http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 6 (2): 310-324, 2007
ISSN 1812-5638

A Visual System Theoretic Cost Criterion and its Application to Clustering and Fuzzy Modeling

[1,2]Shitong Wang, [2]Fu-lai Chung, [1]Min Xu, [1]Zhaohong Deng and [3]Dewen Hu
[1]School of Information Engineering, Southern Yangtze University, Wuxi, China
[2]Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
[3]School of Automation, National Defense University of Science and Technology, Changsha, China

Abstract: We all know that our eyes can inherently and effectively recognize/classify objects under complex conditions. Hence, we believe that an efficient clustering approach not only depends on the principles of physical systems by which the data are generated but also on the manner that human eyes sense the structure of the data. In this study a visual system theoretic cost criterion function is proposed and based upon which a new clustering algorithm is derived. The new cost criterion is visual sampling and Weber's law is applied. The new criterion function can be made "kernelized" so that developed based on a visual system modeling of the multi-dimensional data where the visual system theories like different kernel functions can be used under different practical requirements. Furthermore, it evaluates the tightness of intra-group's data distribution and the separable degree among groups simultaneously. The experimental results demonstrate that the new clustering algorithm is especially suitable for nonlinearly separable datasets.

Key words: Nonparametric clustering, pattern recognition, visual system, Weber's law

INTRODUCTION

Clustering has long been a hot research topic in various disciplines and it has been widely applied in the past decades, e.g., exploratory pattern analysis, data grouping, decision making, document retrieval, image segmentation and pattern classification. Recently, more and more researchers are interested in its application in large datasets (Anderberg, 1973; Jain *et al*., 1988). Clustering algorithms attempt to organize unlabeled input vectors into clusters or "natural groups" such that data points within a cluster are more similar to each other than those belonging to different clusters, i.e., to maximize the intra-cluster similarity while minimizing the inter-class similarity. Various clustering approaches have been proposed (Arabie *et al*., 1996; Dubes *et al*., 1976; Shitong *et al*., 2006; Marr, 1982; Koenderint, 1984; Romeny *et al*., 1993, 1997) and they can be divided into two types: hard vs. fuzzy/probabilistic. In fuzzy/probabilistic clustering, a given data sample does not necessarily belong to only one cluster but can have varying degrees of memberships/probabilities to several clusters.

There are two basic approaches in fuzzy/ probabilistic clustering algorithms, namely, parametric and nonparametric. In parametric clustering, we assume a predefined distribution for the data set and calculate the sufficient statistics or fuzziness that describes the data set in a compact way. For example, in probabilistic clustering, the sufficient statistics are the sample mean and the sample covariance matrix, which describe the assumed normal distribution perfectly. In fuzzy clustering, the membership functions heavily depend on the distance measures. Unfortunately, if the data set is not distributed in accordance with our choice, then statistics/fuzziness can be very misleading. Although various parametric clustering algorithms have earned great success in many fields, their performances are often sensitive to the initial conditions and estimating the parameters is not a trivial task.

The nonparametric clustering algorithms on the other hand divide the data set into groups of points that have strong internal similarities. In order to measure the similarities, a cost criterion function is typically used and the grouping maximizing (or minimizing) the cost criterion is determined. This kind of algorithms requires a cost criterion function to evaluate how well the clustering fits the data and an algorithm to optimize the cost function. For example, the valley seeking clustering (Zheng, 1998) and the information theoretic clustering (Romeny *et al*., 1993) belong to this category. A good cost criterion should well evaluate the tightness of intra-group's data distribution and the separable degree among groups simultaneously. However, the valley seeking and information theoretic clustering algorithms only consider the separable degree among groups and the reported experimental results demonstrate their successes.

Corresponding Author: Shitong Wang, School of Information Engineering, Southern Yangtze University, Wuxi, China

our eyes can inherently and effectively recognize/classify objects under complex conditions. Thus, we believe that an efficient clustering approach not only depends on the principles of physical systems by which the data are generated but also on the manner that human eyes sense the structure of the data. In this study a visual theoretic clustering algorithm is proposed. It is based on a new visual system theoretic cost criterion which can be made "kernelized" so that different kernel functions can be used under different practical requirements. The experimental results demonstrate that the new clustering algorithm is especially suitable for nonlinear separable datasets.

VISUAL SYSTEM MODELING OF DATA

According to statistics, more than 80% of information obtained by humans comes from the eyes. With the development of physiology and psychology, people have gained more and more knowledge about how the eyes function, which makes it possible for us to imitate the visual mechanism of eyes. In early 1980s, Marr proposed the concept of visual computational theory (Coren *et al*., 1994). For the past two decades, this theory has been prevailing in computer vision research. It is well known that sensations are non-uniformly distributed on retina. The fruitful research results in the fields of physiology and psychology have revealed that the visual system of advanced creatures is an active perception process based on visual sampling and eye motion (Poggio, 1990). Visual systems have found wide applications in computer science, especially in the field of image processing. Being motivated by the strong and inherent clustering ability of human eyes, we propose to model multi-dimensional data by visual system theories and describe the modeling results below.

Given a dataset consisting of N d-dimensional data points $X = \{\vec{x}_i \in R^d \mid i = 1,2,\cdots,N\}$ and according to the visual system theories (Poggio, 1990; Shitong *et al*., 2002; Nyquist, 1928; Xuanli *et al*., 1999), an "image" representation of the distribution of these N data points can be expressed as

$$f(\vec{x},\theta) = c\sum_{i=1}^{N} K(\vec{x},\vec{x}_i,\theta) \tag{1}$$

where $K(\vec{x},\vec{x}_i,\theta)$ is a kernel function with parameter θ to denote the visual sampling frequency (Poggio, 1990) or the window size, c is a constant whose value depends on $K(\vec{x},\vec{x}_i,\theta)$. The two most important kernel functional forms are Sinc and Gaussian. Such a modeling of multi-dimensional data is called visual sampling (Poggio, 1990) which is originated from the classical (1-D) signal sampling principle, i.e.,

$$s(t) = \sum_{k=-\infty}^{\infty} s(kT)\frac{\sin\pi(\frac{t}{T}-k)}{\pi(\frac{t}{T}-k)} \tag{2}$$

where s(kT) denotes the sample of signal s(t) at time kT. Here, we are given a d-dimensional data set X. If each datum of the dataset is considered as a sampling point, we will obtain a sample image or image representation of the dataset based, i.e.,

$$f(\vec{x},T) = \sum_{i=1}^{N}\prod_{k=1}^{d}\frac{\sin(\frac{x_k - x_{ik}}{T})}{(\frac{x_k - x_{ik}}{T})} \tag{3}$$

where x_{ik} denotes the K^{th} component of the i^{th} sample and T is the visual sampling frequency. The equivalence of (1) and (3) is now obvious. Based on (1), a new cost criterion function is derived

The Weber's law (Shitong *et al*., 2000; Jifeng *et al*., 1988; Nyquist, 1928) is another milestone of visual system research. It says that if the luminance l_o of an object is just noticeably different from the luminance l_s of its surround, then their ratio is

$$\frac{|l_s - l_o|}{l_o} = \text{constant.} \tag{4}$$

The Weber's law quantitatively shows that a fixed-proportion increase in stimulus intensity s is sufficient to produce a just noticeable change in sensation. That is, the change of stimulus intensity Δs should follow ks, where k is the Weber fraction (or Weber ratio). According to the Goddess neural theory (Poggio, 1990), the reason that we sense such a noticeable change (Poggio, 1990; Bezdek, 1992) is that all neurons within the reception field $s+\Delta s$ are activated. When clustering is explained using visual behaviors, the Weber's law may be used to tell us that we can simply change the visual sampling frequency or window size in a fixed proportion. In other words, the value of θ in (1) should follow the Weber's law, i.e., $\theta \leftarrow \theta + k\theta$, where k is the Weber fraction. Thus, each "image" representation of the multi-dimensional data set has a life cycle that we can sense this "image" representation in a window interval (θ_1,θ_2). If we have several possible results with various sets of (θ_1,θ_2), we consider that the most reasonable/satisfactory result with the maximal (θ_1,θ_2) should be most stable from the viewpoint of visual systems and Weber's law. Thus, we should take it as the final one.

