2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

Optimal Cluster Selection Based on Fisher Class Separability Measure

Xudong Wang and Vassilis L. Syrmos
Department of Electrical Engineering
University of Hawaii
Honolulu, HI 96822
Email: syrmos@hawaii.edu

Abstract— **In this paper, a novel hierarchical clustering algorithm is proposed, where the number of clusters is optimally determined according to the Fisher class separability measure. The clustering algorithm consists of two phases: (1) Generation of sub-clusters based on the similarity metric; (2) Merging of sub-clusters based on the Fisher class separability measure. The proximity matrices are constructed. Each sub-cluster comprises patterns close to each other in proximity metric. The trellis diagram is used for searching of sub-clusters. Connections between consecutive layers in the trellis diagram are weighted by the similarity metric. The threshold for the merge of sub-clusters is numerically designed according to Fisher class separability measure. The proposed algorithm can pre-process the data for the supervised learning. It also can be applied for the optimal determination of basis functions for radial basis function (RBF) networks.**

I. Introduction

Clustering is a discovery process that group or segment a collection of data into subsets or “clusters”, such that the intracluster similarity and intercluster dissimilarity are both maximized. Clustering problems arise in many applications, such as data mining and knowledge discovery [1], data compression and vector quantization [2], and pattern recognition (e.g. regression and classification problems) [3]. Existing clustering algorithms, such as K-means [4], PAM [5], CLARANS [6], CURE [7], and DBSCAN [8] are designed to find clusters that fit some static models. In [4, 5, 6], clusters are assumed to be hyper-ellipsoid (or globular) and of similar sizes. DBSCAN assumes that all points within genuine clusters are density reachable and points across different clusters are not. CURE is designed for the large database clustering. The CURE algorithm is more robust to outliers and identifies clusters having non-spherical shapes and wide variance in size.

Among clustering formulations that are based on minimizing a formal cost (or loss) function, K-means clustering algorithm is the most widely used and studied. Given a set of n data points, $\{x_i \in \Re^d\}$ and a pre-determined integer k, the problem is to determine a set of k points in $\Re^d$, called as centers, so that the squared Euclidean distance from each data point to its nearest center is minimized. This measure is often called the *squared-error distortion* [4, 9] and this type of clustering falls into the general category of variance-based clustering [10]. There are many K-means clustering algorithms when the size and the number of clusters are known in advance. One of the most popular heuristics for solving k-means problem is based on a simple iterative scheme for finding a locally minimal solution. This algorithm is often called the *k-means algorithm*. However, there are problems with such a technique: k-means requires the number of clusters to be specified beforehand. Determining the number of clusters is not easy. Minimal Spanning Tree Clustering Algorithm was proposed for the unknown cluster problem. This hierarchial method starts by considering each data point to be a cluster. Next, the two clusters with the minimum distance between them are fused to form a single cluster. This process is repeated until all data points are grouped into the desired clusters. Agglomerative hierarchical algorithms [4] start with all the data points as a separate cluster. Each step of the algorithm involves merging two clusters that are the most similar. After each merge, the total number of clusters decreases by one. These steps can be repeated until the desired number of clusters is obtained or the distances between pairs of clusters satisfy the certain threshold distance. However, for a large set of data points, the searching procedure is complex and time-consuming.

In this paper, we present a novel hierarchical clustering algorithm that measures the similarity of pair of sub-clusters and creates the merging criterion (proximity threshold), which is numerically determined by the Fisher class separability measure. The proposed clustering algorithm has the bottom-up structure. In the clustering process, each sub-cluster is denoted as “leaves” and the merging process of sub clusters is called as the formation of “branches”. The proximity matrices and trellis diagram [11] are used to generate the sub-clusters. Two sub clusters are merged if and only if the proximity (closeness) satisfies the merging criterion. The diagram of hierarchical clustering is illustrated in Figure 1. It consists of three layers: (1) data point layer; (2) sub-clusters layer; (3) final cluster layer.

The rest of paper is organized as following. Section II gives an overview of the clustering analysis. The proposed hierarchical clustering algorithm is presented in Section III. The effectiveness of the proposed algorithm is evaluated in Section IV. Section V contains conclusions and directions for the future work.

This research has been funded by The Boeing Company under contract # 436481.

