Bayesian SOM based Fuzzy Clustering

JunHyeog Choi*, SungHae Jun**, KeeWook Rim***, JungHyun Lee****

*Division of Computer Science, Kimpo College
**Dept. of Statistics, Chongju Univ.
***Knowledge & Industrial Engineering Dept., SunMoon Univ.
****Dept. of Computer Science & Engineering, Inha Univ.

*San 14-1, Ponae-ri, Wolgot-myeon, Gimpo-si, Gyonggi, 415-873, Rep. of Korea
**Kalsan-ri, Tangjeong-myeon, Asna-si, Chungnam, 336-708, Rep. of Korea
***36, Naedok-dong, Sangdang-gu, Cheongju-si, Chungbuk, 360-764, Rep. of Korea
****253, Younghyun-dong, Nam-gu, Incheon, 402-751, Rep. of Korea
email : {jhchoi@kimpo.ac.kr, shjun@cju.ac.kr, rim@sunmoon.ac.kr, jhlee@inha.ac.kr }

***Abstract* – There have been introduced lots of clustering algorithms. Among them, K-means and hierarchical clustering methods especially require an initial number of clusters to be determined. But this determination of cluster size is subjective and the accuracy of the clustering results depends on the initial cluster size. In this paper, we'll determine the optimal number of clusters using fuzzy set from Bayesian learning based SOM (Self Organizing Maps). Unlike the general fuzzy logic approach, these membership functions of fuzzy set are not subjective but objective because of Bayesian learning from given training data. This approach contributes to the performance of clustering results. We'll propose a Bayesian SOM based Fuzzy Clustering to explain this theory in this paper.**

I. INTRODUCTION

We live in an age of uncertainty. The same issue can be explained differently by different points of view, and the same point of view can have different strengths. This kind of uncertainty is one of problems to be solved in artificial intelligence. We need a solution for the uncertainty that can't be proved either to be true or false. One of the trial to solve the uncertainty is the fuzzy theory of Zadeh[9]. If X is a collection of objects denoted generally by x, then a fuzzy set A in X is a set of x and its membership function. The membership function expresses value from 0 to 1 as degree of truth that maps X to A. But in this theory, we have a question how to determine the membership function. Currently, the answer is to decide the membership function subjectively. But it can make the problems more uncertain in the artificial intelligence area that should solve the uncertain problems. In this paper, we'll propose an objective method to determine the membership function with experience using Bayesian inference. In section 2, we'll introduce fuzzy logic for uncertainty and fuzzy clustering, and propose Bayesian SOM(Self Organizing Maps) based fuzzy system. In section 3, experiments and some results for the proposed methods will be introduced. Finally, section 4 will give conclusion of this paper and the direction of future study.

II. BAYESIAN SOM BASED FUZZY CLUSTERING

A. Clustering and number of cluster

The cluster is a set of adjacent objects in training data. Objects in the same cluster have close similarity and objects in other clusters have dissimilarity. We use distance as a measure of similarity between objects. The first problem to consider in clustering is to determine the number of clusters. K-means method requires an initial number of clusters and hierarchical clustering technique also requires an optimal number of clusters for stopping clustering process. But it is hard to find any objective algorithm to determine the initial cluster size, and most of them are determined subjectively. So we'll propose a Bayesian SOM based Fuzzy Clustering algorithm to determine the optimal cluster size.

B. Fuzzy Set and the number of cluster

Use Let X is a nonempty set and x is an element of X. A fuzzy set A is defined as follows.

$$A = \{(x, \mu_A(x)) \mid x \in X\} \quad (1)$$

Where $\mu_A(x)$ is a membership function that expresses a degree of inclusion x into A. In this paper, fuzzy set is used to determine the number of clusters in clustering. This membership function of fuzzy set for clustering is computed repeatedly by Bayesian SOM from given training data. That is, X becomes a set of all possible cluster sizes and A becomes a fuzzy set of appropriate cluster size. And $\mu_A(x)$ becomes a membership function for each possible cluster size. Therefore the element with the largest membership function in fuzzy set A is determined as an optimal cluster size.

