We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 13

Fuzzy c-Means Clustering, Entropy Maximization, and Deterministic and Simulated Annealing

Makoto Yasuda

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48659

1. Introduction

Many engineering problems can be formulated as optimization problems, and the deterministic annealing (DA) method [20] is known as an effective optimization method for such problems. DA is a deterministic variant of simulated annealing (SA) [1, 10]. The DA characterizes the minimization problem of cost functions as the minimization of Helmholtz free energy which depends on a (pseudo) temperature, and tracks the minimum of free energy while decreasing temperature and thus it can deterministically optimize the function at a given temperature [20]. Hence, the DA is more efficient than the SA, but does not guarantee a global optimal solution. The study on the DA in [20] addressed avoidance of the poor local minima of cost function of data clustering. Then it was extensively applied to various subjects such as combinational optimization problems [21], vector quantization [4], classifier design [13], pairwise data clustering [9] and so on.

On the other hand, clustering is a method which partitions a given set of data points into subgroups, and is one of major tools for data analysis. It is supposed that, in the real world, cluster boundaries are not so clear that fuzzy clustering is more suitable than crisp clustering. Bezdek[2] proposed the fuzzy c-means (FCM) which is now well known as the standard technique for fuzzy clustering.

Then, after the work of Li et al.[11] which formulated the regularization of the FCM with Shannon entropy, Miyamoto et al.[14] discussed the FCM within the framework of the Shannon entropy based clustering. From the historical point of view, however, it should be noted that Rose et al.[20] first studied the statistical mechanical analogy of the FCM with the maximum entropy method, which was basically probabilistic clustering.

To measure the "indefiniteness" of fuzzy set, DeLuca and Termini [6] defined fuzzy entropy after Shannon. Afterwards, some similar measures from the wider viewpoints of the indefiniteness were proposed [15, 16]. Fuzzy entropy has been used for knowledge retrieval from fuzzy database [3] and image processing [31], and proved to be useful.

Tsallis [24] achieved nonextensive extension of the Boltzmann-Gibbs statistics. Tsallis postulated a generalization form of entropy with a generalization parameter q, which, in a

limit of $q \to 1$,reaches the Shannon entropy. Later on, Menard et.al.[12] derived a membership function by regularizing FCM with the Tsallis entropy.

In this chapter, by maximizing the various entropies within the framework of FCM, the membership functions which take the familiar forms of the statitical mechanical distribution functions are derived. The advantage to use the statistical mechanical membership functions is that the fuzzy c-means clustering can be interpreted and analyzed from a statistical mechanical point of view [27, 28]

After that, we focus on the Fermi-Dirac like membership function, because, as compared to the Maxwell-Boltzmann-like membership function, the Fermi-Dirac-like membership function has extra parameters α_ks (α_k corresponds to a chemical potential in statistical mechanics[19], and k denotes a data point), which make it possible to represent various cluster shapes like former clustering methods based on such as the Gaussian mixture[7], and the degree of fuzzy entropy[23]. α_ks strongly affect clustering results and they must be optimized under a normalization constraint of FCM. On the other hand, the DA method, though it is efficient, does not give appropriate values of α_ks by itself and the DA clustering sometimes fails if α_ks are improperly given. Accordingly, we introduce SA to optimize α_ks because, as pointed above, both of DA and SA contain the parameter corresponding to the system temperature and can be naturally combined as DASA.

Nevertheless, this approach causes a few problems. (1)How to estimate the initial values of α_ks under the normalization constraint ? (2)How to estimate the initial annealing temperature? (3)SA must optimize a real number α_k[5, 26]. (4)SA must optimize many α_ks[22].

Linear approximations of the Fermi-Dirac-like membership function is useful in guessing the initial α_ks and the initial annealing temperature of DA.

In order to perform SA in a many variables domain, α_ks to be optimized are selected according to a selection rule. In an early annealing stages, most α_ks are optimized. In a final annealing stage, however, only α_ks of data which locate sufficiently away from all cluster centers are optimized because their memberships might be fuzzy. Distances between the data and the cluster centers are measured by using linear approximations of the Fermi-Dirac-like membership function.

However, DASA suffers a few disadvantages. One of them is that it is not necessarily easy to interpolate membership functions obtained by DASA, since their values are quite different each other. The fractal interpolation method [17] is suitable for these rough functions [30].

Numerical experiments show that DASA clusters data which distribute in various shapes more properly and stably than single DA. Also, the effectiveness of the fractal interpolation is examined.

2. Fuzzy c-means

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}(\mathbf{x}_k = (x_k^1, \ldots, x_k^p) \in R^p)$ be a data set in a p-dimensional real space, which should be divided into c clusters $C = \{C_1, \ldots, C_c\}$. Let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_c\}(\mathbf{v}_i = (v_i^1, \ldots, v_i^p))$ be the centers of clusters and $u_{ik} \in [0,1](i = 1, \ldots, c; k = 1, \ldots, n)$ be the membership function. Also let

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} (d_{ik})^m \quad (m > 1) \tag{1}$$

be the objective function of the FCM where $d_{ik} = \|\mathbf{x}_k - \mathbf{v}_i\|^2$. In the FCM, under the normalization constraint of

$$\sum_{i=1}^{c} u_{ik} = 1 \ \ \forall k, \tag{2}$$

the Lagrange function L_{FCM} is given by

$$L_{FCM} = J - \sum_{k=1}^{n} \eta_k \left(\sum_{i=1}^{c} u_{ik} - 1 \right), \tag{3}$$

where η_k is the Lagrange multiplier. Bezdek[2] showed that the FCM approaches crisp clustering as m decreases to $+1$.

3. Entropy maximization of FCM

3.1. Shannon entropy maximization

First, we introduce the Shannon entropy into the FCM clustering. The Shannon entropy is given by

$$S = -\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik}. \tag{4}$$

Under the normalization constraint and setting m to 1, the fuzzy entropy functional is given by

$$\delta S - \sum_{k=1}^{n} \alpha_k \delta \left(\sum_{i=1}^{c} u_{ik} - 1 \right) - \beta \sum_{k=1}^{n} \sum_{i=1}^{c} \delta(u_{ik} d_{ik}), \tag{5}$$

where α_k and β are the Lagrange multipliers, and α_k must be determined so as to satisfy Eq. (2). The stationary condition for Eq. (5) leads to the following membership function

$$u_{ik} = \frac{\mathrm{e}^{-\beta d_{ik}}}{\sum_{j=1}^{c} \mathrm{e}^{-\beta d_{jk}}}. \tag{6}$$

and the cluster centers

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik} \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}}. \tag{7}$$

3.2. Fuzzy entropy maximization

We then introduce the fuzzy entropy into the FCM clustering.

The fuzzy entropy is given by

$$\hat{S} = -\sum_{k=1}^{n} \sum_{i=1}^{c} \{ \hat{u}_{ik} \log \hat{u}_{ik} + (1 - \hat{u}_{ik}) \log(1 - \hat{u}_{ik}) \}. \tag{8}$$

The fuzzy entropy functional is given by

$$\delta \hat{S} - \sum_{k=1}^{n} \alpha_k \delta \left(\sum_{i=1}^{c} \hat{u}_{ik} - 1 \right) - \beta \sum_{k=1}^{n} \sum_{i=1}^{c} \delta(\hat{u}_{ik} d_{ik}), \tag{9}$$

where α_k and β are the Lagrange multipliers[28]. The stationary condition for Eq. (9) leads to the following membership function

$$\hat{u}_{ik} = \frac{1}{\mathrm{e}^{\alpha_k + \beta d_{ik}} + 1}, \tag{10}$$

and the cluster centers

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} \hat{u}_{ik} \boldsymbol{x}_k}{\sum_{k=1}^{n} \hat{u}_{ik}}. \tag{11}$$

In Eq. (10), β defines the extent of the distribution [27]. Equation (10) is formally normalized as

$$\hat{u}_{ik} = \frac{1}{\mathrm{e}^{\alpha_k + \beta d_{ik}} + 1} \Big/ \sum_{j=1}^{c} \frac{1}{\mathrm{e}^{\alpha_k + \beta d_{jk}} + 1}. \tag{12}$$

3.3. Tsallis entropy maximization

Let $\tilde{\boldsymbol{v}}_i$ and $\tilde{u}_{ik}$ be the centers of clusters and the membership functions, respectively.