0-7803-9098-9/05/$25.00 ©2005 AACC

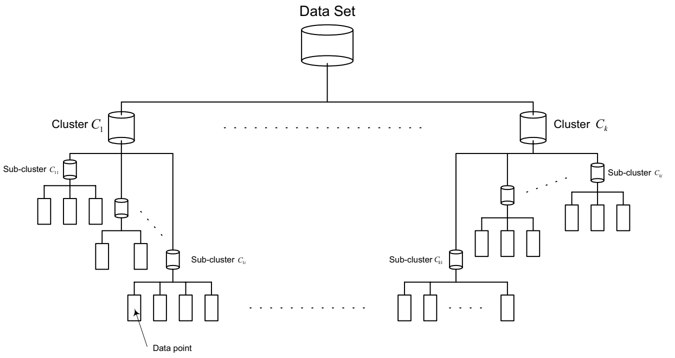

Fig. 1. Diagram representation of hierarchical clustering

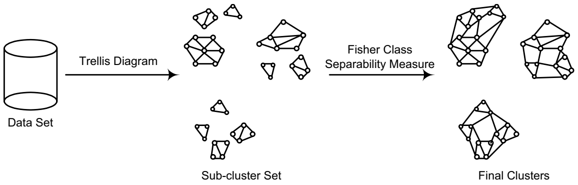

Fig. 2. Overall framework of hierarchical clustering algorithm

II. Clustering Analysis

The goal of cluster analysis is to group a set of data such that the intracluster similarity is maximized, and to arrange the clusters into a natural hierarchy. This involves two phases: (1) grouping data points into a candidate sub-cluster; (2) merging the sub clusters according to similarity metric. At each level of the hierarchy, clusters or data points within the same group are more similar to each other than those in different groups. Fundamental to clustering techniques is the choice of distance or similarity measure between data points or sub clusters. In this paper, a proximity matrix is constructed and provided as input to the clustering process.

Given a set of n data points, $\{x_i \in \Re^d\}$, the proximity matrix $D \in \Re^{n\times n}$, has nonnegative entries. The diagonal elements are set to be equal to zero. The off-diagonal elements are computed as

$$D_{ij} = \Phi(\|x_i - x_j\|) \quad \forall i \neq j \tag{1}$$

where Φ is nonlinear transfer function. In order to avoid more emphasis on larger differences between data points than smaller ones, the Gaussian type function is used to calculate off-diagonal elements.

$$\Phi(\|x_i - x_j\|) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \tag{2}$$

where σ is chosen according to the span of data points. The similarity measure between two different data points is constrained into the range from 0 to 1; and the more close two data points are, the greater similarity they have. As we can see, the proximity matrix is symmetric. The similarity measure defined in (1) is suitable for the *Quantitative variables*.

For nonquantitative variables (e.g. categorical data), the similarity measure in (1) may not be appropriate. In [12], the similarity measure for nonquantitative variable is presented. With unordered categorical variables, the degree-of-difference between pairs of variables can be delineated as the following. If the variable is assumed to have M distinct values, the proximity matrix is defined as a symmetric matrix in $\Re^{M\times M}$ with nonnegative elements as

$$L_{ij} = \{ \begin{array}{cc} 0 & i = j \\ > 0 & i \neq j \end{array}$$

The most common choice for off-diagonal elements is

$$L_{ij} = 1 \quad \forall i \neq j$$

For the clustering algorithm, the appropriate proximity matrix is crucial according to the different types of data. In this paper, quantitative data is considered for the design of clustering algorithm.

III. Hierarchical Clustering Algorithm Based on Fisher Class Separability Measure

The widely used K-mean clustering algorithm starts the clustering process with the pre-defined number of clusters. However, it is not easy to determine the appropriate number of clusters without any prior knowledge about the data set being clustered. Here, we propose a novel hierarchical clustering algorithm based on Fisher class separability measure. The clustering process consists of two phases. In the first phase, the proximity matrix is generated for the quantitative data, Trellis Diagram is constructed so that each layer contains all data points and the connection between the consecutive layers is weighted by the proximity matrix, and paries of data points with large similarity measure are chained to form a sub-cluster. During the second phase, the centroid of data points in each sub-cluster is calculated, and the similarity between two sub clusters is measured by the proximity between the corresponding centroids. The similarity is compared with the certain threshold to determine if the merge of two sub clusters occur or not. The merging process will be repeated until the similarity measure between two closet clusters does not meet the merging criterion. The effectiveness of the merging can be evaluated by Fisher class separability measure. The clustering framework is illustrated in Figure 2.