A VISUAL SYSTEM THEORETIC COST CRITERION

Nonparametric clustering is to partition the given dataset into several groups according to a cost criterion. Suppose there are two groups of d-dimensional data points ω_1 and ω_2. According to the visual system modeling described in previous section, the corresponding two images are

$$f_1(\vec{x},\theta) = c\sum_{\vec{x}_i\in\omega_1} K(\vec{x},\vec{x}_i,\theta) \tag{5}$$

and

$$f_2(\vec{x},\theta) = c\sum_{\vec{x}_i\in\omega_2} K(\vec{x},\vec{x}_i,\theta) \tag{6}$$

respectively. If $K(\vec{x},\vec{x}_i,\theta)$ is taken as the Sinc kernel function $S(\vec{x},\vec{x}_i,T)$, then let

$$S(\vec{x},\vec{x}_i,T) = \prod_{k=1}^{d}\frac{\sin(\frac{x_k - x_{ik}}{T})}{(\frac{x_k - x_{ik}}{T})} \tag{7}$$

where T is the visual sampling frequency. If $K(\vec{x},\vec{x}_i,\theta)$ is taken as the Gaussian kernel function $G(\vec{x},\vec{x}_i,\sigma)$, then let

$$G(\vec{x},\vec{x}_i,\sigma) = (2\pi\sigma^2)^{-\frac{d}{2}}\exp(-\frac{\|\vec{x}-\vec{x}_i\|^2}{2\sigma^2}) \tag{8}$$

where σ is called the size of the observable window.

We define the following squared error function D to measure the distance between the above two images

$$D(\vec{x}) = \int (f_1(\vec{x},\theta) - f_2(\vec{x},\theta))^2 dx \tag{9}$$

$$= \int [f_1^2(\vec{x},\theta) + f_2^2(\vec{x},\theta) - 2f_1(\vec{x},\theta)f_2(\vec{x},\theta)]dx \tag{10}$$

When the Gaussian kernel function is taken, we have

$$\begin{aligned} D(\vec{x}) &= \int [f_1^2(\vec{x},\theta) + f_2^2(\vec{x},\theta) - 2f_1(\vec{x},\theta)f_2(\vec{x},\theta)]dx \\ &= (2\pi\sigma^2)^{-d}[\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_1} G(\vec{x}_i,\vec{x}_j,2\sigma) + \sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_2} G(\vec{x}_i,\vec{x}_j,2\sigma) \\ &\quad - 2\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_2} G(\vec{x}_i,\vec{x}_j,2\sigma)] \\ &= D_1 + D_2 - 2\times D_{12} \end{aligned} \tag{11}$$

where

$$D_1 = (2\pi\sigma^2)^{-d}\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_1} G(\vec{x}_i,\vec{x}_j,2\sigma) \tag{12}$$

$$D_2 = (2\pi\sigma^2)^{-d}\sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_2} G(\vec{x}_i,\vec{x}_j,2\sigma) \tag{13}$$

$$D_{12} = (2\pi\sigma^2)^{-d}\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_2} G(\vec{x}_i,\vec{x}_j,2\sigma) \tag{14}$$

Here, the squared error function D gives a new visual system theoretic cost criterion. Similarly, if the Sinc function is taken, we may define the squared error function D as

$$\begin{aligned} D(\vec{x}) &= \sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_1} S(\vec{x}_i - \vec{x}_j,2T) + \sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_2} S(\vec{x}_i - \vec{x}_j,2T) \\ &\quad -2\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_2} S(\vec{x}_i - \vec{x}_j,2T) = D_1 + D_2 - 2\times D_{12} \end{aligned} \tag{15}$$

where

$$D_1 = \sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_1} S(\vec{x}_i - \vec{x}_j,2T), \tag{16}$$

$$D_2 = \sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_2} S(\vec{x}_i - \vec{x}_j,2T) \tag{17}$$

$$D_{12} = \sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_2} S(\vec{x}_i - \vec{x}_j,2T), \tag{18}$$

and

$$S(\vec{x}_i - \vec{x}_j,2T) = \prod_{k=1}^{d}\frac{\sin(\frac{x_{ik} - x_{jk}}{2T})}{(\frac{x_{ik} - x_{jk}}{2T})} \tag{19}$$

where x_{ik} denotes the kth component of the ith sample and T is the visual sampling frequency. We define sin/x=1 if x=0.

As D measures the distance between two images or classes, the derivations above can be similarly applied to other kernel functions to suit the practical requirements.

In order to obtain good clustering results, we need to optimize D by maximizing the intra-class measures D_1 and D_2 while minimizing the inter-class measure D_{12}. As is well known, the traditional approach to clustering is to measure the hard/soft distance between two groups (to be generalized to multi-group in the next section), i.e., to compute the mean distance among all the data points of the two groups by

$$D_{avg}(\omega_1,\omega_2) = \frac{1}{N_1N_2}\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_2}\|\vec{x}_i - \vec{x}_j\| \tag{20}$$

where N_1 and N_2 denote the number of data points of the two datasets ω_1 and ω_2 respectively. Such a measure is good enough for datasets of sufficient separable degree. For groups that are rather close to each other and are nonlinearly separable, this measure may fail.

For our visual system theoretic cost criterion D in (11) or (15), the measure D_{12} actually maps $D_{avg}(\vec{x}_1,\vec{x}_2)$ into a high-dimensional space using kernel functions such as Sinc and Gaussian as shown in (14) and (18), respectively. In other words, D_{12} only measures the soft distance or the separable degree between two groups. A good cost criterion should well evaluate the tightness of intra-group's data distribution and the separable degree among groups simultaneously. The goal of optimizing the clustering criterion is to make each group as tight as possible while the separable degree among different groups as large as possible. Obviously, the new criterion function D in (11) or (15) is a reasonable realization because D_1 and D_2 reflect the tightness or looseness of intra-group's data distribution, while D_{12} indicates the separable degree or closeness among groups.

VISUAL SYSTEM THEORETIC CLUSTERING

Multi-group modeling: The formulation presented earlier only deals with the case of two groups. We extend it to handle multiple groups and present the corresponding clustering procedure. Suppose there are C clusters/ groups in the dataset. The cost criterion function D can be redefined for C clusters as

$$\hat{D} = D_1 + D_2 + \cdots + D_C - 2\sum_{i=1}^{C-1}\sum_{j>i}^{C} D_{ij} \tag{21}$$

Thus, $\sum_{i=1}^{C-1}\sum_{j>i}^{C} D_{ij}$ measures the overall separable degree among all C×(C-1)/2 cluster pairs. Obviously, the reformulation for multiple groups is rational. If the sinc function in (7) is taken as kernels, we can easily have

$$\begin{aligned}
\hat{D} &= D_1 + D_2 + \cdots + D_C - 2\sum_{i=1}^{C-1}\sum_{j>i}^{C} D_{ij} \\
&= \sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_1} S(\vec{x}_i-\vec{x}_j,2T) + \cdots + \sum_{\vec{x}_i\in\omega_K}\sum_{\vec{x}_j\in\omega_K} S(\vec{x}_i-\vec{x}_j,\ 2T) \\
&\quad + \cdots + \sum_{\vec{x}_i\in\omega_C}\sum_{\vec{x}_j\in\omega_C} S(\vec{x}_i-\vec{x}_j,\ 2T) \\
&\quad - 2\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_2} S(\vec{x}_i-\vec{x}_j,2T) - \cdots - 2\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_K} S(\vec{x}_i-\vec{x}_j,\ 2T) \\
&\quad - \cdots - 2\sum_{\vec{x}_i\in\omega_1}\sum_{\vec{x}_j\in\omega_C} S(\vec{x}_i-\vec{x}_j,\ 2T) \\
&\quad - 2\sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_3} S(\vec{x}_i-\vec{x}_j,2T) - \cdots - 2\sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_K} S(\vec{x}_i-\vec{x}_j,\ 2T) \\
&\quad - \cdots - 2\sum_{\vec{x}_i\in\omega_2}\sum_{\vec{x}_j\in\omega_C} S(\vec{x}_i-\vec{x}_j,\ 2T) \\
&\quad \vdots \\
&\quad - 2\sum_{\vec{x}_i\in\omega_{C-1}}\sum_{\vec{x}_j\in\omega_C} S(\vec{x}_i-\vec{x}_j,\ 2T)
\end{aligned} \tag{22}$$

according to (15). Here, N_K is the number of the data points in the K^{th} cluster ω_K and K=1,2,...,C and, which is the total number of data points in the

$$\sum_{K=1}^{C} N_K = N$$

dataset.