The Tsallis entropy is defined as

$$\tilde{S} = -\frac{1}{q-1} \left(\sum_{k=1}^{n} \sum_{i=1}^{c} \tilde{u}_{i\,k}^{q} - 1 \right), \tag{13}$$

where $q \in \boldsymbol{R}$ is any real number. The objective function is rewritten as

$$\tilde{U} = \sum_{k=1}^{n} \sum_{i=1}^{c} \tilde{u}_{ik}^{q} \tilde{d}_{ik}, \tag{14}$$

where $\tilde{d}_{ik} = \|\boldsymbol{x}_k - \tilde{\boldsymbol{v}}_i\|^2$.

Accordingly, the Tsallis entropy functional is given by

$$\delta \tilde{S} - \sum_{k=1}^{n} \alpha_k \delta \left(\sum_{i=1}^{c} \tilde{u}_{ik} - 1 \right) - \beta \sum_{k=1}^{n} \sum_{i=1}^{c} \delta(\tilde{u}_{ik}^{q} \tilde{d}_{ik}). \tag{15}$$

The stationary condition for Eq. (15) yields to the following membership function

$$\tilde{u}_{ik} = \frac{\{1 - \beta(1-q)\tilde{d}_{ik}\}^{\frac{1}{1-q}}}{\tilde{Z}}, \tag{16}$$

where

$$\tilde{Z} = \sum_{j=1}^{c} \{1 - \beta(1-q)\tilde{d}_{jk}\}^{\frac{1}{1-q}}. \tag{17}$$

In this case, the cluster centers are given by

$$\tilde{\boldsymbol{v}}_i = \frac{\sum_{k=1}^{n} \tilde{u}_{ik}^{q} \boldsymbol{x}_k}{\sum_{k=1}^{n} \tilde{u}_{ik}^{q}}. \tag{18}$$

In the limit of $q \to 1$, the Tsallis entropy recovers the Shannon entropy [24] and $\tilde{u}_{ik}$ approaches u_{ik} in Eq.(6).

4. Entropy maximization and statistical physics

4.1. Shannon entropy based FCM statistics

In the Shannon entropy based FCM, the sum of the states (the partition function) for the grand canonical ensemble of fuzzy clustering can be written as

$$Z = \prod_{k=1}^{n} \sum_{i=1}^{c} \mathrm{e}^{-\beta d_{ik}}. \tag{19}$$

By substituting Eq. (19) for $F = -(1/\beta)(\log Z)$[19], the free energy becomes

$$F = -\frac{1}{\beta} \sum_{k=1}^{n} \log \left\{ \sum_{i=1}^{c} \mathrm{e}^{-\beta d_{ik}} \right\}. \tag{20}$$

Stable thermal equilibrium requires a minimization of the free energy. By formulating deterministic annealing as a minimization of the free energy, $\partial F / \partial \boldsymbol{v}_i = 0$ yields

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik} \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}}. \tag{21}$$

This cluster center is the same as that in Eq. (7).

4.2. Fuzzy entropy based FCM statistics

In a group of independent particles, the total energy and the total number of particles are given by $E = \sum_l \epsilon_l n_l$ and $N = \sum_l n_l$, respectively, where ϵ_l represents the energy level and n_l represents the number of particles that occupy ϵ_l. We can write the sum of states, or the partition function, in the form:

$$Z_N = \sum_{\sum_l \epsilon_l n_l = E, \sum_l n_l = N} e^{-\beta \sum_l \epsilon_l n_l} \tag{22}$$

where β is the product of the inverse of temperature T and k_B (Boltzmann constant). However, it is difficult to take the sums in (22) counting up all possible divisions. Accordingly, we make the number of particles n_l a variable, and adjust the new parameter α(chemical potential) so as to make $\sum_l \epsilon_l n_l = E$ and $\sum_l n_l = N$ are satisfied. Hence, this becomes the grand canonical distribution, and the sum of states (the grand partition function) Ξ is given by[8, 19]

$$\Xi = \sum_{N=0}^{\infty} (e^{-\alpha})^N Z_N = \prod_l \sum_{n_l=0}^{\infty} (e^{-\alpha - \beta\epsilon_l})^{n_l}. \tag{23}$$

For particles governed by the Fermi-Dirac distribution, Ξ can be rewritten as

$$\Xi = \prod_l (1 + e^{-\alpha - \beta\epsilon_l}). \tag{24}$$

Also, n_l is averaged as

$$\langle n_l \rangle = \frac{1}{e^{\alpha + \beta\epsilon_l} + 1} \tag{25}$$

where α is defined by the condition that $N = \sum_l \langle n_l \rangle$ [19]. Helmholtz free energy F is, from the relationship $F = -k_B T \log Z_N$,

$$F = -k_B T \left(\log \Xi - \alpha \frac{\partial}{\partial \alpha} \log \Xi \right) = -\frac{1}{\beta} \left\{ \sum_l \log(1 + e^{-\alpha - \beta \epsilon_l}) + \alpha N \right\}. \tag{26}$$

Taking that

$$E = \sum_l \frac{\epsilon_l}{e^{\alpha + \beta \epsilon_l} + 1} \tag{27}$$

into account, the entropy $S = (E - F)/T$ has the form

$$S = -k_B \sum_l \{ \langle n_l \rangle \log \langle n_l \rangle + (1 - \langle n_l \rangle) \log(1 - \langle n_l \rangle) \}. \tag{28}$$

If states are degenerated to the degree of ν_l, the number of particles which occupy ϵ_l is

$$\langle N_l \rangle = \nu_l \langle n_l \rangle, \tag{29}$$

and we can rewrite the entropy S as

$$S = -k_B \sum_l \{ \frac{\langle N_l \rangle}{\nu_l} \log \frac{\langle N_l \rangle}{\nu_l} + \left(1 - \frac{\langle N_l \rangle}{\nu_l} \right) \log \left(1 - \frac{\langle N_l \rangle}{\nu_l} \right) \}, \tag{30}$$

which is similar to fuzzy entropy in (8). As a result, u_{ik} corresponds to a grain density $\langle n_l \rangle$ and the inverse of β in (10) represents the system or computational temperature T.

In the FCM clustering, note that any data can belong to any cluster, the grand partition function can be written as

$$\Xi = \prod_{k=1}^{n} \prod_{i=1}^{c} (1 + e^{-\alpha_k - \beta d_{ik}}), \tag{31}$$

which, from the relationship $F = -(1/\beta)(\log \Xi - \alpha_k \partial \log \Xi / \partial \alpha_k)$, gives the Helmholtz free energy

$$F = -\frac{1}{\beta} \sum_{k=1}^{n} \left\{ \sum_{i=1}^{c} \log(1 + e^{-\alpha_k - \beta d_{ik}}) + \alpha_k \right\}. \tag{32}$$

The inverse of β represents the system or computational temperature T.