A. Clustering Algorithm Phase I

Given a set of n data points $\{x_i\}_{i=1}^n$, the proximity matrix is calculated using similarity measure as (1). A $n \times n$ Trellis Diagram is constructed as Figure 3, with the interconnection between two consecutive layers weighted by the similarity measure. Before the clustering process, an index vector, $\bar{m} \in \Re^n$, is defined to record the status of each point. The value of m_i is binary, where '0' denotes that i^{th} data point

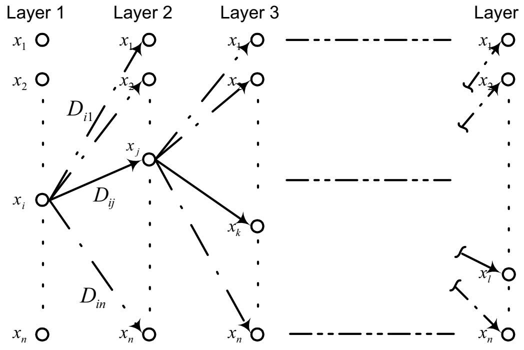

Fig. 3. A $n \times n$ Trellis Diagram with weights input from proximity matrix

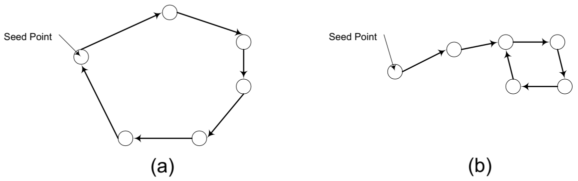

Fig. 4. Diagram representation of a closure chain

is free to be clustered, and '1' denotes that i^{th} data point has been grouped into a certain sub-cluster.

The procedure of clustering the data into the sub-cluster is illustrated as following.

- As following the Trellis Diagram, free data point, x_i, in the first layer is picked as to be a seed for a sub-cluster.
- Seek the largest weight from x_i in the first layer to data points in the second layer. Assume the connection from x_i to x_j has the largest weight. There are two situations here.
  - If x_j is free data point, the searching continues until the closure loop is created as Figure 4.
  - If x_j belongs to a sub-cluster, then the chain of data points from the seed point to x_j is attached to the sub-cluster as shown in Figure 5.
- The searching procedure is repeated until all data points have been clustered.

After all data being clustered, a sparse matrix, $C_L \in \Re^{n\times n}$, is generated, in which the nonzero elements in each row represents the data points in the same sub-cluster. The matrix C_L has the form as

$$C_L = \begin{pmatrix} 1 & i & j & \cdots & k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ l & m & h & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

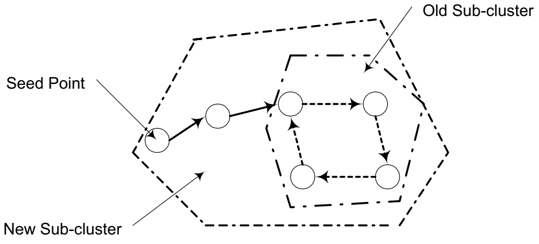

Fig. 5. Diagram representation of attaching data point (data chain) to a existing sub-cluster

TABLE I

HIERARCHICAL CLUSTERING ALGORITHM

```
Initialization, $i$ = 0, $m$ = $0_{n\times 1}$, $C_L$ = $0_{n\times n}$
Do $i \longleftarrow i+1$
    if $x_i$ is free (e.g. $m_i$ = 0)
        $C_L(i,1) = i$
        $l \longleftarrow i$, flag $\longleftarrow$ 0, $j \longleftarrow$ 0
        Do $j \longleftarrow j+1$
            $h$ = $\arg\max_k(D_{lk})$, $k = 1,\cdots,n$
            $l \longleftarrow h$
            if Closure loop or $x_h$ belongs to a sub-cluster
                flag $\longleftarrow$ 1
            else
                $C_L(i,h) = h$
            end
        until flag = 1 or $j > n$
    else
        $C_L(i,:)$ = $0_{1\times n}$
    end
until $i > n$
```

The value of each nonzero element in matrix C_L represents the position of data point in data set $\{x_i\}_{i=1}^n$. The total number of sub-clusters is denoted as n_{sb}, which is equal to the number of nonzero elements in the first column of the matrix C_L. The batching clustering algorithm can be summarized as Table I.

