Paper:

Proposal of Dynamic Generation of Subordinate Clusters based on Bayesian Information Criterion for Must-link Constrained K-means

Shota Shimizu[1], Shun Sakayauchi, Hiroki Shibata[2], Yasufumi Takama[3]

Graduate School of System Design, Tokyo Metropolitan University, Japan
E-mail: [1]shimizu-shota@ed.tmu.ac.jp, [2]hshibata@tmu.ac.jp, [3]ytakama@tmu.ac.jp

This paper proposes a constrained K-means clustering method that dynamically generates subordinate clusters based on Bayesian Information Criterion. COP K-means, which considers a pair-wise constraints in partition-based clustering, have difficulty handling the case that a Must-link is given to instances located far away from each other. To address this problem, the proposed method generates subordinate clusters that have a Must-link to a master cluster during a clustering process. The final clustering result is obtained by merging the subordinate clusters. In addition, this paper proposes to determine whether to generate subordinate clusters or not based on the Bayesian Information Criterion. This paper evaluates the effectiveness of the proposed method with synthetic datasets.

Keywords: Clustering, Constrained Clustering, Bayesian Information Criterion (BIC)

1. Introduction

This paper proposes a constrained K-means clustering method that dynamically generates subordinate clusters based on Bayesian Information Criterion (BIC). The purpose of clustering is to divide similar data into groups without external criteria. Clustering is generally used for analyzing or mining data without supervised data (labels). However, such an unsupervised learning does not always satisfy the user needs. Therefore constrained clustering methods, which belongs to semi-supervised learning, have been proposed to incorporate user knowledge and feedback into clustering process [1].

Constrained clustering methods are classified in terms of the type of constraints. This paper focuses on the clustering based on pair-wise constraints. Pair-wise constraints are classified into two kinds: Must-link that specifies that two instances must belong to the same cluster and Cannot-link that specifies that two instances must belong to different clusters. COP K-means [1] introduces these pair-wise constraints into K-means [2]. Although it can partition data without changing feature space, it cannot get good partition when two instances that are located far away from each other have a Must-link. Such a situation is supposed to happen when there is a gap between distance on feature space and distance on metric space.

To solve above-mentioned problem, this paper proposes to introduce dynamic generation mechanism of subordinate clusters into COP K-means clustering. Subordinate clusters have its master cluster, to which those are merged when obtaining a final clustering result. Subordinate cluster is generated when a distance between two instances that have a Must-link are far away from each other. A final clustering result is obtained by merging subordinate clusters into those master clusters. If the number of resultant clusters is greater than specified number, agglomerative hierarchical clustering is further applied.

One of the challenges of the proposed method is how to judge two instances are far away from each other. Although setting a distance threshold as a predetermined parameter is one possible approach, different thresholds should be appropriately set for different datasets. To determine whether to generate subordinate clusters without a predetermined distance threshold, this paper proposes to employ BIC. The proposed method determines whether to generate a subordinate cluster based on the difference of BIC when a subordinate cluster is generated for an instance and when the instance is assigned to an existing cluster.

This paper is organized as follows. Section 2 introduces related works about constrained clustering methods. Section 3 describes the proposed clustering algorithm. Section 4 shows the experimental result with synthetic datasets.

2. Related Works

2.1. COP K-means clustering

COP K-means is a constrained clustering based on K-means clustering. This method has two kinds of pair-wise constraints. One is Must-link which makes two instances belong to the same cluster. The other one is Cannot-link which makes two instances not belong to the same cluster. The constraints was introduced to reflect user background knowledge or user

intention. There are cases that the iteration in algorithm never ends because an instance may have multiple constraints that conflict each other. In addition clustering with Cannot-link is generally NP-complete [4]. Futhermore COP K-means cannot appropriately handle a Must-link between two instances that are located far away from each other, because it cannot generate a cluster of non-convex shape.

2.2. X-means Clustering

X-means [5] is also an extended version of K-means, which addresses the problem that K-means needs the number of clusters as an input from a user. X-means consists of two main processes: Improve-Parameters that performs normal K-means clustering and Improve-Structure that estimates the number of clusters. In Improve-Structure process, it judges whether to split a cluster or not based on BIC. That is, it generates clusters dynamically without a parameter. It has been shown that X-means can get better clustering result than K-means.