4.3. Correspondence between Fermi-Dirac statistics and fuzzy clustering

In the previous subsection, we have formulated the fuzzy entropy regularized FCM as the DA clustering and showed that its mechanics was no other than the statistics of a particle system (the Fermi-Dirac statistics). The correspondences between fuzzy clustering (FC) and the Fermi-Dirac statistics (FD) are summarized in TABLE 1. The major difference between fuzzy clustering and statistical mechanics is the fact that data are distinguishable and can belong to multiple clusters, though particles which occupy a same energy state are not distinguishable. This causes a summation or a multiplication not only on i but on k as well in fuzzy clustering. Thus, fuzzy clustering and statistical mechanics described in this paper are not mathematically equivalent.

- **Constraints:** (a) Constraint that the sum of all particles N is fixed in FD is correspondent with the normalization constraint in FC. Energy level l is equivalent to the cluster number

	Fermi-Dirac Statistics	Fuzzy Clustering
Constraints	(a)$\sum_l n_l = N$ (b)$\sum_l \epsilon_l n_l = E$	(a)$\sum_{i=1}^{c} u_{ik} = 1$
Distribution Function	$\langle n_l \rangle = \frac{1}{e^{\alpha+\beta\epsilon_l}+1}$	$u_{ik} = \frac{1}{e^{\alpha_k+\beta d_{ik}}+1}$
Entropy	$S = -k_B \sum_l \left\{ \frac{\langle N_l \rangle}{\nu_l} \log \frac{\langle N_l \rangle}{\nu_l} \right.$ $\left. + \left(1 - \frac{\langle N_l \rangle}{\nu_l}\right) \log \left(1 - \frac{\langle N_l \rangle}{\nu_l}\right) \right\}$	$S_{FE} = -\sum_{k=1}^{n} \sum_{i=1}^{c} \{ u_{ik} \log u_{ik}$ $+(1-u_{ik}) \log(1-u_{ik}) \}$
Temperature(T)	(given)	(given)
Partition Function(Ξ)	$\prod_l \left(1 + e^{-\alpha-\beta\epsilon_l}\right)$	$\prod_{k=1}^{n} \prod_{i=1}^{c} \left(1 + e^{-\alpha_k-\beta d_{ik}}\right)$
Free Energy(F)	$-\frac{1}{\beta} \left\{ \sum_l \log(1 + e^{-\alpha-\beta\epsilon_l}) + \alpha N \right\}$	$-\frac{1}{\beta} \sum_{k=1}^{n} \left\{ \sum_{i=1}^{c} \log(1 + e^{-\alpha_k-\beta d_{ik}}) + \alpha_k \right\}$
Energy(E)	$\sum_l \frac{\epsilon_l}{e^{\alpha+\beta\epsilon_l}+1}$	$\sum_{k=1}^{n} \sum_{i=1}^{c} \frac{d_{ik}}{e^{\alpha_k+\beta d_{ik}}+1}$

Table 1. Correspondence of Fermi-Dirac Statistics and Fuzzy Clustering.

i. In addition, the fact that data can belong to multiple clusters leads to the summation on k. (b) There is no constraint in FC which corresponds to the constraint that the total energy equals E in FD. We have to minimize $\sum_{k=1}^{n} \sum_{i=1}^{c} d_{ik} u_{ik}$ in FC.

- **Distribution Function:** In FD, $\langle n_l \rangle$ gives an average particle number which occupies energy level l, because particles can not be distinguished from each other. In FC, however, data are distinguishable, and for that reason, u_{ik} gives a probability of data belonging to multiple clusters.
- **Entropy:** $\langle N_l \rangle$ is supposed to correspond to a cluster capacity. The meanings of S and S_{FE} will be discussed in detail in the next subsection.
- **Temperature:** Temperature is given in both cases [1].
- **Partition Function:** The fact that data can belong to multiple clusters simultaneously causes the product over k for FC.
- **Free Energy:** Helmholtz free energy F is given by $-T(\log \Xi - \alpha_k \partial \log \Xi / \partial \alpha_k)$ in FC. Both S and S_{FE} equal $-\partial F / \partial T$ as expected from statistical physics.
- **Energy:** The relationship $E = F + TS$ or $E = F + TS_{FE}$ holds between E, F, T and S or S_{FE}.

4.4. Meanings of Fermi-Dirac distribution function and fuzzy entropy

In the entropy function (28) or (30) for the particle system, we can consider the first term to be the entropy of electrons and the second to be that of holes. In this case, the physical limitation that only one particle can occupy an energy level at a time results in the entropy that formulates the state in which an electron and a hole exist simultaneously and exchanging them makes no difference. Meanwhile, what correspond to electron and hole in fuzzy clustering are the probability of fuzzy event that a data belongs to a cluster and the probability of its complementary event, respectively.

Fig.2 shows a two-dimensional virtual cluster density distribution model. A lattice can have at most one data. Let M_l be the total number of lattices and m_l be the number of lattices which

[1] In the FCM, however, temperature is determined as a result of clustering.

have a data in it (marked by a black box). Then, the number of available divisions of data to lattices is denoted by

$$W = \prod_l \frac{M_l!}{m_l!(M_l - m_l)!} \tag{33}$$

which, from $S = k_B \log W$ (the Gibbs entropy), gives the form similar to (30)[8]. By extremizing S, we have the most probable distribution like (25). In this case, as there is no distinction between data, only the numbers of black and white lattices constitute the entropy. Fuzzy entropy in (8), on the other hand, gives the amount of information of weather a data belongs to a fuzzy set (or cluster) or not, averaged over independent data $\mathbf{x}_k$.

Changing a viewpoint, the stationary entropy values for the particle system seems to be a request for giving the stability against the perturbation with collisions between particles. In fuzzy clustering, data reconfiguration between clusters with the move of cluster centers or the change of cluster shapes is correspondent to this stability. Let us represent data density by $\langle\rangle$. If data transfer from clusters C_a and C_b to C_c and C_d as a magnitude of membership function, the transition probability from $\{\ldots, C_a, \ldots, C_b, \ldots\}$ to $\{\ldots, C_c, \ldots, C_d, \ldots\}$ will be proportional to $\langle C_a\rangle\langle C_b\rangle(1 - \langle C_c\rangle)(1 - \langle C_d\rangle)$ because a data enters a vacant lattice. Similarly, the transition probability from $\{\ldots, C_c, \ldots, C_d, \ldots\}$ to $\{\ldots, C_a, \ldots, C_b, \ldots\}$ will be proportional to $\langle C_c\rangle\langle C_d\rangle(1 - \langle C_a\rangle)(1 - \langle C_b\rangle)$. In the equilibrium state, the transitions exhibit balance (this is known as the principle of detailed balance[19]). This requires

$$\frac{\langle C_a\rangle\langle C_b\rangle}{(1 - \langle C_a\rangle)(1 - \langle C_b\rangle)} = \frac{\langle C_c\rangle\langle C_d\rangle}{(1 - \langle C_c\rangle)(1 - \langle C_d\rangle)}. \tag{34}$$

As a result, if energy d_i is conserved before and after the transition, $\langle C_i\rangle$ must have the form

$$\frac{\langle C_i\rangle}{1 - \langle C_i\rangle} = e^{-\alpha - \beta d_i} \tag{35}$$

or Fermi-Dirac distribution

$$\langle C_i\rangle = \frac{1}{e^{\alpha + \beta d_i} + 1}, \tag{36}$$

where α and β are constants.

Consequently, the entropy like fuzzy entropy is statistically caused by the system that allows complementary states. Fuzzy clustering handles a data itself, while statistical mechanics handles a large number of particles and examines the change of macroscopic physical quantity. Then it is concluded that fuzzy clustering exists in the extreme of Fermi-Dirac statistics, or the Fermi-Dirac statistics includes fuzzy clustering conceptually.