B. Clustering Algorithm Phase II

During the second phase, the centroid of each sub-cluster is calculated as

$$m_{ij} = \frac{\sum_{x_k \in C_{ij}} x_k}{n_{ij}} \tag{3}$$

where n_{ij} is the number of data points in sub-cluster C_{ij}, and the subscription of C_{ij} denotes the j^{th} sub-cluster that will be merged into the i^{th} final cluster, C_i. The n_{ij} satisfies

$$n = \sum_i \sum_j n_{ij}$$

From the phase I, we can get n_{sb} sub-clusters and n_{sb} centroids; and the similarity between pairs of sub-clusters is measured using (1) with input variables taken as centroids.

$$D_{ijkl} = \Phi(\|m_{ij} - m_{kl}\|) \tag{4}$$

Thus, the proximity matrix for sub-cluster set is a symmetric matrix in $\Re^{n_{sb}\times n_{sb}}$.

A similarity threshold, SM_T, is defined as

$$\text{SM}_T = \Phi(\|x_T\|) \tag{5}$$

where x_T is a vector in the observation space. Since the function $\Phi(\cdot)$ is deterministic, the similarity threshold totally depends on the value of x_T. A pair of sub-clusters are merged if their similarity measure is greater than the SM_T. As we can see, the similarity measure reflects how close the pair of sub-clusters are. The choice of similarity threshold, SM_T or x_T, really affects the number of final clusters, denoted as k. With different choices of x_T, we may get different number of final clusters. It is assumed that the effect of x_T on k can be described by an unknown nonlinear function as

$$k = \varphi(x_T) \tag{6}$$

Thus, it is crucial to design an appropriate x_T for the sub-cluster merge to yield optimal clustering result so that the data characteristics can be revealed to the most extend.

Assume the clustering process is completed, k final clusters are generated with n_i data points in i^{th} cluster. In order to evaluate the clustering result, an objective function is defined. The objective function will be optimized with the optimal choice of x_T. According to the assumption, since the number of clusters is known and data points are grouped into corresponding clusters, the problem is turned to be a supervised learning. We can exploit some useful properties of supervised learning such as *locality property* [13], which is that *Patterns that belong to the same class are close to each other and those in different classes are relatively farther away, according to some distance metric*. The locality property can be evaluated using Fisher class separability measure, which is derived from the Fisher discriminant rule [14]. Next, the Fisher class separability measure is briefly presented.

Let μ_i be the mean of cluster C_i defined by

$$\mu_i = \frac{1}{n_i} \sum_{x \in C_i} x \tag{7}$$

where n_i is the number of data points in the cluster C_i. Let μ be the mean of all data points given by

$$\mu = \frac{1}{n} \sum_{i}^{n} x_i \tag{8}$$

The within cluster scatter matrix is a measure of how compact the cluster is. The within cluster scatter matrix is defined as

$$S_W(k) = \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T \tag{9}$$

The between cluster scatter matrix measures the separation between clusters. It is defined as

$$S_B(k) = \sum_{i=1}^{k} n_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{10}$$

The locality property can be translated into an objective function $J(k)$ in terms of the between cluster scatter matrix and within cluster scatter matrix as follows.

$$J(k) = \psi(S_B, S_W, k) \tag{11}$$

By applying the Fisher class separability measure, the objective function is defined as

$$J(k) = \frac{\text{tr}(S_B)}{\text{tr}(S_W)} \tag{12}$$

The term $\text{tr}(S_B)$ can be written as

$$\begin{aligned} \text{tr}(S_B) &= \sum_{i=1}^{k} n_i (\mu_i - \mu)^T (\mu_i - \mu) && (13) \\ &= \sum_{i=1}^{k} n_i \mu_i^T \mu_i - n \mu^T \mu && (14) \end{aligned}$$

and the term $\text{tr}(S_W)$ is written as

$$\begin{aligned} \text{tr}(S_W) &= \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)^T (x - \mu_i) && (15) \\ &= \sum_{i=1}^{n} x_i^T x_i - \sum_{i=1}^{k} n_i \mu_i^T \mu_i && (16) \end{aligned}$$