Initialization by grouping algorithm: Like many other clustering algorithms, the new clustering algorithm based on the visual system theoretic cost criterion is also sensitive to initialization. Many works have revealed that the grouping algorithm is an effective approach for initializing the dataset (Romeny *et al*., 1993; 1997; Zheng, 1998), so our algorithm here adopts this algorithm to initialize the dataset. The implementation details can be found by (Romeny *et al*., 1993) and here we merely sketch the main idea of the grouping algorithm. Assume that the size of a cluster is N_K and the starting data point is randomly chosen for this cluster, i.e., this class comprises just one data point at the very beginning. New data point (outside this cluster) will be included only when it is closest to all the data points in this cluster. Such a process is repeated to find the next data point to be included in this cluster, until the number of data points in this cluster reaches N_K. Accordingly, all the initial clusters can be constructed in this way. In fact, many experimental results demonstrate that the selection procedure in the grouping algorithm is very effective.

Meanwhile, we need to emphasize that the grouping algorithm can only make a rough assignment of data points in the dataset, i.e., only initiated the cluster distribution. An actual clustering process is needed afterwards. A new clustering algorithm based on the proposed cost criterion is proposed to make finer adjustments on the dataset. The grouping algorithm can help to avoid the local minima in the clustering space and save the running time and enhance the performance of the new clustering algorithm.

Clustering based on the visual system theoretic cost criterion: The new clustering algorithm based on the visual system theoretic cost criterion function in (21) and (22) can be described as follows:

Based on the initial clustering of the data points by the grouping algorithm with cluster sizes $N_1,...,N_K,...,N_C$ for the C clusters, the new clustering algorithm above consists of two major repeated processes (nested loops in Fig.1). For the inner loop with some sampling frequency T, this new clustering algorithm attempts to reach the current optimal clustering with C classes by finding the maximum value of the cost criterion function $\check{D}$. The result is saved in Clustering Set. For the outer loop, the visual sampling frequency is increased according to the Weber's law. It terminates when the data

```
InitializeByGroupingAlgorithm(NC, ClusterLabels)
/* NC is the size of each initial group, i.e. N1,..., NK,..., NC */
InitializeSamplingFrequency(T)
/* The visual sampling frequency T should be set to a small value initially */
ClusteringSet=Nil
/* Places for storing the clustering results */
REPEAT
    REPEAT
        { FOR K=1 to C
        /* C is the number of clusters */
            { FOR j=1 to NK
                Change the clustering label of the data points of the corresponding group gradually;
                Compute the corresponding cost criterion D̂;
                IF (the cost criterion D cannot become larger) THEN
                    Restore the previous state of the data points;
                ELSE
                    Regard the current clustering as the current optimal clustering;
            }
        }
    UNTIL (there is no more change among groups with the current T)
    CC=Current optimal clustering result for the given T
    Clustering Set=Clustering Set ∪ {(CC,T)}
    T=T+k×T
    /* k is the Weber fraction taking the value 0.05 in this study */
UNTIL (there is no more reasonable clustering result)
/* It means that with the change of T, no other satisfactory clustering results can be obtained, or the number of clusters is less than C occasionally */
Choose two clustering results (CC,T1) and (CC,T2) such that the width of the window interval (T1,T2) is maximum and the clustering results
        within the interval are most satisfactory;
CC=Final clustering result.
```

Fig. 1: Clustering based on the visual system theoretic cost criterion

is clustered into less than C classes, i.e., one or more clusters are empty, or there is no other satisfactory clustering result obtained. Using the terminology of visual systems and the Weber's law, this termination condition means that when T is increased up to some threshold, we will not be able to sense C classes from the data. Finally, this algorithm picks the clustering results (CC,T_1) and (CC,T_2) such that the width of the window interval (T_1,T_2) is maximum while the clustering results within the interval are most satisfactory. The corresponding optimal clustering result is considered as the output of the clustering algorithm.

The proposed algorithm can be considered as a new class of nonparametric clustering algorithms because it originates from visual system theories. As compared with the information theoretic clustering recently proposed by (Romeny *et al.*, 1993), the parameter tuning process of the new algorithm is more systematic and theoretically sound. One might have to carry out a lengthy trial and error process to determine an appropriate variance parameter σ for the information theoretic clustering algorithm.

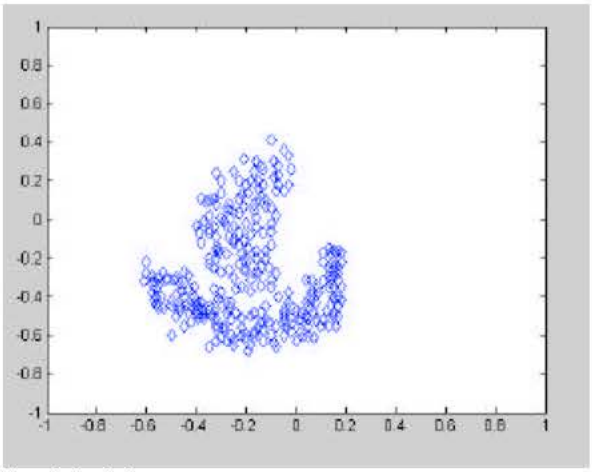
Next, let us analyze the time complexity of this new clustering algorithm. According to Fig. 1, the algorithm will stop after limited cycles because it will continue only when there is an increment in the value of the cost criterion $\hat{D}$ and meanwhile C clusters of data can be sensed. Suppose after t iterations of the inner repeat loop, the algorithm manages to obtain the maximum $\hat{D}$ value for a certain T. Thus, the time complexity of the inner cycle is $O(C\times N\times t)$. For the many experiments being carried out, t generally takes no more than 8. Accordingly, if we assume that the maximal sampling frequency interval in all clustering results is (T_1,T_2), the time complexity of the whole clustering algorithm is $O(C\times N\times t\times(T_1-T_2)/k)$, where k is the Weber fraction.

SIMULATIONS

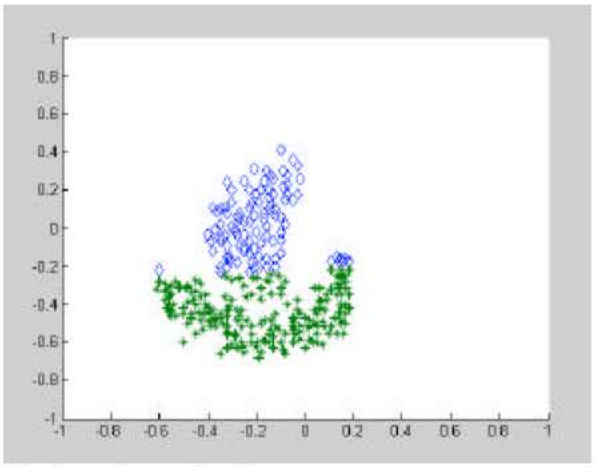
Testing on benchmarking data: We demonstrate the effectiveness of the new clustering algorithm in a simple dataset for benchmarking. The dataset which is obviously nonlinearly separable Fig. 2a, where the two clusters are represented by symbols "∗" and "◇" respectively. Fig. 2b gives the clustering result of the famous fuzzy c-means (FCM) clustering algorithm, which is obviously unfit for this dataset. The reason for such a poor clustering result is that FCM is inherently only suitable for convex datasets. Based upon the Weber's law, we obtain the visual sampling frequency intervals as (Table 1 and Fig. 2c). We decide that the visual

Table 1: Parameter analysis of the new algorithm

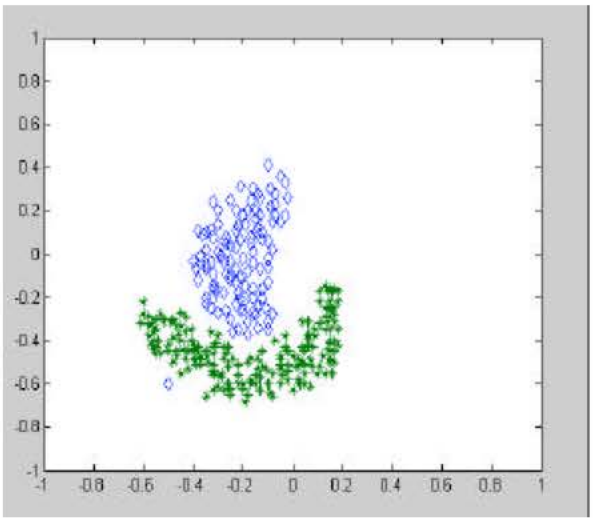
Observable visual Sampling frequency or window size (T)	No of iterations in the inner repeat loop t	Clustering result
< 0.058	≥7	Not ideal
0.058-0.080	7	Ideal
0.081-0.09	7	Almost ideal
> 0.09	≥7	Not ideal