4.5. Tsallis entropy based FCM statistics

On the other hand, $\tilde{U}$ and $\tilde{S}$ satisfy

$$\tilde{S} - \beta\tilde{U} = \sum_{k=1}^{n} \frac{\tilde{Z}^{1-q} - 1}{1 - q}, \tag{37}$$

which leads to

$$\frac{\partial\tilde{S}}{\partial\tilde{U}} = \beta. \tag{38}$$

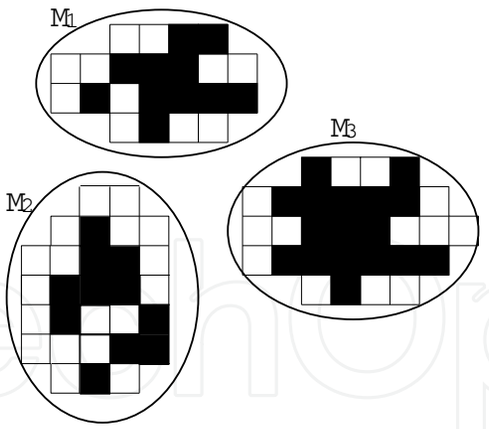


Figure 1. Simple lattice model of clusters. $M_1, M_2, \ldots$ represent clusters. Black and white box represent whether a data exists or not.

Equation (38) makes it possible to regard β^{-1} as an artificial system temperature T [19]. Then, the free energy can be defined as

$$\tilde{F} = \tilde{U} - T\tilde{S} = -\frac{1}{\beta}\sum_{k=1}^{n}\frac{\tilde{Z}^{1-q}-1}{1-q}. \tag{39}$$

$\tilde{U}$ can be derived from $\tilde{F}$ as

$$\tilde{U} = -\frac{\partial}{\partial\beta}\sum_{k=1}^{n}\frac{\tilde{Z}^{1-q}-1}{1-q}. \tag{40}$$

$\partial\tilde{F}/\partial\tilde{\boldsymbol{v}}_i = 0$ also gives

$$\tilde{\boldsymbol{v}}_i = \frac{\sum_{k=1}^{n}\tilde{u}_{ik}^{q}\boldsymbol{x}_k}{\sum_{k=1}^{n}\tilde{u}_{ik}^{q}}. \tag{41}$$

5. Deterministic annealing

The DA method is a deterministic variant of SA. DA characterizes the minimization problem of the cost function as the minimization of the Helmholtz free energy which depends on the temperature, and tracks its minimum while decreasing the temperature and thus it can deterministically optimize the cost function at each temperature.

According to the principle of minimal free energy in statistical mechanics, the minimum of the Helmholtz free energy determines the distribution at thermal equilibrium [19]. Thus, formulating the DA clustering as a minimization of (32) leads to $\partial F/\partial\mathbf{v}_i = 0$ at the current temperature, and gives (10) and (11) again. Desirable cluster centers are obtained by calculating (10) and (11) repeatedly.

In this chapter, we focus on application of DA to the Fermi-Dirac-like distribution function described in the Section 4.2.

5.1. Linear approximation of Fermi-Dirac distribution function

The Fermi-Dirac distribution function can be approximated by linear functions. That is, as shown in Fig.1, the Fermi-Dirac distribution function of the form:

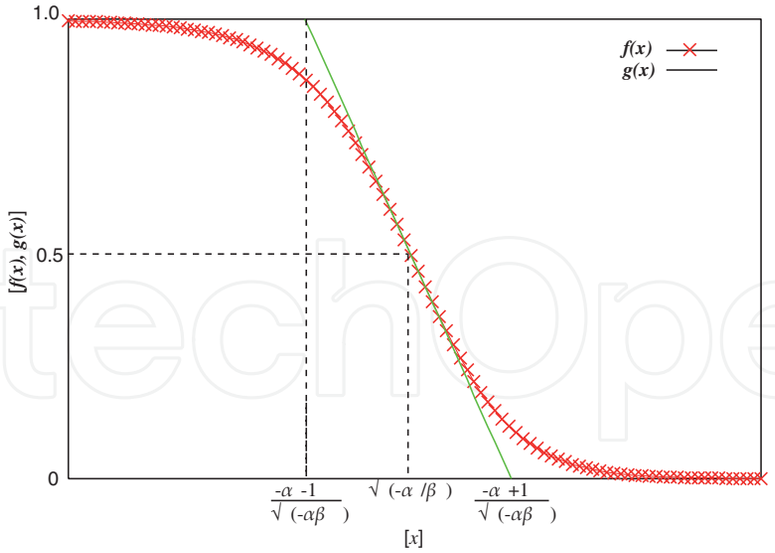


Figure 2. The Fermi-Dirac distribution function $f(x)$ and its linear approximation functions $g(x)$.

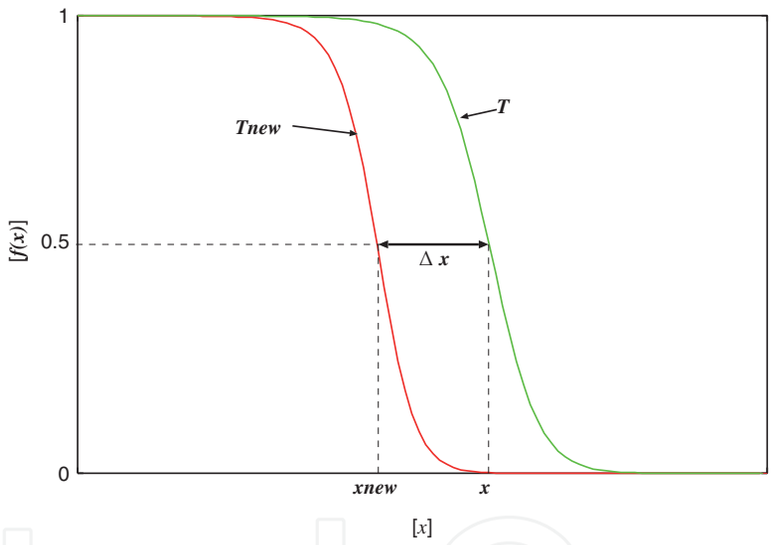


Figure 3. Decreasing of extent of the Fermi-Dirac distribution function from x to x_{new} with decreasing the temperature from T to T_{new}.

$$f(x) = \frac{1}{e^{\alpha+\beta x^2}+1} \tag{42}$$

is approximated by the linear functions

$$g(x) = \begin{cases} 1.0 & \left(x \le \dfrac{-\alpha-1}{\kappa}\right) \\ -\dfrac{\kappa}{2}x - \dfrac{\alpha}{2} + \dfrac{1}{2} & \left(\dfrac{-\alpha-1}{\kappa} \le x \le \dfrac{-\alpha+1}{\kappa}\right), \\ 0.0 & \left(\dfrac{-\alpha+1}{\kappa} \le x\right) \end{cases} \tag{43}$$

where $\kappa = \sqrt{-\alpha\beta}$. $g(x)$ satisfies $g(\sqrt{-\alpha/\beta}) = 0.5$, and requires α to be negative.

In Fig.2, $\Delta x = x - x_{new}$ denotes a reduction of the extent of distribution with decreasing the temperature from T to T_{new} ($T > T_{new}$). The extent of distribution also narrows with increasing α. α_{new} ($\alpha < \alpha_{new}$) which satisfies $g(0.5)_{\alpha} - g(0.5)_{\alpha_{new}} = \Delta x$ is obtained as

$$\alpha_{new} = -\left\{\sqrt{-\alpha} + \sqrt{-\alpha\beta_{new}}\left(\frac{1}{\sqrt{\beta}} - \frac{1}{\sqrt{\beta_{new}}}\right)\right\}^2, \tag{44}$$

where $\beta = 1/T$ and $\beta_{new} = 1/T_{new}$. Thus, taking that T to the temperature at which previous DA was executed and T_{new} to the next temperature, a covariance of α_k's distribution is defined as

$$\Delta\alpha = \alpha_{new} - \alpha. \tag{45}$$

5.2. Initial estimation of α_k and annealing temperature

Before executing DA, it is very important to estimate the initial values of α_ks and the initial annealing temperature in advance.

