Volume 3, No. 6, June 2012

Journal of Global Research in Computer Science

REVIEW ARTICLE

Available Online at www.jgrcs.info

ISSN-2229-371X

NOVEL INITIALIZATION TECHNIQUE FOR K-MEANS CLUSTERING USING SPECTRAL CONSTRAINT PROTOTYPE

Mrs.S. Sujatha and Mrs. A. Shanthi Sona
Associate Professor School of IT&Science, Dr.G.R.Damodaran College of Science, Coimbatore.
sujathas05@gmail.com
Tiruppur Kumaran College for Women Tirupur.
srteam36@rediffmail.com

Abstract---Clustering is a general technique used to classify collection of data into groups of related objects. One of the most commonly used clustering techniques in practice is K-Means clustering. The major limitation in K-Means is its initialization technique. Several attempts have been made by many researchers to solve this particular issue, but still there is no effective technique available for better initialization in K-Means. In general, K-Means follows randomly generated initial starting points which often result in poor clustering results. The better clustering results of K-Means technique can be accomplished after several iterations. However, it is very complicated to decide the computation limit for obtaining better results. In this paper, a novel approach is proposed for better initialization technique for K-Means using Spectral Constraint Prototype (K-Means using SCP). The proposed method incorporates constraints as vertices. In order to incorporate the constraints as vertices, SCP approach is used. The proposed approach is tested on the UCI Machine learning repository. The proposed initialization provides better clustering accuracy with lesser execution time.

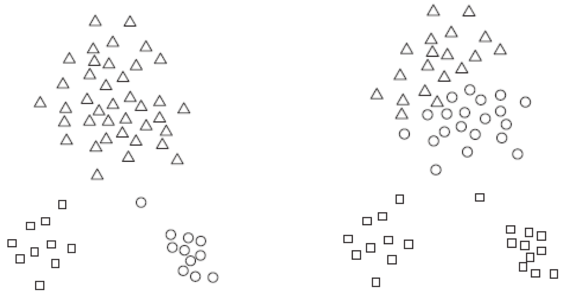
Keywords---Spectral Co-clustering, Semi-Unsupervised Gene Selection, K-Means, Initial Centroids, Spectral Constraint Prototype

INTRODUCTION

Cluster analysis is a technique which groups or clusters the available data into a meaningful or valuable cluster [1]. If meaningful groups are the objective, then the clusters are supposed to capture the expected structure of the data. In certain cases, on the other hand, cluster analysis is only a valuable basis for other purposes, for instance, data summarization. Whether for understanding or effectiveness, cluster analysis has significantly played a key role in a broad category of fields, such as psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. There have been many applications of cluster analysis to practical problems.

Prototype-based clustering approaches generate a one-level partitioning of the data objects. There are several techniques, but two of the most well-known are K-Means and K-medoid. K-Means defines a prototype in terms of a centroid [2], which is frequently the mean of a group of points, and is typically applied to objects in a continuous n-dimensional space. K- medoid defines a prototype in terms of a medoid, which is the most representative point for a group of points, and can be applied to a wide range of data since it requires only a proximity measure for a pair of objects. While a centroid almost never corresponds to an actual data point, a medoid, by its definition, must be an actual data point. In this section, focused solely on K-means, which is one of the oldest and most widely used clustering algorithms.

When random initialization of centroids is used, different runs of K-Means typically produce different total SSEs [3]. This is illustrated with the set of two- dimensional points shown in Figure 1.1, which has three natural clusters of points. Figure 1.1(a) shows a clustering solution that is the global minimum of the SSE for three clusters, while Figure 1.1(b) shows a suboptimal clustering that is only a local minimum. Choosing the proper initial centroids [4] is the key step of the basic K-Means procedure. A common approach is to choose the initial centroids randomly, but the resulting clusters are often poor.



(a) Optimal clustering (b) Suboptimal clustering

Figure 1.1 Three optimal and non-optimal clusters

The following procedure is another approach to selecting initial centroids. Select the first point at random or take the centroid of all points. Then, for each successive initial centroid, select the point that is farthest from any of the initial centroids already selected. In this way, a set of initial centroids obtained that is guaranteed to be not only randomly selected but also well separated. Unfortunately, such an approach can select outliers, rather than points in dense regions (clusters). Also, it is expensive to compute the farthest point from the current set of initial centroids. To overcome these problems, this approach is often applied to a sample of the points. Since outliers are rare, they tend not to show up in a random sample. In contrast, points from every dense region are likely to be included unless the sample size is very small. Also, the computation involved in finding the initial centroids is greatly reduced because the sample size is typically much smaller than the number of points.