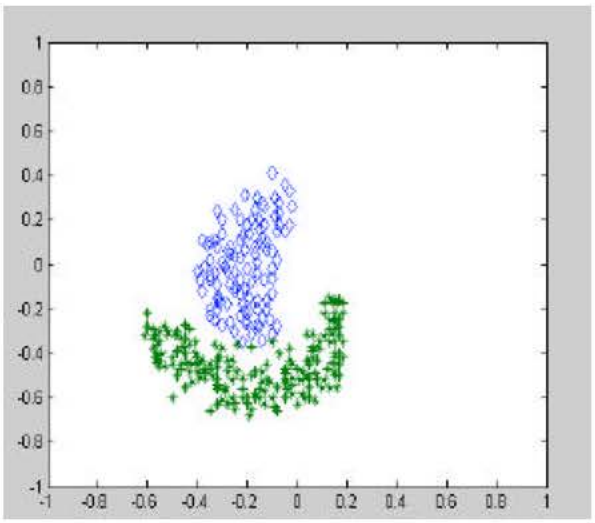
(a) Original dateset



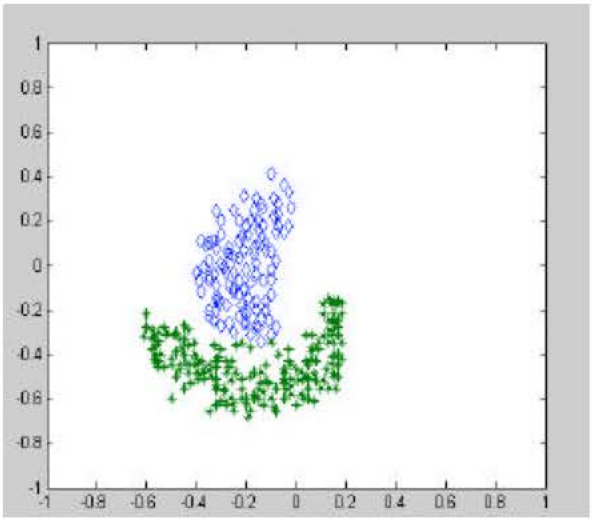
(b) Clustering results of FCM



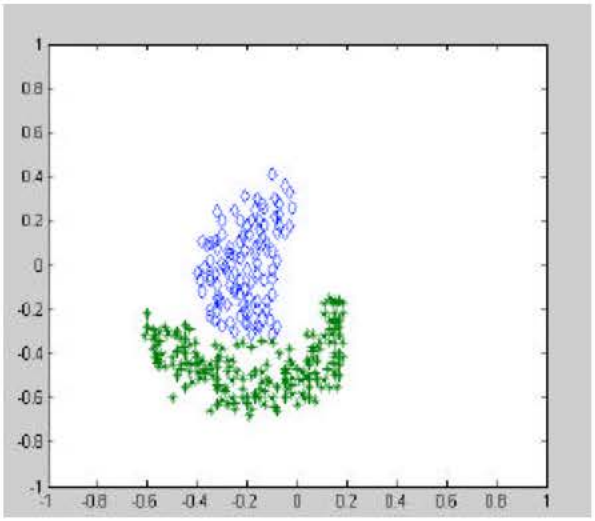
T < 0.058



T = 0.081



T = 0.082



T = 0.083-0.084

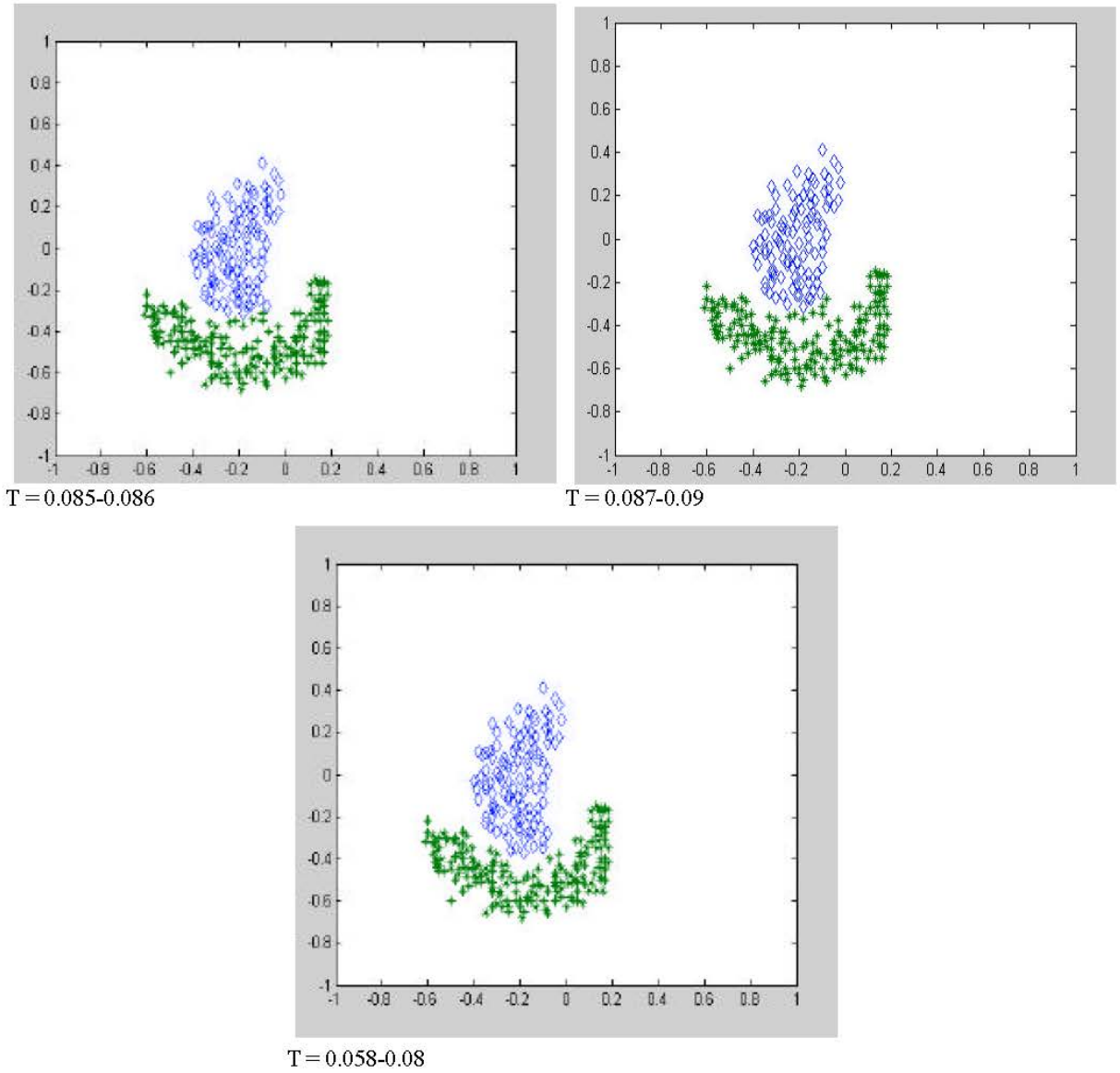
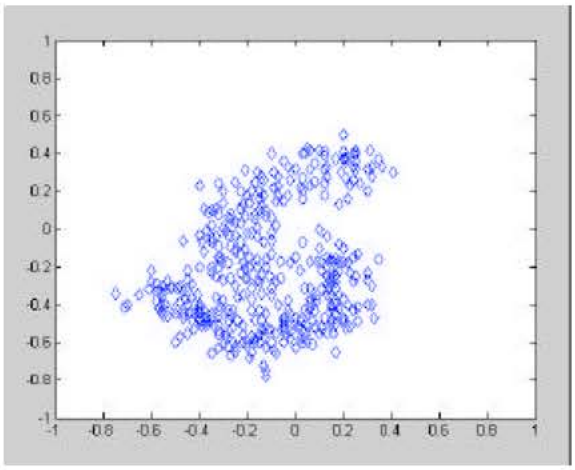


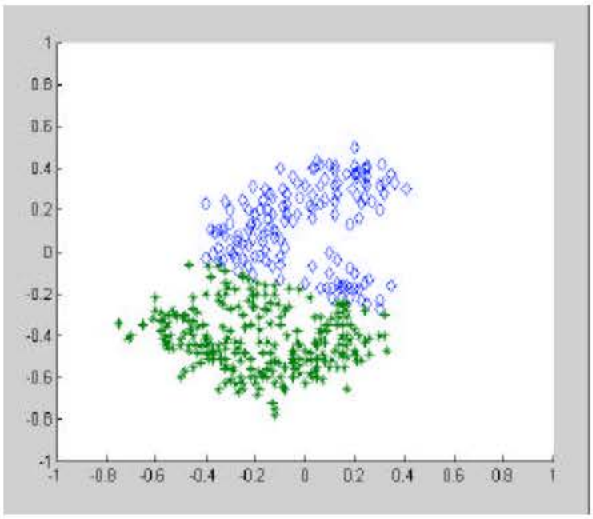
Fig. 2: Clustering results of the new clustering algorithm

optimal sampling frequency interval is (0.058, 0.080). With the sampling frequency shifting away from the optimal interval, the clustering results will become worse.

