

AN EXPANSION OF X -MEANS FOR AUTOMATICALLY DETERMINING THE OPTIMAL NUMBER OF CLUSTERS — PROGRESSIVE ITERATIONS OF K -MEANS AND MERGING OF THE CLUSTERS —

Tsunenori Ishioka
Research Division
National Center for University Entrance Examinations
2-19-23 Komaba, Meguro-ku,
Tokyo 153-8501, Japan
email: tunenori@rd.dnc.ac.jp

ABSTRACT

We expand a non-hierarchical clustering algorithm that can determine the optimal number of clusters by using iterations of k -means and a stopping rule based on Bayesian Information Criterion (BIC). The procedure requires merging the clusters that a k -means iteration has made to avoid unsuitable division caused by the division order. By using this additional merging operation, the case of adequate clustering was increased for various types of simulation runs. With no prior information about the number of clusters, our method can get the optimal clustering based on information theory instead of on a heuristic method. The computational complexity of our method is $\mathcal{O}(N \log k)$ for the sample size N and the number of final clusters, k .

KEY WORDS

Non-hierarchical clustering, Information criterion, BIC, Feedback operation, Computer simulation

1 Introduction

Clustering is an important technique in the data-mining area [1]. If data mining is considered a knowledge discovery from amount of data, we can obtain useful information from the number of clusters into how many clusters the whole data can be divided, as well as the available maximum sample size and the computational complexity.

Four methods have been developed for finding the number of optimal clusters when no prior information is available about the number of clusters.

1. A method that finds the optimal number of clusters heuristically by using appropriate information criterion based on a different setting of cluster numbers.
2. A method using the minimum volume ellipsoid (MVE) estimator [7].
3. A method that starts with a cluster division of more than the optimal solution. A suitable number of clusters is determined by merging near clusters and/or removing spurious clusters [8].

4. A method that repeats two divisions according to a k -means method until the division is not judged to be appropriate after classifying it into a sufficiently small number of clusters that are made by the first k -means method [11].

The first method is the simplest and most authentic, but the function that evaluates the goodness-of-fit to the model, such as an information criterion, does not necessarily become convex against the number of clusters. In addition, since a lot of clusterings are possible for a huge amount of data, many function values associated with each clustering should be evaluated. Therefore, this is not realistic from the point of view of the computational amount.

Hardy [3] surveyed seven typical evaluation criteria, two of which can be applied for hierarchical clustering methods with various datasets. However, varying the number of clusters requires much computation, because we have to use k -means repeatedly.

The second method [7] determines a single ellipsoid and removes it one by one from the whole, using a Kolmogorov-Smirnov goodness-of-fit test. Each cluster is the basis of assumption of being with an ellipsoid. However, designing validity measures that perform well on a variety of data sets for this method is well known to be difficult, and it is very weak for noise contamination [9].

In the third method [8], not all data is classified exclusively; the data considered as an outlier is eliminated from a cluster. In data mining or data detection, we do not prefer this method because an outlier gives us important information.

The fourth method is called the x -means because the number of final clusters is unfixed. In the context of recent research into data mining, several high-performance techniques for the k -means, the basis of the x -means, have been developed along with self-organizing maps [12, 13]. Pelleg [10] showed a great reduction in effort for updating cluster centers by storing sufficient statistics in kd -trees; Huang [5] presents a clustering technique for large datasets with categorical values; BIRCH, proposed by Zhang [14], can typically find good clusters with a single scan of data and can improve the quality further with a few additional

