Controlled-sized Clustering Based on Optimization

Yasunori Endo
Faculty of Eng., Info. & Sys.
University of Tsukuba
Tsukuba, Ibaraki, 305-8573, Japan
Email: endo@risk.tsukuba.ac.jp

Sachiko Ishida
Department of Risk Engineering
Graduate School of Sys. & Info. Eng.
University of Tsukuba
Tsukuba, Ibaraki, 305-8573, Japan
Email: s1520570@u.tsukuba.ac.jp

Naohiko Kinoshita
Research Fellow of JSPS
University of Tsukuba
Tsukuba, Ibaraki, 305-8573, Japan
Email: kinoshita@risk.tsukuba.ac.jp

Abstract—**Clustering is one of unsupervised classification method, that is, it classifies a data set into some clusters without any external criterion. Typical clustering methods, e.g. *k*-means (KM) and fuzzy *c*-means (FCM) are constructed based on optimization of the given objective function. Many clustering methods as well as KM and FCM are formulated as optimization problems with typical objective functions and constraints. The objective function itself is also an evaluation guideline of results of clustering methods. Considering together with its theoretical extensibility, there is the great advantage to construct clustering methods in the framework of optimization. From the viewpoint of optimization, some of the authors proposed an even-sized clustering method based on optimization (ECBO), which is with strengthened constraints of cluster size, and constructed some variations of ECBO. The constraint considered in ECBO is that each cluster size is K or $K+1$. ECBO is based on KM and its algorithm is constructed as iterative optimization. The belongingness of each object to clusters are calculated by the simplex method in each iteration. The numerical experiments show that ECBO has higher classification accuracy than other similar clustering methods. It is considered that ECBO has the advantage in the viewpoint of clustering accuracy, cluster size, and optimization framework than other similar methods. However, the constraint of cluster sizes of ECBO is strict so that it may be inconvenient in case that the partition results, of which each cluster size need not be strictly even, but uneven, is desirable. Moreover, it is expected that new clustering algorithms of which each cluster size can be controlled can deal with more various datasets. In this paper, we first propose two new clustering algorithms based on ECBO. Each cluster size can be controlled in the proposed algorithms. Next, we estimate the new clustering algorithms through some numerical experiments.**

I. Introduction

Clustering is one of the data mining techniques and it classifies a dataset into some clusters automatically. *K*-member clustering (KMC) is one of the clustering method and it classifies a dataset into some clusters of which the size is at least *K*. The following three methods are known as typical KMC methods: greedy *k*-member clustering (GKC) [1], one-pass *k*-means (KM) algorithm for *K*-anonymization (OKA) [2], and clustering-based *K*-anonymity (CBK) [3].

978-1-5090-4917-2/17$31.00 © 2017 IEEE

Conventional KMC methods including GKC, OKA, and CBK have some problems for classifying a dataset into some clusters of which the size is at least *K*. The clusters by GKC and OKA have sometimes no sense of unity. The cluster number is not maximized under the constraint that the size of each cluster is or more than *K* by CBK.

To solve the problem of CBK, two-division clustering for *K*-anonymity of cluster maximization (2DCKM) was proposed and extended by one of the authors which is referred to as extended two-division clustering for *K*-anonymity of cluster maximization (E2DCKM) [4]. Both 2DCKM and E2DCKM are based on CBK, then they obtain final cluster division from iteration of classification of one cluster into two clusters and adjustment of each cluster size. These KMC methods classify a dataset into clusters of which the size is at least *K*. However, the classification accuracy of the above methods is not so high. One of the reason is that those methods is not based on optimization.

A clustering algorithm based on graph theory, which classifies a dataset into some even-sized clusters, was proposed [5]. However, it is also not based on optimization.

