Two-Step Classification Based on Scale Space*

Ming TANG†, Jing XIAO and SongDe MA
National Laboratory of Pattern Recognition, P.O.Box 2728, Beijing 100080, P.R. China

Abstract

A new two-step classification scheme based on nonparametric estimation of density function and scale-space filtering is presented in this paper. This scheme is able to combine traditional supervised classification techniques with clustering. After nonparametric estimation of the underlying density function, this scheme utilizes scale-space filtering and a novel classification algorithm to extract the intrinsic **basic** *structure of the data. Then, depending on applications, one of the traditional clustering or classification techniques may be employed to obtain a final* **high level** *data structure.*

Keywords classification, clustering

1 Introduction

The Gaussian mixture model approach (GMM) has been widely employed as a clustering method. The main problem of GMM is that the cluster validity is subject to a strong assumption that the form of the underlying density function is known *a priori*. In addition, the local centroid (mean) estimation is not robust to noise. In order to overcome the above drawbacks, Wilson and Spann [5] presented a paradigm shift for evaluation of the clustering problem. Their shift makes the estimate of valid structure within a data set be robust to both noise and spatial scale changes of the data. Roberts [2] went a step further to propose a method of unsupervised classification using scale-space filtering, and showed that GMM fails for data sets which are not multivariate Gaussian while the scale-space based method is considerably more robust.

In general, a clustering algorithm may be required to cluster the data according to a preferred number of partitions. Although the scale-space based clustering method [2] can also be directly extended to control the number of resulting clusters and [2] indeed provided a successful example, the precision of such extension may not be high in general, and the resulting clusters may not be what is wanted. In addition, the classification rule adopted in [2] is not general and precise enough for data whose dimension is larger than one. The hill-clustering method proposed by Tsai and Chen [4] can only be used in one dimensional histogram and is not precise enough in peak positioning if employed as the classification rule.

To overcome these drawbacks, this paper presents a new two-step classification scheme. This scheme is able to combine traditional supervised classification techniques with clustering. After nonparametric estimation of the underlying density function, our scheme utilizes scale-space filtering and a novel classification algorithm to extract the intrinsic *basic* structure of the data. The novel classification algorithm can precisely operate in the feature histogram of *any dimensions*. Finally, depending on applications, various clustering or classification techniques can be employed to obtain a *high level* data structure.

The rest of this paper is organized as follows. Section 2 briefly describes the scale-space based clustering method [2] at first, then presents a general and precise classification algorithm as our classification rule. In section 3, a new two-step classification scheme is presented. Experimental results and the conclusion are given in Sections 4 and 5, respectively.

Due to the limitation of the paper length, the description is very brief in this paper. Refer to [3] for the strict and full description of our method and the experiments.

2 Scale-Space Filtering Based Clustering

2.1 Theory

A nonparametric estimation of the probability density function of n ($n \geq 1$) dimensional noisy data is the weighted combination of a set of basic functions $f_i(\vec{x})$

$$\hat{p}_\nu(\vec{x}) = \sum_{i=1}^{N} w_i f_i(\vec{x})$$

*Specially thank Dr. YanXi Liu of CMU for her help in preparing part of the experimental data. The first author would also thank Prof. GongQing Zhang for his help in preparing the text. This work is partially supported by Beijing Institute of Environment Feature.

†Email: tangm@nlpr.ia.ac.cn

0-7695-0750-6/00 $10.00 © 2000 IEEE

where $\vec{x}$ is the n-dimensional datum. If Parzen-window approach is used to estimate densities, N denotes the number of samples, and $w_i = w$ is a normalizing contant. If the approach to finding a set of basic functions to expand the density function is employed, N is the number of the basic functions used, and w_i is also a normalizing factor and is concerned with all samples and i-th basic function.