C. Bayesian approach to SOM based fuzzy system

SOM[3] is a clustering algorithm and unsupervised neural networks. SOM requires a size of feature maps though it does not determine the number of clustering as K-means and hierarchical clustering methods do. Each node of the output layer achieves clustering by competitive learning from training data. Each object crisply belongs to only one exclusive cluster after the last training. And the clustering result is only one type because the weights have fixed values in nodes of SOM after final training. This result is usually not optimal[1],[7],[8] and it is impossible to repeat the different experiments to determine membership function of fuzzy set. In this paper, we'll get a fuzzy set with repeated experiments by using Bayesian inference[2],[4] that consists of prior probability distribution, posterior probability distribution, and likelihood distribution to SOM. The proposed Bayesian SOM updates parameters of probability distribution without having the fixed values of weights on each node of output layer. This strategy makes it possible to create the membership function by performing repeated experiments with same data to get different results. The proposed method doesn't always offer same results for the same training data because it uses a random number from the last updated distribution for clustering. The membership function of fuzzy set is determined by Bayesian learning[5] based SOM that computes a posterior by combining prior and likelihood. The proposed algorithm in this paper is composed of four phases. The first one is an initial phase. In this phase, standardization for the input data is performed to use the Euclidean distance that computes distance input data and weights which can be used as a measurement to determine winner node. Generally SOM is normalized from 0 to 1 while the proposed algorithm changes the input data to the standardized data that follow the Gaussian distribution of mean 0 and variance 1 to combine proposed distribution. And the size of feature maps is decided in this phase. This decision can be subjective. But it is relatively objective compared to the K-means or hierarchical clustering that requires an initial number of clusters in advance. For example, if the size of the feature maps is determined as 5×5, the results of clusters are from 1 to 25 optimally. Of course if the size of feature maps grows, the number of the optimal cluster can grows accordingly. But the advantage of SOM is that it allows objective clustering without exact information on clustering. And we determine prior probability distribution of the weights of the nodes in feature maps. A Gaussian distribution with mean 0 and variance 1 is used because input data are standardized and we use an Euclidean distance as a similarity measure. In second phase, the distance between input data and weights is measured and the winner node is determined as minimum distance node. Next phase is the parameters updating step of weights distribution. Yet the parameters update of weights distribution is limited to the winner node. This learning is repeated until the given stopping rules are satisfied. Generally the given sopping rules are determined by the number of iteration of learning data and the tolerance that has update range for the parameters of weights distribution. In the last phase, we find fuzzy set about optimal number of clusters through repeated experiments using the final updated weights distribution of feature maps. We summarized the proposed algorithm in this paper as follows.

(Bayesian SOM based Fuzzy System algorithm)

Phase1: Initialize, (n: data size)

1.1 Normalization of input vectors

$x_i^{normal} = \left(\frac{x_{i1}-\mu_1}{\sigma_1}, \cdots, \frac{x_{ip}-\mu_{1p}}{\sigma_p}\right) = \left(x_{i1}^{normal}, \ldots, x_{ip}^{normal}\right)$

$x_i^{normal} \sim N(0,1)$, $i = 1, \ldots n$; likelihood

1.2 Prior of weights

1.2.1 determine the distribution type of $f(\cdot)$

$w \sim f(\theta)$, optionally, $\theta \sim g(\varphi)$: φ is the hyper-parameter of θ

Phase2: Determine winner node

(m: feature map dimension)

2.1 Weights sampling from current prior

2.2 Compute the Euclidean distance of x_i^{normal} and w_j

$$dist(x_i^{normal}, w_j) = \sqrt{(x_{i1}^{normal} - w_{j1})^2 + \cdots + (x_{ip}^{normal} - w_{jp})^2}$$

P: the dimension of input vector

$i = 1, \ldots, n$, $j = 1, \ldots, m$

2.3 Determine winner node

w_k is winner node,

if $dist(x, w_k) < dist(x, w_j)$

Phase3: Update distribution of weights

3.1 Compute posterior of winner node using Bayes' rule

3.2 Replace current posterior with new prior

Repeat phase2 and phase3 until given conditions are satisfied

Phase4: Extract Fuzzy Set for number of Clusters

4.1 Repeat experiments until given number

4.2 Determine the membership function of fuzzy set

III. THE DESIGN AND RESULTS OF EXPERIMENTS

We performed the Bayesian SOM based Fuzzy Clustering with the existing popular data for clustering and discrimination. We used Microsoft Visual C++ 6.0 as a compiler in Microsoft Windows 98 on P-II 350Mhz 1 CPU.

