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 5 (5): 779-791, 2006
ISSN 1812-5638

Visual Sampling Based Clustering Algorithm VSC

[1,2,3]Wang Shitong, [2]F.L. Chung, [1,3]Guo Wei and [1,3]Han Bin
[1]School of Information, Southern Yangtse University, Wu Xi, Jiang Su, China
[2]Department Computing, Hong Kong Polytechnic University, Hong Kong, China
[3]National Key Laboratory of Computer Science, Institute of Software, CAS SINICA, China

Abstract: This study attempts to achieve two goals: (1) The novel visual sampling based clustering algorithm VSC is proposed, based on the visual sampling principle. The clustering algorithm VSC incorporates the visual sampling principle together with the famous Weber law such that it has two distinctive advantages: (a) it is insensitive to initial conditions and very effective for convex datasets; (b) the reasonable cluster number can be effectively determined by the new Weber-law-based clustering validity index. Our experimental results demonstrate its success. (2) The link relationship between our algorithm VSC and algorithm SCA. Both theoretical analysis and experimental results show that in many cases, our algorithm VSC here has almost the same clustering results as algorithm SCA. This fact reveals that our algorithm can be utilized to overcome the drawback of SCA, i.e., the parameter γ therein is very difficult to be well determined.

Key words: Clustering, visual sampling, weber law, clustering validity index, attractors, fixed points

MOTIVATIONS

Clustering analysis plays an important role in many engineering fields such as pattern recognition, system modeling, image processing, data mining and so on. Clustering algorithms try to partition a set of unlabeled input data into a number of subsets (clusters) such that data in the same cluster are more similar to each other than to data in the other clusters. Clustering algorithms can also be divided into two types: hard vs fuzzy/probabilistic. Hard clustering assigns each datum to exactly one cluster. In fuzzy/probabilistic clustering, a given pattern does not necessarily belong to only one cluster but can have varying degrees of memberships/probability to several clusters. Many clustering algorithms could be found in the related books (Bezdek *et al*., 1992; Nadler *et al*., 1993).

There are two basic approaches in hard/fuzzy clustering algorithms, which we call parametric and nonparametric. In parametric clustering we assume a predefined distribution for the data set and calculate the sufficient statistics or fuzziness which will describe the data set in a compact way. For example, in probabilistic clustering, for a normal distribution, the sufficient statistics are the sample mean and the sample covariance matrix, which will describe the 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 a great success in many fields, their performances are often sensitive to the initial conditions and estimating the parameters is not a trivial work

The present nonparametric approach to clustering divides the data set into groups of points which have strong internal similarities. In order to measure the similarities we use a criterion function and seek the grouping that maximizes (or minimizes) the criterion. This kind of algorithms requires a cost function to evaluate how well the clustering fits to the data and an algorithm to minimize the cost function. For example, valley seeking clustering (Fugunaga, 1990) and information theoretic clustering (Gokcay *et al*., 2002) belong to the approach.

In this study, we will introduce the visual sampling theorem and the Weber law into clustering analysis in the first time. As we all know, eyes can inherently effectively recognize/classify objects under complex environments. Thus an efficient clustering algorithm should depend not only on the principle of physical system by which the data are generated but also on the manner in which human eyes sense the structure of the data. In this study, with the help of the visual sampling theorem and the Weber law, we will present a new visual sampling based clustering algorithm VSC, which is the first attempt to reach this goal, to best of the author's knowledge. This new clustering algorithm is a nonparametric clustering one. The contributions of our work here are two fold:

Corresponding Author: Wang Shitong, School of Information, Southern Yangtse University, WuXi, JiangSu, China
Tel: 865108591251

- The proposed clustering algorithm is rooted on the new principle and it has strong clustering capability for convex datasets and its robustness for initial conditions. The Weber law helps us to build a new clustering validity index CVI, which can effectively determine the reasonable number of the cluster centers in the given dataset.
- Recently, the new similarity based robust clustering algorithm SCA is proposed by Yang *et al.* (2004). Algorithm SCA's advantage exists in its very good robustness. However, one of its distinctive drawbacks is the difficulty in reasonably choosing parameter γ. (Yang *et al.*, 2004), γ may be 1, 5, 10, 15..... When γ takes a big value, the corresponding computational complexity will greatly increase. The other drawback is that this algorithm still does not escape from the classical framework of all current clustering algorithms, i.e, the number of the cluster centers must be predefined. In this study, we will theoretically and experimentally show the link relationship between algorithm VSC and SCA, that is, in many cases, our algorithm VSC here has almost the same clustering results as algorithm SCA. Due to its nonparametric property, our algorithm VSC has the distinctive advantage over algorithm SCA, i.e., avoiding the difficulty in determining the reasonable parameter γ in algorithm SCA.

THE VISUAL SAMPLING PRINCIPLE, THE WEBER LAW AND CLUSTERING

As the statistic data states, more than 80% information obtained by people comes from the eyes, so eyes are the very organ that people study in the most depth and know the best up to now. Human eyes have inherent and strong clustering capability. With the development of physiology and psychology, people have gained more and more knowledge about how the eyes function, which makes it become possible for us to imitate the clustering mechanism of eyes such that the shortcomings of the conventional clustering algorithms are avoided. At the beginning of 1980s, Marr (1982) proposed the concept of visual computational theory and during the past 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. Visual systems have found wide applications in computer science, especially in the field of image processing. The visual sampling theorem and the Weber law are two milestones of visual system research. The visual sampling theorem originates from the classical sampling principle, which plays a very important role in signal processing. This principle can be stated in the following formula:

$$s(t) = \sum_{-\infty}^{\infty} s(kT) \frac{\sin \pi(\frac{t}{T} - k)}{\pi(\frac{t}{T} - k)} \quad (1)$$

Where s(kT) denotes the sample of signal s(t) at time kT. Since it was proposed in the late 1940s (Shitong *et al.*, 2002), its several variants have been obtained.

All signals are 1-dimensional in signal processing. However, data are often multi-dimensional in visual systems, so, we must extend the classical sampling principle to multi-dimensional cases. This is called the visual sampling principle (Zheng, 1998). That is to say, given a d dimensional dataset $X = \{\vec{x}_i \in R^d : (i = 1, 2, ..., n)\}$, if each datum of the dataset is considered as a sampling point, then we will obtain a sampling image of the dataset based on the visual sampling theory, i.e.,

$$f(\vec{x}, T) = \sum_{i=1}^{N} \prod_{j=1}^{d} \frac{\sin(\frac{x_j - x_{ij}}{T})}{(\frac{x_j - x_{ij}}{T})} \quad (2)$$

where x_{ij} denotes the jth component of the ith sample and T is the sampling frequency.