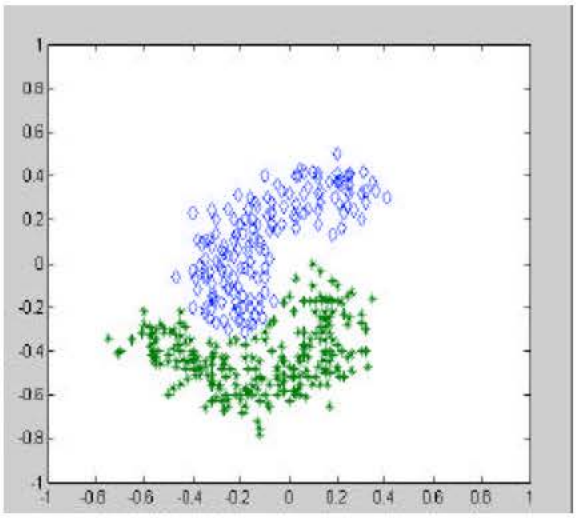
Testing on complex data: This dataset further demonstrates the clustering capability of this new algorithm. From a perception point of view, the dataset given in (Romeny et al., 1993) and shown in Fig. 3a, should be nonlinearly separated into two groups, although its boundary seems a little bit unclear. FCM classifies the dataset into two classes, however, its result is obviously unreasonable as shown in Fig. 3b. Figure 3c demonstrates that our algorithm works well for the dataset with the range of the visual sampling frequency (0.087,0.09). By comparing the results here with those reported in (Romeny *et al*., 1993), we can easily see that the proposed algorithm is at least comparable with the information theoretic clustering algorithm. However, a lengthy trial and error process to determine the variance parameter can be avoided here because we can tune the visual sampling frequency parameter based on the Weber's law. A pretty rational explanation for the capability of this new clustering algorithm for nonlinearly



(a) Original dataset



(b) Clustering results of fuzzy c-means



(c) Clustering results of the new algorithm

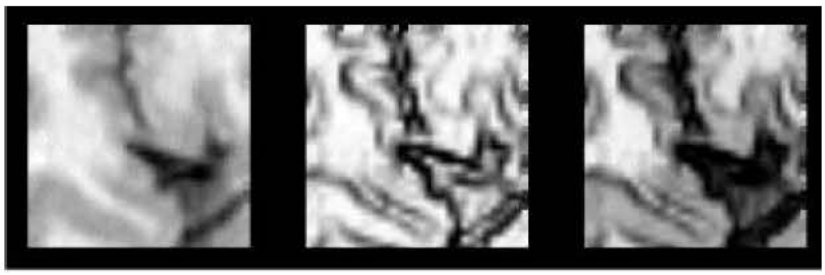
Fig. 3: Clustering performance of a complex dataset

separable datasets is that the high-dimensional mapping of Sinc functions in the new cost criterion makes the new criterion effective for clustering nonlinearly separable datasets.

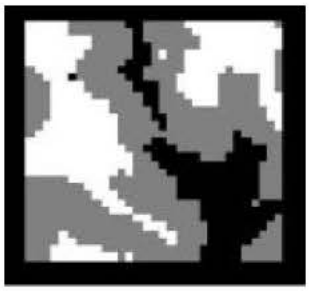
Testing on MRI Data

MRI image: In order to further compare the performance of our algorithm with that of the information theoretic clustering, we have picked the same brain MRI image from (Romeny *et al*., 1993), which is a three-class (white matter, gray matter and cerebro-spinal fluid) identification of brain tissue from Magnetic Resonance Imaging (MRI). To date, this is still a challenging problem. Figure 4a shows the original test image, the corresponding information potential image and its feature image used in (Romeny *et al*., 1993). Figure 4b shows the pixel assignments for this small image block using the information theoretic clustering. Figure 4c reports our clustering result. After carefully checking of the test image and Fig. 4b and c, we can easily see that the result of our new algorithm has comparable performance with the information theoretic clustering approach. In fact, our algorithm is slightly better than that the information theoretic clustering in some locally small regions.

Complex images: In order to examine the power of our algorithm here, we run this algorithm on another three more complex images. One is another MRI image, as shown in Fig. 5a. Another two satellite images are shown in Fig. 6a and 7a. All these three images contain three-class information therein: white, gray and black. For each image, we demonstrate two experimental results. One is for the new clustering algorithm using the sinc kernel function, the other is for the new clustering algorithm



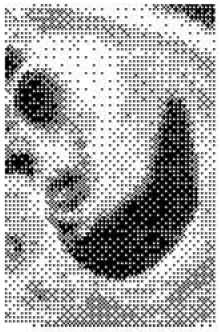
(a) Test image, information potential image, feature image



(b) Clustering result of information theorectic clustering

(c) Clustering results of the new algorithm

Fig. 4: Clustering performance of MRI data

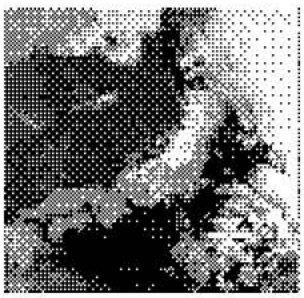


MRI-3 (a)

Gaussian (b)

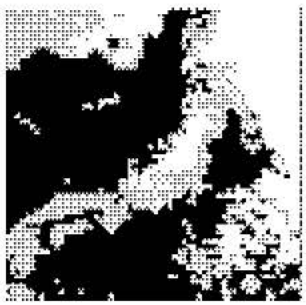
Sinc (c)

Fig. 5: The clustering results for MRI-3



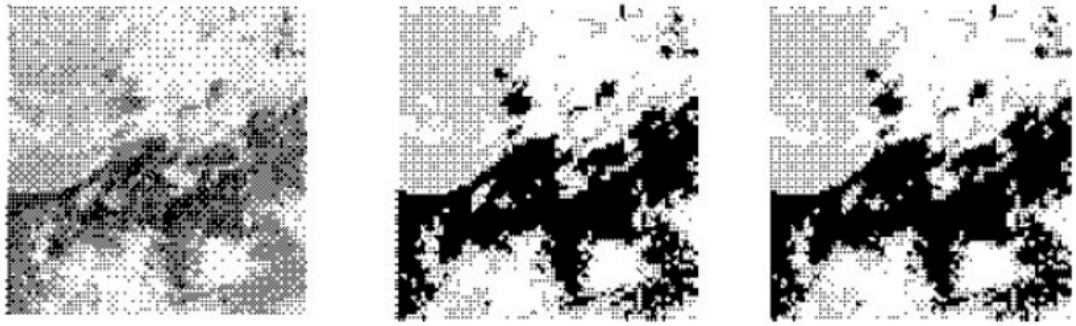
Satellite imaget (a)

Gaussian (b)



Sinc (c)

Fig. 6: The clustering results for satellite image 1



Satellite image 2 (a) Gaussian (b) Sinc (c)

Fig. 7: The clustering results for satelite image 2

using the Gaussian kernel function. These experimental results are shown in Fig. 5(b,c), 6(b,c) and Fig. 7(b,c), respectively. Figure 5-7 shows that the above new clustering algorithm exhibits very satisfactory clustering performance.

APPLICATION TO FUZZY MODELING

Tsk fuzzy system: We will discuss how to apply the above unified criterion to fuzzy modeling using the TSK fuzzy system. We first introduce the classical TSK fuzzy system in brief. The classical TSK fuzzy system takes the following forms of fuzzy rules:

$$R^i: \text{ If } x_1 \text{is } A_1^i, x_2 \text{is } A_2^i, \cdots, x_d \text{is } A_d^i$$
$$\text{then } y^i = P_0^i + P_1^i x_1 + \cdots + P_d^i x_d$$

where A_j^i is the fuzzy set, P_j^i is a real parameter, y^i is the output of the ith fuzzy rule, I = 1,2...,m, m is the number of fuzzy rules. That is to say, in the TSK fuzzy system, the If part of one fuzzy rule is fuzzy, while the then part is definite, i.e., the output of one fuzzy rule is the linear combination of corresponding inputs. For the k th input vector $x_k = (x_{k1}, x_{k2}, \ldots, x_{kd})$, the output of the TSK fuzzy system is equal to the weighted mean y^i:

$$y_k = \frac{\sum_{i=1}^{m} w^i y_k^i}{\sum_{i=1}^{m} w^i} \tag{24}$$

where w^i is defined as follows:

$$w^i = \prod_{j=1}^{d} \mu_{A_j^i}(x_j) \tag{25}$$

where d is the dimensional number of inputs and $\mu_{A_j^i}(x_j)$ is the membership of the fuzzy set A_j^i .