Typical clustering methods, e.g. KM and fuzzy *c*-means (FCM) are constructed based on optimization of the given objective function [6]. Many clustering methods as well as KM and FCM are formulated as optimization problems with typical objective functions and constraints. The objective function itself is also an evaluation guideline of results of clustering methods. Considering together with its theoretical extensibility, there is the great advantage to construct clustering methods in the framework of optimization. From the viewpoint of optimization, some of the authors proposed an even-sized clustering method based on optimization (ECBO) [7], [8], which is with more strengthened constraints of cluster size than KMC, and constructed some variations of ECBO. The constraint considered in ECBO is that each cluster size is K or $K+1$. Here we have to notice that the existence of the cluster number

c obviously depends on the dataset size n and K. For example, in case that $n = 10$ and $K = 6$, the cluster number c which satisfies the conditions does not exist. Conversely, K exists for any c ($c < n$). Therefore, a condition of n and K for preventing the case is needed. ECBO is based on KM and its algorithm is constructed as iterative optimization. The belongingness of each object to clusters are calculated by the simplex method in each iteration. The numerical experiments show that ECBO has higher classification accuracy than E2DCKM [7], [8].

It is considered that ECBO has the advantage in the viewpoint of clustering accuracy, cluster size, and optimization framework than other KMC methods. However, the constraint of cluster sizes of ECBO is strict so that it may be inconvenient in case that the partition results, of which each cluster size need not be strictly even, but uneven, is desirable. For example, in case that $n = 10$ and $c = 3$, each cluster size is not 3 or 4, but it is in the range of $[2, 5]$.

In this paper, we first propose two new clustering algorithms based on ECBO, which can control each cluster size. One is the basic algorithm which is referred to as COntrolled-sized Clustering Based on Optimization (COCBO), and another is an extended COCBO. COCBO is based on the KM so that the clustering results strongly depend on the initial value. In KM, KM++ algorithm were proposed to solve the problem [9]. Inspiring KM++, we extend COCBO by introducing the idea of KM++ into COCBO to solve the problem. Next, we estimate the new clustering algorithms through some numerical experiments for artificial and real datasets.

II. Even-sized Clustering Based on Optimization

Let $x \in \Re^p$ be an object, and $X = \{x\}$ be a set of objects. $v \in \Re^p$ and V is a cluster center and a set of cluster centers, respectively. C_i is the i-th cluster. Moreover, let $U = (u_{ki})_{k=1,\ldots,n,\ i=1,\ldots,c}$ be a partition matrix of membership grades. $u_{ki} = 1$ iff x_k is in C_i and $u_{ki} = 0$ iff x_k is not in C_i.

Even-sized clustering classifies datasets into some clusters of which the object number are almost even. Depending on the size of dataset, it is not possible to classify the dataset into completely evenly, then each cluster size is defined as K or $K + 1$. Here, K is a given constant. As mentioned above, the existence of c depends on the dataset size n and the given number K for cluster size. We mention the condition of K later.

Even-sized Clustering Based on Optimization (ECBO) [7], [8] is one of even-sized clustering and it is based on KM. The difference of ECBO from the conventional even-sized clustering is that ECBO classifies datasets in the framework of optimization, that is, ECBO minimizes an objective function under constraints on membership grades and even cluster sizes.

The objective function and the constraints are as follows:

$$\text{minimize} \quad J_{\mathrm{ECBO}}(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \|x_k - v_i\|^2 \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{c} u_{ki} = 1 \quad (k = 1, \cdots, n) \tag{2}$$

$$K \leq \sum_{k=1}^{n} u_{ki} \leq K + 1 \quad (i = 1, \cdots, c) \tag{3}$$

(1) and (2) are the same objective function and constraint as KM. (3) is the constraints on even cluster size.

These equations are linear with u_{ki}, hence the optimal solution of u_{ki} is obtained by the simplex method. The cluster centers v_i can be calculated in the same way of KM.

Before starting the ECBO algorithm, we have to give a constant K or a cluster number c. The relation between the dataset size n, c, and K in ECBO was considered in Ref. [8]. Here, we reconsider more precise relation.

The relation between n, c and K is $K = \left\lfloor \frac{n}{c} \right\rfloor$. If n and c are given, K exists. On the other hand, even if n and K are given, c does not always exist. Thus, it is necessary to satisfy the following relation between n and K:

$$0 < n \leq (K + 1)\frac{n - (n \bmod K)}{K} \tag{4}$$

If (4) holds true,

$$c = \frac{n - (n \bmod K)}{K} \tag{5}$$

We show the ECBO algorithm as Algorithm 1.