General speaking, a sampling image can also be expressed using other kernel functions like the following:

$$f(x, T) = \sum_{i=1}^{n} (\prod_{j=1}^{d} \frac{\sin(\frac{x_i - x_{ij}}{T})}{\frac{x_i - x_{ij}}{T}})^{\gamma} \quad (3)$$

$$f(x, T) = \sum_{i=1}^{n} \exp(-\frac{\| x - x_i \|^2}{T}) \quad (4)$$

$$f(x, T) = \sum_{i=1}^{n} (\exp(-\frac{\| x - x_i \|^2}{T}))^{\gamma} \quad (5)$$

where, γ is a constant. Or, a sampling image can be expressed using the following generalized visual sampling function:

$$f(x, T) = \sum_{i=1}^{n} g(x, x_i, T) \quad (6)$$

where, g (x, x_i, T) denotes some kernel function.

For a given d dimensional dataset, $X = \{\vec{x}_i \in R^d : (i = 1, 2, ..., n)\}$, we may rationally assume the clustering process for the dataset to be a sampling process. Thus, we can establish its sampling image using (6). Furthermore, according to the scale space theory (Romeny *et al*., 1997; Zheng, 1998), we may even utilize multi-frequency sampling processes in our new visual-sampling clustering algorithm to mimic the non-uniform sampling behavior of visual systems.

The Weber law (Coren *et al*., 1994) quantitatively shows that a fixed-proportion increase in stimulus intensity s is sufficient to produce a just noticeable change in sensation. That is to say, the change δs of stimulus intensity should follow: $\delta s = k \times s$, where k is a Weber fraction. According to the Goddess neural theory (Poggio, 1990), the reason that we can sense such a noticeable change is that all neurons within the reception field are $s + \delta s$ activated.

If we view a clustering process as a visual sampling one, the Weber law tells us that we can only change the sampling frequency in a fixed proportion and each sampling image has a life cycle, i.e., we only sense this sampling image in some sampling frequency interval (T_1, T_2). This important feature will be used to define the new cluster validity indexes in this study.

NEW VISUAL-SAMPLING BASED CLUSTERING ALGORITHM

On the new algorithm VSC: Given a d dimensional dataset $X = \{\vec{x}_i \in R^d : (i = 1,2,..., n)\}$, in terms of (6), we can obtain the corresponding sampling image. This sampling image consists of several sub-images and each sub-image actually denotes the scope of its corresponding attractor, i.e., the maximum value of $f(\vec{x}, T)$ at some point. The scope $S(\vec{x}^*)$ of the sub-image corresponding to attractor $\vec{x}^*$ may be determined by solving the following gradient system:

$$\frac{d\vec{x}}{dt} = \nabla_{\vec{x}} f(\vec{x}, T) \tag{7}$$

and then

$$S(\vec{x}^*) = \{\vec{x}_0 \in R^d : \lim_{t\to\infty} \vec{x}(t, \vec{x}_0) = \vec{x}^*\} \tag{8}$$

where $\vec{x}(t, \vec{x}_0)$ is the solution of the following gradient system:

$$\begin{cases} \dfrac{d\vec{x}}{dt} = \nabla_{\vec{x}} f(\vec{x}, T) \\ \vec{x}(0, \vec{x}_0) = \vec{x}_0 \end{cases} \tag{9}$$

So by solving (9), for any datum, $\vec{x}_0$ once T is given, we can easily determine which sub-image the datum $\vec{x}_0$ belongs to, thus, the dataset can be effectively clustered in this way. In other words, if the dataset is sampled with the sampling frequency T, its clustering result may be determined by finding out the attractors in (9) and the number of clusters may also be automatically determined, i.e., the number of clusters is equal to that of attractors in (9).

Especially, in discrete case, (7) can be represented as

$$\vec{x}(t+1) = \vec{x}(t) + \beta \nabla_{\vec{x}} f(\vec{x}(t), T) \tag{10}$$

$$\nabla_x f(x(t), T) = \frac{\partial \sum_{i=1}^{n} g(x, x_i, T)}{\partial x} = \sum_{i=1}^{n} \frac{\partial g(x, x_i, T)}{\partial x} \tag{11}$$

where t denotes the iteration number and β denotes the speed rate, which often takes some small positive value. Given an initial condition, in order to obtain the stable solution of (7), we can run (10) repeatedly until $|\vec{x}(t+1) - \vec{x}(t)| < \varepsilon$, where is ε some given positive real number. Thus, in discrete case, the dataset may also be effectively clustered by running (10) repeatedly.

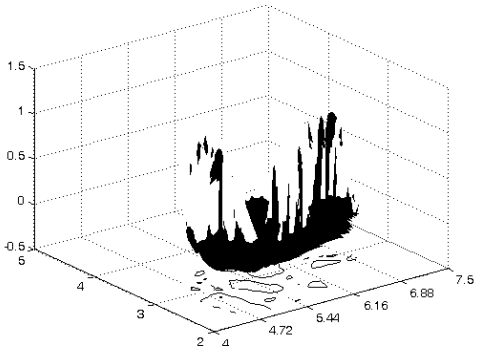
Obviously, a cluster center corresponds to the attractor of its sub-image. Since an attractor corresponds to some maximum value of $f(\vec{x}, T)$ thus, given T, all cluster centers of the dataset can be obtained by solving the following differential equation:

$$\frac{\partial f(\vec{x}, T)}{\partial \vec{x}} = 0$$

Figure 1a and b show the sampling images of the dataset consisting of the first 100 2-dimensional records in the IRIS dataset with two different sampling frequencies, respectively. Figure 1a is the sampling image under a lower frequency while Fig. 1b is that with a higher sampling frequency. Obviously, Fig. 1b is more smooth than Fig. 1a, that is to say, there are more attractors (maximum) with the lower sampling frequency than those with the higher sampling frequency but the scope of every attractor under the higher sampling frequency is wider.

When the sampling frequency is low enough, then the obtained sampling image can not reflect the structure of the dataset completely such that there are n clusters in the dataset, where n is the number of data in the dataset. That is to say, every datum in the dataset represents one cluster. When the sampling frequency is high enough, only one cluster exists! That is to say, the whole dataset

(a)



(b)

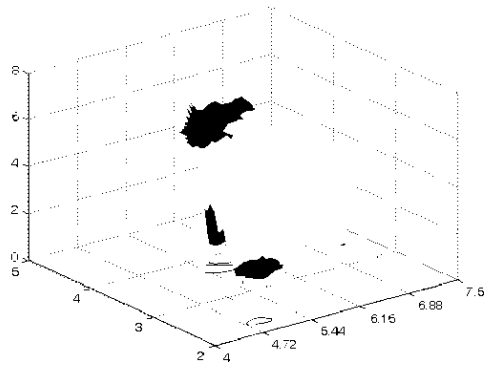


Fig. 1: Sampling images obtained under two different sampling frequencies

falls into one cluster. With the increase of the sampling frequency, we will obtain a tree of the sampling images under different sampling frequencies, then we may select the most stable sampling images from the tree to reflect the structure (clusters) of the dataset.

In terms of the Weber law, there is a life cycle for every structure. That is to say, while clustering, for any structure of the dataset, we can *only* sense it in some sampling frequency interval (T_1, T_2). If the sampling frequency is less than T_1, this structure does not exist and even decomposes several smaller structures. If the sampling frequency is more than T_2, the number of attractors may decrease, but the scope of each attractor gets bigger, so this structure may be merged together with other structures into some bigger structure. We call (T_1, T_2) the life cycle of this structure.