Although the mentioned initialization algorithms can help finding good initial centers for some extent, they are quite complex and some use the K-Means algorithm as part of their algorithms, which still need to use the random method for cluster center initialization. The novel method to find the initial centroids of K-Means is presented in this paper. The methodology and the experimental results are presented in the following sections.

RELATED WORKS

Kohei Arai et al., [5] stated about Hierarchical K-means: an algorithm for centroids initialization for K-means. Initial starting points those generated randomly by K-Means often make the clustering results reaching the local optima. The better results of K-Means clustering can be achieved after computing more than one times. However, it is difficult to decide the computation limit, which can give the better result. In this paper, a new approach is proposed to optimize the initial centroids for K-means [6]. It utilizes all the clustering results of K-Means in certain times, even though some of them reach the local optima. Then, transform the result by combining with Hierarchical algorithm in order to determine the initial centroids for K-means. The experimental results show how effective the proposed method to improve the clustering results by K-means.

Manjunath Aradhya et al., [7] analyzed the rapid advance of computer technologies in data processing, collection, and storage has provided unparalleled opportunities to expand capabilities in production, services, communications, and research. However, the immense quantities of high-dimensional data renew the challenges to the state-of-the-art data mining techniques. Feature selection is an effective technique for dimension reduction and an essential step in successful data mining applications. It is a research area of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. Its direct benefits include the building simpler and more comprehensible models, improving data mining performance, and helping prepare, clean, and understand data. Briefly introduced the key components of feature selection, and review its developments with the growth of data mining. Then overview FSDM and the papers of FSDM10, which showcases of a vibrant research, held of some contemporary interests, new applications, and ongoing research efforts. Then examine nascent demands in the data-intensive applications and identify some potential lines of research that require multidisciplinary efforts.

A. M. Fahim et al. [8] proposed an efficient method for assigning data-points to clusters. The original K-Means algorithm is computationally very expensive because, all iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose- one similar to the K-Means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original K-Means algorithm. Hence there is no guarantee for the accuracy of the final clusters.

METHODOLOGY

One of the most popular clustering methods is K-Means clustering algorithm. It generates k points as initial centroids arbitrarily, where k is a user specified parameter. Each point is then assigned to the cluster with the closest centroid [9], [10], [11]. Then the centroid of each cluster is updated by taking the mean of the data points of each cluster. Some data points may move from one cluster to other cluster. Again new centroids are calculated and assign the data points to the suitable clusters. The assignment is repeated and update the centroids, until convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroids remain the same.

Although K-Means has the great advantage of being easy to implement, it has some drawbacks. The quality of the final clustering results of the K-Means algorithm highly depends on the arbitrary selection of the initial centroids. In the original K-Means algorithm, the initial centroids are chosen randomly and hence Different clusters are obtained for different runs for the same input data [12]. Moreover, the K-Means algorithm is computationally very expensive also.

The proposed method consists of two steps namely Spectral Co-clustering [13] and Incorporating Constraints as vertices [14]. The steps involved in this procedure are as follows.

Spectral Co-clustering:

Spectral biclustering can be carried out in the following three steps: data normalization, Bistochastization and seeded region growing clustering. The raw data can be arranged in one matrix. In this matrix, denoted by A, the rows and columns represent the data and the different conditions respectively. Then the data normalization is performed as follows. Take logarithm of the data. Carry out five to ten cycles of subtracting either the mean or median of the rows and columns and then perform five to ten cycles of row-column normalization.

Define $\bar{A}_i = (1/m)\sum_{j=1}^{m} A_{ij}$ to be the average of ith row, $\bar{A}_i = (1/n)\sum_{j=1}^{n} A_{ij}$ to be the average of th column, and $\bar{A}_{..} = (1/mn)\sum_{j=1}^{n}\ \sum_{j=1}^{m} A_{ij}$ to be the average of the whole matrix, where m is the number of rows and n the number of columns.