Algorithm 1 ECBO

```
Step 1. Give the constants $K$ or $c$.
Step 2. Give the initial cluster centers $V$ randomly.
Step 3. Update $U$ by the simplex method with fixing $V$.
Step 4. Update $V$ by $v_i = \sum_{k=1}^{n} u_{ki}x_k / \sum_{k=1}^{n} u_{ki}$ with fixing
    $U$.
Step 5. If $V$ changes from previous $V$, go back to Step
    3. Otherwise, stop.
```

We often obtain more natural results by ECBO than KMC methods because ECBO classifies datasets in the framework of optimization. However, the constraint (3) is so strict that it may be inconvenient in case that the partition results, of which each cluster size need not be strictly even, but uneven, is desirable. Moreover, it is expected that new clustering algorithms of which each cluster size can be controlled can deal with more various dataset. Therefore, we propose new clustering algorithms to solve the problems in the following section.

III. Proposed Methods

We propose two new clustering algorithms, one is COCBO and another is COCBO++.

A. COntrolled-sized Clustering Based on Optimization (COCBO)

COntrolled-sized Clustering Based on Optimization (COCBO) is based on ECBO and the constraint of cluster size (3) is changed.

The objective function and constraints are as follows:

$$\text{minimize} \quad J_{\text{COCBO}}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\|x_k - v_i\|^2 \qquad (6)$$

$$\text{s.t.} \quad \sum_{i=1}^{c} u_{ki} = 1 \quad (k = 1,\cdots,n) \qquad (7)$$

$$\underline{K} \leq \sum_{k=1}^{n} u_{ki} \leq \overline{K} \quad (i = 1,\ldots,c) \qquad (8)$$

$$\underline{K} \leq K, \quad \overline{K} \geq K + 1$$

(6) and (7) are the same objective function and constraint as KM. (8) is the constraints on cluster size. We cannot change or control the cluster size K of (3) in ECBO. On the other hand, we can control all cluster sizes of (8) in COCBO to change $\underline{K}$ and $\overline{K}$. If the cluster size K is given, the cluster number c is calculated by (5).

We show the COCBO algorithm as Algorithm 2.

Algorithm 2 COCBO

```
Step 1. Give the constants $K$ or $c$. Set $\underline{K}$ and $\overline{K}$ of (8).
Step 2. Give the initial cluster centers $V$ randomly.
Step 3. Update $U$ by the simplex method with fixing $V$.
Step 4. Update $V$ by $v_i = \sum_{k=1}^{n} u_{ki}x_k / \sum_{k=1}^{n} u_{ki}$ with fixing $U$.
Step 5. If $V$ changes from previous $V$, go back to Step 3. Otherwise, stop.
```

B. COntrolled-sized Clustering Based on Optimization++ (COCBO++)

In Algorithm 2 of COCBO, initial values of V are randomly given same as KM. Because COCBO is based on ECBO and ECBO is based on KM, COCBO has the same problem of KM: initial-value-dependence (i.v.d.). That is, results of COCBO depend on the initial value of V strongly. In KM, KM++ algorithm were proposed to solve the i.v.d. problem [9]. Inspiring KM++, we extend COCBO by introducing the idea of KM++ into COCBO to solve the problem.

The objective function and the constraints of COCBO++ are the same as COCBO, i.e., (6), (7), and (8). The difference between COCBO++ and COCBO is the way to choose the initial values of V. In COCBO, V are given randomly. On the other side, V in COCBO++ are chosen following Algorithm 3. After choosing the initial V, Algorithm 2 is done.