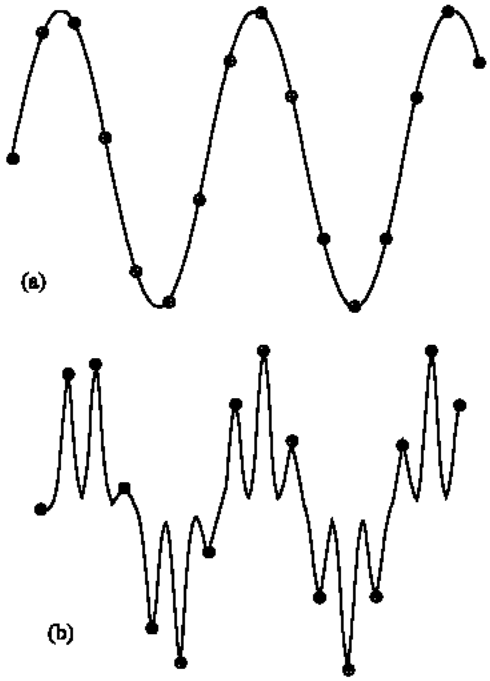


Fig. 8: Two predicted curves for the same dataset

Applying the unified criterion to fuzzy modeling: In general, we take MSE measure to tune parameters, i.e.

$$MSE = \frac{1}{2N} \sum_{ai=pi=1}^{N} (y_{ai} - y_{pi})^2 \tag{26}$$

where N is the total number of samples, y_{ai} is the predicted output of the TSK fuzzy system, while y_{pi} is the actual output from the original dataset. Since MSE is only the sum of all the square errors for each sample point, it does not consider the mutual relationship among samples as a whole, thus under some cases it may happen that even if MSE is very small, the generalization capability of the system is very poor, (Fig. 8). Figure 8 shows the two

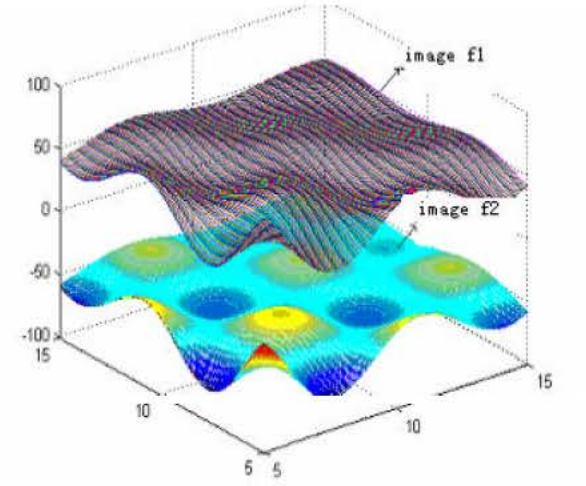


Fig. 9: Visual explanation of fuzzy modeling

predicted curves for the same sample dataset represented by red circles. Obviously, with the same MSE for all the sample data, the curve in Fig. 8a is much better than the curve in Fig. 8b, due to the fact that the curve in Fig. 8a well reflects the change characteristic of the sample data and that this change trend in the curve is well in line with human visual behavior for the sample data. Therefore, when training sample data, we should consider not only the data itself but also its visual characteristic.

Now, let us give a new explanation for the TSK fuzzy modeling in terms of visual systems and the Web law. Suppose the original dataset to be modeled forms an image f_1 and the outputs of the fuzzy system to be designed forms the other image f_2. Thus, fuzzy system modeling should attempt to make the image f_2 approximate to the image f_1, (Fig. 9). With this explanation, we can skillfully introduce the visual system into fuzzy modeling.

Suppose the original dataset $z_{pi} = (x_i, y_{pi})$, where xi $\in R^d$, $y_{pi} \in R$, and i = 1,2,...., N. Let

$$\begin{aligned} &\{z_{pi} \mid z_{pi} = (x_i, y_{pi}), i = 1,2,\cdots,N\} \quad z_{pi} \in R^s \quad s = d+1 \\ &\{z_{ai} \mid z_{ai} = (x_i, y_{ai}), i = 1,2,\cdots,N\} \quad z_{ai} \in R^s \quad s = d+1 \end{aligned} \tag{27}$$

where y_{pi} and y_{ai} are the actual output and the predicted output of the TSK fuzzy system for x_i, respectively. In terms of the visual sampling theorem, we can easily construct two corresponding images:

$$f_p(\bar{z},\sigma) = (2\pi\sigma^2)^{-\frac{d}{2}} \sum_{pi=1}^{N} G(\bar{z} - \bar{z}_{pi}, 2\sigma^2) \tag{28}$$

$$f_a(\bar{z},\sigma) = (2\pi\sigma^2)^{-\frac{d}{2}} \sum_{ai=1}^{N} G(\bar{z} - \bar{z}_{ai}, 2\sigma^2) \tag{29}$$

where

$$G(\bar{z} - \bar{z}_k, 2\sigma^2) = \exp\left(-\frac{\|\bar{z} - \bar{z}_k\|^2}{2\sigma^2}\right) \tag{30}$$

We use the following square error function E to measure the distance between the above two images:

$$\begin{aligned} E = \int (f_p(\bar{z},\sigma) - f_a(\bar{z},\sigma))^2 dx = \int [f_p^2(\bar{z},\sigma) \\ + f_a^2(\bar{z},\sigma) - 2f_p(\bar{z},\sigma) f_a(\bar{z},\sigma)] dx \end{aligned} \tag{31}$$

Substituting (35) and (36) into (38), we have

$$\begin{aligned} E = (2\pi\sigma^2)^{-d} [\sum_{pi=1}^{N}\sum_{pj=1}^{N} G(\bar{z}_{pi} - \bar{z}_{pj}, 2\sigma^2) + \\ \sum_{ai=1}^{N}\sum_{aj=1}^{N} G(\bar{z}_{ai} - \bar{z}_{aj}, 2\sigma^2) - 2\sum_{pi=1}^{N}\sum_{ai=1}^{N} G(\bar{z}_{pi} - \bar{z}_{ai}, 2\sigma^2)] \end{aligned} \tag{32}$$

In fuzzy modeling, we will use E as the error function instead of MSE. Unlike MSE, E embodies ont only the sample data itself but also its visual characteristic. Our experimental results will demonstrate that fuzzy modeling with E has very good approximation performance. What is more important, our work here reveals the successful application of visual systems in fuzzy modeling.

TSK Fuzzy modeling algorithm: Here, we derive the TSK fuzzy modeling algorithm as follows:

1. Intialize (m, ξ, σ)

/* when m is the rule number, ξ is the minimal error and σ is the size of the observable window.*/

2. Cluster the input sample dataset using clusterith FCM and then obtain the clustering centers

$c^i = (c^i_1, c^i_2, \cdots, c^i d), i = 1,2,\cdots,m$

/* actually , all $_iC$ determine the intial partition for the sample space*/

3. Construct the initial TSK fuzzy system according to the above clustering results:

R^i: If $x_1$is $A^i_1, x_2$is $A^i_2, \cdots, x_d$is A^i_d

then $y^i = P^i_0 + P^i_1 x_1 + \cdots + P^i_d x_d$

$$A^i_j(x_j) = \exp\left(-\frac{(x_j - c^i_j)^2}{\delta^i_j}\right) \tag{33}$$

/* where $A^i_j$is the fuzzy set, δ^i_j is the widh of the membership function,

P^i_j is the parameter, y^i is the output from the ith rule, i = 1,2,...,m */

4. Tune parameters in the fuzzy system, using the following update rule:

$$(1)\quad P^i_j(k+1) = P^i_j(k) - \gamma_1 \cdot \frac{\partial E}{\partial P^i_j} \tag{34}$$