From Fig.1, distances between a data point and cluster centers are averaged as

$$L_k = \frac{1}{c}\sum_{i=1}^{c} \|\mathbf{x}_k - \mathbf{v}_i\|, \tag{46}$$

and this gives

$$\alpha_k = -\beta(L_k)^2. \tag{47}$$

With given initial clusters distributing wide enough, (47) overestimates α_k, so that α_k needs to be adjusted by decreasing its value gradually.

Still more, Fig.1 gives the width of the Fermi-Dirac distribution function as wide as $2(-\alpha + 1)/(\sqrt{-\alpha\beta})$, which must be equal to or smaller than that of data distribution (=2R). This condition leads to

$$2\frac{-\alpha + 1}{\sqrt{-\alpha\beta}} = 2R. \tag{48}$$

As a result, the initial value of β or the initial annealing temperature is roughly determined as

$$\beta \simeq \frac{4}{R^2} \quad \left(T \simeq \frac{R^2}{4}\right). \tag{49}$$

5.3. Deterministic annealing algorithm

The DA algorithm for fuzzy clustering is given as follows:

1 *Initialize*: Set a rate at which a temperature is lowered T_{rate}, and a threshold of convergence test δ_0. Calculate an initial temperature $T_{high}(= 1/\beta_{low})$ by (49) and set a current temperature $T = T_{high}(\beta = \beta_{low})$. Place c clusters randomly and estimate initial α_ks by (47) and adjust them to satisfy the normalization constraint (2).
2 Calculate u_{ik} by (12).

3 Calculate $\mathbf{v}_i$ by (11).
4 Compare a difference between the current objective value $J_{m=1} = \sum_{k=1}^{n}\sum_{i=1}^{c} d_{ik}u_{ik}$ and that obtained at the previous iteration $\hat{J}$. If $\|J_{m=1} - \hat{J}\|/J_{m=1} < \delta_0 \cdot T/T_{high}$ is satisfied, then return. Otherwise decrease the temperature as $T = T * T_{rate}$, and go back to **2**.

6. Combinatorial algorithm of deterministic and simulated annealing

6.1. Simulated annealing

The cost function for SA is

$$E(\alpha_k) = J_{m=1} + S_{FE} + K\sum_{k=1}^{n}\left(\sum_{i=1}^{c} u_{ik} - 1\right)^2, \tag{50}$$

where K is a constant.

In order to optimize each α_k by SA, its neighbor α_k^{new} (a displacement from the current α_k) is generated by assuming a normal distribution with a mean 0 and a covariance $\Delta\alpha_k$ defined in (45).

The SA's initial temperature $T_0(= 1/\beta_0)$ is determined so as to make an acceptance probability becomes

$$\begin{gathered}\exp\left[-\beta_0\{E(\alpha_k) - E(\alpha_k^{new})\}\right] = 0.5 \\ (E(\alpha_k) - E(\alpha_k^{new}) \geq 0)\end{gathered} \tag{51}$$

By selecting α_ks to be optimized from the outside of a transition region in which the membership function changes form 0 to 1, computational time of SA can be shortened. The boundary of the transition region can be easily obtained with the linear approximations of the Fermi-Dirac-like membership function. From Fig.1, data which have distances bigger than $\sqrt{-\alpha_k/\beta}$ from each cluster centers are selected.

6.2. Simulated annealing algorithm

The SA algorithm is stated as follows:

1 *Initialize*: Calculate an initial temperature $T_0(= 1/\beta_0)$ from (51). Set a current temperature T to T_0. Set an iteration count t to 1. Calculate a covariance $\Delta\alpha_k$ for each α_k by (45).
2 Select data to be optimized, if necessary.
3 Calculate neighbors of current α_ks.
4 Apply the Metropolis algorithm to the selected α_ks using (50) as the objective function.
5 If $max < t$ is satisfied, then return. Otherwise decrease the temperature as $T = T_0/\log(t+1)$, increment t, and go back to **2**.

6.3. Combinatorial algorithm of deterministic and simulated annealing

The DA and SA algorithms are combined as follows:

1 *Initialize*: Set a threshold of convergence test δ_1, and an iteration count l to 1. Set maximum iteration counts max_0, max_1, and max_2.
2 Execute the DA algorithm.
3 Set $max = max_0$, and execute the SA algorithm.
4 Compare a difference between the current objective value e and that obtained at the previous iteration $\hat{e}$. If $\|e - \hat{e}\|/e < \delta_1$ or $max_2 < l$ is satisfied, then go to **5**. Otherwise increment l, and go back to **2**.
5 Set $max = max_1$, and execute the SA algorithm finally, and then stop.

7. Experiments 1

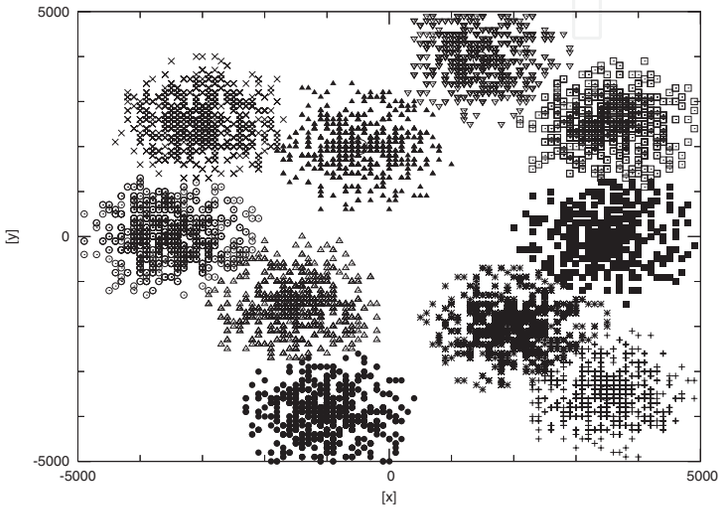


Figure 4. Experimental result 1. (Fuzzy clustering result using DASA. Big circles indicate centers of clusters.)

To demonstrate effectiveness of the proposed algorithm, numerical experiments were carried out. DASA's results were compared with those of DA (single DA).

We set $\delta_0 = 0.5$, $\delta_1 = 0.01$, $T_{rate} = 0.8$, $max_0 = 500$, $max_1 = 20000$, and $max_2 = 10$. We also set R in (48) to 350.0 for experimental data 1∼3, and 250.0 for experimental data 4 [2].

In experiment 1, 11,479 data points were generated as ten equally sized normal distributions. Fig.4 shows a fuzzy clustering result by DASA. Single DA similarly clusters these data.

In experiment 2-1, three differently sized normal distributions consist of 2,249 data points in Fig.5-1 were used. Fig.5-1(0) shows initial clusters obtained by the initial estimation of α_ks and the annealing temperature. Fig.5-1(1)∼(6a) shows a fuzzy clustering process of DASA. At the high temperature in Fig.5-1(1), as described in 4.3, the membership functions were widely distributed and clusters to which a data belongs were fuzzy. However, with decreasing of the temperature (from Fig.5-1(2) to Fig.5-1(5)), the distribution became less and less fuzzy. After executing DA and SA alternately, the clusters in Fig.5-1(6a) were obtained. Then, data to be optimized by SA were selected by the criterion stated in the section 4, and SA was executed. The final result of DASA in Fig.5-1(6b) shows that data were desirably clustered. On the contrary, because of randomness of the initial cluster positions and hardness of good estimation of the initial α_ks, single DA becomes unstable, and sometimes gives satisfactory

[2] These parameters have not been optimized particularly for experimental data.

