Seeding method based on Independent Component Analysis for k-means Clustering

Takashi Onoda
System Engineering Lab.
Central Research Institute Electric Power Industry
Tokyo, JAPAN
Email: onoda@criepi.denken.or.jp

Miho Sakai
Department of Computational Intelligence and Systems Science
Tokyo Institute of Technology
Yokohama, JAPAN
Email: sakai@ntt.dis.titech.ac.jp

Seiji Yamada
Digital Content and Media Sciences Research Division
National Institute of Informatics
Tokyo, JAPAN
Email: seiji@nii.ac.jp

***Abstract*—The k-means clustering method is a widely used clustering technique for the Web because of its simplicity and speed. However, the clustering result depends heavily on the chosen initial clustering centers, which are chosen uniformly at random from the data points. We propose a seeding method based on the independent component analysis for the k-means clustering method. We evaluate the performance of our proposed method and compare it with other seeding methods by using benchmark datasets. We applied our proposed method to a Web corpus, which is provided by ODP. The experiments show that the normalized mutual information of our proposed method is better than the normalized mutual information of k-means clustering method and k-means++ clustering method. Therefore, the proposed method is useful for Web corpus.**

***Keywords*-k-means; k-means++; independent component analysis; seeding;**

I. Introduction

Clustering is one of the classic problems in machine learning and computational geometry. In the popular k-means formulation, one is given an integer k and a set of n data points in $\mathbf{R}^d$. The goal is to choose k centers so as to minimize the sum of the squared distances between each point and its closest center.

Solving this problem exactly is NP-hard, even with just two clusters [1], but twenty-five years ago, Lloyd[2] proposed a local search solution that is still very widely used today (see for example [3], [4]). Indeed, a recent survey of data mining techniques states that it ”is by far the most popular clustering method used in scientific and industrial applications”[5].

Usually referred to simply as k-means, Lloyd's method begins with k arbitrary centers, typically chosen uniformly at random from the data points. Each point is then assigned to the nearest center, and each center is recomputed as the center of mass of all points assigned to it. These two steps (assignment and center calculation) are repeated until the process stabilizes.

One can check that the sum of the squared distances between each point and its closest center is monotonically decreasing, which ensures that no clustering is repeated during the course of the method. Since there are at most k^n possible clusterings, the process will always terminate. In practice, very few iterations are usually required, which makes the method much faster than most of its competitors.

Unfortunately, the empirical speed and simplicity of the k-means clustering method come at the price of accuracy. There are many natural examples for which the method generates arbitrarily bad clusters. Furthermore, these examples do not rely on an adversarial placement of the starting centers, and the ratio can be unbounded with high probability even with the standard randomized seeding technique.

In this paper, we propose a way of initializing k-means by choosing Independent Component Analysis based starting centers. We also provide preliminary experimental data showing that in practice, our proposed method really does outperform k-means and k-means++ in terms of both accuracy and speed.

II. Related Works

In this section, we formally define the k-means clustering problem, as well as the k-means clustering and k-means++ clustering methods.

For the k-means problem, we are given an integer k and a set of n data points $\chi \subset \mathbf{R}^d$. We wish to choose k centers $\mathcal{C}$ so as to minimize the potential function,

$$\phi = \sum_{\mathbf{x} \in \chi} \min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{x} - \mathbf{c}\|^2.$$

Choosing these centers implicitly defines a clustering for each center, we set one cluster to be the set of data points that are closer to that center than to any other. As noted above, finding an exact solution to the k-means problem is NP-hard. Throughout the paper, we will let $\mathcal{C}_{opt}$ denote the optimal clustering for a given instance of the k-means problem, and we will let ϕ_{opt} denote the corresponding potential. Given a clustering $\mathcal{C}$ with potential ϕ, we also let $\phi(\mathcal{A})$ denote the contribution of $\mathcal{A} \subset \chi$ to the potential (i.e., $\phi(\mathcal{A}) = \sum_{x \in \mathcal{A}} \min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{x} - \mathbf{c}\|^2$).

