A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science

Department of Computational Biology
Graduate School of Frontier Sciences, The University of Tokyo

Kazuki Ichikawa

Introduction

K-means clustering has been widely used to gain insights into biological systems from large-scale life science data, such as gene expression data monitored by microarrays, histone modification data and nucleosome positioning. Despite comparatively low computational complexity of *K*-means clustering, accelerating *k*-means clustering algorithms is still necessary to process the growing volume of biological data due to the recent progress in data collection by next-generation sequencing.

The basic concept of *k*-means clustering is simple.

1. It first selects *k* cluster centroids in some manner.
2. Subsequently, *k*-means clustering repeats the process of assigning individual points to their nearest centroids(assigning step) and updating each of *k* centroids as the mean of points assigned to the centroid(updating step) until no further changes occur on the *k* centroids

Euclidean distance and Pearson correlation distance have been widely used for large-scale biological data processing. Euclidean distance is sensitive to scaling, while correlation is unaffected by scaling. To address this problem, the standardized Euclidean distance, which is not sensitive to scaling, is frequently used.

Despite the importance of the Pearson correlation distance and standardized Euclidean distance optimization methods tailored to these distances are largely unexplored. In general, several efficient *k*-means clustering algorithms have been proposed for processing Euclidean distances by utilizing the triangle inequality for Euclidean distance or by analyzing the correlation coefficient between the centroids. Thus, we can manage to use optimization methods for the Euclidean distance to yield a *k*-means clustering result with the standardized Euclidean distance, which concords with a result with the Pearson correlation distance.

In this research, we examine the property of the Pearson correlation distance and propose novel ways of avoiding unnecessary computation so as to boost *k*-means clustering for the Pearson correlation distance, and demonstrate that our method outperforms applications of pruning methods for the Euclidean distance to the standardized Euclidean distance.

Methods

Definition. Two d dimensional vectors $\boldsymbol{x} = (\boldsymbol{x}[1], \dots, \boldsymbol{x}[d])$, $\boldsymbol{y} = (\boldsymbol{y}[1], \dots, \boldsymbol{y}[d])$, we define Pearson's correlation coefficient:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{\boldsymbol{x}[i] - \bar{\boldsymbol{x}}}{\sigma_x}\right)\left(\frac{\boldsymbol{y}[i] - \bar{\boldsymbol{y}}}{\sigma_y}\right)$$

Main approach to problem of our program

We can accelerate the assigning step in *k*-means clustering if we can test whether the nearest centroid for a point remains unchanged without recalculating the distances between the point and all centroids. Suppose that after the updating step, the centroid $\boldsymbol{c}_p$ nearest to $\boldsymbol{x}$ moves to $\boldsymbol{c}_p{}'$ for $p = 1, \dots, k$, and any other centroid $\boldsymbol{c}_q$ $(q = 1, \dots, k, q \neq p)$ moves to $\boldsymbol{c}_q{}'$.

We ask if $\boldsymbol{x}$ is still closest to cluster $\boldsymbol{c}_p{}'$ after the updating step:

$$\mathrm{dis}(\boldsymbol{c}_p{}',\boldsymbol{x}) \leq \mathrm{dis}(\boldsymbol{c}_q{}',\boldsymbol{x}) \qquad \text{for } q = 1,\dots,k\ (q \neq p).$$

To check this test efficiently for any point $\boldsymbol{x}$ without recalculating the new distances on both sides of the inequality, we will develop an efficient method to estimate an upper bound and a lower bound by using existing distances.

If $\quad \mathrm{dis}(\boldsymbol{c}_p,\boldsymbol{x}) + \text{an\_upper\_bound} \leq \mathrm{dis}(\boldsymbol{c}_q,\boldsymbol{x}) + \text{a\_lower\_bound} \quad \text{for } q = 1,\dots,k\ (q \neq p),$

we can confirm $\mathrm{dis}(\boldsymbol{c}_p{}',\boldsymbol{x}) \leq \mathrm{dis}(\boldsymbol{c}_q{}',\boldsymbol{x})\ \ (q \neq p)$ without calculating the new distances, while retaining the final solution.

Results

We compared our program "BoostKCP" with other available pruning methods using real biological datasets. We used human nucleosome positioning signals at genomic positions around transcription starting sites, gene expression data.

We compared BoostKCP with Lloyd's, Elkan's, and Hamerly's algorithms using nucleosome positioning data for dimension (d = 10, 20, 50, 101, 201, 501, 1001 and 2001) and for number of clusters k=10, 20, and 30. BoostKCP outperformed Lloyd's and Hamerly's algorithms for all combinations of parameter values, and was also faster than Elkan's algorithm in most cases. The acceleration rates of average elapsed time by BoostKCP in comparison with each of Lloyd's, Hamerly's are 2.5~6.8. The acceleration rates in comparison with Elkan's are 0.9~2.6. (Fig.1)

We also applied BoostKCP and Elkan's algorithms to gene expression data, a set of 54,613 vectors of dimension 180. BoostKCP clearly outperformed Elkan's algorithm for each k (= 10, 20, 30, …, 70) and the acceleration rate increased for larger values of k. (Fig.2)

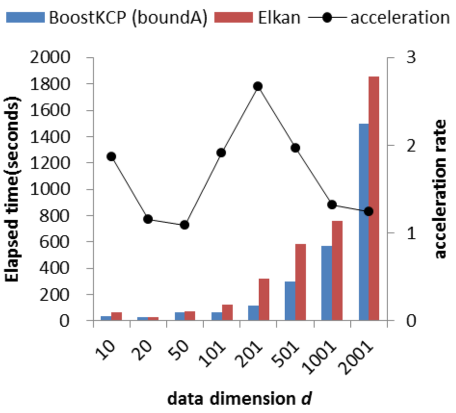

Fig. 1. Average elapsed time by BoostKCP and Elkan's algorithm with Human nucleosome positioning signals

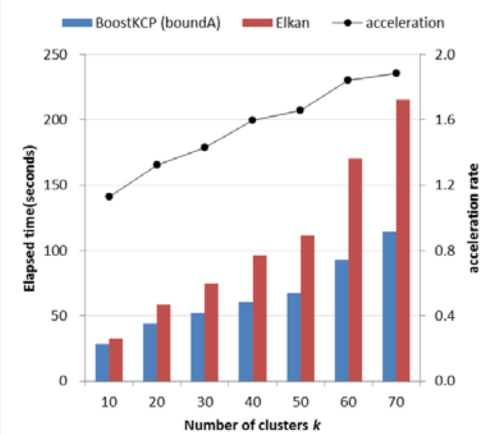

Fig. 2. Average elapsed time by BoostKCP and Elkan's algorithm with gene expression data of Glioma samples

Conclusion and Discussions

K-means clustering with the Pearson correlation distance and standardized Euclidean distance has proven useful in obtaining novel insights from large-scale biological data, but it is likely to be a computationally heavy task, demanding a method of accelerating the computational performance for higher dimensional biological data.

In this research, we proposed BoostKCP, a new pruning method that has proven useful for reducing the computational time. BoostKCP outperformed Lloyd's algorithm, Hamerly's algorithm, and the state-of-the-art Elkan's algorithm.

Publication

Kazuki Ichikawa and Shinichi Morishita. A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) (in press)