Considering the extreme situations for $k = 1$ and $k = n$ as

$$\text{tr}(S_B) = \{ \begin{array}{cc} \text{c} > 0 & k = 1 \\ 0 & k = n \end{array}$$

$$\text{tr}(S_W) = \{ \begin{array}{cc} 0 & k = 1 \\ \text{c} > 0 & k = n \end{array}$$

However, it is quite hard to find the explicit function to delineate relationships between S_B (S_W) and k. Based on the clustering result from Phase I, it can be seen that the upper bound of k is equal to n_{sb}, which is normally much less than n. The optimal number of final cluster, k_{opt} will be in $(1, n_{sb}]$.

In this paper, the numerical analysis is used to find the optimal solution for k. Referring to (6), numerical determination of k_{opt} is equivalent to look for the optimal choice of x_T in the observation space. The objective function is locally maximized so that the clusters are as separated as possible and each cluster is as compact as possible. The optimal choice of x_T is described as

$$x_T = \arg\max_{x \in \Re^d} J(\varphi(x)) \tag{17}$$

The merging process of Phase II is illustrated as following:

- ◇ Select an appropriate x_T and compute the SM_T.
- ◇ Merge sub-clusters whose similarity measures are larger than SM_T.
- ◇ Stop the merging process until none of the similarity measures between pairs of sub-clusters is larger than SM_T.

