approx 4.69\%$, whereas the error ratio for class 4 is $4/192 \approx 2.08\%$, for the sake of that there are four data points belong to class 4 are misjudged to be in class 1. The clustering accuracy of the proposed visualization method for this dataset is then equal to 80.21, i.e., $1 - 38/192 \approx 80.21\%$.

5. Conclusion

In this paper, we develop a novel methodology for applications of data clustering and visualization, which is based on the SOM approach. The main process of our approach can be summarized as following. If the dataset is unlabeled, we calculate the best number of clusters in advance, and then assign the cluster labels to the neurons. After completed the cluster labels assignment, we apply the proposed method to enhance the effects of visualization. On the other hand, when the dataset is labeled, our visualization method will directly represent cluster structures in an enhanced visual form. The experimental results show that the proposed visualization method efficiently and effectively demonstrates the data distribution, inter-neuron distances, and cluster boundary. Therefore, our proposed visualization scheme is not only intuitively easy understanding of the clustering results, but also having good visualization effects on unlabeled data set.

References

Blake, C. L., & Merz, C. J. (1998). *UCI repository of machine learning databases*. Department of Information and Computer Science, University of California at Irvine, CA. <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Brugger, D., Bogdan, M., & Rosenstiel, W. (2008). Automatic cluster detection in Kohonen's SOM. *IEEE Transactions on Neural Networks, 19*(3), 442–459.

Card, S. K., & Mackinlay, J. (1997). The structure of the information visualization design space. In *Processing of information visualization*, pp. 92–99.

Chakma, J., & Umemura, K. (2003). Factor controlled hierarchical SOM visualization for large set of data. *IEICE Transactions on Information and Systems, 86*(9), 1796–1803.

Chen, C. C. (2005). *Computational mathematics*, Univ. Tsing Hua, Institute of Information Systems and Applications. <http://www.cs.nthu.edu.tw/∼cchen/ISA5305//isa5305.html>.

DeSieno, D. (1988). Adding a conscience to competitive learning. In *Proceeding of IEEE international conference on neural networks*, vol. 1, pp. 117–124.

Diri, B., & Albayrak, S. (2008). Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications, 34*, 628–634.

Fernandez, E. A., & Balzarini, M. (2007). Improving cluster visualization in self-organizing maps: Application in gene expression data analysis. *Computers in Biology and Medicine, 37*(12), 1677–1689.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann.

Jain, A. K., Murt, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Kiang, M. Y. (2001). Extending the Kohonen self-organizing map networks for clustering analysis. *Computation Statistics & Data Analysis, 38*, 161–180.

Kohonen, T. (1990). The self-organizing feature map. *Proceedings of the IEEE, 78*(9), 1464–1480.

Lampinen, J., & Oja, E. (1992). Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision, 2*, 261–272.

Murtagh, F. (1995). Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters, 16*, 399–408.

Pampalk, E., Rauber, A., & Merkl, D. (2002). Using Smoothed data histograms for cluster visualization in self-organizing maps. In *Proceedings of the international conference of artificial neural networks*, Madrid, Spain: Springer Lecture Notes in Computer Science.

PÄolzlbauer, G., Rauber, A., & Dittenbach, M. (2005). Graph projection techniques for self-organizing maps. In: *Proceedings-European symposium on artificial neural networks bruges (ESANN'2005)*.

Petrilis, D., & Halatsis, C. (2008). Two-level clustering of web sites using self-organizing maps. *Neural Processing Letters, 27*(1), 85–95.

Ressom, H., Wang, D., & Natarajan, P. (2003). Clustering gene expression data using adaptive double self-organizing map. *Physiological Genomics, 14*, 35–46.

Ritter, H., Martinetz, T., & Schulten, K. (1992). *Neural Computation and Self-Organizing Maps An Introduction*. Addison-Wesley.

Shieh, S. L., & Liao, I. E. (2009). A new clustering validity index for cluster analysis based on a two-level SOM. *IEICE Transactions on Information and Systems, E92-D*(9), 1668–1674.

Shieh, S. L., Liao, I-En, Hwang, Kuo-Feng, & Chen, H. Y. (2009). An efficient initialization scheme for SOM algorithm based on reference point and filters. *IEICE Transactions on Information and Systems, E92-D*(3), 422–432.

Su, M. C., Liu, T. K., & Chang, H. T. (1999). An efficient initialization scheme for the self-organizing feature maps. In *Proceeding IEEE international joint conference of neural networks, Washington, DC*, pp. 1906–1910.

Tangsripairoj, S., & Samadzadeh, M. H. (2006). Organizing and visualizing software repositories using the growing hierarchical self-organizing map. *Journal of Information Science and Engineering, 22*, 283–295.

Taşdemir, E., Merényi, E. (2008). Cluster analysis in remote sensing spectral imagery through graph representation and advanced SOM visualization. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5255, LNAI, pp. 259–271.

Ultsch, A. (2003). *U∗-matrix: A tool to visualize clusters in high dimensional data*. Technical Report 36, Germany: Computer Science Department, Philipps-University Marburg.

Ultsch, A., & Siemon, H. P. (1990). Kohonen's self organizing feature maps for exploratory data analysis. In *Proceeding of international neural network conference*, pp. 305–308.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis, 3*(2), 111–126.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks, 11*(3), 586–600.

Wu, S., & Chow, T. W. S. (2004). Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition, 37*, 175–188.

Wu, S., & Chow, T. W. S. (2005). PRSOM: A new visualization method by hybridizing multidimensional scaling and self-organizing map. *IEEE Transactions on Neural Networks, 16*(6), 1362–1380.

Yin, H. (2002). ViSOM: A novel method for multivariate data projection and structure visualization. *IEEE Transactions on Neural Networks, 13*(1), 237–243.

Yuan-chao, Liu, Chong, Wu, & Ming, Liu (2011). Research of fast SOM clustering for text information. *Expert Systems with Applications, 38*(8), 9325–9333.