Our purpose is to detect the genuine data peaks of $\hat{p}_\nu(\vec{x})$ with a specified filter. [5] and [2] showed that, for the data of any dimensions, Gaussian filter is a proper choice

$$\hat{p}_\sigma(\vec{x}) = \hat{p}_\nu(\vec{x}) * G_\sigma(\vec{x})$$

where $G_\sigma(\vec{x})$ is Gaussian.

Therefore, the genuine data structure can be approximately extracted from $\hat{p}_\sigma(\vec{x})$ with some proper σ.

2.2 Cluster Validity

Let $\pi(\sigma)$ be the number of peaks of $\hat{p}_\sigma(\vec{x})$. Suppose that in the evolution of $\hat{p}_\sigma(\vec{x})$ with the increment of σ, after $\sigma \geq \sigma_1$, $\pi(\sigma_1) = \pi(\sigma) = \pi(\sigma_2)$, i.e., $\pi(\sigma)$ is stable over the range of σ_1 to σ_2, where $\sigma \in [\sigma_1, \sigma_2]$ and $|\sigma_1 - \sigma_2| > v_t$, v_t is a threshold and may be determined empirically. And also suppose that there does not exist any such stable interval if $\sigma < \sigma_1$. Then a *valid estimation of the density function* of the noisy data is $\hat{p}_{\sigma_1}(\vec{x})$.

2.3 Classification Algorithm

As the form of $\hat{p}_{\sigma_1}(\vec{x})$ may be extremely complicated in practice, generally speaking, any single formula (e.g., the classification formula used in [2]) to determine the cluster for each datum may cause serious error in the case of high dimensional data. Therefore, a general and precise classification algorithm has to be designed to label all data.

The *essential difference* between the following **Algorithm 1** and other approaches to dealing with the similar problem is that it considers the density function $\hat{p}_{\sigma_1}$ to be discrete, but others (e.g., [1]) consider it to be continuous. The formulae for the latter are elegant, but do not work well in high dimensional space.

Before describing our classification algorithm formally, we first illustrate its basic strategy. The notation of intervals will be employed to express the set of discrete samples in an interval. Fig.1 gives an example of density function upon the discrete sample set $[B_l, B_r]$. According to such interval notation, B_l and B_r are samples, too. Our classification algorithm will follow the procedure below to label all $x \in [B_l, B_r]$. Note that $[B_l, B_r] - \{x_i | i = 1, \ldots, 8\} \neq \phi$. Firstly, find out all its local maximum points, i.e., x_1 and x_8. Therefore, the number of the clusters is 2 and the centers of the clusters are x_1 and x_8, respectively. With the gradient increment method, if point x can reach the cluster center x_i, x is labelled with x_i ($i = 1, 8$). Therefore, $\forall x \in [B_l, x_2]$ and $\forall x \in [x_7, B_r]$ can be labelled. Secondly, suppose $Y = \{x | \nabla\hat{p}_\sigma(x) = 0, x \text{ does not belong to any peak}\}$; then $Y = (x_2, x_3] \bigcup (x_4, x_5] \bigcup (x_6, x_7)$. Finding all maximal connected subsets of Y, $C_1 = (x_2, x_3]$, $C_2 = (x_4, x_5]$ and $C_3 = (x_6, x_7)$ are obtained, where each subset has a unique value of density function. As $\hat{p}_\sigma(y_1) > \hat{p}_\sigma(y_2) > \hat{p}_\sigma(y_3)$, where $y_1 \in C_1$, $y_2 \in C_2$ and $y_3 \in C_3$, our algorithm will first label $x \in C_1$, then $x \in C_2$, and at last $x \in C_3$. Therefore, $\forall x \in C_1$ are labelled firstly with what x_2 has been labelled. As $\forall x \in (x_3, x_4]$ can reach x_3 with gradient increment method, and x_3 has been labelled, $\forall x \in (x_3, x_4]$ are labelled with what x_3 has been labelled. Next step, $\forall x \in C_2$ are labelled with what x_4 has been labelled, and then $\forall x \in (x_5, x_6]$ are labelled with what x_5 has been labelled. As x_6 and x_7 have been labelled, for any $x \in C_3$, if $|x - x_6| > |x - x_7|$, x is labelled with what x_7 has been labelled, otherwise x will be labelled with what x_6 has been labelled. Therefore, all sample points are labelled with x_1 or x_8.