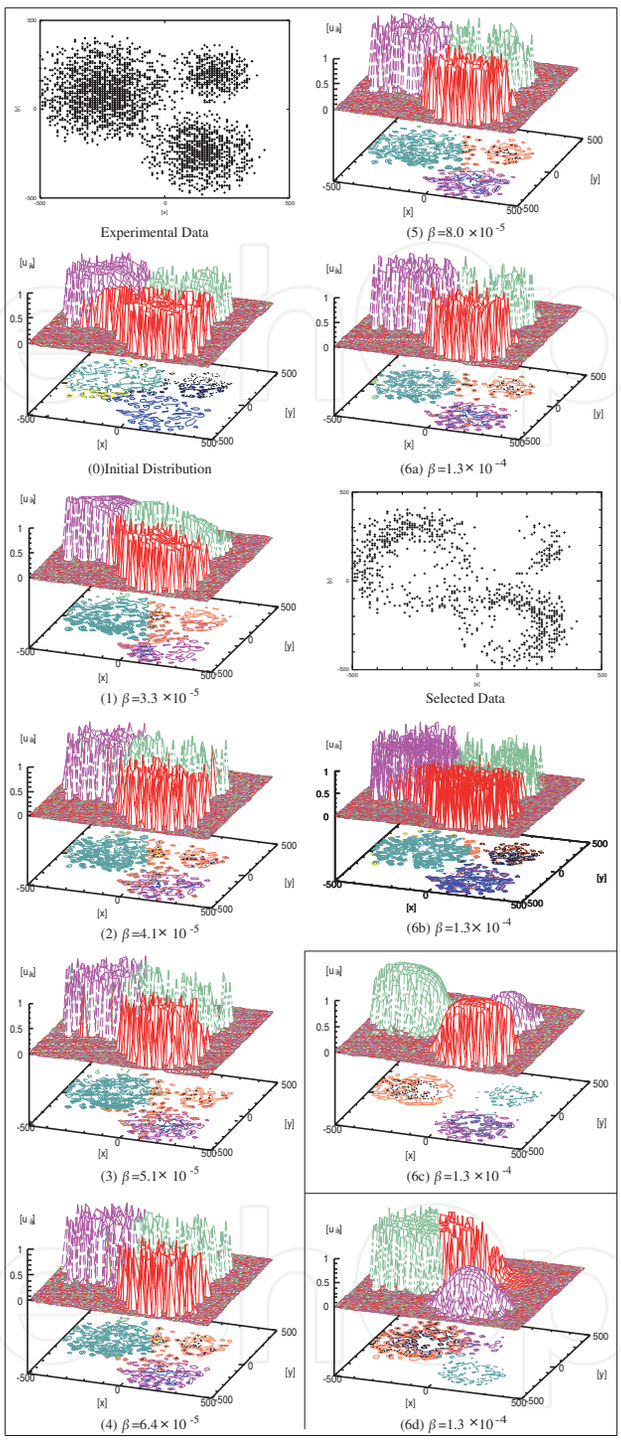


Figure 5-1. Experimental result 2-1. (Fuzzy clustering result by DASA and single DA. "Experimental Data" are given data distributions. "Selected Data" are data selected for final SA by the selection rule. (1)∼(6a) and (6b) are results using DASA. (6c) and (6d) are results using single DA (success and failure, respectively). Data plotted on the xy plane show the cross sections of u_{ik} at 0.2 and 0.8.)

results as shown in Fig.5-1(6c) and sometimes not as shown in Fig.5-1(6d). By comparing Fig.5-1(6b) to (6c), it is found that, due to the optimization of α_ks by SA, the resultant cluster shapes of DASA are far less smooth than those of single DA.

Changes of the costs of DASA ($J_{m=1} + S_{FE}$ for DA stage and (50) for SA stage (K was set to 1×10^{15} in (50)), respectively) are plotted as a function of iteration in Fig.5-2, and the both costs decreases with increasing iteration. In this experiment, the total iteration of SA stage was about 12,500, while that of DA stage was only 7. Accordingly, the amount of simulation time DASA was mostly consumed in SA stage.

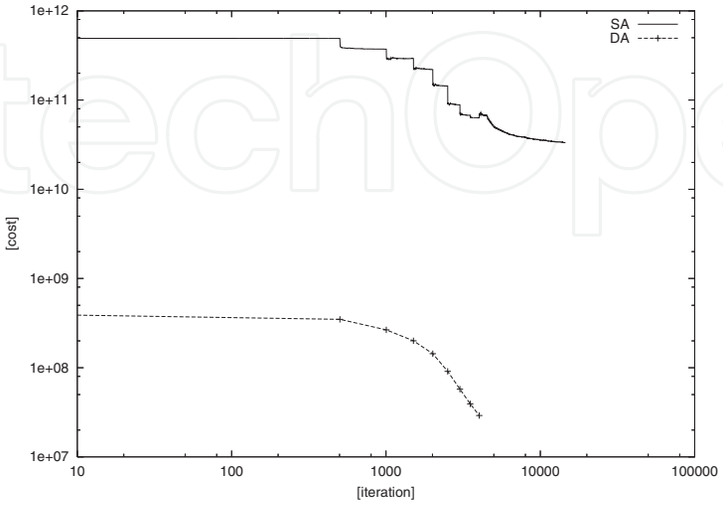


Figure 5-2. Experimental result 2-1. (Change of the cost of DASA as a function of iteration. $J_{m=1} + S_{FE}$ for DA stage and $J_{m=1} + S_{FE} + K\sum_{k=1}^{n}\left(\sum_{i=1}^{c} u_{ik} - 1\right)^2$ for SA stage, respectively.)

In experiment 2-2, in order to examine effectiveness of SA introduced in DASA, experiment 2 was re-conducted ten times as in Table 1, where *ratio* listed in the first row is a ratio of data optimized at SA stage. "UP" means to increase *ratio* as $1.0 - 1.0/t$ where t is a number of execution times of SA stage. Also, "DOWN" means to decrease *ratio* as $1.0/t$. Results are judged "Success" or "Failure" from a human viewpoint [3]. From Table 1, it is concluded that DASA always clusters the data properly if *ratio* is large enough ($0.6 < ratio$), whereas, as listed in the last column, single DA succeeds by 50%.

	DASA					DA
ratio	0.3	0.6	1.0	UP	DOWN	
Success	6	9	10	6	7	5
Failure	4	1	0	4	3	5

Table 2. Experimental result 2-2. (Comparison of numbers of successes and failures of fuzzy clustering using DASA for $ratio = 0.3, 0.6, 1.0, 1.0 - 1.0/t$(UP), $1.0/t$(DOWN) and single DA. (t is a number of execution times of SA stage))

In experiments 3 and 4, two elliptic distributions consist of 2,024 data points, and two horseshoe-shaped distributions consist of 1,380 data points were used, respectively. Fig.5 and 6 show DASA's clustering results. It is found that DASA can cluster these data properly. In experiment 3, a percentage of success of DASA is 90%, though that of single DA is 50%. In experiment 4, a percentage of success of DASA is 80%, though that of single DA is 40%.

[3] No close case was observed in this experiment.