2.3. U-k-means Clustering

U-k-means [6] is a supervised learning schema for K-means, whose purposes are to decrease influence of initialization and to find optimal number of clusters. The objective function of U-k-means consists of those of K-means and entropy terms. In U-k-means clustering process, the number of clusters is determined automatically on the basis of the initial number determined from the number of data. Similar to X-means, it also generates clusters dynamically. It has been shown that U-k-means is robust to different cluster volume and shapes.

3. Proposed Method

This paper proposes the method to solve the problem that COP K-means cannot appropriately handle a Must-link between two instances that are located far away from each other. The proposed method dynamically generates subordinate clusters during clustering process. Subordinate clusters have its master cluster, to which those are merged when obtaining a final clustering result. BIC is used to determine whether to generate subordinate clusters or not. By introducing subordinate clusters, two instances that relate to but are far away from each other can belong to the same cluster without giving much impact on overall clustering result.

3.1. Dynamic Generation of Subordinate Clusters

The criteria for generating subordinate clusters for a constrained data is based on the difference in the entropy between when the data is assigned to an existing cluster, into which the data has a Must-link, and when a subordinate cluster is generated with that data.

The difference of the entropy ΔI is defined as Eq.(1), in which I_{G} is the entropy of the model when a subordinate cluster is generated and I_{A} is the entropy of the model when the data is assigned to a master cluster.

$$\Delta I = I_{\mathrm{G}} - I_{\mathrm{A}}. \quad \cdots \quad (1)$$

Given the constrained data $d \in D \subset \mathbb{R}^n$ and a cluster C that includes a Must-linked data to d, I_{G} is defined as

$$\begin{aligned} I_{\mathrm{G}} &= \sum_{x \in C} \log p(X = x; \theta_C) + \log p(X = d; \theta_{C_{\mathrm{s}}}) \\ &\quad - q \log(|C| + 1) \\ &= L(C) - q \log(|C| + 1), \end{aligned} \quad (2)$$

where $p(X;\theta)$ is the occurrence probability function, for which in this paper employs multivariate normal distribution. θ_C denotes the parameters of cluster C: $\theta_C = (\mu_C, \Sigma_C)$. C_{s} denotes a subordinate cluster that contains d only at its generation, q denotes the number of parameters in the model, $|C|$ denotes the number of data in C, and $L(C)$ denotes log-likelihood of C. $\log p(X = d; \theta_{C_{\mathrm{s}}}) = 0$ considering that the centroid of C_{s} is equal to d.

I_{A} is defined as

$$\begin{aligned} I_{\mathrm{A}} &= \sum_{x \in C'} \log p(X = x; \theta_{C'}) - \frac{q}{2} \log(|C| + 1) \\ &= L(C') - \frac{q}{2} \log(|C| + 1), \end{aligned} \quad (3)$$

where $C' = C \cup \{d\}$. From Eq.(2) and Eq.(3), ΔI can be rewritten as

$$\begin{aligned} \Delta I &= I_{\mathrm{G}} - I_{\mathrm{A}} \\ &= L(C) - L(C') - \frac{q}{2} \log(|C| + 1). \end{aligned} \quad (4)$$

If $\Delta I > 0$, a subordinate cluster containing d is generated; otherwise d is assinged to C.

Once a subordinate cluster C_{s} is generated, it is regarded as a normal cluster in the subsequent clustering process. However, as C_{s} has only one data, $\Sigma_{C_{\mathrm{s}}} = 0$. As a result, the log-likelihood $L(C_{\mathrm{s}})$ cannot be calculated.

To solve this problem, this paper sets the parameter of a subordinate cluster assuming that it has the same variance as its master cluster C_{m}.

When deciding whether or not to assign d to C_{s}, I_{G} is modified from Eq.(2) as

$$I_{\mathrm{G}} = |C_{\mathrm{m}}| \log p(X = x; (x, \Sigma_{C_{\mathrm{m}}})) - q \log(|C_{\mathrm{s}}| + 1), \quad (5)$$

where x is the data contained in C_{s}, which is regarded as its centroid. That is, $L(C_{\mathrm{s}})$ is calculated as if there are $|C_{\mathrm{m}}|$ data located at the same position as x. Note that the penalty term (last term in Eq.(5)) uses the actual number of data in C_{s}: $|C_{\mathrm{s}}| = 1$.

