2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

Independent Component Analysis based Seeding method 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*-independent component analysis; seeding; k-means clustering;**

I. Introduction

Search engines are useful tools for retrieving information from the Web. When a user inputs his or her query into them, it returns a list of results ranked in order of relevance to the query. The user starts at the top of the list and follows it down, examining one result at a time, until the relevant information has been found.

While search engines are definitely good for certain search tasks such as finding the home page of an organization, they may be less effective for satisfying vague queries. The results on different subtopics or meanings of the input query will come together in a list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. On the other hand, there is no way to estimate what is relevant to the user given that the queries are usually very short and their interpretation is inherently fuzzy in the absence of context.

An approach for clustering the entire Web is different to one for retrieving information from the Web. This clustering approach shows the results, which are associated with clusters that consist of similar items manually or automatically (see Figure 1). Even the largest Web directories cover only a small fragment of the existing web pages. Furthermore, these directories may be of little help for a particular user query, or for a particular aspect of the query. In fact, Web directories are most often used to affect the output of a direct search in response to common user queries.

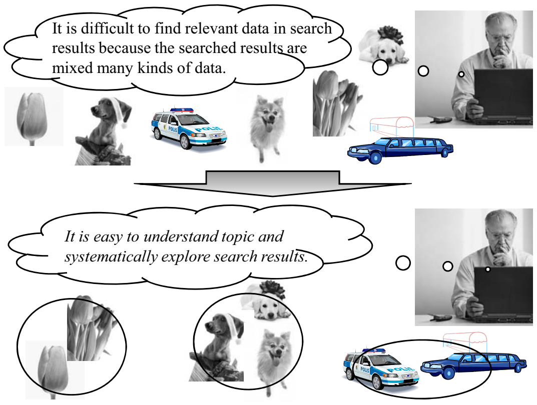
It is difficult to find relevant data in search results because the searched results are mixed many kinds of data.
It is easy to understand topic and systematically explore search results.

Figure 1. Effect of clustering for serch results

The clustering the entire Web has attracted considerable commercial interest and it is also an active research area. In this research area, a large number of papers have been published and focusing on specific issues and systems. The clustering the entire Web is clearly related to document clustering. However, our focus is on easy clustering methods and optimal clusters. Especially, we are interested in the simplest and quickest clustering method. Therefore we deal with the k-means clustering method in our research. We also discuss how to solve the problem of ”seeding” in the k-means clustering method.

The rest of the paper is organized as follows. Section II discusses related work and the k-means clustering, KKZ clustering method and k-means++ clustering methods. Sec-

978-0-7695-4513-4/11 $26.00 © 2011 IEEE
DOI 10.1109/WI-IAT.2011.29

tion III discusses a problem of these clustering methods and introduces our proposed method. Section IV presents experimental results from comparing the performance of the proposed method with those of k-means clustering, KKZ clustering method and k-means++ clustering methods. Section V concludes this research.

II. Related Works

Clustering is a classic problem 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 $chi \subset \mathbf{R}^2$. k is the number of cluster centers. The goal is to choose k centers $\mathcal{C}$ to minimize the sum of the squared distances between each point and its closest center.

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

Solving this problem is NP-hard, even with just two clusters[1], but 25 years ago, Lloyd[2] proposed a local search solution that is still widely used today. A recent survey of data mining techniques states that it "is by far the most popular clustering algorithm used in scientific and industrial applications"[3].

In this section, we formally define k-means clustering method, KKZ clustering method and k-means++ clustering method.

A. k-means clustering method

The k-means clustering method is simple and fast, which locally improves an arbitrary k-means cluster method. It works as follows.

1) Arbitrarily choose k initial centers $\mathcal{C} = \mathbf{c}_1, \cdots, \mathbf{c}_k$,
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 the mass of all points in a set C_i of cluster i $\mathbf{c}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x}$
4) Repeat Steps 2) and 3) until $\mathbf{c}_i$ 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; therefore, the method makes local improvements to an arbitrary cluster until it is no longer possible to do so.

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

B. KKZ clustering method

KKZ method was proposed by Katsavounidis et al. [4]. This method calculates all distance between data and find the data which have a wide distance. The data are selected as initial cluster 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 KKZ clustering method[4].

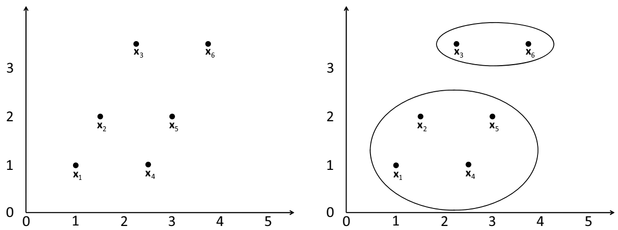

Figure 2. Given Data & Global Optimal Clustering Case

1a) Choose initial centers $\mathbf{c}_1$ and $\mathbf{c}_2$. The distance between $\mathbf{c}_1$ and $\mathbf{c}_2$ is the widest of all distance between a data point and the other data point.
1b) For all data, $D(\mathbf{x}_j), j \in \{1, \cdots, n\}$ are calculated.
1c) Choose the next center $\mathbf{c}_i$, selecting $\mathbf{c}_i = \mathbf{x}' \in \chi$ with the widest distance $D(\mathbf{x}')$.
1d) Repeat Step 1b) until we have chosen a total of k centers.