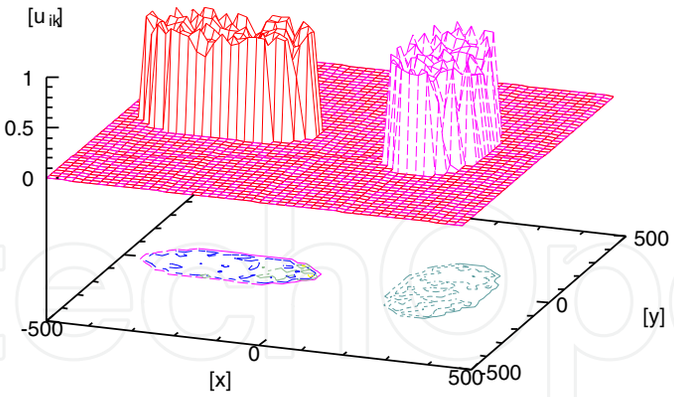


Figure 6. Experimental result 3. (Fuzzy clustering result of elliptic distributions using DASA. Data plotted on the xy plane show the cross sections of u_{ik} at 0.2 and 0.8.)

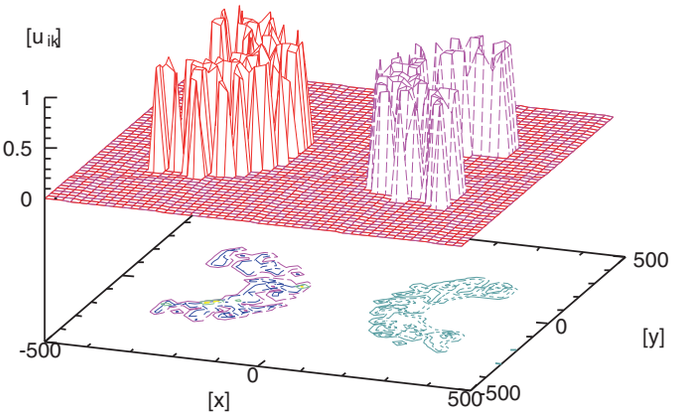


Figure 7. Experimental result 4. (Fuzzy clustering result of horseshoe-shaped distributions using DASA. Data plotted on the xy plane show the cross sections of u_{ik} at 0.2 and 0.8.)

These experimental results demonstrate the advantage of DASA over single DA. Nevertheless, DASA suffers two disadvantages. First, it takes so long to execute SA repeatedly that, instead of (10), it might be better to use its linear approximation functions as the membership function. Second, since α_ks differ each other, it is difficult to interpolate them.

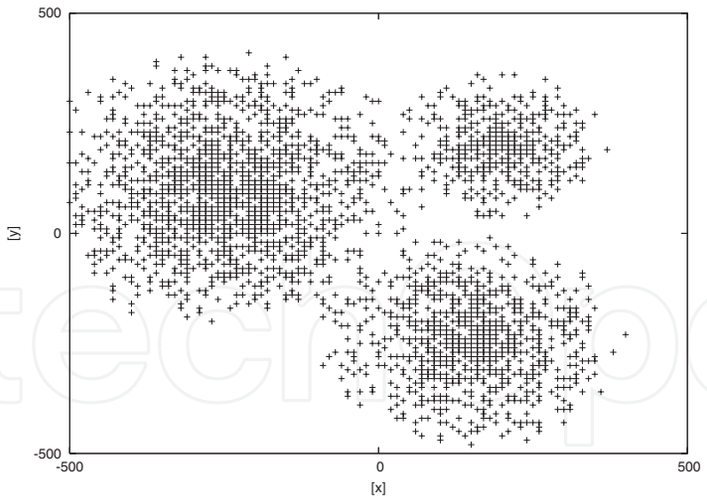
8. Experiments 2

8.1. Interpolation of membership function

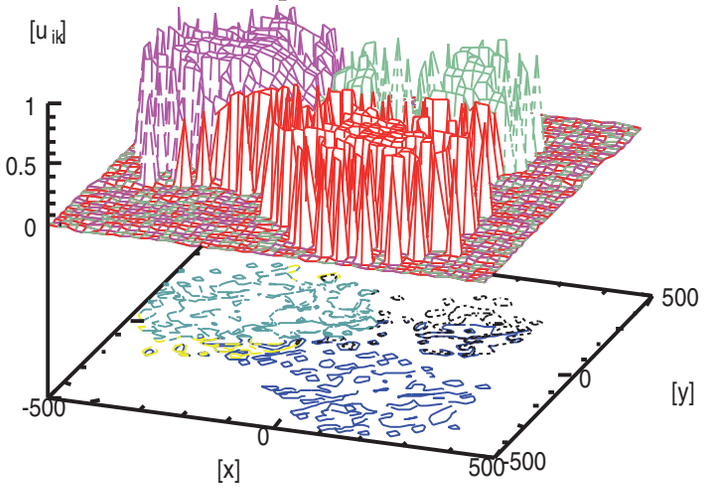
DASA suffers a few disadvantages. First, it is not necessarily easy to interpolate α_k or u_{ik}, since they differ each other. Second, it takes so long to execute SA repeatedly.

A simple solution for the first problem is to interpolate membership functions. Thus, the following step was added to the DASA algorithm.

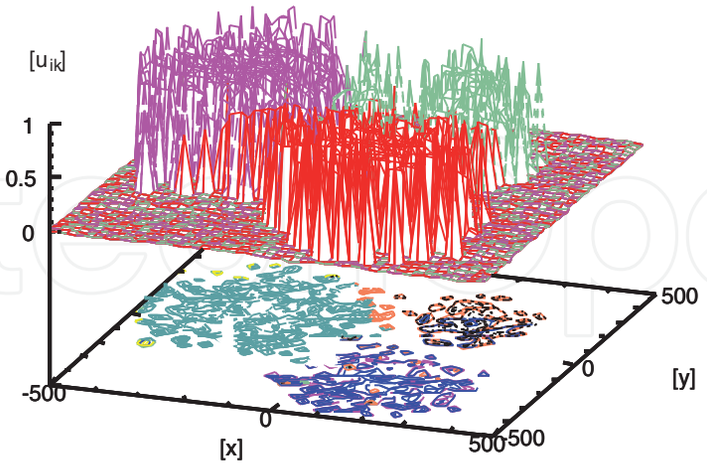
6 When a new data is given, some neighboring membership functions are interpolated at its position.



(a)Experimental data.



(a)Initial distribution.



(b)Final distribution.

Figure 8. Experimantal data and membership functions obtained by DASA.(Data plotted on the xy plane show the cross sections of u_{ik} at 0.2 and 0.8)

To examine an effectiveness of interpolation, the proposed algorighm was applied to experimental data shown in Fig.8(a). For simplicity, the data were placed on rectangular grids on the xy plane.

First, some regional data were randomly selected from the data. Then, Initial and final memberhip functions obtained by DASA are shown in Figs.8(b) and (c) respectively.

After that, remaining data in the region were used as test data, and at each data point, they were interpolated by their four nearest neighboring membership values. Linear, bicubic and fractal interpolation methods were compared.

Prediction error of linear interpolation was 6.8%, and accuracy was not enough. Bicubic interpolation[18] also gave a poor result, because its depends on good estimated gradient values of neighboring points. Accordingly, in this case, fractal interpolation[17] is more suitable than smooth interpolation methods such as bicubic or spline interpolation, because the membership functions in Figs.8(c) are very rough.

The well-known InterpolatedFM (Fractal motion via interpolation and variable scaling) algorithm [17] was used in this experiment. Fractal dimension was estimated by the standard box-counting method [25]. Figs.9(a) and 3(b) represent both the membership functions and their interpolation values. Prediction error (averaged over 10 trials) of fractal interpolation was 2.2%, and a slightly better result was obtained.