$$\frac{\partial E}{\partial P_j^i} = -(2\pi\sigma^2)^{-s}\frac{1}{\sigma^2}[\sum_{ai=1}^{N}\sum_{aj=1}^{N} G(\bar{z}_{ai}-\bar{z}_{aj},2\sigma^2)(y_{ai}-y_{aj})$$
$$(\frac{\partial y_{ai}}{\partial P_j^i}-\frac{\partial y_{aj}}{\partial P_j^i})+2\sum_{pi=1}^{N}\sum_{ai=1}^{N} G(x_{pi}-x_{ai},2\sigma^2)(y_{pi}-y_{ai})\frac{\partial y_{ai}}{\partial P_j^i}] \quad (35)$$

$$\frac{\partial y_{ai}}{\partial P_j^i} = \frac{w^i}{\sum_{k=1}^{m} w^k}(x_{aj})_j \quad (36)$$

$$\frac{\partial y_{aj}}{\partial P_j^i} = \frac{w^i}{\sum_{k=1}^{m} w^k}(x_{aj})_j \quad (37)$$

$$(2)\ c_j^i(k+1) = c_j^i(k) - \gamma_2 \cdot \frac{\partial E}{\partial c_j^i}$$

$$\frac{\partial E}{\partial c_j^i} = -(2\pi\sigma^2)^{-s}\frac{1}{\sigma^2}[\sum_{ai=1}^{N}\sum_{aj=1}^{N} G(\bar{z}_{ai}-\bar{z}_{aj},2\sigma^2)(y_{ai}-y_{aj})$$
$$(\frac{\partial y_{ai}}{\partial c_j^i}-\frac{\partial y_{aj}}{\partial c_j^i})+2\sum_{pi=1}^{N}\sum_{ai=1}^{N} G(x_{pi}-x_{ai},2\sigma^2)(y_{pi}-y_{ai})\frac{\partial y_{ai}}{\partial c_j^i}] \quad (38)$$

$$\frac{\partial y_{ai}}{\partial c_j^i} = \frac{y_{ai}^i\sum_{k=1}^{m} w^k - \sum_{k=1}^{m}(w^k y_{ai}^k)}{(\sum_{k=1}^{m} w^k)^2}\cdot\frac{2w^i[(x_{ai})_j - c_j^i]}{\delta_j^i} \quad (39)$$

$$\frac{\partial y_{aj}}{\partial c_j^i} = \frac{y_{aj}^i\sum_{k=1}^{m} w^k - \sum_{k=1}^{m}(w^k y_{aj}^k)}{(\sum_{k=1}^{m} w^k)^2}\cdot\frac{2w^i[(x_{aj})_j - c_j^i]}{\delta_j^i} \quad (40)$$

$$(3)\ \delta_j^i(k+1) = \delta_j^i(k) - \gamma_3 \cdot \frac{\partial E}{\partial \delta_j^i}$$

$$\frac{\partial E}{\partial \delta_j^i} = -(2\pi\sigma^2)^{-s}\frac{1}{\sigma^2}[\sum_{ai=1}^{N}\sum_{aj=1}^{N} G(\bar{z}_{ai}-\bar{z}_{aj},2\sigma^2)(y_{ai}-y_{aj})$$
$$(\frac{\partial y_{ai}}{\partial \delta_j^i}-\frac{\partial y_{aj}}{\partial \delta_j^i})+2\sum_{pi=1}^{N}\sum_{ai=1}^{N} G(x_{pi}-x_{ai},2\sigma^2)(y_{pi}-y_{ai})\frac{\partial y_{ai}}{\partial \delta_j^i}] \quad (41)$$

$$\frac{\partial y_{ai}}{\partial \delta_j^i} = \frac{y_{ai}^i\sum_{k=1}^{m} w^k - \sum_{k=1}^{m}(w^k y_{ai}^k)}{(\sum_{k=1}^{m} w^k)^2}\cdot\frac{[(x_{ai})_j - c_j^i]^2 w^i}{(\delta_j^i)^2} \quad (42)$$

$$\frac{\partial y_{aj}}{\partial \delta_j^i} = \frac{y_{aj}^i\sum_{k=1}^{m} w^k - \sum_{k=1}^{m}(w^k y_{aj}^k)}{(\sum_{k=1}^{m} w^k)^2}\cdot\frac{[(x_{aj})_j - c_j^i]^2 w^i}{(\delta_j^i)^2} \quad (43)$$

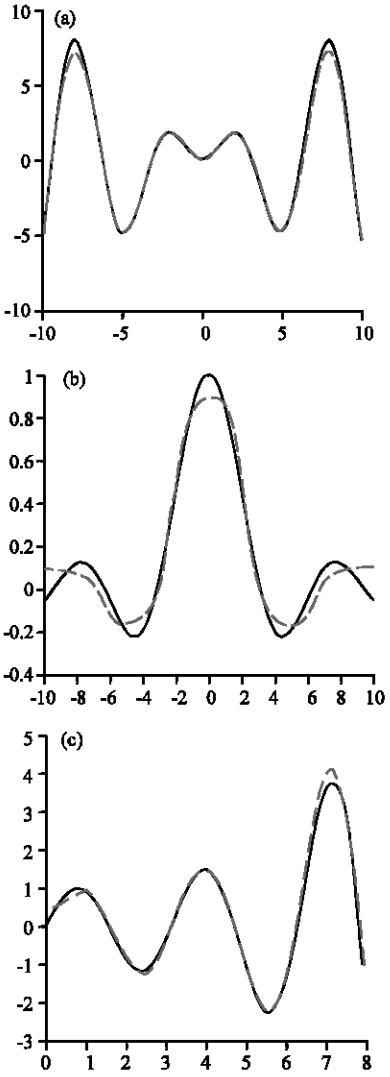


Fig. 10: 2-dimensional examples

/* I corresponds to the ith rule, j denotes the jth dimension, k denotes the iteration number,
and $\gamma_1,\gamma_2,\gamma_3$ are the learning rates*/
5. Compute E, if E<ζ (the given threshold), then terminate else go to step 2

Simulations: We will use several examples to demonstrate the good approximation performances of the above fuzzy modeling algorithm. All training samples are uniformly taken for these examples. Table 2 lists the experimental results for 7 examples. With an appropriate σ (we may determine it using the Weber law) and comparatively less iteration number and number of fuzzy rules, the above fuzzy modeling algorithm reaches very