Similarly, I_{A} is modified from Eq.(3) as

$$\begin{aligned} I_{\mathrm{A}} =& |C_{\mathrm{m}}| \log p(X=x;(\mu'_{C'_{\mathrm{s}}}, \Sigma_{C_{\mathrm{m}}})) \\ &+ \log p(X=d;(\mu'_{C'_{\mathrm{s}}}, \Sigma_{C_{\mathrm{m}}})) - \frac{q}{2}\log(|C_{\mathrm{s}}|+1), \end{aligned} \tag{6}$$

$$\mu'_{C'_{\mathrm{s}}} = \frac{|C_{\mathrm{m}}|x+d}{|C_{\mathrm{m}}|+1}, \tag{7}$$

where $C'_{\mathrm{s}} = C_{\mathrm{s}} \cup \{d\}$.

From Eq.(5) and Eq.(6), ΔI is modified from Eq.(4) as

$$\begin{aligned} \Delta I =& |C_{\mathrm{m}}| \log p(X=x;(x, \Sigma_{C_{\mathrm{m}}})) \\ &- |C_{\mathrm{m}}| \log p(X=x;(\mu'_{C'_{\mathrm{s}}}, \Sigma_{C_{\mathrm{m}}})) \\ &- \log p(X=d;(\mu'_{C'_{\mathrm{s}}}, \Sigma_{C_{\mathrm{m}}})) \\ &- \frac{q}{2}\log(|C_{\mathrm{s}}|+1). \end{aligned} \tag{8}$$

3.2. Clustering with Dynamic Generation of Subordinate Clusters

Algorithm 1
Clustering with Dynamic Generation of Subordinate Clusters

Require: dataset D, Must-link Constraint $C_{\mathrm{M}} \subset D \times D$, number of clusters n, iteration step $s=0$, maximum iteration step $s_{\max}$, clustering state $S=$ true

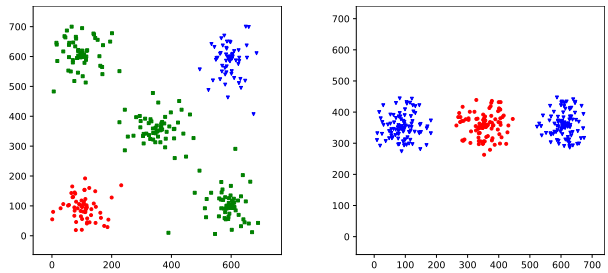
1: Let C_1, ..., C_n be the initial cluster centroids
2: Initial assignment by COP K-means
3: while $s < s_{\max} \wedge S$ do
4: if $d \in D$ does not have a constraint then
5: assign d to the closest cluster C_j
6: else
7: calculate ΔI
8: if $\Delta I > 0$ then
9: generate subordinate cluster C_{s} whose centroid is d
10: else
11: assign d to closest cluster C_j satisfying d's constraints
12: end if
13: end if
14: if $s \geq 1 \wedge \{C_1,\ldots,C_k\}_s = \{C_1,\ldots,C_{k'}\}_{s-1}$ then
15: $S =$ false
16: end if
17: $s = s+1$
18: end while
19: merge subordinate cluster C_{s} to its master cluster C_{m}
20: if $|\{C_1,\ldots,C_k\}_s| > n$ then
21: agglomerative hierarchical clustering
22: end if

This section describes the clustering algorithm that introduces dynamic generation of subordinate clusters into COP K-means. Algorithm 1 is the pseudo code of the proposed method. The algorithm needs a dataset D, a set of Must-links $C_{\mathrm{M}} \subset D \times D$, and the number of cluster n as its input. The algorithm returns n clusters as a final result.

The proposed method also needs the maximum iteration step $s_{\max}$ as a hyper-parameter like COP K-means.

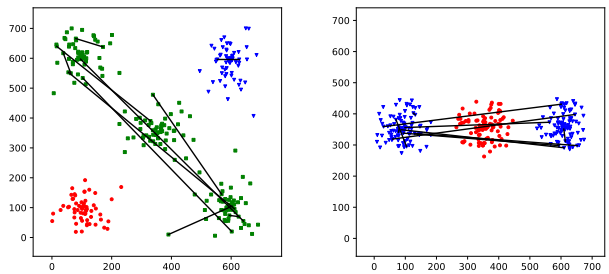
In L1, an initial centroid of each cluster is determined by sampling 10 data from dataset and calculating those arithmetic mean. When a subordinate cluster is generated for d, d is always assigned to that cluster during clustering process. When calculating ΔI in L7, C in Eq.(2) and Eq.(3) (C_j in L11) is the closest cluster for d among those satisfying d's constraints. After the iteration ends, generated subordinate clusters are merged to its master cluster (L19). If the number of resultant clusters is greater than n, agglomerative hierarchical clustering is further applied (L21).