A. The k-means clustering method

The k-means clustering method is a simple and fast method that attempts to locally improve an arbitrary k-means clustering. It works as follows.

1) Arbitrarily choose k initial centers $\mathcal{C} = \mathbf{c}_1, \ldots, \mathbf{c}_k$.
2) For each $i \in \{1, \ldots, k\}$, set the cluster $\mathbf{c}_i$ to be the set of points in χ that are closer to $\mathbf{c}_i$ than they are to $\mathbf{c}_j$ for all $j \neq i$.
3) For each $i \in \{1, \ldots, k\}$, set $\mathbf{c}_i$ to be the center of mass of all points in $\mathbf{C}_i$: $\mathbf{c}_i = \frac{1}{|\mathbf{C}_i|} \sum_{x \in \mathbf{C}_i} \mathbf{x}$.
4) Repeat Steps 2) and 3) until $\mathcal{C}$ no longer changes.

It is standard practice to choose the initial centers uniformly at random from χ. For Step 2), ties may be broken arbitrarily, as long as the method is consistent. Steps 2) and 3) are both guaranteed to decrease ϕ, so the method makes local improvements to an arbitrary clustering until it is no longer possible to do so. To see that Step 3) does in fact decreases ϕ, it is helpful to recall a standard result from linear algebra.

The k-means clustering method is attractive in practice because it is simple and it is generally fast. Unfortunately, it is guaranteed only to find a local optimum, which can often be quite poor.

B. The k-means++ clustering method

The k-means clustering method begins with an arbitrary set of cluster centers. The k-means++ clustering method proposes for specifically choosing these centers. At any given time, let $D(\mathbf{x})$ denote the shortest distance from a data point $\mathbf{x}$ to the closest center we have already chosen. Then, the following clustering method is defined as k-means++ clustering method[6].

1a) Choose an initial center $\mathbf{c}_1$ uniformly at random from χ.

1b) 1b) Choose the next center $\mathbf{c}_i$ by the following.

   a) Find a real value y uniformly at random. The value satisfies the following equation

$$0 \leq y \leq \sum_{\mathbf{x} \in \chi} D(\mathbf{x})^2.$$

   b) Find $\mathbf{x}_i$ satisfying the following equation. Select $\mathbf{x}_i$ as a cluster center $\mathbf{c}_i$.

$$D(\mathbf{x}_{i-1})^2 \leq y \leq D(\mathbf{x}_i).$$

1c) Repeat Step 1b) until we have chosen a total of k centers.

Step 2)-4) proceed as with the standard k-means clustering method. We call the weighting used in Step 1b) simply "D^2 weighting".

III. Proposed Method

This section describes a problem for k-means clustering and k-means++ clustering methods. Then, we proposes k-means combined with Independent Component Analysis (ICA) based seeding method.