The new visual sampling based clustering algorithm VSC can be summarized as follows.

Visual sampling based clustering algorithm VSC

- For the given dataset, let t = 1. Initialize the sampling frequency $T^{(t)}$, where the initial $T^{(t)}$ T is low enough such that every datum in the dataset is an attractor of the sampling image, that is to say, each data point is an independent cluster.
- Construct the sampling image of the dataset using (2) ((6) in general). Compute all the attractors by solving the differential function $\frac{\partial f(\bar{x},T^{(t)})}{\partial \bar{x}}=0$ and then obtain the scope of each sub-image (i.e, each attractor) using (8) and (9).
- Find out all the attractors p under the sampling frequency $T^{(t)}$ such that the scope of such a p completely covers the scopes of at least two attractors under the sampling frequency $T^{(t-1)}$. If such an attractor p exists, then merge all the data points within the scopes of the corresponding attractors under the sampling frequency $T^{(t-1)}$ into a bigger cluster, which falls into the scope of the attractor p under the sampling frequency $T^{(t)}$.
- If the sampling frequency is high enough such that only one attractor exists, i.e., all the data points of the dataset fall into one cluster, then stop,
  else $T^{(t-1)} = T^{(t)} + k \times T^{(t)}$, t = t + 1, go to step 2, where k is the Weber fraction.

According to the Weber law in visual systems, for a sampling frequency T, only when the change δT of the sampling frequency T satisfies $\delta T \geq kT$ can the visual systems sense the change of such a scale, i.e, the sampling frequency T, where k is the Weber fraction and k = 0.03 for 1-dimensional datasets (Coren *et al.*, 1994). Our many experiments show that for most multi-dimensional datasets, $k \in [0.03,0.07]$ is appropriate. In this study, we take k = 0.05 for all the experimental studies.

On the new clustering validity index CVI: Obviously, for the given dataset, the new clustering algorithm VSC can find out all possible clusters at all possible levels of sampling frequencies which change according to the Weber law. In other words, for the given dataset, all these clusters can form a sampling image tree according to all possible levels of sampling frequencies. Accordingly, two critical problems appear. One is that what is the appropriate definition of the clustering validity index whose role is to decide upon the reasonable number of clusters in the dataset. The other is that how can we select the appropriate clusters from all possible clusters which the above new algorithm VSC produces. Fortunately, the Weber law can help us answer the first problem. In terms of the Weber law, there is a life cycle (T_1, T_2) for every sampling image. This means that within the life cycle (T_1, T_2), all the clusters occurring in the sampling image keep unchangeable and beyond this life cycle, the corresponding clusters in the sampling image will probably be merged or decomposed. Therefore, the

Weber law implies a new rational clustering validity index, called CVI here, i.e, if the life cycle of a sampling image is, (T_1, T_2) then its corresponding clustering validity index CVI is defined as

$$\ln T_2 - \ln T_1 \tag{12}$$

In general, CVI may be determined by users or experts. Once CVI, as the threshold, is given, after the new visual clustering algorithm terminates, we may obtain the final clustering results from the sampling images which have the largest CVI or whose CVIs are greater than the given threshold, thus, the second problem is accordingly settled down.

Comparison with related works: Compared with the conventional parametric clustering algorithms such as FCM, the distinctive advantage of our algorithm VSC is its insensitivity to initial conditions. Compared with the current nonparametric clustering algorithms such as information theoretic clustering (Gokcay *et al*., 2002), there is no any criterion function existing in the new clustering algorithm, thus the time-consuming computation existing in information theoretic clustering may be avoided.

Pedrycz *et al*. (2002) proposed another new clustering algorithm called granular clustering, which organizes finding about data in the form of a collection of information granules-hypeboxes. However, our clustering algorithm VSC should also belong to another new type of granular clustering, since it finds out the granular signature of a dataset based on the visual sampling principle and the Weber law. This is the big discrepancy between two algorithms. It seems to be very difficult that which is better while clustering, however, the new clustering algorithm VSC here seems to mimic the clustering behavior of human beings better.

There are several definitions (Arabie *et al*., 1996; Bezdak *et al*., 1992; Nadler *et al*., 1993; Pedrycz *et al*., 2002; Mali *et al*., 2002) on the clustering validity index. However, because CVI originates from the new visual theoretic viewpoint, so, it is completely different from them. CVI deeply reflects the expansion of the information granules.

EXPERIMENTAL RESULTS

Several experiments are done to demonstrate algorithm VSC's strong clustering capability for convex datasets and the experimental results are appealing.

Example 1: The artificial dataset is shown in Fig. 2. In this experiment, we take (5) as the visual sampling function, thus, (11) becomes

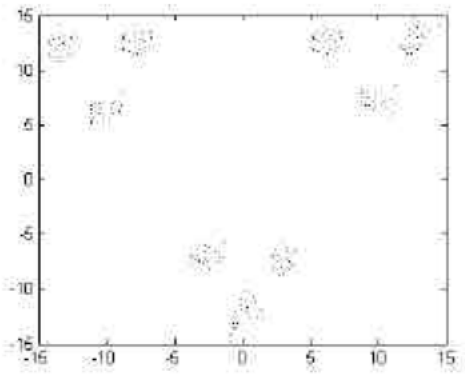


Fig. 2: The artificial dataset

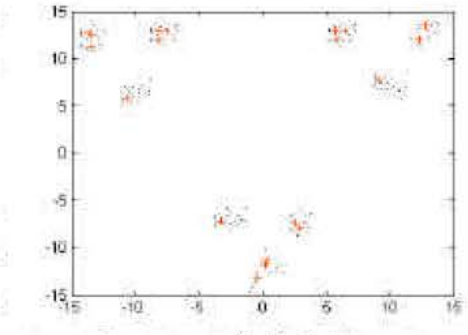
$$\nabla_x f(x(t),T) = \sum_{i=1}^{n} \frac{2\gamma}{T}(x_i - x)\exp(-\frac{\gamma \| x - x_i \|^2}{T}) \tag{13}$$

Fig. 3-5 demonstrate the clustering results of algorithm VSC with various sampling frequencies for three cases ($\gamma = 1, 2, 5$), respectively. In these figures, samples points are depicted using · and attractors (i.e., cluster centers) are depicted using +. Fig. 3a demonstrates 19 cluster centers with T = 0.5, Fig. 3b demonstrates 9 cluster centers with T = 2. The same clustering result keeps in Figure 3c when T = 8. When T = 30, the number c of the cluster centers decreases into 3, the corresponding clustering result Fig. 3d. Therefore, Fig. 3a-d clearly reflect the changes of the clustering results with the increase of the sampling frequency T. When $\gamma = 2, 5$, the similar changes can be seen in Fig. 4 and 5.