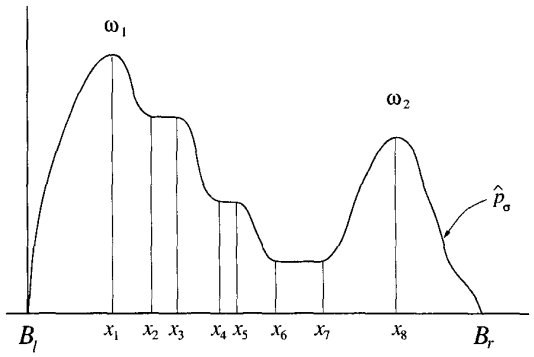

Figure 1. The basic strategy adopted in Algorithm 1. See text.

The points labelled with x_1 are accepted to belong to ω_1, other points belong to ω_2.

Algorithm 1. (Labeling every point with a proper peak)

Suppose m is the peak of discrete function f, and that M is the set of all m's. X is the domain of f. While the gradient increment method is used at $\vec{x}$, only the values of $f(\vec{x})$ at $\vec{x}$'s $3^n - 1$ closest neighbors are compared to decide the next point to go.

Step 1. Starting with every $\vec{x} \in X$, employ the approach of gradient increment to find out its corresponding local maximum point $\vec{x}_{max}$. The path of gradient increment is saved as $H_{\vec{x}}$. And $\vec{x}$ is only located in *one* increment path.

Step 2. For $\forall \vec{x} \in m$, label $\vec{x}$ with m. If $\vec{x}_{max} \in m$, label with m all points in $H_{\vec{x}}$. Remove all of them from X. If $X = \Phi$, stop.

Step 3. Find $Y = \{\vec{x} | \vec{x} \in X$, the maximal increment of $f(\vec{x})$ among $(3^n - 1)/2$ directions is $0\}$ $(Y \neq \phi)$.

Step 4. Find $S = \{s | s \subseteq Y, \forall \vec{x} \in s, f(\vec{x})$'s are identical.$\}$.

Step 5. Find $C = \{c | c \subseteq s, s \in S, c$ is connected. For any $\vec{x} \in s$ and $\vec{x} \notin c, c \bigcup \{\vec{x}\}$ is disconnected$\}$.

Step 6. Select out c which satisfies the following *condition*: $\vec{x} \in c, \vec{x}' \in c', c \in C, c' \in C, c \neq c', f(\vec{x}) \geq f(\vec{x}')$. Set $P = \{\vec{x} | \vec{x} \notin c, \vec{x}$ is a neighbor of $\vec{y} \in \partial c, \vec{x}$ has been labelled.$\}$.

Step 7. For every $\vec{y} \in c$,

$$\vec{x} = \arg \min_{\vec{x}_i \in P} \|\vec{x}_i - \vec{y}\|$$

label $\vec{y}$ with $m_{\vec{x}}$.

Step 8. For each $\vec{x} \in \partial c$, $H_{\vec{y}}$ is a path of gradient increment from $\vec{y}$ to $\vec{x}$. If $\vec{r} \in H_{\vec{y}}$ is not labelled, label $\vec{r}$ with $m_{\vec{x}}$.

Step 9. Remove c from C. If C is not empty, go to **Step 6**. Otherwise, stop.

After running this algorithm with $f = \hat{p}_{\sigma_1}$, the data labelled with the same m belong to the same cluster.

3 Scale-Space Based Two-Step Classification

We have improved the scale-space based method of clustering [2] via inducing a general and precise classification algorithm (**Algorithm 1**). But when one wants to control the number of resulting partitions, what should he do? The direct extension of scale-space based clustering to do this is unsatisfactory [3]. It will introduce too many errors. In order to control the number of resulting classes while reduce the error, we propose the following algorithm.