Algorithm 3 Initial Selection of V

```
Step 1. $i = 1$. Give a $x$ as a cluster center $v_i$.
Step 2. $i := i + 1$. Choose a $x$ as a cluster center $v_i$ with probability $D(x)^2 / \sum_{x \in X} D(x)^2$. $D(x)$ is the shortest distance from $x$ to the closest center we have already chosen.
Step 3. If $i < c$, go back to Step. 2. Otherwise, stop.
```

Algorithm 4 COCBO++

```
Step 1. Give the constants $K$ or $c$. Set $\underline{K}$ and $\overline{K}$ of (8).
Step 2. Give the initial cluster centers $V$ by Algorithm 3.
Step 3. Update $U$ by the simplex method with fixing $V$.
Step 4. Update $V$ by $v_i = \sum_{k=1}^{n} u_{ki}x_k / \sum_{k=1}^{n} u_{ki}$ with fixing $U$.
Step 5. If $V$ changes from previous $V$, go back to Step 3. Otherwise, stop.
```

TABLE I
Comparing partitions U and V

	partition V	
partition U	same group	different group
same group	a	b
different group	c	d

IV. Numerical Experiments

A. Adjusted Rand Index

In this paper, we use Adjusted Rand Index (ARI) [10]. ARI is well-known as the standard performance evaluation criteria of clustering results. and it means the accuracy rate of classification. We show the formula as Table I and (9).

$$\text{ARI} = \frac{{}_nC_2(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{{}_nC_2{}^2 - [(a+b)(a+c) + (c+d)(b+d)]} \qquad (9)$$

The ARI shows the proximity between two partitions. The more similar the two clusters are, the closer to 1 the ARI is. Conversely, the more different the two clusters are, the closer to 0 the ARI is. If the two partitions are the same, ARI = 1.

B. Artificial Dataset

The artificial dataset consists of a small circle of 155 objects and a big circle of 100 objects (Fig. 1). Each object is generated randomly. We classify the dataset by KM, ECBO, COCBO, and COCBO++ with $c = 2$, and repeat 10 times with different initial values by one-time.

We show the partition results when the ARI is maximum, the relation between the value of the ARI and the objective function, and one between the value of the ARI and the variance as Table II and Fig. 2 to 8.

"Centroid" in figures means a cluster center of each cluster. "Margin" in tables and figures means $\overline{K} - K$ or $K - \underline{K}$. In these numerical experiments, $\overline{K} - K = K - \underline{K}$. "$J$" in tables means the objective function.

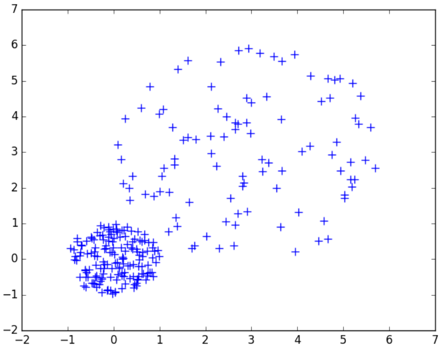

Fig. 1. Artificial Dataset (n = 255)