4. Experimental Results



(a) dataset A (b) dataset B

Fig. 1. : Synthetic dataset



(a) dataset A with 10 constraints (b) dataset B with 10 constraints

Fig. 2. : Examples of constraints

This section shows experimental results using synthetic datasets. Figure 1 shows the synthetic datasets, each dataset includes 300 points whose color represents the supervised labels. 5 - 50 pair-wise constraints are given for each dataset: a constraint is generated by connecting the data with the same label. Figure 2 shows examples of the dataset given 10 constraints.

The maximum iteration step $s_{\max}$ is set to 100 for all experiments. This paper evaluates the quality of generated clusters by Normalized Mutual Information (NMI) at an average of 30 times.

4.1. Result

Figure 3 shows the average NMI against the number of constraints by COP K-means in each dataset. Figure 4 shows the maximum NMI against the number of constraints by COP K-means. Figure 5 shows the average NMI against the number of constraints by the proposed method. Figure 6 shows the maximum NMI against the number of constraints by the proposed method. Figure 8 shows the average number of clusters during the clustering process of the proposed method.

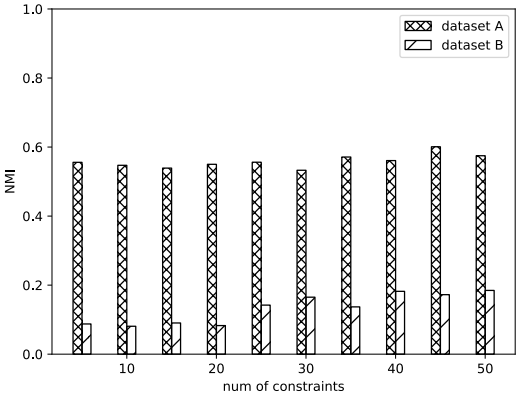


Fig. 3. : The average NMI of COP K-means

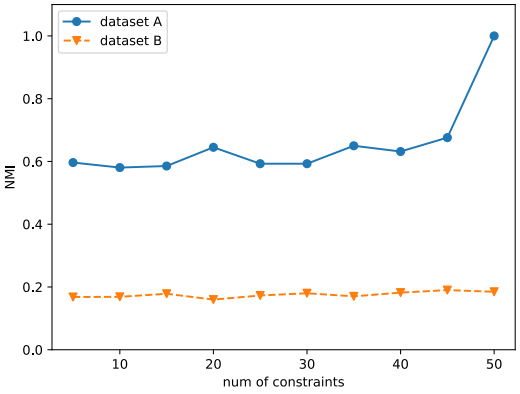


Fig. 4. : The maximum NMI of COP K-means

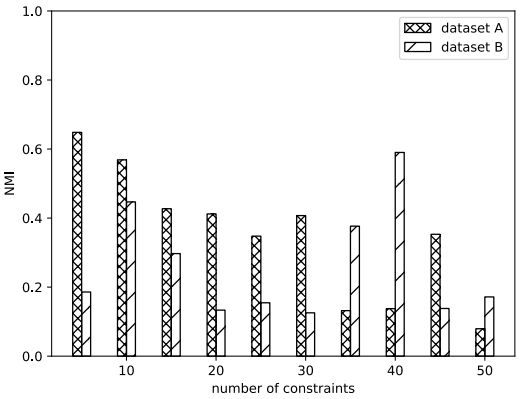


Fig. 5. : The average NMI of proposed method

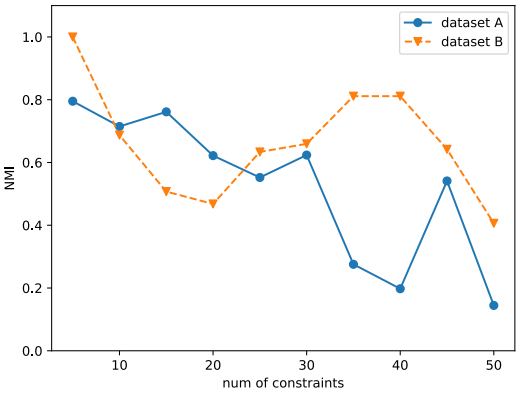
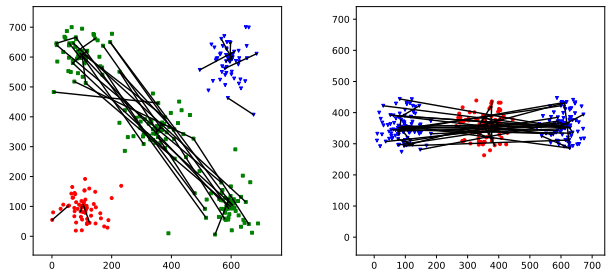


