An Alternative Technique of Selecting the Initial Cluster Centers in the k-means Algorithm for Better Clustering

Sisir Kumar Rajbongshi
Department of Information Technology
Gauhati University
Guwahati-781014, Assam, India

Anjana Kakoti Mahanta
Department of Computer Science
Gauhati University
Guwahati-781014, Assam, India

ABSTRACT
Although k-means works well in many cases it offers no accuracy guarantee and it has no idea to select ideal cluster representatives. This article presents a technique in which the initial cluster representatives in the standard k-means algorithm are chosen intelligently. Comparison of the quality of the clusters produced by the standard k-means algorithm, k-means using Furthest-First, and k-means using the proposed initialization technique have investigated. Experiment result shows that the quality of the clusters improves with the proposed algorithm in most of the cases.

General Terms
Data Mining Technique, Clustering Technique, Cluster.

Keywords
Cluster representative, cluster quality, Furthest-First Technique, centroid.

1. INTRODUCTION

In the field of Data Mining, clustering plays an important role. Clustering is a useful technique for discovery of data distribution and patterns in the underlying data, is an important task in machine learning and pattern recognition [1]. If meaningful groups are the goal, then the clusters should capture the natural structure of the data. Typically in clustering there is no one perfect solutions to the problem, but algorithms seek to minimize a certain mathematical criterion (which varies between algorithms).

There have been many applications of cluster analysis to practical problems. They play an important role in how people analyze and describe the world. They have been widely applied in Information Retrieval, understanding Earth's climate, Psychology and Medicine and in Business.

Most of the initial clustering techniques were developed by the Statistic or Pattern Recognition fraternities, where the goal was to cluster modest number of instances. However, within the data mining fraternity the focus has been on clustering large datasets [1]. Developing clustering algorithms to cluster rapidly growing datasets effectively and efficiently has been identified challenge. We use the term "centre-based clustering" to refer to the family of algorithms such as k-means, since they use a number of "centers" to represent and/ or partition the input data. Each center also called "centroid" defines with a central point. Center-based clustering algorithms begin with a guess about the solution, and then refine the positions of centers until reaching a local optimum [2]. Converging to bad local optima is related to sensitivity to initialization, and is a primary problem of data clustering. In this article, we address the problem of efficient data clustering on small and large datasets. Our work is in the context of k-means clustering which was developed by MacQueen [3]. There has been an extensive study on clustering algorithms in the literature. Comprehensive survey on this subject can be obtained from the book and papers [1, 2, 4, 5, 6, 7, 8]. In this discussion, we limit ourselves to the improvements over k-means. In [4], a lot of variants of k-means presented, where the initial cluster representatives were selected intelligently than randomly. The Paper [5] proposed a fast and exact out of core k-means which requires only one or small number of passes on entire dataset. Another article [6], presents a way of initializing k-means by choosing random starting centers with very specific probabilities. Domingos and Hultel [7] proposed a faster version of k-means using sampling based on similar statistical bound. The algorithm consists of a number of runs of k-means with sample; the sample size is increased in the iterations to maintain the less bound from the multi pass k-means. All these variants of the standard k-means algorithms are concerned on the improvement over the algorithm.

Lots of the above efforts are concerned on how to improve the quality of clusters produced by the k-means algorithm. Developing a variant of k-means algorithm with this goal is the focus of our work. The main problem with k-means algorithm is that it has no idea about ideal initial cluster representatives as it chooses the initial representatives randomly from the dataset. Therefore, an interesting question is, "can we have a technique of selecting the initial cluster centers in the k-means-algorithm which selects the initial centers intelligently than randomly in the standard k-means algorithm, and can produce the better results than the k-means algorithm". The Furthest First (FF) is a technique [4], which selects the initial centers in the k-means intelligently. But this technique has the tendency of selecting outliers as cluster centers. To overcome this problem, we propose an approach in which the initial cluster representatives selected by furthest-first technique are pulled towards the middle of the dataset. The main idea of the technique is not only to select the initial centers well separated from each other but also to avoid outliers.

The outline of the rest of the paper is as follows- section 2 presents the main ideas of the k-means algorithm, Furthest-First initialization technique and the proposed initialization technique; section 3 provides the implementation details; section 4 lists the Statistical data obtained from the experiment results. The discussions on the experiments carried out are provided in the section 5 and the conclusions are provided in the section 6.

2. ALGORITHM DESCRIPTIONS

This section formally defines the standard k-means problem, Furthest-First technique and the proposed initialization technique.

2.1 The k-means Algorithm

The k-means algorithm is simple and fast clustering algorithm that attempts to locally improve an arbitrary k-means clustering. The algorithm is as follows:

- Arbitrarily choose k points as the initial centroids.
- Repeat
    - From k cluster by assigning each point to its closest centroid.
- Re-compute the centroids of each cluster
- Until centroids do not change.