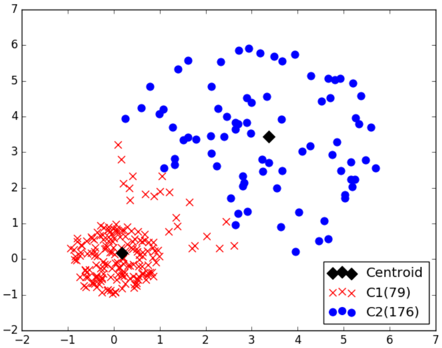

Fig. 2. Partition result of KM

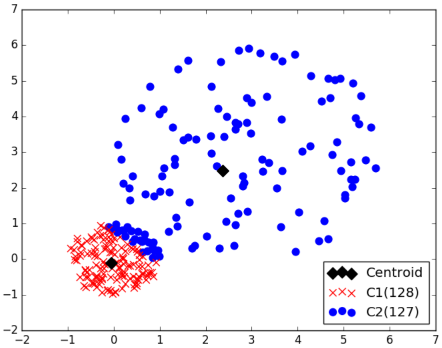

Fig. 3. Partition result of ECBO

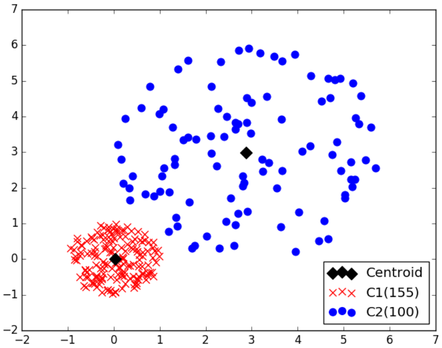

Fig. 4. Partition result of COCBO and COCBO++

TABLE II
Partition results of each method for the artificial dataset

Method	Partition result	Margin	min J
KM	(79,176)	—	486.63
ECBO	(127,128)	—	786.42
COCBO	(100,155)	27	557.01
COCBO++	(100,155)	27	557.01

Method	Variance of J	max ARI
KM	1.52	0.69
ECBO	7.71	0.62
COCBO	52454.44	1.00
COCBO++	1.29E^{-26}	1.00

In KM, the boundary of clusters is pulled to the center of the large circle cluster to divide both the regions equally, and finally, the partition result looks unnatural. In ECBO, the boundary of clusters is pulled to the center of the cluster in which more objects belongs to make both the cluster sizes even, and finally, ARI = 0.62.

On the other side, both the COCBO and COCBO++ classify the dataset into the desirable clusters and ARI = 1.00.

In comparison with the values of the objective functions of ECBO and COCBO, COCBO is smaller than ECBO. The reason is that the constraint of cluster sizes of COCBO is more loosened than ECBO and consequently, COCBO can classifies the dataset into more coherent clusters than ECBO.

The variance of the objective function in COCBO++ is almost 0. The reason is that the way of choice of the initial values is revised and consequently, the partition results are averagely stable in each trial.

C. Real Dataset: Fisher's Iris Dataset

Fisher's Iris dataset [11] consists of 50 objects from each of three species of Iris. Each object has four attributes, i.e., the length and the width of the sepals and petals.

We show the partition results when the ARI is maximum, the relation between the value of the ARI and the objective function, and one between the value of the ARI and the variance as Table III and Fig. 9 to 12.

All cluster sizes of the dataset are equally 50. Therefore, we thought the partition result and the ARI of ECBO is better than COCBO because the constraint of cluster size of ECBO is strict before the numerical experiment. However, the experiment shows that the

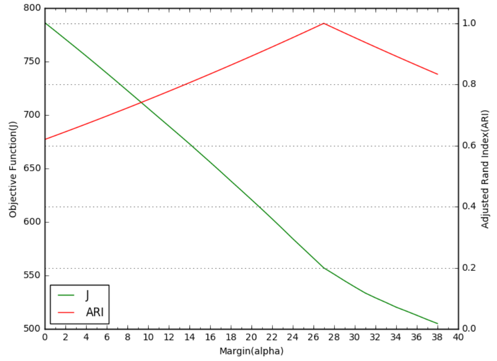

Fig. 5. The relation between the margin, J, and ARI by COCBO for the artificial dataset

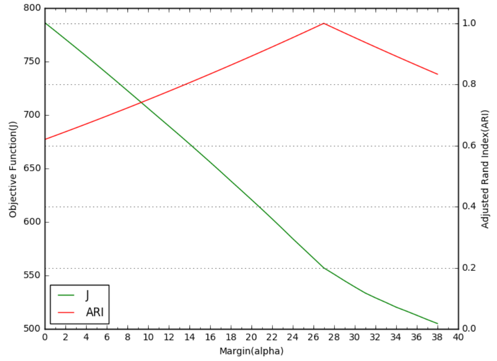

Fig. 6. The relation between the margin, J, and ARI by COCBO++ for the artificial dataset

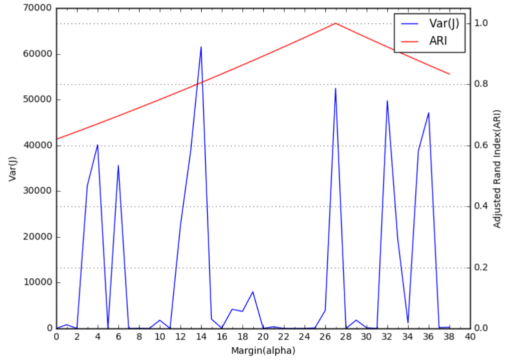

Fig. 7. The relation between the margin, the variance of J, and ARI by COCBO for the artificial dataset

ARI of COCBO is better than ECBO, even though 0.01.