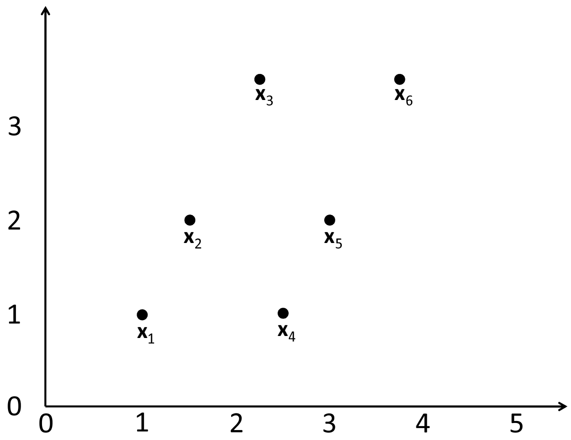

Fig. 1. Given Data

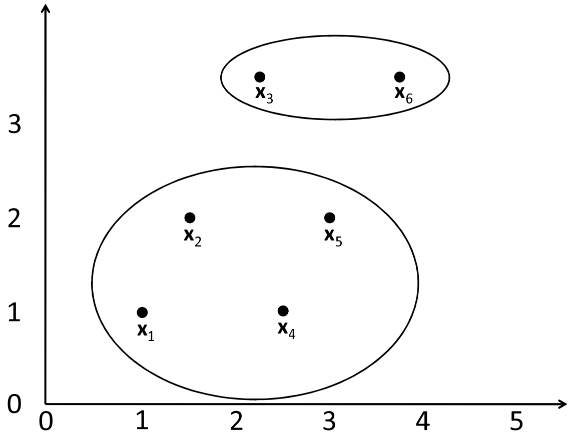

Fig. 2. Global Optimal Clustering Case

A. Problem for k-means and k-means++ clustering methods

We have 6 points data which consist of $\mathbf{x}_i,\ i = 1, \ldots, 6$ and these points are divided into two clusters. Figure 1 shows these 6 points. And Figure 2 shows the global optimal clustering result for these 6 points data. The first cluster consists of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$ and the other cluster consists of $\{\mathbf{x}_3, \mathbf{x}_6\}$. We assume that clustering methods can find the global optimal clusters. However, the k-means clustering method generates bad clusters if $\mathbf{x}_2$ and $\mathbf{x}_5$ are chosen as initial cluster centers $\mathbf{c}_1$ and $\mathbf{c}_2$. Figure 3 shows local optimal clusters, which are bad clusters. The k-means++ clustering method was developed to avoid this bad clustering.

However, the k-means++ clustering method sometimes generates bad clusters because it depends on choice of the initial center $\mathbf{c}_1$. The initial center $\mathbf{c}_1$ is chosen uniformly at random from χ.

B. The k-means combined with ICA based seeding method

The k-means clustering method begins with an arbitrary set of cluster centers. The k-means++ clustering method begins

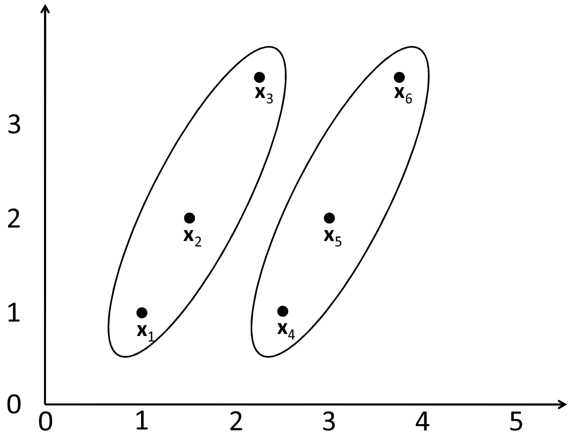

Fig. 3. Local Optimal Clustering Case

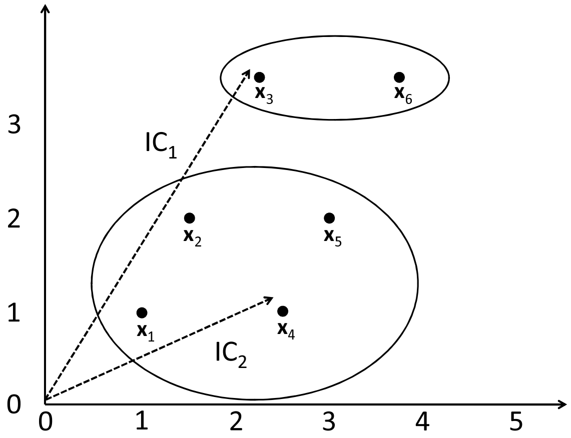

Fig. 4. The Concept of our proposed method

with a small arbitrary set of cluster centers. As stated above, we propose a method for specifically choosing these centers. At any given time, we can get independent components(ICs) from given data $\mathbf{x}$. Then, we define the following seeding method.

1a) Extract k independent components $\mathbf{IC}_m,\ m = 1, \ldots, k$ from given data $\mathbf{x}$.

1b) Choose an initial center $\mathbf{c}_1$, selecting $\mathbf{c}_1 = \mathbf{x}' \in \chi$ with minimum $\frac{\mathbf{IC}_1 \cdot \mathbf{x}'}{|\mathbf{IC}_1||\mathbf{x}'|}$.