A. Design of experiments

Use The prior probability distribution for each output node is Gaussian distribution as (2).

$$w \sim N(\mu, \sigma_w^2) \ , \quad f(w) = \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{(w-\mu)^2}{2\sigma_w^2}\right) \tag{2}$$

Mean, μ and variance, σ_w^2 are 0 and 1 respectively. We used them to match the scale of input data and weights because the proposed algorithm computes and uses the Euclidean distance between each weight from this distribution and standardized input data. The likelihood distribution of input value, x is a random sample from Gaussian distribution as (3).

$$x \sim N(w, \sigma_x^2) \ , \quad l(x \mid w) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-w)^2}{2\sigma_x^2}\right) \tag{3}$$

Mean and variance in (3) are 0 and 1 because input data are standardized in advance. We compute a posterior probability distribution as (4) using Bayes' rule. Also weights of output node are generated from current posterior probability distribution.

$$\begin{aligned} P(w \mid x) &\propto f(w) l(x \mid w) \\ &\propto \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{(w-\mu)^2}{2\sigma_w^2}\right) \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-w)^2}{2\sigma_x^2}\right) \\ &\propto \exp\left(-\frac{(w - \frac{\sigma_x^2 \mu + \sigma_w^2 x}{\sigma_x^2 + \sigma_w^2})^2}{2\frac{\sigma_x^2 \sigma_w^2}{\sigma_x^2 + \sigma_w^2}}\right) \end{aligned} \tag{4}$$

The posterior probability distribution is used in repeat training as an updated prior distribution. We get the final updated distribution by repeating such Bayesian learning.

B. Membership function determination using repeated experiments: Experiments and results using Class Identification data

Use Glass Identification[10] data by German for the proposed algorithm are composed of 214 items. Input variables are composed of 1 refractive index variable and 8 natural disposition variables (Sodium, Magnesium, Aluminum, Silicon, Potassium, Calcium, Barium and Iron). 1 target variable represents 6 class types. The dimension of feature maps is determined as 9×5 considering the 9 input variables. We created fuzzy set by 10 and 100 repeated experiments in final

model and compared the convergence of membership function according to the number of repetition. Table 1 shows initial weights from prior probability distribution of equation (2).

Table 1. Initial weights of each node in 9×5 feature maps

Node	w_{i1}	w_{i2}	w_{i3}	w_{i4}	w_{i5}	w_{i6}
1	1.2635	1.1588	1.3018	1.6172	0.2551	1.4735
2	0.0486	0.0199	0.5069	0.8673	0.1423	0.7257
3	1.1491	1.3966	1.1179	0.5209	0.4254	0.2286
…	…	…	…	…	…	…
43	0.5203	0.2377	0.3195	0.6826	1.1466	2.7261
44	1.3180	0.3247	0.7505	1.5026	0.1939	2.0893
45	1.0545	0.1036	0.3307	0.5557	0.2043	0.6425

Table 2 shows the mean and variance of Guassian distribution that each of the final node updated by Bayesian SOM-based Fuzzy Clustering belongs to. The repeated experiments perform clustering by generating weight of each node from this distribution. We get a different result from each experiment because each experiment uses different weights from distribution. The fuzzy set is generated with the different result of each experiment heuristically.

Table 2. Final updated distribution of each node in 9×5 feature maps (mean(M), variance(V))

Node		w_{i1}	w_{i2}	w_{i3}	w_{i4}	w_{i5}
1	M	0.18	0.96	0.53	0.53	0.30
	V	0.23	0.57	0.11	0.11	0.18
2	M	0.79	0.98	0.84	0.11	0.21
	V	0.06	0.03	0.05	0.25	0.22
3	M	0.09	0.30	1.03	0.70	0.03
	V	0.33	0.19	0.03	0.53	0.29
		…	…	…	…	…
43	M	0.25	0.12	0.13	0.68	0.60
	V	0.20	0.35	0.35	0.52	0.50
44	M	0.60	0.51	0.53	0.49	0.16
	V	0.09	0.48	0.11	0.48	0.24
45	M	0.35	1.11	0.20	1.01	0.84
	V	0.43	0.02	0.38	0.57	0.05

The result of 10 repeated experiments using proposed algorithm in final updated model is shown in Table 3 and Fig. 1. Cluster numbers of 5, 6 and 8 have 20%, the highest percentage.