V. Conclusion

In this paper, we proposed two new clustering algorithms based on ECBO, which can control each cluster size. One is COCBO, and another is COCBO++. COCBO++ is the extended algorithm of COCBO by introducing the idea of KM++ into COCBO to solve the i.v.d. problem. Next, we estimated the new clustering algorithms through some numerical experiments of artificial and real datasets.

The proposed methods obtained better partition results, smaller values of the objective functions, and larger values of the ARI than KM and ECBO stably. In particular, when we set adequate $\underline{K}$ and $\overline{K}$, all proposed methods obtained ARI = 1 no matter initial values and

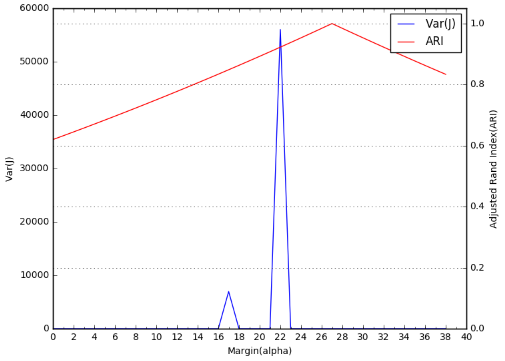

Fig. 8. The relation between the margin, the variance of J, and ARI by COCBO++ for the artificial dataset

TABLE III
Partition results of each method for Fisher's Iris dataset

Method	Partition result	Margin	min J
KM	(35,50,65)	—	80.45
ECBO	(50,50,50)	—	81.40
COCBO	(49,50,51)	1	81.18
COCBO++	(48,50,52)	2	80.94

Method	Variance of J	max ARI
KM	977.14	0.75
ECBO	1143.89	0.79
COCBO	0.01	0.80
COCBO++	0.05	0.79

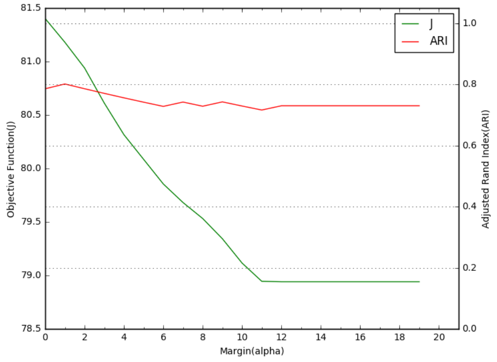

Fig. 9. The relation between the margin, J, and ARI by COCBO for Fisher's Iris dataset

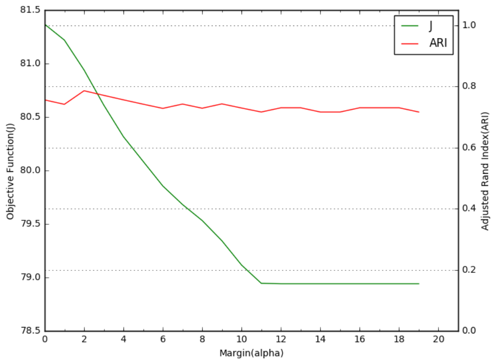

Fig. 10. The relation between the margin, J, and ARI by COCBO++ for Fisher's Iris dataset

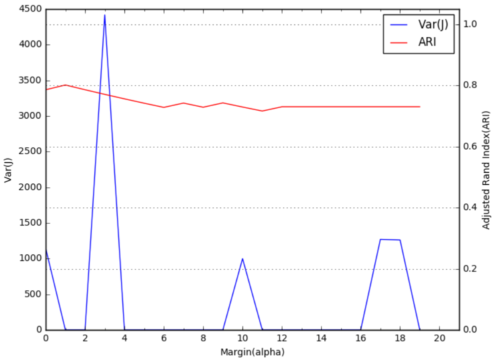

Fig. 11. The relation between the margin, the variance of J, and ARI by COCBO for Fisher's Iris dataset