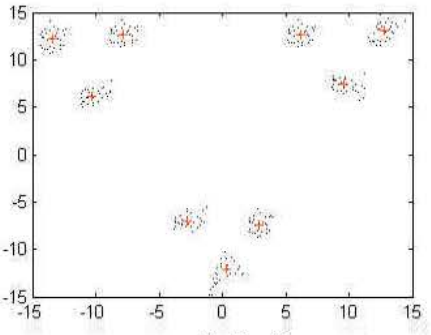
Figure 6a and b clearly indicates the change trend of numbers of cluster centers with the increase of the sampling frequency. When $1 < T \leq 10$, algorithm VSC is stable with 9 cluster centers; and when $16 \leq T \leq 150$, algorithm VSC is stable with 3 cluster centers. However, when T>160, the number of the obtained cluster centers will become 1.

The strong clustering capability and applicability of algorithm VSC may also be observed from the following 3 experiments, where we adopt (2) (i.e., function Sinc rather than function exp) as the visual sampling function in algorithm VSC.

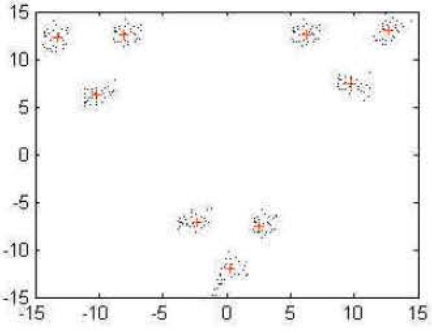
Example 2: This experiment is done using the real dataset. For simplicity and visualization, we select the first 2-dimensional 100 records in the IRIS dataset as the test dataset. And this dataset is a linearly separable dataset, (Fig. 7). In terms of the definition of CVI, algorithm VSC reaches its maximum, i.e, 0.537, when the number c of cluster centers is 2. Figure 8 demonstrates the corresponding clustering result.



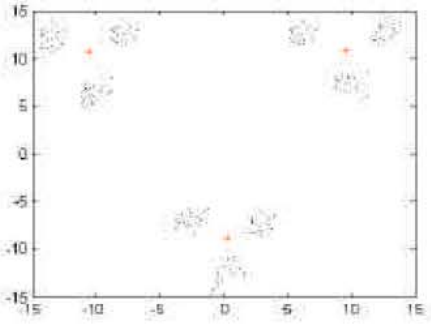
(a) T = 0.5



(b) T = 2

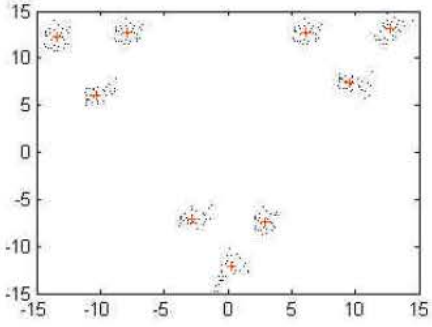


(c) T = 8

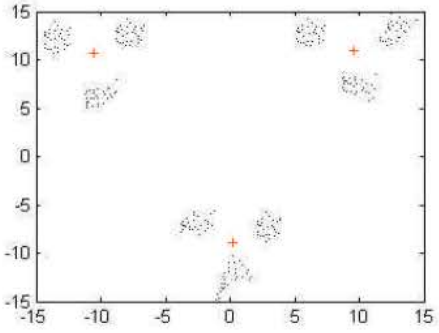


(d) T = 30

Fig. 3: The clustering results under different T (γ = 1)

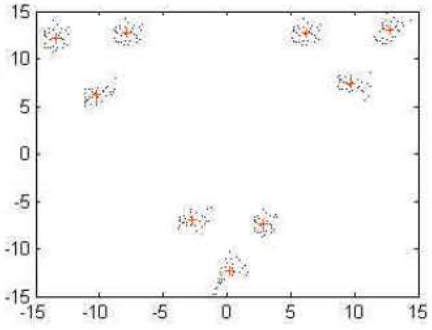


(a) T = 3

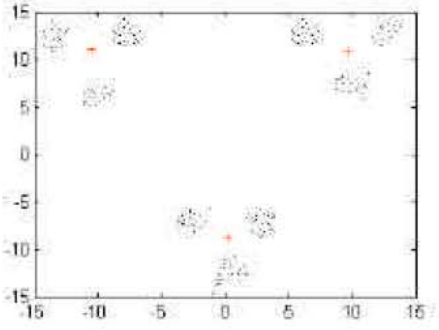


(b) T = 70

Fig. 4: The clustering results under different T (γ = 2)



(a) T = 18



(b) T = 118

Fig. 5: The clustering results under different T (γ = 5)

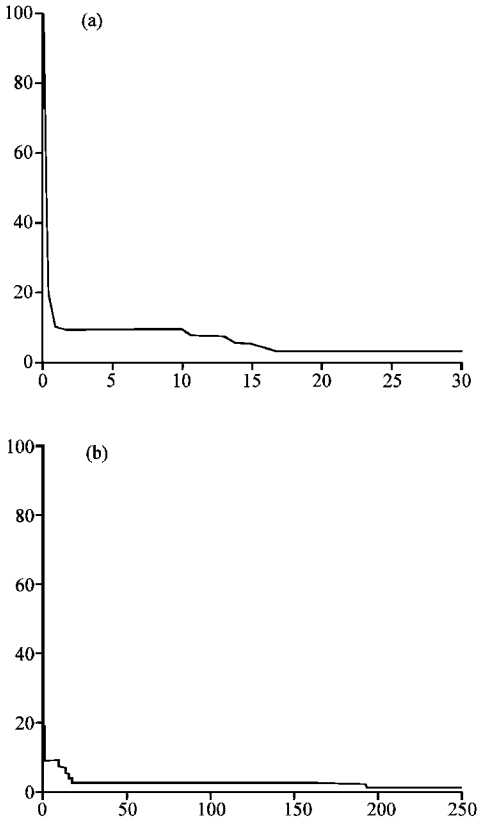


Fig. 6: Number of cluster centers vs sampling frequencies

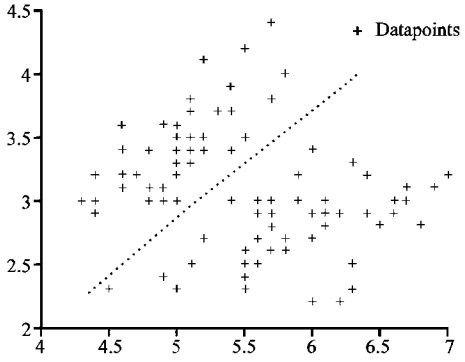


Fig. 7: The first 2-dimensional IRIS dataset

Example 3: In many applications, we may be very interested in the clustering results at some appropriate granular levels. The new clustering algorithm VSC can easily be used to achieve this goal. Figure 9 shows such an artificial dataset. Obviously, c = 3 and c = 9 are possibly appropriate numbers of cluster centers of this dataset. Figure 10 shows the change of the numbers of the corresponding cluster centers with respect to the sampling frequency T. Table 1 gives their corresponding