Table 3. Result of 10 repeated experiments in 9×5 feature

No. of clusters	No. of results	Percentage
4	1	10
5	2	20
6	2	20
7	1	10
8	2	20
10	1	10
11	1	10

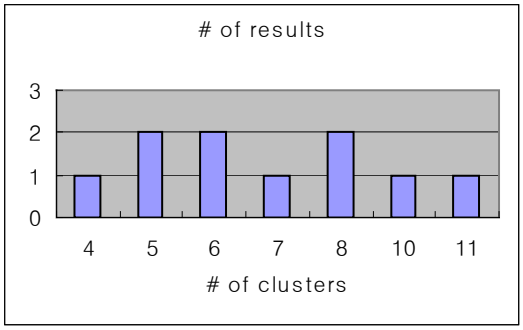

Fig. 1. Result of 10 repeated experiments in 9×5 feature maps

The fuzzy set of 10 repeated experiments can be represented for the collection of objects, X={1,2,3,⋯} as follows.

$$A = \{(4, 0.1), (5, 0.2), (6, 0.2), (7, 0.1), (8, 0.2), (10, 0.1), (11, 0.1)\}$$

According to the above fuzzy set, the membership functions are determined as $\mu_A(4) = 0.1$, $\mu_A(5) = 0.2$, $\mu_A(6) = 0.2$, $\mu_A(7) = 0.1$, $\mu_A(8) = 0.2$, $\mu_A(10) = 0.1$, and $\mu_A(11) = 0.1$. The possibility of optimal cluster size using membership function in training data is found in 5, 6, and 8 with the highest value of 20%. Therefore the optimal numbers of clusters through 10 repeated experiments using proposed algorithm are 5, 6, and 8. Table 4 and Fig. 2 show the clustering results by 100 repeated experiments. The highest percentage of the cluster size is found in number 6 with 21% from the experimental results.

Table 4. Result of 100 repeated experiments in 9×5 feature maps

No. of clusters	No. of results	Percentage
2	1	1.0
3	4	4.0
4	6	6.0
5	18	18.0
6	21	21.0
7	16	16.0
8	16	16.0
10	10	10.0
11	5	5.0
12	1	1.0
13	2	2.0

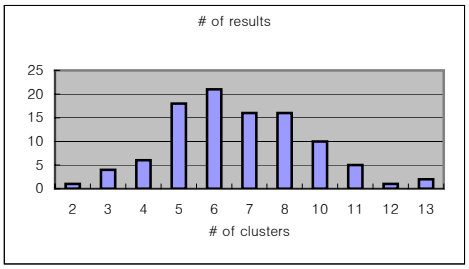

Fig. 2. Result of 100 repeated experiments in 9×5 feature maps

The fuzzy set of 100 repeated experiments can be represented for the collection of objects, X={1,2,3,⋯} as follows,

$$B = \{(2,0.01),(3,0.04),(4,0.06),(5,0.18),(6,0.21),(7,0.16),(8,0.16),(10,0.1),(11,0.05),(12,0.01),(12,0.02)\}$$

According to the above fuzzy set, the membership functions are determined as $\mu_B(2) = 0.01$, $\mu_B(3) = 0.04$, $\mu_B(4) = 0.06$, $\mu_B(5) = 0.18$, $\mu_B(6) = 0.21$, $\mu_B(7) = 0.16$, $\mu_B(8) = 0.16$, $\mu_B(10) = 0.1$, $\mu_B(11) = 0.05$, $\mu_B(12) = 0.01$, and $\mu_B(13) = 0.02$.

The possibility of optimal cluster size using membership function in training data is found in 6 with the highest value of 21%. Therefore the optimal number of cluster for Glass Identification data is 6. In conclusion, the larger the number of repeated experiments is, the easier the determination of fuzzy membership function is.

C. Membership function determination using repeated experiments: Experiments and results using Iris data

We used Iris data[10] by Fisher for another experiment. These data are composed of 4 input variables with sepal length, sepal width, petal length and petal width for an external shape of flower and 1 target variable with setosa, versicolor and virginica for class label. 100 samples from 150 training data using re-sampling technique were used for repeated experiments. We made different fuzzy sets with different feature map sizes and compare them each other. 100 repeated experiments were performed in total. Table 3.5 shows the results of 100 repeated experiments in 4×10 feature maps size. The Cluster number of 3 has 34%, the highest percentage.