In the first step, k-points are chosen randomly as the centroids and points are assigned to the initial centroids, which are all in the larger group of points. After points are assigned to the "closest" centroid, the centroid is updated. Euclidian distance is often used. In the second step, points are assigned to the updated centroids and the centroids are updated again. When the k-means algorithm terminates, the centroids would have identified the natural grouping of points. For some combinations of proximity functions and types of centroids, k-means always converge to a solution i.e. k-means reaches a state in which no points are shifting from one cluster to another and hence the centroids do not change. Choosing the proper initial centroids should be the key step of the k-means algorithm which it often violates. Random selection is the common procedure but resulting clusters can produce poor quality clusters.

2.2 The Furthest-First Technique

Because of the problems with using randomly selected initial centroids, which even repeated runs may not overcome, alternative techniques have to be employed for initialization. Furthest-First (FF) is such a technique which selects initial cluster representatives in the k-means algorithm intelligently. The procedure selects the first point at random. Then, for each successive initial centroid, the point is selected that is furthest from the initial centroids already selected. In this way, we obtain a set of initial centroids that is guaranteed to be not only randomly selected but also well separated. The Furthest-First initialization technique is as follows-

- Repeat
- Pick the first center randomly
- Next center is the point furthest from the first center
- Third center is the point which is furthest from both the previous centers, in general the next center $= \arg\ \max_x\ \min_c\ \mathrm{d}(X, C)$
- Until k centers are selected

After picking up the first center randomly this algorithm selects the next center furthest from the first center. The distance is measured by Euclidian Distance. The step 4 indicates that the next is that point which has the maximum value among the points containing the distance value from it to the nearest center. The initial representatives selected by this technique are well separated from each other. Though using the Furthest-First initialization technique we can obtain a set of initial centroids well separated, such approach has the tendency to select the outliers, rather than points in dense regions (clusters) and as a result poor quality of clusters can result. In particular, when outliers are present, the resulting cluster centroids may not be as representatives as they otherwise would be and thus, the sum of squared errors (**SSE**) will be higher as well. As a result the quality of the cluster becomes lower. Also, it is expensive at the time of converges in the k-means algorithm.

To overcome such problems, the following approach is proposed in which the centroids selected by the Furthest-First method are pulled towards the middle of the set, which can produce better quality (lower **SSE)** clustering.

2.3 The Proposed Initialization Technique

The proposed initialization technique is as follows:

- Input: p, k
- Select the first center randomly
- Get a point furthest from the first center and shrink the point towards the first center by $p\%$ -this point is the second center.
- Get the point which is furthest from both the previous centers. Select the next center by shrinking that point by $p\%$ towards the closest center. In general: next center is the point X_i obtained by shrinking the point X by $p\%$, where X is given by $= \arg\ \max_X\ \min_C (X, C)$
- Repeat the above step until k centers are not selected.

It has explained that the Furthest-First technique has high probability to select outliers as the cluster representatives. The proposed technique brings the outliers to the dense space and also they are well separated. The point X_i obtained by shrinking the point X towards the closest center by $p\%$ is nothing but the point which divides the line joining the point X to the closest center in the ratio $100 - p : p$.

3. IMPLEMENTATION DETAILS

All the implementations of the standard k-means algorithm, k-means with Furthest-First initialization and k-means with the proposed initialization technique are implemented using data structure in C++. In the implementations, the input data matrices are taken in such a way that each row represents an instance. Usually in a standard dataset, instances are also in the form rows.

3.1 Implementation of F measure for quality measure

The datasets that have been used in our experiments are the classified datasets. The class labels of each data points are known. One of the quality measure techniques of the cluster is the F measure technique. The F measure is a measure that combines the **precision** and **recall** ideas from information retrieval. For cluster j and i, $recall(i, j) = \frac{n_{ij}}{n_i}$ and

$precision(i, j) = \frac{n_{ij}}{n_j}$, where n_{ij} is the numbers of members of class i in cluster j, n_j is the number of members in cluster j, n_i is the number of members in class i. Then the F measure of cluster j and class i is given by-

$$F(i, j) = \frac{2 \times recall(i, j) \times precision(i, j)}{precision(i, j) + recall(i, j)}$$

An overall value of F measure is computed by computing the weighted average of all values for the F measure is given by the following equation:

$F = \sum_i \frac{n_i}{n} \max_j \{F(i, j)\}$, Where, the maximum is taken over all the clusters at all levels and n is the total numbers of instances.

3.2 Datasets

We evaluated the quality of the clusters obtained from the standard k-means algorithm, k-means using Furthest-First initialization technique and k-means using the proposed technique on three standard large datasets which are available globally.

The first dataset [8], Shuttle Dataset, contains 9 attributes all of which are numerical. It consists of 1000 points; approximately 80% of the data belong to class 1.

The second dataset [9], Synthetic Control Chart Time dataset contains 600 examples of control charts synthetically generated. There are six different classes of control charts. The data stored in an ASCII file in where 600 rows, 60 columns with a single chart per line.