Fig. 6. : The maximum NMI of proposed method



(a) dataset A with 40 constraints (b) dataset B with 40 constraints

Fig. 7. : synthetic dataset

It is observed that the average NMI of dataset A by the proposed method tends to be large when the number of constraints is small. It is also observed that the maximum NMI of both datasets by the proposed method tend to be large when the number of constraints is small. In particular, NMI reached to 1.0 when 5 constraints are given for dataset B. The results of the proposed method on dataset B are better than those of COP K-means, which suggests that the proposed method is effective when different clus-

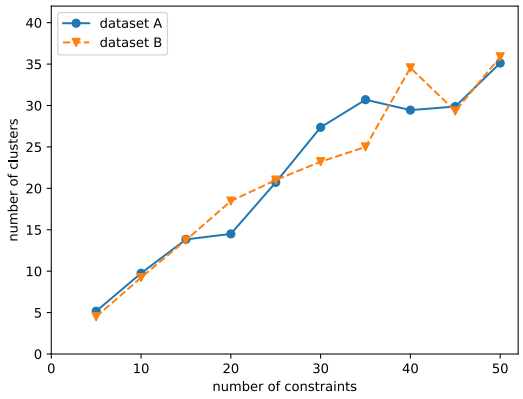


Fig. 8. : Number of clusters

ters overlap as shown in Figure 2 (b). However, it was also observed that the performance of the proposed method was sensitive to the given constraints, which need to be improved in the future.

When the number of constraints is 35 - 40, the result of both datasets by the proposed method showed opposite tendency: NMI of dataset B increased while that of dataset A decreased. It is supposed that the position of constraints may cause the results. Figure 7 shows both datasets with 40 constraints. In dataset A, red or blue instances that locate relatively close to the center region of the data space can be assigned to the green cluster because they have with no constraints. On the other hand, in dataset B, blue instances relatively close to the center of the data space may not be assigned to the red cluster because they have constraints. This suggests that how constraints are given may also affects clustering quality.

Figure 8 shows that the number of subordinate clusters generated during clustering process was proportional to the number of constraints. However, when the number of constraints was large, the number of generated clusters was smaller than that of constraints. This results show the effect of Eq.(8) on suppressing the generation of subordinate clusters.

5. Conclusion

This paper proposed a constrained K-means clustering method that dynamically generates subordinate clusters based on BIC. The proposed method addresses the problem when a Must-link is given to instances located far away from each other by generating subordinate clusters.

This paper evaluates the quality of the clustering result by NMI. The result shows that NMI tends to be large when the number of constraints is small. In addition, the maximum NMI reached to 1.0 when the number of constraints was 5 for dataset B, which mean two instance groups located far away from each other can belong to the same cluster completely.

The number of subordinate clusters generated during clustering process was proportional to the number of constraints. How to further suppress the generation of subordinate clusters is one of the future works.

As other future plans, it is important to investigate how initialization affects clustering. Experimental results suggested given constraints has an impact on the performance of the proposed method. Therefore, it is also important to consider how to give the constraints.

References:

[1] W. Kiri, C. Claire, R. Seth, and S. Stefan, “Constrained K-Means Clustering with Background Knowledge,” Proc. of the Eighteenth International Conference on Machine Learning, pp. 577-584, 2001.

[2] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” Proc. of the fifth Berkeley symposium on mathematical statistics and probability, pp. 281-297, 1967.

[3] S. Gideon, “Estimating the Dimension of a Model,” The Annals of Statistics, Vol.6, No.2, pp. 461-64. 1978.

[4] I. Davidson and S. S. Ravi, “Clustering with Constraints: Feasibility Issues and the k-Means Algorithm,” Proc. of the 2005 SIAM International Conference on Data Mining, pp. 138-149, 2005.

[5] D. Pelleg and A. Moore, “X-means : Extending K-means with Efficient Estimation of the Number of Clusters,” Proc. of the 17th International Conference on Machine Learning, pp. 727-734. 2000.

[6] K. P. Sinaga and M. Yang, “Unsupervised K-Means Clustering Algorithm,” IEEE Access, vol. 8, pp. 80716-80727, 2020.