Table 5. The result of 4×10 feature maps

No. of clusters	No. of results	Percentage(%)
2	20	20.0
3	34	34.0
4	17	17.0
5	16	16.0
6	9	9.0
7	3	3.0
8	1	1.0

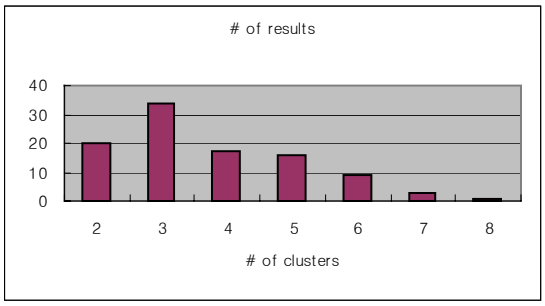

Fig. 3. The result of 4×10 feature maps

The fuzzy set of 100 repeated experiments can be represented for the collection of objects, X={1,2,3,⋯} as follows:

$$C = \{(2,0.2),(3,0.34),(4,0.17),(5,0.16),(6,0.09),(7,0.03),(8,0.01)\}$$

The initial number of clustering for Iris data is determined as 3 that has the largest membership function of 0.34. The experimental results of 4×20 feature maps size with the same training data are shown in Table 6 and Fig. 4.

Table 6. The result of 100 repeated experiments by 100 samples in 4×20 Maps

No. of clusters	No. of results	Percentage(%)
2	9	9.0
3	28	28.0
4	24	24.0
5	19	19.0
6	14	14.0
7	2	2.0
8	3	3.0
9	1	1.0

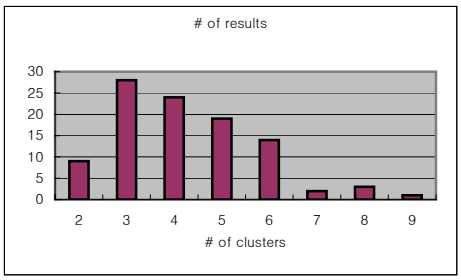

Fig. 4. The result of 100 repeated experiments by 100 samples in 4×20 Maps

The fuzzy set of 100 repeated experiments can be represented for the collection of objects, X={1,2,3,⋯} as follows.

$$D = \{(2,0.09), (3,0.28), (4,0.24), (5,0.19), (6,0.14), (7,0.02), (8,0.03), (9,0.01)\}$$

The initial number of clustering for Iris data is determined as 3 that has the largest membership function value of 0.28. We can find that the proportion of cluster numbers of more than 4 have increased more than 2s or 3s comparing with 4×10 feature maps size. That is, the larger the dimension of feature maps is, the more the number of clusters in final updated model increases. But we find that the order of optimal number of clusters is maintained.

IV. CONCLUSION AND FUTURE WORK

This paper proposed to determine initial number of clusters in clustering using fuzzy system. We proposed objective method of membership function determination of fuzzy set using Bayesian learning unlike other previous subjective methods. This heuristic membership function is used to determine optimal number of clusters. In this paper, we used Bayesian inference using conjugate prior probability distribution. But Bayesian inference with MCMC(Markov Chain Monte Carlo) can be used for complex domain to get more exact results. Computing time needs to be considered in that situation. This is future work.

Acknowledgements

This paper was supported in part by a grants from the Kimpo College.

Reference

[1] Bishop, C. M., Svensen, M., Williams, C. K. I.: GTM : a principled Alternative to the Self Organizing Map. ICANN 96, Volume1112, Bochum, Germany, pp. 165-170, 1996.

[2] Gelman, A., Carlin, J. B., Stern, H. S., Rudin, D. B.: Bayesian Data Analysis, Chapman & Hill, 1995.

[3] Teuvo Kohonen : Self Organizing Maps, Second Edition. Springer, 1997.

[4] Martin A. Tanner: Tools for Statistical inference, Springer, 1996.

[5] Neal, R. M.: Bayesian Learning for Neural Networks. Springer, 1996.

[6] Utsugi, A.: Topology selection for self-organizing maps. Network : Computation in Neural Systems, Vol. 7, No. 4, pp. 727-740, 1996.

[7] Utsugi, A.: Hyperparameter selection for self-organizing maps. Neural Computation, Vol. 9, No. 3, pp. 623-635, 1997.

[8] UCI Machine Learning Repository, http://www1.ics.uci.edu/~mlearn

[9] Zadeh, L.: Fuzzy Sets. Information and Control, 1965.

[10] H. J. Zimmermann: Fuzzy Set Theory and its Applications, Third Edition, 1996.