Table 1: Clustering validity indexes of the artificial dataset

	CVI of c = 3	CVI of c = 9
Algorithm VSC	1.122	1.757

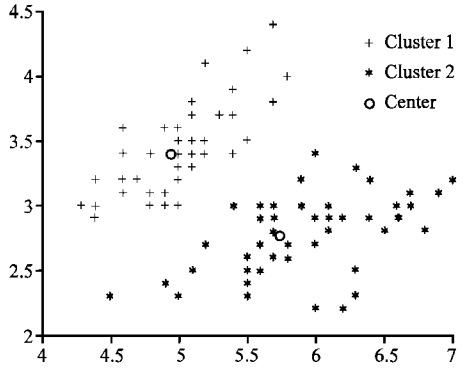


Fig. 8: The clustering result of algorithm VSC

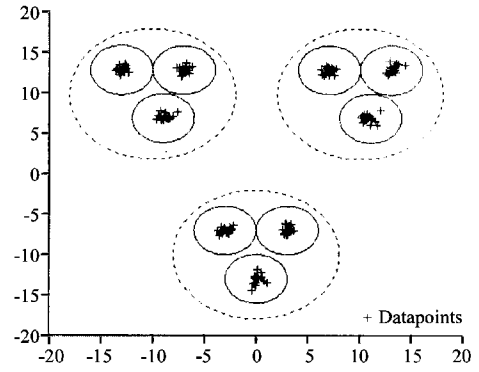


Fig. 9: An artificial dataset

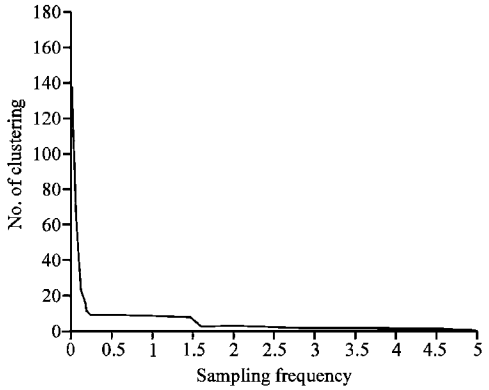
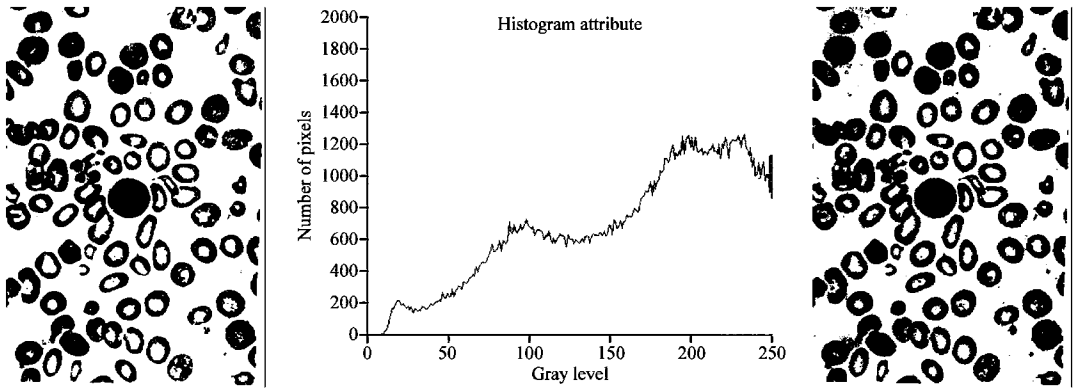


Fig. 10: Number of cluster centers vs. the sampling frequency T

CVI on the dataset. If we analyse the dataset in a bigger granule, the dataset should be classified into 3 clusters. If we want to see more details from the dataset, the dataset should be classified into 9 clusters in a lower granule.



(a) (b) (c)

Fig. 11: Segmentation on the medical image processing using algorithm VSC

Example 4: A test image is used here to demonstrate the power of the new algorithm VSC. We select a 335×500 medical cell image (Fig. 11a). Based on the histograms of the image (Fig. 11b), we consider all the histograms as the dataset. For the visualization purpose, we ignore these pixels with the gray level that is equal to 255. Based on our many experiments, c = 3 is an appropriate choice for most of the histogram datasets of the medical images. As we know, according to the conventional method, we often take the gray level lying in the lowest valley between two peaks of the histograms as the segmentation threshold (Jifeng *et al*., 1988). This segmentation threshold, which corresponds to the second large attractor of the histogram dataset, can easily be determined using the new clustering algorithm VSC. Generally speaking, the sampling frequency T takes more than 200. Figure 8c shows the segmentation result on this medical image. Figure 11c, shows that the new clustering algorithm is very effective for medical image segmentation.

THE LINK RELATIONSHIP BETWEEN ALGORITHM SCA AND OUR ALGORITHM VSC

Overview of SCA: At present, most clustering algorithms attempt to acquire the reasonable clustering results by minimizing the total *dissimilarity* measure. FCM and PCM are two typical examples among them. Yang *et al*. (2004) proposed the novel similarity-based clustering algorithm SCA, which escapes from the above conventional framework and achieves the reasonable clustering result by maximizing the total *similarity* measure. Romeny *et al*. (1993), defined the similarity measure $S(x_j, p_i)$ between x_j and the ith cluster center p_i

$$S(x_j, p_i) = \exp(-\frac{\| x_j - p_i \|^2}{\beta}) \quad (14)$$

where β is the normalized term. Based on this measure, algorithm SCA utilizes the following objective function:

$$J_s(p) = \sum_{i=1}^{c}\sum_{j=1}^{n} f(S(x_j, p_i)) = \sum_{i=1}^{c}\sum_{j=1}^{n} (\exp(-\frac{\| x_j - p_i \|^2}{T}))^{\gamma} \quad (15)$$

where, $\gamma(>0)$ is an important parameter which can determine the locations of peaks in the objective function; $p = \{p_1,..., p_c\}$ is the vector of all cluster centers. All the cluster centers can by obtained by maximizing (15). In order to achieve this goal, by setting $\partial J_s(p)/\partial p = 0$, i.e.,

$$\frac{\partial J_s(p)}{\partial p_i} = \sum_{j=1}^{n} \frac{2\gamma}{\beta} (x_j - p_i) \exp(-\frac{\gamma}{\beta}\left\| x_j - p_i \right\|^2) = 0 \quad (16)$$

we can derive the following iterative equations:

$$p_i = \frac{\sum_{j=1}^{n} \exp(-\frac{\gamma \| x_j - p_i \|^2}{\beta})\ x_j}{\sum_{j=1}^{n} \exp(-\frac{\gamma \| x_j - p_i \|^2}{\beta})} \quad (17)$$

For the given dataset, with the initial cluster centers, we can finally obtain the cluster centers by using (17). In summary, algorithm SCA may be described as follows:

Similarity-based Clustering Algorithm (SCA)

Step 1: Initialize $p^{(0)} = \{p_1^0, p_2^0, \cdots, p_c^0\}$ and give $\varepsilon > 0$ and let iteration counter $l = 0_o$.

Step 2: Calculate the cluster centers $p^{(l+1)}$ using (17) and let $l = l + 1$.

Step 3: Repeat Step 2, until $\| p^{(l+1)} - p^{(l)} \| < \varepsilon$.