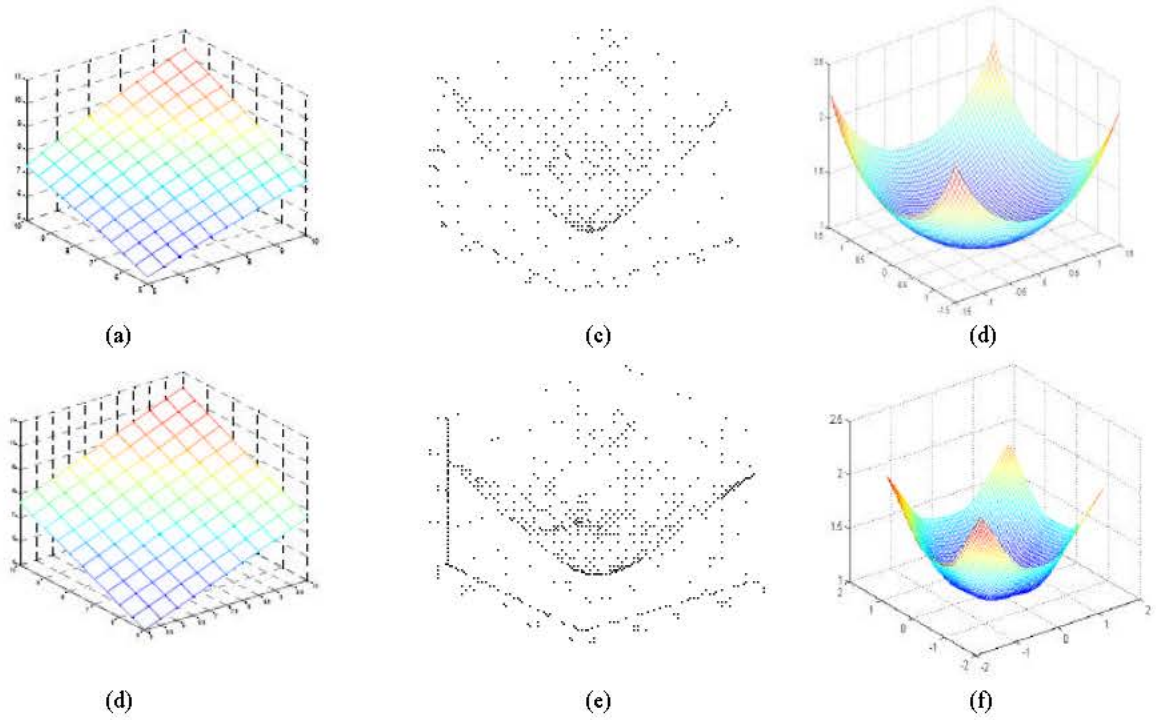


Fig.11: 3-dimensional examples

Table 2: Experimental results for 7 examples

Function f	No. of Training samples	Domain of x	No. of Rules	σ	Iteration number	E
(a) $f_1(x) = x \sin x$	101	$10 \le x \le 10$	20	1	100	0.043995
(b) $f_2(x) = \frac{\sin x}{x}$	67	$10 \le x \le 10$	5	0	200	0.0019682
(c) $f_3(x) = e^{(-0.162x)^2} \sin(2x)$	42	$0 \le x \le 8$	12	0.5	50	0.0019682
(d) $f_4 = \frac{1}{\ln(x_1 x_2)} + \sqrt{x_1 x_2}$	49	$5 \le x_1 \le 10$ $5 \le x_2 \le 10$	2	5	20	0.010591
(e) $f_5 = (x_1^2 + x_2^2)^{\frac{2}{3}}$	64	$-1 \le x_1 \le 1$ $-1 \le x_2 \le 1$	10	0.5	70	0.0008313
(f) $f_6 = \frac{x_1 x_2}{\sin(x_1)\sin(x_2)}$	81	$-1.5 \le x_1 \le 1.5$ $-1.5 \le x_2 \le 1.5$	9	0.5	100	0.00050483
(g) $f_7 = \frac{1}{\ln(x_1 x_2)} + \sqrt{x_1 x_2} + \frac{1}{(x_1 + x_3)^2}$	40	$5 \le x_1 \le 10$ $5 \le x_2 \le 10$ $5 \le x_3 \le 10$	9	2.5	100	0.029995

good approximation and generalization results. Figure 10 and 11 demonstrate that all the curves of the built TSK fuzzy systems are well in line with all the original functions, respectively, where the blue curves correspond to the original functions and the red curves correspond to the predicted curves of the fuzzy systems in Fig. 10 and the left figures correspond to the original functions and the right figures correspond to the predicted surfaces. Since it is very difficult to draw the high-dimensional function and its output curve of the fuzzy system, we use

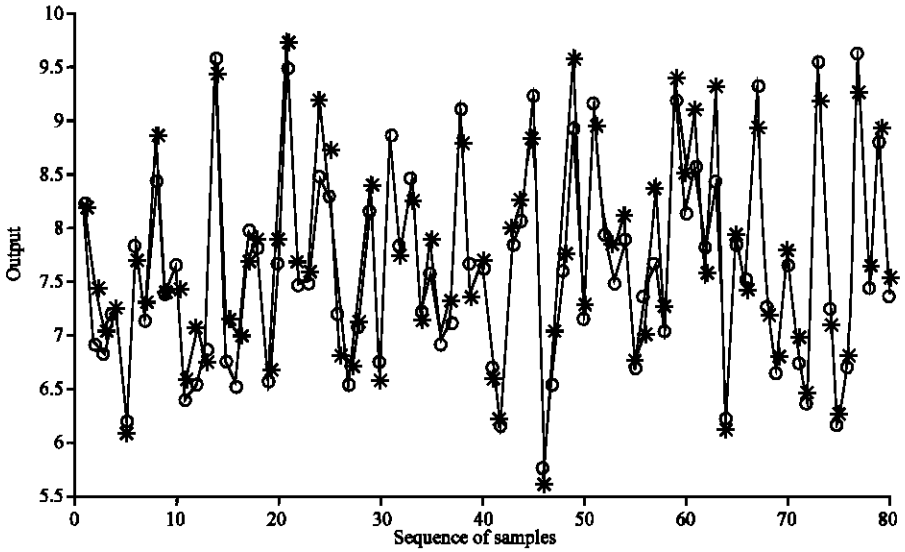


Fig. 12: One 4-dimensional example

Fig. 12 to show this case where all the red * points denote the actual outputs of samples and all the circles denote the predicted outputs. Fig. 12 also demonstrates its very good performance. Therefore, these experimental results effectively verify the successful application of the unified visual system theoretic criterion in fuzzy modeling.

CONCLUSIONS

In this study based on the visual system theories, a new cost criterion is defined and kernelized so that we can choose the appropriate kernel function for different application requirements. Our simulations demonstrate that the new clustering algorithm is especially suitable for nonlinearly separable datasets. Our study reveals that human visual behaviors can be well integrated with clustering, which offers a new direction for studying clustering and perhaps other data modeling tasks like fuzzy systems.

Many works of this study merit further investigations. First of all, running the current version of the new clustering algorithm is still time-consuming. For large datasets, we need to investigate how the computational complexity can be reduced. Like many other clustering algorithms, obtaining good performance requires knowing the number of clusters in advance and this is always a difficult task and is seldom predictable in most practical applications. On the other hand, the human visual system senses the objects and classifies them almost instantly. Hence, we are working on how to determine the number of clusters automatically and efficiently and will report the result hopefully in the near future.

ACKNOWLEDGMENTS

This research is supported by the Hong Kong PolyU CRG (grant No. G-T912) and Hong Kong RGC Competitive Earmarked Research Grant (grant No. 5065/98E), National Science Foundation of China (grant No. 60225015), Natural Science Foundation of JiangSu Province (grant No. BK2003017), National Key Lab. Of Pattern Recognition at Institute of Automation of CAS SINICA, The JiangSu Key Lab. Of Information Processing, National Key Lab. Of Computer Science at Institute of Software of CAS SINICA (Grant No. SYSKF0406), 2005 Key Project of Ministry of Education of China (MOE) and The 2004 Outstanding Teacher Grant of MOE.

REFERENCES

Anderberg, M.R., 1973. Cluster Analysis for Applications. New York: Academic Press, Inc.

Arabie, P. *et al.*, 1996. Clustering and Classification. River Edge, NJ: World Scientific Publishing.

Bezdek, J.C. *et al.*, (Eds). 1992. Fuzzy models for pattern recognition, IEEE Press, New York.

Coren, S. *et al.*, 1994. Sensation and Perception. 4th Edn. Fort Worth, TX: Cold Spring Harcourt Brace College Publishers.

Dubes, R. *et al*., 1976. Clustering techniques: The user's dilemma. Pattern Recognition, 8: 247-260.

Gonzalez, R.C. *et al*., 1992. Digital Image Processing. Addison-Wesley.

Jain, A.K. *et al*., 1998. Algorithms for clustering Data. Englewood Cliffs, New Jersey: Prentice-Hall.

Jifeng, W. *et al*., 1988. Image recognition by computers. Railway Publishing House of China pp: 74-75. (In Chinese)

Koenderint, J.J., 1984. The structure of images. Biological Cybernetics, 50: 363-370.

Marr, D., 1982. Vision, A Computational Investigation into the Human Representation. San Francisco: W.H. Freeman.

Mali, K. *et al*., 2002. Clustering of Symbolic and its Validation. In: Advances in Soft Computing---AFSS 2002 (Pal, N.R. Eds.), Springer, pp: 339-345.

Nyquist, H., 1928. Certain topics in telegraph transmission theory. AIEE Transactions, 47: 43-51.

Poggio, T., 1990. A Theory of How the Brain Might Work. In: The Brain. Harbor Laboratory Press, pp: 899-910.

Romeny, B.M.H. *et al*., 1993. A Multiscale Geometric Model of Human Vision. In: Hendee, W.R. Wells, P.N.T. (Eds) The Perception of Visual Information. New York: Springer-Verlag, pp: 73-114.

Romeny, T.H.B. *et al*., 1997. (Eds) Scale-Space Theory in Computer Vision.Berlin Heidelberg: Springer-Verlag.

Shitong, W. *et al*., 2006. Note on the relationship between probabilistic/fuzzy clustering. Soft Camputing (accepted).

Shitong, W. *et al*., A new integrated clustering algorithm GFC and switching regressions. Intl. J. Pattern Recognition and Artificial Intelligence, 16: 433-446.

Shitong, W., 1998. 2000. Fuzzy Systems, Fuzzy Neural Networks and their Programming. 2nd Edn., Shanghai Press of Science and Technology,

Shitong, W. *et al*., A new Gaussian noise filter based on interval type 2 fuzzy systems. Soft Computing (accepted)

Xuanli, L.X., *et al*., 1991. A validity Measure for fuzzy clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 13: 1107-1116.

Zheng, N.N., 1998. Computer Vision and Pattern Recognition. National Defense Industry Publishing House (In Chinese).