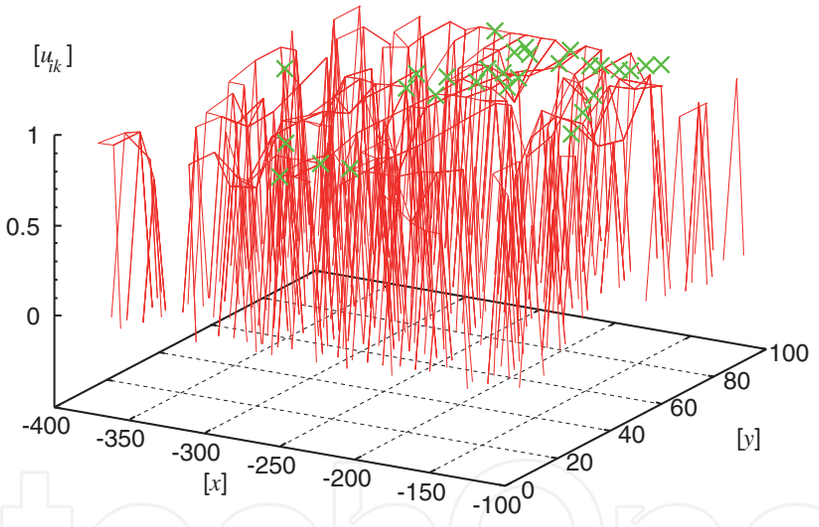


Figure 9. Plotted lines show the membership functions obtained by DASA . The functions are interpolated by the InterpolatedFM algorithm. Crosses show the interpolated data.

9. Conclusion

In this article, by combining the deterministic and simulated annealing methods, we proposed the new statistical mechanical fuzzy c-means clustering algorithm (DASA). Numerical experiments showed the effectiveness and the stability of DASA.

However, as stated at the end of **Experiments**, DASA has problems to be considered. In addition, a major problem of the fuzzy c-means methodologies is that they do not give a number of clusters by themselves. Thus, a method such as [28] which can determine a number of clusters automatically should be combined with DASA.

Our future works also include experiments and examinations of the properties of DASA, especially on an adjustment of its parameters, its annealing scheduling problem, and its applications for fuzzy modeling[29].

However, DASA has problems to be considered. One of them is that it is difficult to interpolate membership functions, since their values are quite different. Accordingly, the fractal interpolation method (InterpolationFM algorithm) is introduced to DASA and examined its effectiveness.

Our future works include experiments and examinations of the properties of DASA, a comparison of results of interpolation methods (linear, bicubic, spline, fractal and so on), an interpolation of higher dimensional data, an adjustment of DASA's parameters, and DASA's annealing scheduling problem.

Author details

Makoto Yasuda
Gifu National College of Technology, Japan

10. References

[1] E. Aarts and J. Korst, "Simulated Annealing and Boltzmann Machines", Chichester: John Wiley & Sons, 1989.
[2] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", New York: Prenum Press, 1981.
[3] B.P.Buckles and F.E.Petry, "Information-theoretical characterization of fuzzy relational database", IEEE Trans. Systems, Man and Cybernetics, vol.13, no.1, pp.74-77, 1983.
[4] J. Buhmann and H. K*ü*hnel, "Vector quantization with complexity costs", IEEE Trans. Information Theory, vol.39, no.4, pp.1133-1143, 1993.
[5] A.Corana, M.Marchesi, C.Martini, and S.Ridella, "Minimizing multimodal functions of continuous variables with the simulated annealing algorithm", ACM Trans. on Mathematical Software, vol.13, no.3, pp.262-280, 1987.
[6] A. DeLuca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory", *Information and Control*, vol.20, pp.301–312, 1972.
[7] A.P.Dempster, N.M.Laird, and D.B.Rubin, "Maximum likelihood from incomplete data via the EM algorithms", Journal of Royal Stat. Soc., Series B, vol.39, pp.1-38, 1977.
[8] W. Greiner, L. Neise, and H. St*ö*cker, "Thermodynamics and Statistical Mechanics", New York: Springer-Verlag, 1995.
[9] T. Hofmann and J. Buhmann, "Pairwise data clustering by deterministic annealing," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.19, pp.1-14, 1997.
[10] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by simulated annealing", Science, vol.220, pp.671-680, 1983.
[11] R.-P. Li and M. Mukaidono, "A Maximum entropy approach to fuzzy clustering", Proc. of the 4th IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE/IFES'95), pp.2227-2232, 1995.
[12] M. Menard, V. Courboulay, and P. Dardignac, "Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatistics", Pattern Recognition, vol.36, pp.1325–1342, 2003
[13] D. Miller, A.V. Rao, K. Rose, and A. Gersho, "A global optimization technique for statistical classifier design", IEEE Trans. Signal Processing, vol.44, pp.3108-3122, 1996.

[14] S. Miyamoto and M. Mukaidono, "Fuzzy c-means as a regularization and maximum entropy approach", Proc. of the 7th Int. Fuzzy Systems Association World Congress, vol.II, pp.86-92, 1997.
[15] N.R. Pal and J.C. Bezdek, "Measuring fuzzy uncertainty", IEEE Trans. Fuzzy Systems, vol.2, no.2, pp.107-118, 1994.
[16] N.R. Pal, "On quantification of different facets of uncertainty", Fuzzy Sets and Systems, vol.107, pp.81-91, 1999.
[17] H.-O. Peitgen, et.al., *The science of fractal images*, Springer-Verlag, 1988
[18] W.H. Press, S.A. Teukolsky, W.T. Vetteriling, and B.P. Flannery, *Numerical Recipes in C++*, Cambridge University Press, 2002.
[19] L. E. Reichl, *A Modern Course in Statistical Physics*, New York: John Wiley & Sons, 1998.
[20] K. Rose, E. Gurewitz, and B.C. Fox, "A deterministic annealing approach to clustering", Pattern Recognition Letters, vol.11, no.9, pp.589-594, 1990.
[21] K. Rose, E. Gurewitz, and G.C. Fox, "Constrained clustering as an optimization method", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.15, no.8, pp.785-794, 1993.
[22] P.Siarry, "Enhanced simulated annealing for globally minimizing functions of many-continuous variables", ACM Trans. on Mathematical Software, vol.23, no.2, pp.209-228, 1997.
[23] D.Tran and M.Wagner, "Fuzzy entropy clustering", Proc. of the 9th IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE2000), vol.1, pp.152-157, 2000.
[24] C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics*, Journal of Statistical Phys., vol.52, pp.479–487, 1988.
[25] R. Voss, *Random fractals: characterization and measurement*, Plenum Press, 1986.
[26] P.R.Wang, "Continuous optimization by a variant of simulated annealing", Computational Optimization and Applications, vol.6, pp.59-71, 1996.
[27] M. Yasuda, T. Furuhashi, and S. Okuma, *Statistical mechanical analysis of fuzzy clustering based on fuzzy entropy*, IEICE Trans. Information and Systems, Vol.ED90-D, No.6, pp.883-888, 2007.
[28] M. Yasuda and T. Furuhashi, *Fuzzy entropy based fuzzy c-means clustering with deterministic and simulated annealing methods*, IEICE Trans. Information ans Systems, Vol.ED92-D, No.6, pp.1232-1239, 2009.
[29] M. Yasuda and T. Furuhashi, *Statistical mechanical fuzzy c-means clustering with deterministic and simulated annealing methods*, Proc. of the Joint 3rd Int. Conf. on Soft Computing and Intelligent Systems, in CD-ROM, 2006.
[30] M. Yasuda, *Entropy based annealing approach to fuzzy c-means clustering and its interpolation*, Proc. of the 8th Int. Conf. on Fuzzy Sysmtes and Knowledge Discovery, pp.424-428, 2011.
[31] S.D. Zenzo and L. Cinque, "Image thresholding using fuzzy entropies", IEEE Trans. Systems, Man and Cybernetics-Part B, vol.28, no.1, pp.15-23, 1998.