Remark 1: Algorithm SCA is very sensitive to the parameter γ. Different γ will result in the shift of the cluster centers. As we may know, clustering analysis heavily depend on the *unknown* density distribution of the given dataset. In order to get a good density estimate of the dataset, the authors in (Yang *et al*., 2004) suggested the following formula to analyse the effect of γ:

$$\overline{J}_s(x_k) = \sum_{j=1}^{n} (\exp - \frac{\|x_j - x_k\|^2}{\beta})^{\gamma} \qquad (18)$$

In their study, they first used the sample variance estimate to determine β, then applied the so-called Correlation Comparison Algorithm (CCA) to get the estimate of γ. According to their suggestion, γ may be taken as 1, 5, 10, 15 and 20 etc. Now, let us rewrite (18) into (19):

$$\begin{aligned} \overline{J}_s(x_k) &= \sum_{j=1}^{n} (\exp - \frac{\|x_j - x_k\|^2}{\beta})^{\gamma} \\ &= \sum_{j=1}^{n} \exp(-\gamma \frac{\|x_j - x_k\|^2}{\beta}) \\ &= \sum_{j=1}^{n} \exp(- \frac{\|x_j - x_k\|^2}{(\beta/\gamma)}) \end{aligned} \qquad (19)$$

It is well known that we can utilize the Parzon window formula

$$\frac{1}{n} \sum_{j=1}^{n} \exp(-\frac{\|x_j - x\|^2}{\sigma})$$

to estimate the density of the sample x in the dataset, where σ represents the unknown variance and has the serious influence on the density estimate. Generally speaking, estimating σ is not trivial task at all. After comparing the Parzon window formula with the above (19), we can easily see that the authors in (Yang *et al*., 2004) in fact transformed such an uneasy task into another one, that to say, how to determine γ effectively. In other words, how to effectively estimate the unknown variance of the given dataset to be clustered has not yet been solved. Moreover, even if their suggestion on γ's value is reasonable, for a large value of γ, for example, γ = 10, 15, when we use other kernel functions instead of exp functions to define the similarity measure, the corresponding computational burden will become very big. The above statements actually indicate its drawback of algorithm SCA.

Remark 2: Since we can not solve (16) directly, so, algorithm SCA actually uses the fixed-point iterative method to approximate the real cluster centers. The obtained c fixed points actually correspond to the c peaks of the objective function. However, due to its monotone property, to large extent, such peaks may perhaps be the same as the cluster centers obtained by our algorithm VSC.

The link relationship between SCA and VSC: Based on the following two theorems, we will establish the link relationship between algorithm SCA and algorithm VSC in this subsection.

Theorem 1: Given the d dimensional dataset $X = \{x_1, x_2, ..., x_n\}$, let $a = (a_1, a_2, ..., a_d)$, where $a_i = \min\{x_{ji}\}$, x_{ji} denotes the ith component of the jth sample x_j, i = 1, ..., d j = 1, 2, ..., n, let $b = (b_1, b_2, ..., b_d)$, $b_i = \max\{x_{ji}\}$, i = 1,..., d j = 1, 2, ..., n, then the maximum of (5) satisfies

$$\begin{aligned} &\max f(x, T) = \max \{f(x_1, T), f(x_2, T), ..., f(x_n, T)\} \\ &x \in [a, b] (= [a_1, b_1] \times [a_2, b_2] \times .... \times [a_d, b_d] \end{aligned} \qquad (20)$$

In other words, the maximum of f (x, T) on $[a_1, b_1] \times [a_2, b_2] \times .... \times [a_d, b_d]$ is equal to its peak among all the corresponding values at $x_1, x_2, ..., x_n$.

Proof: Let us extract a subsequence with a enough length: $x^{(1)} < x^{(2)} < ... < x^{(m)}$ from the dataset $X = \{x_1, x_2, ..., x_n\}$ and let $L^{(i)}$ be the number of $x^{(i)}$ occurring in X, thus, we can rewrite (5) as

$$f(x,T) = \sum_{i=1}^{m} L^{(i)} \exp(-\frac{\| x - x^{(i)} \|^2}{T})^{\gamma}$$

Since $\exp(-\| x - x^{(i)} \|^2/T)^{\gamma}$ is a strit concave function, so f (x, T) is also one and can be expressed by summing up the corresponding m concave functions at the intervals $[x^{(1)}, x^{(i+1)}]$ (i = 1, 2,… …, m-1). In terms of the characteristic of the concave functions, we have

$$\max f(x, T) = \max (f(x^{(1)}, T), f(x^{(i+1)}, T)) \quad x \in [x^{(1)}, x^{(i+1)}]$$

Thus, it is easy to know that this theorem holds true.

The visual sampling image f (x, T) in essence is a density estimate for sample *x*. Therefore, theorem 1 shows that the maximum of the obtained density estimate is just equivalent to its peak among all the corresponding values at $x_1, x_2, ..., x_n$.

Theorem 2: Given the d dimensional dataset $X = \{x_1, x_2,..., x_n\}$, with the fixed sampling frequency T, assume algorithm VSC obtains the *c* cluster centers $X^* = \{x_1^*, x_2^*, ..., x_c^*\}$, let

$$f(x_k^*,T) = \sum_{i=1}^{n}(\exp(-\frac{\| x_k^* - x_i \|^2}{T}))^{\gamma}, k = 1,2,\cdots,c \quad (21)$$

Suppose algorithm SCA obtains the c fixed points $p = \{p_1, ..., p_c\}$ from the above dataset£¬then the objective function (12) becomes

$$J_s(p) = \sum_{i=1}^{c}\sum_{j=1}^{n}(\exp(-\frac{\| x_j - p_i \|^2}{T}))^{\gamma}$$

Thus, we have

$$J_s(p) = f(x_1^*,T) + f(x_2^*,T) + \cdots + f(x_c^*,T) \quad (22)$$

Proof: In terms of theorem 1 and algorithm VSC, for the obtained cluster centers $x_1^*, x_2^*, ..., x_c^*$, we easily know that $f(x_1^*, T), f(x_2^*, T) + ... + f(x_c^*, T)\}$ is equivalent to the sum of all the first, second, …, the cth maximums among $f(x_1, T), f(x_2, T), ..., f(x_n, T)$.

Since $(\exp(-\| x_j - p_i \|^2/T))^{\gamma}$ in (15) is a concave function, so, $J_S(p)$ is also one. In terms of theorem 1, the maximum of $J_S(p)$ must be the sum of all the first, second, …, the cth maximums among $f(x_1, T), f(x_2, T), ..., f(x_n, T)$. Thus, this theorem is proved.