The third dataset [10], Wine Recognition Data are the result of a Chemical analysis of Wines grown in the same region in Italy but derived from different cultivars. The analysis determined the quantities of 13 constitutes found in each of the three types of Wines. This dataset contains 13 attributes and 150 data.

4. EXPERIMENTAL RESULTS

The experiments have been carried out to measure the quality of the cluster produced by the standard k-means algorithm, k-means using the Furthest-First (FF) initialization technique and k-means using the proposed initialization technique with the help of the three classified standard as well as large datasets discussed in the above section. The results are tabulated in table 1, 2 and 3. For each dataset, three values of k are considered- 10, 20 and 30. In all the implementations of the datasets, the value of p is 20 for the proposed initialization technique, which means after selection of the first center, next centers are selected by shrinking the further point towards the closest center by 20% is nothing but the point which divides the line joining the point to the closest center in the ratio 100-20: 20

Table 1. Qualities of the clusters for the Shuttle dataset

k	$k-means$	$k-means$ **With FF**	$k-means$ **with the proposed technique**
10	0.591	0.820	0.876
20	0.308	0.711	0.811
30	0.641	0.685	0.723

Table 2. Qualities of the clusters for the Synthetic Control Chart Time Series dataset using

k	$k-means$	$k-means$ **With FF**	$k-means$ **with the proposed technique**
10	0.609	0.844	0.791
20	0.789	0.801	0.890
30	0.806	0.689	0.699

Table 3. Qualities of the clusters for the Wine Recognition dataset

k	$k-means$	$k-means$ **With FF**	$k-means$ **with the proposed technique**
10	0.643	0.777	0.811
20	0.544	0.432	0.665
30	0.389	0.601	0.511

5. DISCUSSTIONS

From the experiments carried out, it should be accepted that the standard k-means algorithm is not the best of clustering method to get the best quality cluster in spite of its simplicity. Random selection of the initial cluster's centers is the main drawback of the algorithm. To overcome this drawback the intelligent selection is the only option to be assured to obtain the ideal clusters. From the experimental result, it can be seen that quality of clusters improves most of the times in k-means using the both of the alternative initialization techniques rather than the standard k means algorithm. It can be observed in the experimental results tabulated that most of cases the quality of the clusters obtained by F-measure are better in k-means using the proposed technique. But in few cases the standard k-means produces better clustering than the k-means with the two alternatives. The probable reason of those may be for character of the dataset or random selection in those cases may be very good. In table 2, k-means with FF produces better clustering in one case. The reason of it may be that the value of p in the proposed technique in that case might not be proper. Therefore, the ideal value p is also is important.

6. CONCLUSIONS

This article has tried to show that selecting the initial centers in the standard k-means algorithm intelligently than randomly is very important to get the better set of clusters. This can be concluded that the random selection of the initial cluster

centers is the main drawback of the k-means algorithm. From the comparisons carried out in this article, it has proved that intelligent selection of the initial centers produces better quality clusters.

The future scopes in the proposed initialization technique are:

- Determining the ideal value of p for a particular dataset.
- Determining the ideal value of p for a particular dataset.
- Examining the results inputting for some other values of k

7. ACKNOWLEDGMENTS

The authors would like to express their sincere thanks to the referees for their valuable comments and suggestions which help a lot for improving the presentation of this paper.

8. REFERENCES

[1] Pujari A. K. Clustering Techniques. Data mining techniques, chapter 5, University Press, pp. 114-130, 2008.

[2] Tan P., Steinbach M. and Kumar V. Introduction to Data Mining, Cluster Analysis: Basic Concepts and Algorithms, Chapter 8, Pearson Education, pp. 487-559, 2009.

[3] J. MacQueen. Some methods for Classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistic and Probability, Volume 1, pp. 281-297, 1967.

[4] Eklan C. Clustering with k-means: faster, smarter and cheaper, University of California, San Diego., April 24, 2004.

[5] Goswami A., Jin R., Agrawal G., Fast and Exact Out-of-Core K-Means Clustering, Department of Computer Science and Engineering Ohio State University, 2004.

[6] Arthur D., Vassilvitskii S.: "k-means++: The advantages of Careful Seeding" 2007 Symposium on Discrete Algorithms (SODA).

[7] Domings P. and Hulten G. A general method for scaling up machine learning algorithms and its application to clustering. In proceedings of the Eighteenth International Conference on Machine learning, 2001.

[8] Shuttle Dataset Available: http://mlr.cs.umass.edu/ml/datasets/stalog+(shuttle)

[9] Synthetic Control Chart Time Series Dataset Available: http://archive.ics.uci.edu/ml/datasets/synthetic+control+chart+time+series

[10] Wine Recognition Datasets Available: http://mlr.cs.umass.edu.edu/ml/datasets/wine