Algorithm 2. (Two-Step classification scheme based on scale-space)

Step 1. Estimate the underlying density function

With the noisy data, make estimation

$$\hat{p}_\nu(\vec{x}) = \sum_{i=1}^{N} w_i f_i(\vec{x})$$

and scale space

$$\hat{p}_\sigma(\vec{x}) = \hat{p}_\nu(\vec{x}) * G_\sigma(\vec{x})$$

Starting with σ_0, evolve $\hat{p}_\sigma(\vec{x})$ with $\Delta\sigma$, the step length of σ, until a valid density estimation, $\hat{p}_{\sigma_1}(\vec{x})$, is established.

Step 2. Find M, the set of all maximum points of $\hat{p}_{\sigma_1}(\vec{x})$.

Step 3. Determine the cluster to which each datum belongs with **Algorithm 1**.

Step 4. Depending on applications, select some classic algorithm of classification(such as c-means algorithm, nonlinear discrimination function method, etc.), and distance measurement(e.g., Euclidean measurement) to partition M into k classes, $|M| \geq k$, where k is a parameter representing the required number of resulting classes and $k = 0$ means $|M|$ classes are required, i.e., $k = |M|$. Correspondingly, $|M|$ clusters, $\omega_i (i = 1, 2, ..., |M|)$, are partitioned into k classes, $c_j (j = 1, 2, ..., k)$, too.

Step 5. For $\forall \vec{x}$, if $\vec{x}$ belongs to ω_i, and ω_i is classified into c_j in **Step 4**, classify $\vec{x}$ into c_j.

When **Algorithm 2** is employed as a clustering method, the cluster validity is determined by the validities of both scale-space based clustering method and the clustering method used in **Step 4**.

4 Experimental Results

Algorithm 2 has been used to classify several data sets. Fig.2 shows an experiment on a 2-dimensional data set of brain pathology. The data are shown in Fig.2(a). It is seen that there is a valid clustering of the data, i.e., 3 clusters. In order to estimate the density function of the data, a histogram is constructed and is shown in Fig.2(d), where the horizontal axis represents the subtraction of the two components of the data. All subtractions are mapped into $[0, 100]$. The reason to map the subtractions into $[0, 100]$ is that the subtraction of the maximum and minimum of subtractions of two components is about 100. An evolved version of Fig.2(d) is shown in (e), where $\sigma = 2.17$, which is the valid estimation of Fig.2(d). Six maximum points are located at 4, 12, 36, 48, 63, 76. To classify the original data in a desirable way, two discrimination points, 42 and 56 are adopted. Consequently, the above 6 points are classified into 3 classes, $\{4, 12, 36\}$, $\{48\}$ and $\{63, 76\}$. The resulting classification of the original data is shown in Fig.2(b).

While the scale-space based clustering method [2] is used to cluster the same data, the data will be clustered into 6 clusters according to the above discussion. This is not valid with respect to the application, although it's valid according to the scale-space based clustering algorithm [2]. On the other hand, if the direct extension is employed, the remaining 3 maximum points are 72, 97, 152, and the final clusters are shown in Fig.2(c). It is noted that the error is considerable and the result is unacceptable. In this example, both the scale-space based clustering method and it's direct extension fails to obtain a valid clustering.