In fact, (22) in theorem 2 establishes the link relationship between algorithm VSC and SCA. In other words, for the given dataset, with the same sampling frequency, due to the existence of this equivalent link relationship and the concave property of the used functions as above, these two clustering algorithms may perhaps have the same or almost the same clustering results in many cases. Our experimental results here always confirm the above claim. However, there is a big discrepancy between them. Algorithm SCA is based on the similarity measure while our algorithm VSC is based on the visual sampling principle. What is more, as it is pointed out in the above, how to choose parameter γ is not a trivial task. However, in our algorithm VSC, the corresponding parameter-the sampling frequency T-may be determined based on the Weber law and meanwhile, this law itself also provides us the new and effective clustering validity index. CVI. This fact clearly indicates our algorithm's advantage over algorithm SCA, that is to say, algorithm VSC may be utilized to replace algorithm SCA such that the drawback of SCA, i.e., the parameter γ therein is very difficult to be well determined, can be circumvented.

Experimental studies: We arrange two experimental results to confirm the claim. For the given dataset, the first experiment attempts to demonstrate that these two algorithms have the same or almost the same clustering results when the sampling frequency and the number of the cluster centers keep the same. The second one aims to show that these two algorithms have the same number of the cluster centers in most cases.

The artificial dataset in Fig. 2 is presented to the first experiment. let γ = 2, Fig. 12-14 demonstrate the clustering results of algorithm VSC and algorithm SCA with different sampling frequencies T. Please note that here we apply algorithm VSC with a fixed T. It is easily seen from these figures that these two algorithms has almost the same clustering results (at least, our eyes can not sense the discrepancy among them). In order to make the experimental results fair and more intuitive, below we will calculate such a discrepancy using two measures Ea and Er.

For the dataset as shown in Fig. 12, we take the clustering results with T = 10 as an example. Figure 13, VSC obtains 9 cluster centers, i.e, 9 attractors $x_1^*, x_2^*,..., x_9^*$; and algorithm SCA also obtains 9 cluster centers algorithm, i.e, 9 fixed points $p_1, p_2,..., p_9$. Table 2 gives the locations of all cluster centers respectively obtained by algorithm VSC and SCA and the distances $\|x_j - p_i\|^2$ among all the corresponding attractors x_i and fixed points p_i. Table 2 shows that all the distances are very small, i.e., x_i is almost the same as p_i and they can even be viewed as the same, after considering the possible computational errors incurred by the computer system itself. Table 3 illustrates all $f(x_i^*, T)$ (i = 1, ..., c (c = 9)) and their total

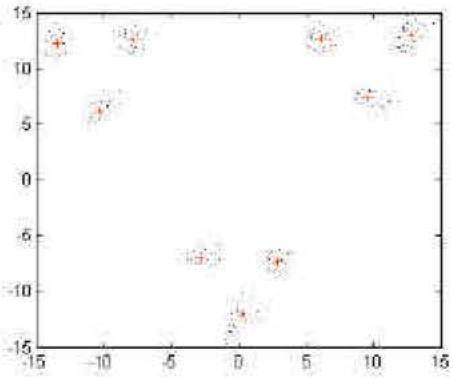
$$\sum_{i=1}^{c} f(x_i^*,T)$$

obtained by VSC and all

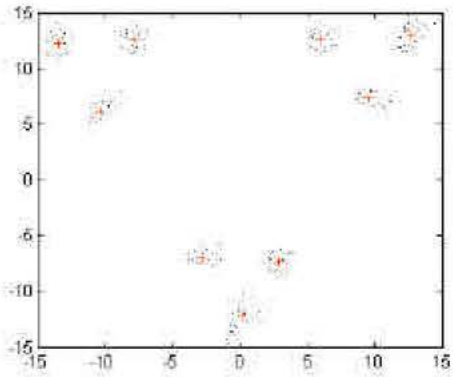
$$g(p_i,T) = \sum_{j=1}^{n}(\exp(-\frac{\| x_j - p_i \|^2}{T}))^{\gamma}$$

(i = 1,..., c (c = 9)) obtained by SCA. Obviously, since

$$g(p_i,T) = \sum_{j=1}^{n}(\exp(-\frac{\| x_j - p_i \|^2}{T}))^{\gamma}, so, J_s(p) = \sum_{i=1}^{c} g(x_i,T) \cdot$$

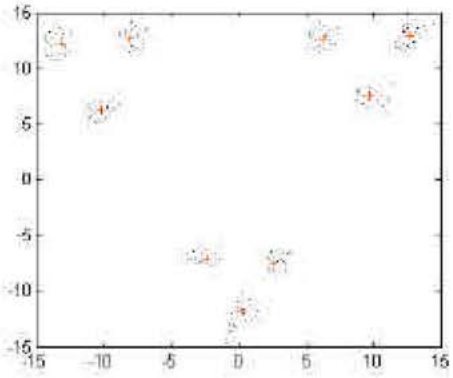


(a) Clustering result of VSC

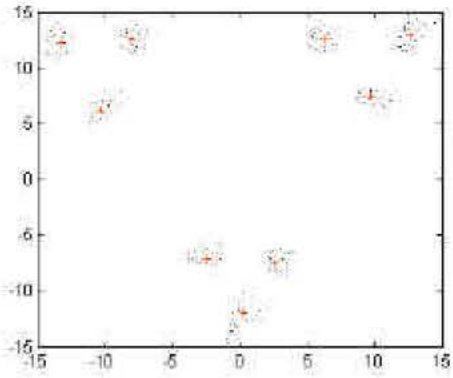


(b) Clustering result of SCA

Fig. 12: Clustering results of VSC and SCA with T = 4

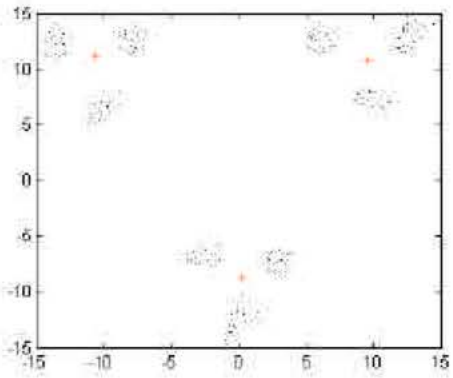


(a) Clustering result of VSC

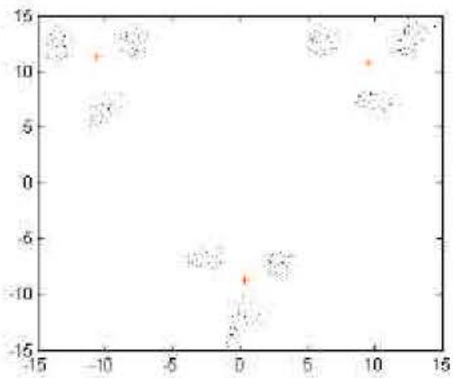


(b) Clustering result of SCA

Fig. 13: Clustering results of VSC and SCA with T = 10



(a) Clustering result of VSC



(b) Clustering result of SCA

Fig. 14: Clustering results of VSC and SCA with T = 40

Table 2: The cluster centers obtained by VSC and SCA

	Center x_i by VSC		Center p_i by SCA		
The ith center	Horizontal axis	Vertical axis	Horizontal axis	Vertical axis	$\| x_i - p_i \|^2$
i = 1	6.1839	12.682	6.1453	12.661	0.00193096
i = 2	12.753	13.025	12.708	12.994	0.002986
i = 3	9.7333	7.3578	9.6892	7.3526	0.00197185
i = 4	2.848	-7.3699	2.8417	-7.4233	0.00289125
i = 5	-2.6605	-6.9798	-2.7054	-7.0104	0.00295237
i = 6	0.21268	-12.247	0.17186	-12.326	0.00790727
i = 7	-7.8486	12.683	-7.8879	12.661	0.00202849
i = 8	-13.41	12.233	-13.447	12.203	0.002269
i = 9	-10.22	6.1863	-10.259	6.1921	0.00155464