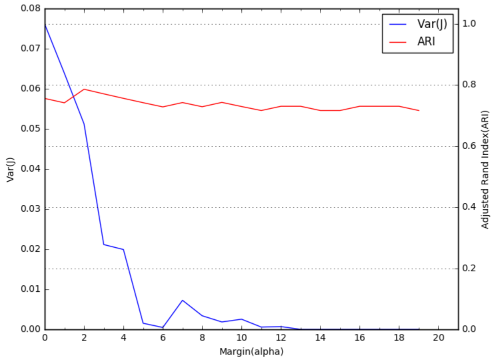

Fig. 12. The relation between the margin, the variance of J, and ARI by COCBO++ for Fisher's Iris dataset

iterative times, while ECBO strongly depends on initial values. The reason is that the constraint of cluster sizes of the proposed methods is more loosened than ECBO.

Moreover in both the artificial and real datasets, the larger the margin is, the more the values of the objective functions decrease toward the right. The reason is that loosening the constraint of cluster sizes makes the values of the objective functions smaller easily. It is obvious the figure of the relation between the margin, the value of the objective function, and the value of the ARI.

It is expected that the proposed clustering algorithms of which each cluster size can be controlled can deal with more various dataset. Hence, in the future, we will deal with not only numerical datasets, but also categorical datasets.

Moreover, we will try to introduce another concept of flexibility except controlled-size. We gave the clustering algorithms flexibility by introducing $\underline{K}$ and $\overline{K}$ into the constraint of cluster sizes in this paper. Instead of the constraints, we will fuzzify the belongingness u_{ki} as flexibility in the forthcoming paper.

Acknowledgment

We would like to thank gratefully and sincerely Professor Emeritus Sadaaki Miyamoto of University of Tsukuba, Japan, and Assistant Professor Yukihiro Hamasuna of Kindai University, Japan, for their advice. This study was supported by JSPS KAKENHI Grant Number JP17K00332.

References

[1] J.-W. Byun, A. Kamra, E. Bertino, N. Li, "Efficient *k*-Anonymization Using Clustering Techniques", Proc. of the 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007), pp.188–200, 2007.

[2] J.-L. Lin, M.-C. Wei, "An efficient clustering method for *k*-anonymization", Proc. of the 2008 international workshop on Privacy and anonymity in information society (PAIS '08), pp.46–50, 2008.

[3] X. He, H. H. Chen, Y. Chen, Y. Dong, P. Wang, Z. Huang, "Clustering-Based *k*-Anonymity", Proc. of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2012), pp.405–417, 2012.

[4] Y. Ogata, Y. Endo, "A Note on the *K*-Member Clustering Problem", The 29th Fuzzy System Symposium (FSS 2013), MB2-2, 2013 (in Japanese).

[5] M.-F. Balcan, V. Nagarajan, E. Vitercik, C. White, "Learning the Best Algorithm for Max-Cut, Clustering, and Other Partitioning Problems", arXiv:1611.04535v1 (2016).

[6] S. Miyamoto, H. Ichihashi, and K. Honda, 'Algorithms for Fuzzy Clustering', Springer, Heidelberg, 2008.

[7] T. Hirano, Y. Endo, N. Kinoshita, Y. Hamasuna, "On Even-sized Clustering Algorithm Based on Optimization", Proc. of Joint 7rd International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on advanced Intelligent Systems (SCIS & ISIS 2014), TP4-3-5-(3), #69, 2014.

[8] Y. Endo, T. Hirano, N. Kinoshita, Y. Hamasuna, "On Various Types of Even-sized Clustering Based on Optimization", The 13th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2016), Springer, LNAI 9880, pp.165–177, 2016.

[9] D. Arthur, S. Vassilvitskii, "*k*-means++: The Advantages of Careful Seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp.1027–1035, 2007.

[10] L. Hubert, P. Arabie, "Comparing Partitions", Journal of Classification 2, pp.193–218, 1985.

[11] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, vol1.7, No.2, pp.179–188, 1936.