Steps 2)-4) proceed as with the standard k-means algorithm.

C. k-means++ clustering method

The k-means method begins with an arbitrary set of cluster centers. k-means++ clustering 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[5].

1a) Choose an initial center $\mathbf{c}_1$ uniformly at random from χ.
1b) Choose the next center $\mathbf{c}_i$, selecting $\mathbf{c}_i = \mathbf{x}' \in \chi$ with probability $\frac{D(\mathbf{x}')^2}{\sum_{\mathbf{x} \in \chi} D(\mathbf{x})^2}$.
1c) Repeat Step 1b) until we have chosen a total of k centers.

Steps 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 the k-means and k-means++ clustering methods. Then, we propose k-means combined with Independent Component Analysis (ICA) based seeding method.

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

We have six data points, which consist of $\mathbf{x}_i, i = 1, \ldots, 6$ and these points, are divided into two clusters. Figure 2 shows these six data points. And Figure 2 shows the global optimal clustering result for these six data points. The first cluster consists of $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$ and the other consists

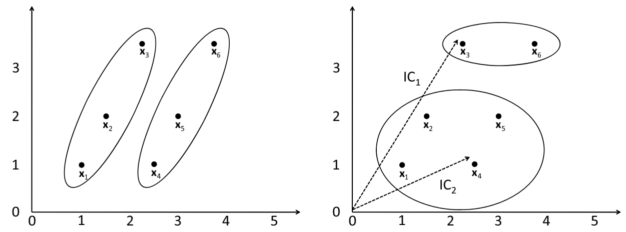

Figure 3. Local Optimal Clustering Case & The Concept of our proposed method

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 the choice of the initial center $\mathbf{c}_1$. The initial center $\mathbf{c}_1$ is chosen uniformly at random from χ.

B. 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 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 obtain 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.

Steps 2)-4) proceed as with the standard k-means clustering method. Figure 3 shows the concept of the k-means clustering method combined with ICA based seeding method.

IV. Experiments

To evaluate the k-means clustering method, the k-means++ clustering method 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 the k-means clustering method, k-means++ clustering method and the proposed method on three datasets of the UCI Machine Learning repository. The first data set, *iris*, consists of 50 samples from three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). The second dataset, *wine*, is the result of a chemical analysis of wines produced 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 wine. The third dataset, *soybean-small*, was for diagnosing four soybean diseases. The dataset consists of 47 samples and 35 attributes.

We used 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 a maximum value of 1 only if the membership ϕ_c perfectly matches ϕ_g and a minimum of zero if the assignments of ϕ_c and ϕ_g are independent. The membership function $\phi_c(\mathbf{x})$ is 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 derived with $NMI(\phi_g, \phi_c) = \frac{MI(\phi_g, \phi_c)}{\max(H(\phi_g), H(\phi_c))}$, 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 k-means clustering method, k-means++ clustering method and the proposed method.

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, II, III, IV 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

Table I
EXPERIMENTAL RESULTS FOR *iris* DATASET

method	NMI	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
KKZ	0.751	-	-	-	-
PCA	0.751	-	-	-	-
ICA	0.751	-	-	-	-

Table II
EXPERIMENTAL RESULTS FOR *wine* DATASET

method	NMI	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
KKZ	0.387	-	-	-	-
PCA	0.429	-	-	-	-
ICA	0.429	-	-	-	-

initial seeding. And Table I shows that the performance of KKZ method is worse than the other methods.

We generally cannot provide true cluster data. NMI with minimum variance is the most important for real-world applications. 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. Also instead of ICA, we used a seeding method based Principal Component Analysis(PCA). Table IV lists the experimental results of the *ODP Web corpus* dataset.

Table III
EXPERIMENTAL RESULTS FOR *soybean-small* DATASET

method	NMI	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
KKZ	0.711	-	-	-	-
PCA	0.711	-	-	-	-
ICA	0.711	-	-	-	-

Table IV
EXPERIMENTAL RESULTS FOR *ODP Web corpus* DATASET

method	NMI	NMI with min. variance	max. NMI	min. NMI	avrg. NMI
k-means	-	0.555	0.589	0.392	0.514
k-means++	-	0.555	0.589	0.425	0.525
KKZ	0.531	-	-	-	-
PCA	0.500	-	-	-	-
ICA	0.638	-	-	-	-

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. Table IV shows that the NMI of the proposed method was 0.638. This value is better than the maximum NMI of k-means clustering method and the maximum NMI of k-means++ clustering method. And also the NMI of the proposed method is better than the NMI of KKZ method and the NMI of PCA based method. Therefore, Table IV shows that the proposed method outperforms k-means clustering method, k-means++ clustering method, KKZ method and PCA bsed method for the *ODP Web corpus* dataset. And the best NMI is achieved by only one initial seeding.

V. CONCLUSIONS

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 method, k-means++ clustering method, KKZ method and PCA based method. 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.

REFERENCES

[1] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering large graphs via the singular value decomposition," *Maching Learning*, 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. Berkhin, "Survey of clustering data mining techniques," Accrue Software, San Jose, CA, Tech. Rep., 2002.

[4] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized lloyd iteration," *IEEE Signal Processing Letters*, vol. 1, no. 10, pp. 144–146, 1994.

[5] 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.