Table 3: $f(x_i^*, T)$ and $g(p_i, T)$ obtained by VSC and SCA

	VSC	SCA	The absolute error	The relative error
The ith center	$f(x_i^*, T)$	$g(p_i, T)$	Ea	Er
i = 1	35.057	35.094	0.037	0.001054311
i = 2	34.215	34.506	0.291	0.008433316
i = 3	35.612	35.51	0.102	0.00287243
i = 4	38.872	38.883	0.011	0.0002829
i = 5	33.215	32.918	0.297	0.009022419
i = 6	40.701	40.711	0.01	0.000245634
i = 7	38.902	39.212	0.31	0.007905743
i = 8	33.195	33	0.195	0.005909091
i = 9	42.107	41.975	0.132	0.003144729

In this table, for the ith cluster center, we also list the absolute error $Ea = | f(x_1^*, T) - g(p_i, T) |$ and the relative error $Er = | f(x_1^*, T) - g(p_i, T) | / g(p_i, T)$ between $f(x_1^*, T)$ and $g(p_i, T)$. In terms of Table 3, we have

$$\sum_{i=1}^{c} f(x_i^*, T) = 331.876, \sum_{i=1}^{c} g(x_i, T) = 331.809,$$

and their absolute error

$$\sum_{i=1}^{c} f(x_i^*, T) - \sum_{i=1}^{c} g(x_i, T) = 0.067$$

and their relative error

$$\frac{\left| \sum_{i=1}^{c} f(x_i^*, T) - \sum_{i=1}^{c} g(x_i, T) \right|}{\sum_{i=1}^{c} g(x_i, T)} = 0.000201923$$

Due to the existence of the computational error incurred by the computer system itself, we may think that

$$\sum_{i=1}^{c} f(x_i^*, T) = \sum_{i=1}^{c} g(x_i, T)$$

that is to say, $J_S(p) = f(x_1^*, T) + ... + f(x_c^*, T)$ in (22) holds. Obviously, these experimental results further confirm theorem 2.

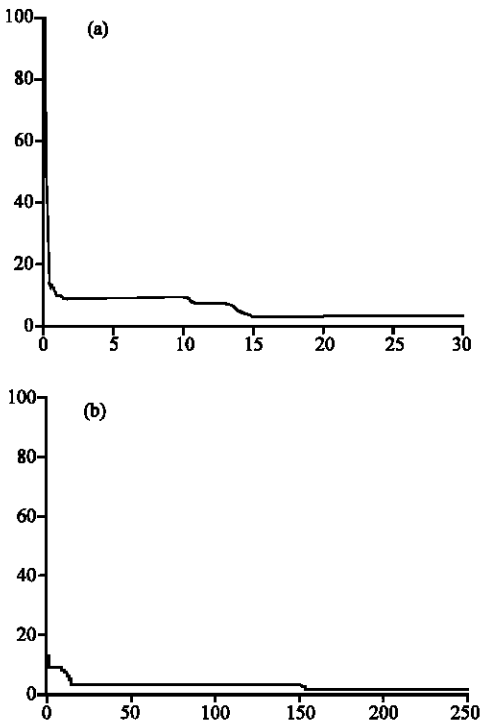


Fig. 15: Number of cluster centers vs sampling frequencies

Now, let us introduce the second experiment and its clustering results. We use the same artificial dataset Fig. 2. in this experiment. Without loss of generality, let $\gamma = 1$. The aim of this experiment is to observe the change trend of the numbers of the cluster centers obtained by algorithm SCA within a wide scope of sampling frequencies T (i.e., changing β in SCA). In order to make the experimental results more clear, we give two figures here, i.e., Fig. 15a and b. Obviously, when $1<T\leq10$, algorithm SCA is stable with 9 cluster centers; and when, algorithm SCA is stable with 3 cluster centers. However, when $T>160$, the number of the obtained cluster centers will become 1. Let us go back to the experimental results obtained by our algorithm VSC, Fig. 6. After comparing Fig. 6 with Fig. 15, we can easily find that with the same sampling frequency, these two clustering algorithms have the same number of the cluster centers in most cases. This fact indicates its support for the above claim from the viewpoint of another angle.

In summary, these two experiments show that under the same sampling frequency, these two clustering algorithms may perhaps have the same or almost the same clustering results in many cases.

CONCLUSIONS

In real life, we often classify various objects effectively under complex conditions by our eyes, so a

ideal clustering algorithm should depend not only on the principle of physical system by which the data are generated but also on the manner in which human being sense the structure of the data. The present study indicates that the visual sampling based clustering algorithm VSC is the first-time attempt toward achieving this aim. Algorithm VSC is robust to initial conditions and very effective for convex datasets. With the new Weber-law-based clustering validity index, algorithm VSC can effectively determine the reasonable cluster number. The link relationship between our algorithm VSC and algorithm SCA. reveals that these two algorithms have the same or almost the same clustering results in many cases, therefore, algorithm VSC can be utilized to replace algorithm SCA such that the drawback of SCA, i.e., the parameter γ therein is very difficult to be well determined, can be avoided.

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 non-uniform sampling and eye motion. The current version of algorithm VSC in the paper is based on uniform sampling., therefore, future work may be concentrated on extending algorithm VSC to non-uniform sampling cases. Another interesting and challenging problem is how we can extend the ideas existing in the paper to non-convex (i.e., nonlinearly separable) datasets.

ACKNOWLEDGMENTS

The present study is supported by the HongKong PolyU CRG (grant No. G-T912) and HongKong 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

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

Bezdek, J.C. *et al.*, 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.

Fugunaga, K., 1990. Introduction to Statistical Pattern Recognition, Academic Press, New York.

Gokcay, E. *et al.*, 2002. Information Theoretic Clustering. IEEE Trans. PAMI, 24: 158-171.

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

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.

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

Nadler, K. *et al.*, 1993. Pattern Recognition Engineering. Wiley, New York.

Pedrycz, W. *et al.*, 2002. Granular Clustering: A granular signature of data. IEEE Trans. SMC (part B), 32: 212-224.

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., P.N.T Wells (Eds.). The Perception of Visual Information. New York: Springer-Verlag, pp: 73-114.

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

Shitong, W. *et al.*, 2002. A new integrated clustering algorithm GFC and switching regressions. Intl. J. Pattern Recognition And Artificial Intelligence, 16: 433-446.

Yang, M.S. *et al.*, 2004. A similarity-based robust clustering method. IEEE Trans. Pattern Analysis and Machine Intelligence, 26: 434-448.

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