1c) Choose the next center $\mathbf{c}_i$, selecting $\mathbf{c}_i = \mathbf{x}' \in \chi$ with minimum $\frac{\mathbf{IC}_i \cdot \mathbf{x}'}{|\mathbf{IC}_i||\mathbf{x}'|}$.

1d) Repeat Step 1c) until we have chosen a total of k centers.

Step 2)-4) proceed as with the standard k-means clustering method. Figure 4 shows the concept of the k-means clustering method combined with ICA based seeding method.

IV. Experiments

To evaluate k-means clustering, k-means++ clustering and the proposed method in practice, we implemented and tested them in matlab. In this section, we discuss the results of these preliminary experiments. We found that the k-means clustering method combined with ICA based seeding method is accurate.

A. Datasets

We evaluated the performance of k-means clustering, k-means++ clustering and the proposed methods on three data sets of UCI Machine Learning repository and a dataset of the *ODP Web corpus*. The first data set, *iris*, consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal. Based on the combination of the four features, Fisher developed a linear discriminant model to determine which species from these four measurements. It is used as a typical test for many other classification techniques.

The second dataset, *wine*, is the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The third dataset, *soybean-small*, is for four soybean disease diagnosis. The dataset is consists of 47 samples and 35 attributes.

The forth dataset is the *ODP Web corpus* dataset for our test experiment. The ODP Web corpus dataset consists of 12 directories, 247 samples, and 344 attributes.

B. Evaluation metrics

We used Normalized Mutual Information as a metric to evaluate the qualities of clustering outputs of different methods. The Normalized Mutual Information measures the consistency of the clustering output compared to the ground truth. It reaches the maximum value of 1 only if the membership ϕ_c perfectly matches ϕ_g and the minimal zero if the assignments of ϕ_c and ϕ_g are independent. The membership function $\phi_c(\mathbf{x})$, the mapping of a point $\mathbf{x}$ to one of the k clusters. The membership $\phi_g(\mathbf{x})$ represents the true cluster label for $\mathbf{x}$. Formally, the Normalized Mutual Information is the following equation.

$$NMI(\phi_g, \phi_c) = \frac{MI(\phi_g, \phi_c)}{\max(H(\phi_g), H(\phi_c))},$$

where $MI(\phi_g, \phi_c)$ denotes

$$MI(\phi_g, \phi_c) = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{g,c}(i, j) \log \frac{p_{g,c}(i, j)}{p_g(i) p_c(j)},$$

$H(\phi_g)$ is

$$H(\phi_g) = \sum_{i=1}^{k} p_g(i) \log \frac{1}{p_g(i)},$$

and $H(\phi_c)$ denotes

$$H(\phi_c) = \sum_{j=1}^{k} p_c(j) \log \frac{1}{p_c(j)}.$$

The $p_g(i)$ is the percentage of points in cluster i according to the ground truth, i.e. $p_g(i) = \frac{\sum_{l=1}^{n} 1(\phi_g(\mathbf{x}_l)-i)}{n}$. Similarly, $p_c(j) = \frac{\sum_{l=1}^{n} 1(\phi_c(\mathbf{x}_l)-j)}{n}$ and $p_{g,c}(i,j)$ is the percentage of points that belong to cluster i in ϕ_g and also cluster j in ϕ_c, i.e. $p_{g,c}(i,j) = \frac{\sum_{l=1}^{n} 1(\phi_g(\mathbf{x}_l)-i)1(\phi_c(\mathbf{x}_l)-j)}{n}$.

The above defined metrics were used to evaluate the accuracy of the clustering methods.