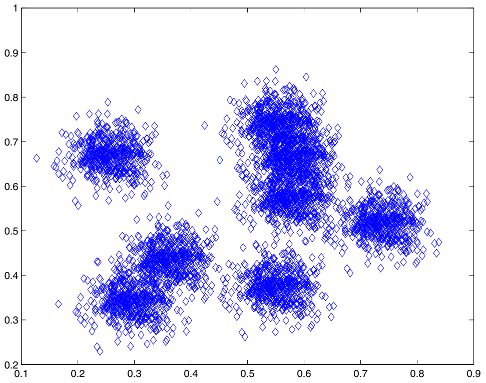

Fig. 6. The data set used for performance evaluation

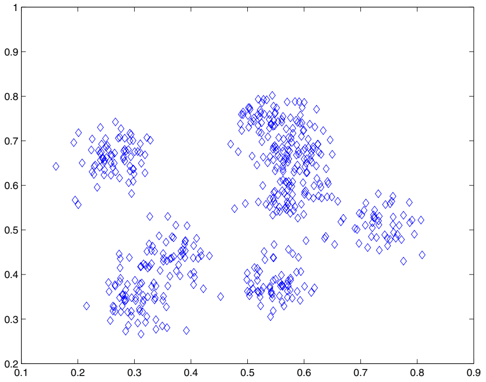

Fig. 7. The 500 data set used for cluster searching

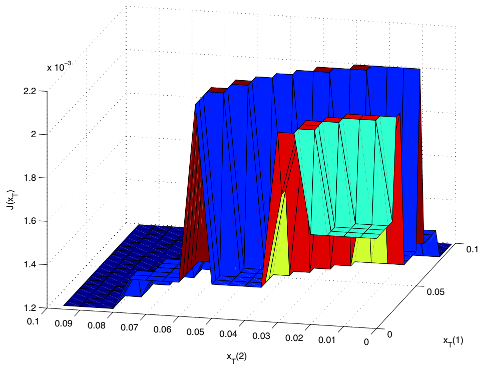

Fig. 8. Fisher class separability measure vs. threshold vector x_T

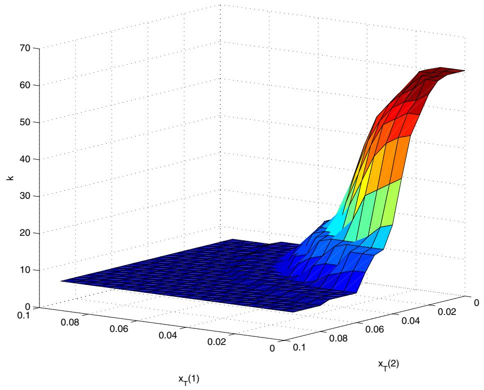

Fig. 9. The number of final cluster vs. threshold vector x_T

C. Computational Complexity Analysis

The overall computational complexity of the proposed hierarchical clustering depends on the amount of time required to perform the two phases of the clustering algorithm. The amount of time required by the two-phase clustering relies on the number of the sub-cluster generated in the phase I and the size of each sub cluster. Without loss of generality and to simplify the analysis, it is assumed that each sub cluster has the same number of data points, m. So the total number of sub clusters is $\frac{n}{m}$. By analyzing the Trellis Diagram, the computation complexity for Phase I is $O((n-m)n)$. And the computation complexity for the phase II is bounded by $O(\frac{n^2}{m^2})$. The overall computational complexity for two-phase clustering algorithm is $O((n-m)n+\frac{n^2}{m^2})$.

IV. Performance Evaluation

A data set, Ψ, containing 8 cloudy points in 2-D as shown in Figure 6, is applied for the performance evaluation of the proposed hierarchical clustering algorithm. The total number of data points in this data set is 4000. A 500 data point set, Ω, is randomly chosen from the original data set as shown in Figure 7. Since the data set is in 2-D, the threshold x_T is denoted as

$$x_T = \begin{pmatrix} x_T(1) & x_T(2) \end{pmatrix}^T$$

The numerical simulation of $J(x_T) \sim x_T$ is demonstrated in Figure 8.

As we can see, the locally maximum value of $J(x_T)$ occurs when the $\|x_T\|$ is around 0.06. The relationship between the number of clusters and the threshold vector x_T is shown in Figure 9. since the maximum value of $J(x_T)$ occurs at $\|x_T\|$ around 0.06, the number of final clusters is determined to be 8 by examining Figure 9. The clustered data set Ω is shown in Figure 10. The Ω data set can be used as 'teacher' for the supervised learning. By applying probabilistic neural network-based (PNN) pattern classifier in [15], the testing patterns are from the data set $\Psi - \Omega$, and the simulation result is demonstrated as in Figure 11. The density estimation from PNN classifier is shown in Figure 12 and Figure 13. As we can see, the 8 cloudy data set Ψ can be correctly clustered by choosing appropriate value of similarity threshold. The density estimation from the PNN classifier correctly reveals the characteristics of data points.

V. Conclusions

This paper presented a novel hierarchical clustering algorithm based on Fisher class separability measure. The Gaussian similarity measure is adopted. The proximity matrix is used to reveal the pairwise similarity between

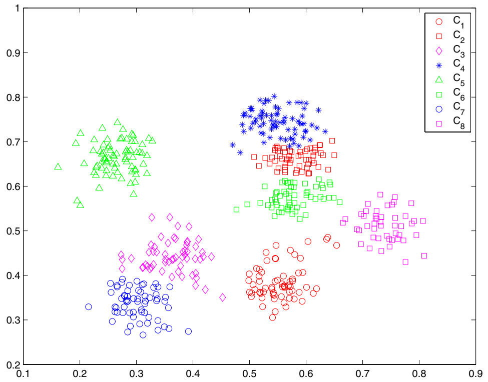

Fig. 10. Clustering result for Ω data set

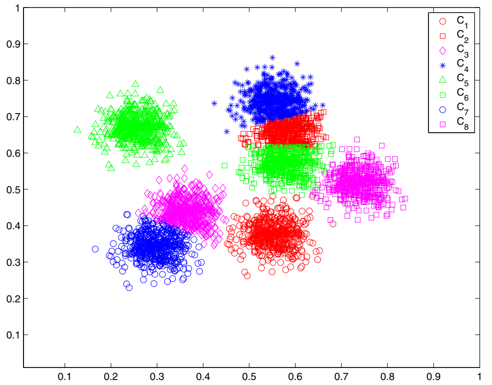

Fig. 11. Classification of testing data points using probabilistic neural network-based classifier