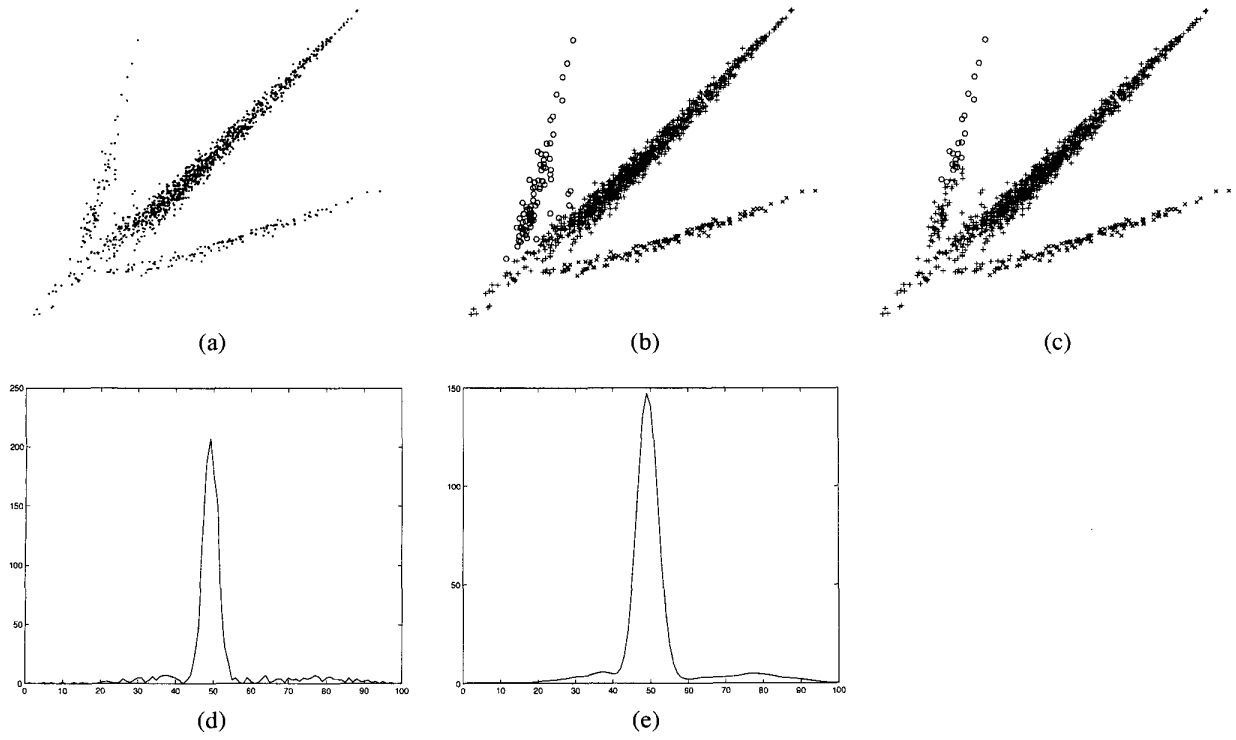

Figure 2. The classification of a set of pathological data of brain. See text for explanation.

For high dimensional data, by means of the knowledge about the data, one may analyze which conventional clustering or classification algorithms may be invoked to determine which maximum points should belong to a class. In such process, methods to reduce the dimension of data, such as KL transformation, may be used.

5 Conclusions

This paper presents a new two-step classification algorithm which can introduce traditional supervised classification techniques into clustering. Strictly speaking, our classification scheme belongs to neither traditional supervised approaches nor traditional unsupervised ones.

References

[1] Devijver, P.A. and Kittler, J., *Pattern Recognition, a Statistical Approach*, Prentice Hall, Englewood Cliffs, London, 1982.

[2] Roberts, S.J., *Parametric and Non-Parametric Unsupervised Cluster Analysis*, Pattern Recognition, Vol.30, No.2, pp261-272, 1997.

[3] Tang, M., Xiao, J. and Ma, S.D., *A New Scheme of Classification based on Scale Space*, Tech. Rep., National Laboratory of Pattern Recognition, Institute of Automation, CAS, 1999.

[4] Tsai, D.M. and Chen, Y.H., *A fast histogram-clustering approach for multilevel thresholding*, Pattern Recognition Letters, Vol 13, pp245-252, 1992.

[5] Wilson, R. and Spann, M., *A New Approach to Clustering*, Pattern Recognition, Vol.23, No.12, pp1,413-1,425, 1990.