C. Experimental results

The k-means and k-means++ clustering methods were each run 100 times with different initializations over all three datasets. The proposed method was run only one time because it can set up a unique initial seeding. Table I lists the experimental results of the iris dataset, Table II lists the experimental results of the wine dataset, and Table III lists the experimental results of the soybean-small dataset. These tables have the averaged NMI, the maximum NMI, the minimum NMI, and the NMI when the clusters achieved minimum variance. Tables I and II show that the proposed method outperforms both the k-means clustering method and the k-means++ clustering method. In Tables I and II, the NMI of our proposed method is as same as the NMI of the maximum performance of the k-means clustering method and the k-means++ clustering method and is achieved by only one initial seeding. Table III shows that the NMI of our proposed method is the same as the NMI of the k-means clustering method and k-means++ clustering method when the clusters achieved minimum variance. This situation shows that the performance of our proposed method is as same as the performance of the k-means clustering method and the k-means++ clustering method for the soybean-small dataset. And the NMI with minimum variance is achieved by only one initial seeding.

The k-means clustering and k-means++ clustering methods were each run 100 times with different initializations for the ODP Web corpus dataset. The proposed method was run only one time because it can set up a unique initial seeding. Table IV lists the experimental results of the ODP Web corpus dataset. The maximum NMI of k-means clustering method was 0.421 and the minimum NMI was 0.341. The maximum NMI of k-means++ clustering method was 0.432 and the minimum NMI was 0.358. Table IV shows that the NMI of our proposed method is better than the NMI of k-means clustering and k-means++ clustering methods when the clusters achieved minimum variance for the ODP Web corpus dataset. The maximum NMI of k-means clustering method is better than the NMI of the k-means clustering method when the clusters achieved minimum variance. However, we generally cannot provide true cluster data. NMI with minimum variance is the most important for real-world applications. Therefore, Table IV shows that the proposed method outperforms both k-means clustering and k-means++ clustering methods for the ODP Web corpus dataset. And the NMI with minimum variance is achieved by only one initial seeding.

TABLE I
EXPERIMENTAL RESULTS FOR *iris* DATASET

method	NMI with min variance	max NMI	min NMI	avrg NMI
k-means	0.751	0.751	0.532	0.703
k-means++	0.751	0.751	0.532	0.749
ICA	0.751	-	-	-

TABLE II
EXPERIMENTAL RESULTS FOR *wine* DATASET

method	NMI with min variance	max NMI	min NMI	avrg NMI
k-means	0.429	0.429	0.387	0.418
k-means++	0.429	0.429	0.387	0.418
ICA	0.429	-	-	-

TABLE III
EXPERIMENTAL RESULTS FOR *soybean-small* DATASET

method	NMI with min variance	max NMI	min NMI	avrg NMI
k-means	0.711	1.000	0.518	0.714
k-means++	0.711	1.000	0.711	0.806
ICA	0.711	-	-	-

TABLE IV
EXPERIMENTAL RESULTS FOR *ODP Web corpus* DATASET

method	NMI with min variance	max NMI	min NMI	avrg NMI
k-means	0.383	0.421	0.341	0.385
k-means++	0.374	0.432	0.358	0.388
ICA	0.385	-	-	-

V. CONCLUSION

We proposed a method that combines k-means clustering method with ICA based seeding method. From our experimental results, our proposed method performed the same as or better than the k-means clustering and k-means++ clustering methods. For our future work, we plan the followings.

1) Apply the proposed method to different Web data and benchmark datasets.
2) Theoretically analyze the computational cost of the proposed method.
3) Develop a method for finding clusters that are on only one independent component.

REFERENCES

[1] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, “Clustering large graphs via the singular value decomposition,” *Mach. Learn.*, vol. 56, no. (1-3), pp. 9–33, 2004.

[2] S. P. Lloyd, “Least squares quantization in pcm,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[3] P. K. Agarwal and N. H. Mustafa, “k-means projective clustering,” in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,*. New York, NY, USA: ACM Press, 2005, pp. 155–165.

[4] R. Herwig, A. Poustka, C. Müller, C. Bull, H. Lehrach, and J. O'Brien, “Large-scale clustering of cdna-fingerprinting data,” *Genome Research*, vol. 9, pp. 1093–1105, 1999.

[5] P. Berkhin, “Survey of clustering data mining techniques,” Accrue Software, San Jose, CA, Tech. Rep., 2002.

[6] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,*. New Orleans, Louisiana, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.