data points. The clustering procedure consists of two phases: (1) Generation of sub-clusters; (2) Merging of sub-clusters. Trellis Diagram is used to generate the sub-cluster set, and the connection between two consecutive layers in Trellis Diagram is weighted by the similarity measure. The Trellis Diagram provides a fast way for searching sub-cluster. Fisher class separability measure is applied for the design of similarity threshold for the merging of sub-clusters so that the data points in the same cluster will have maximum intracluster connectivity while the intercluster similarity is minimized. As mentioned early, the proposed clustering algorithm is applicable for the optimal determination of radial basis function sets and also can be provided as 'teacher' for the supervised learning. In the future work, the design of objective function will be further studied; and the clustering scheme will be improved for the presence of very 'noisy' data.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *"Advances in Knowlodge Discovery and Data Mining"*, AAAI/MIT Press, 1996.

[2] A. Gersho and R. M. Gray, *"Vector Quantization and Signal Compression"*, Boston: Kluwer Academic, 1992.

[3] R. O. Duda and P.E. Hart, *"Pattern Classification and Scene Analysis"*, New York: John Wiley & Sons, 1973.

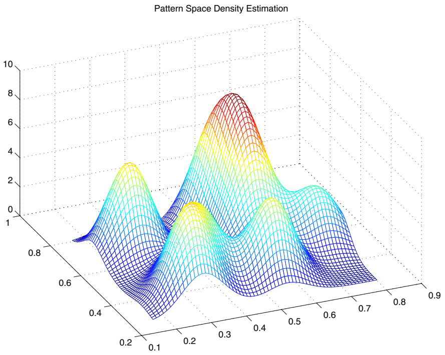

Fig. 12. Density estimation for data set Ψ using probabilistic neural network-based classifier

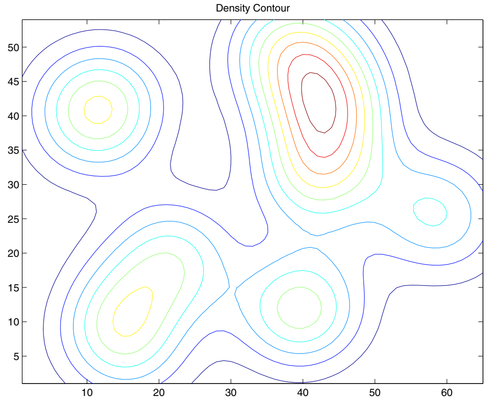

Fig. 13. Density estimation contour for data set Ψ

[4] A. K. Jain and R. C. Dubes, *"Algorithms for Clustering Data"*, Prentice Hall, 1988.

[5] L. Kaufman and P. J. Rousseeuw, *"Finding Groups in Data: an Introduction to Cluster Analysis"*, John Wiley & Sons, 1990.

[6] R. Ng and J. Han, *"Effecient and Effective Clustering Method for Spatial Data Mining"*, in Proc. of the 20th VLDB Conferences, 1994.

[7] S. Guha, R. Rastogi, and K. Shim, *"CURE: An Effiecient Clustering Algorithm for Large Databases"*, in Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data, 1998.

[8] M. Ester, H. P. Kriegel, J. Snader, and X. Xu, *"A Density-based Algorithm for Discovering Clusters in Large Spatial DataBase With Noise"*, in Proc. of the 2nd Int. Conference on Knowledge Discovery and Data Mining, 1996.

[9] A. Gersho and R. M. Gray, *"Vector Quantization and Signal Compression"*, Boston: Kluwer Academic, 1992.

[10] M. Inaba, H. Imai, and N. Katoh, *"Experimental Results of a Randomized Clustering Algorithm"*, in Proc. of 12th Ann. ACM Symp. Computational Geometry, 1996.

[11] H. Skala, *"Trellis Theory"*, Providence, R. I. : American mathematical Society, 1972.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *"The Elements of Statistical Learning: Data Mining, Inference, and Prediction"*, New York: Springer, 2001.

[13] K. Fukunaga, *"Introduction to Statistical Pattern Recognition"*, New York: Academic Press Inc., 1996.

[14] C. M. Bishop, *"Neural Networks for Pattern Recognition"*, OXFORD University Press Inc., 1995.

[15] Xudong Wang and Vassilis L. Syrmos *"Hybrid Probabilistic Neural Networks for Pattern Recognition"*, in Proceedings of 12^{th} Mediterranean Conference on Control and Automation, June, 2004.