Bistochastization may be done as follows. First, a matrix of interactions is defined $K = (K_{ij})$ by $K_{ij} = A_{ij} - \bar{A}_i - \bar{A}_i.j + \bar{A}_{..}$. Then the Singular Value Decomposition (SVD) of the matrix K is computed as given by $= U \Lambda V^T$, where Λ is a diagonal matrix of the same dimension as K and with nonnegative diagonal elements in decreasing order, U and V are $m \times m$ and $n \times n$ orthonormal column matrices. The ith column of the matrix V is denoted by $\vec{v}_1$ and $\vec{v}_2$. Therefore, a scatter plot of experimental conditions of the two best class partitioning eigenvectors $\vec{v}_2$ and $\vec{v}_2$ is obtained. The $\vec{v}_1$ and $\vec{v}_2$ are often chosen as the eigenvectors corresponding to the largest and the second largest eigenvalues, respectively. The main reason is that they can capture most of the variance in the data and provide the optimal partition of different experimental conditions. In general, an s-dimensional scatter plot can be obtained by using eigenvectors $\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots. \overrightarrow{v_s}$ (with largest eigenvalues).

Define $P = [\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots. \overrightarrow{v_s}]^2$ which has a dimension of $n \times s$. The rows of matrix P stand for different conditions, which will be clustered using Seeded Region Growing (SRG). SRG clustering is carried out as follows. It begins with some seeds (initial state of the clusters). At each step of the algorithm, it is considered all as-yet unallocated samples, which border with at least one of the regions. Among them one sample, which has the minimum difference from its adjoining cluster, is allocated to its most similar adjoining cluster. Data can be clustered into several groups with very high accuracy. In the next section, such clustering result is used to select the best initial centroids.

Incorporating Constraints:

Assume that some vertices are believed to belong to the same cluster, one thus expects the co-clustering result to be consistent with the prior knowledge. Initially modeled the prior knowledge with a "must-link" constraint matrix C as

$$[C]_{ij} = \begin{cases} 1 \; if \; vi \; and \; vj \; are \; to \; be \; in \; the \; same \; cluster \\ \qquad\qquad 0 \; otherwise. \end{cases}$$

In the above equation, the vertex v_i and v_j can be either from vertex set V_r or V_c. Then decomposed the constraint matrix C as

$$C = \begin{bmatrix} C_{rr} & C_{rc} \\ C_{cr} & C_{cc} \end{bmatrix}$$

Where C_{rr} denotes the constraints of row vertices that are both in V_r and C_{rc} denotes the constraints of vertices with one in V_r and the other in V_c. C_{cr} and C_{cc} are defined similarly and $C_{cr} = C_{rc}^T$. Given a bipartite graph $G = (V_r, V_c, E)$, the co-clustering constraint is to maximize the following function:

$$\max_{G_1, G_2, \ldots, G_k} \sum_{v_i, v_j \in G_p} [C]_{ij}$$

The global optimization is thus to minimize the following function:

$$\max_{G_1, G_2, \ldots, G_k} \sum_{v_i \in G_p, v_j \in G_q, p \neq q} E_{ij} - \delta \sum_{v_i, v_j \in G_p} [C]_{ij}$$

Where δ is a constraint confidence parameter to regulate the importance of the constraints.

Incorporating Constraints as Additional Links:

Co-clustering constraints can be incorporated by directly modifying the bipartite graph. Given a bipartite graph $G = (V_r, V_c, E)$, if vertex v_1 and v_2 are preferred to be together, added an edge, or increase the edge weight between v_1 and v_2. From a matrix point of view, the adjacency matrix becomes:

$$M' = \begin{bmatrix} \delta C_{rr} & E + \delta C_{rc} \\ (E + \delta C_{rc})^T & \delta C_{cc} \end{bmatrix}$$

Note that in this case, the graph is no longer a bipartite graph since there may be links between any two vertices. In this case, traditional spectral co-clustering [15] cannot solve the problem directly, and carried out spectral partition on the whole graph. Take Ncut [16] as an example. The selected k eigenvectors of the matrix $D^{-1/2}(D - M')D^{-1/2}$ are used to give the solution, where $[D]_{ii} = \sum_j M'_{ij}$.

Incorporating Constraints as Vertices:

The above technique incorporates constraint as an additional link but in this paper the constraints are incorporated as vertices.

Another solution is to model the constraints as pseudo vertices. Intuitively, for each constraint, a pseudo vertex is generated and linked the pseudo vertex with the constrained vertices. More specifically, the following process is performed in order to represent the modified graph more conveniently. Suppose there are r row vertices and c column vertices. The algorithm then generates c pseudo row vertices and r pseudo column vertices. If two row vertices v_i and v_j are constrained to be together, then link v_i to the jth pseudo column vertex, and v_j to the i-th pseudo column vertex, similarly for the constrained column vertices. From a matrix view point, given a bipartite graph $(G = V_r, V_{c,} E)$, it changes to another bipartite graph $\hat{G} = V_r', V_c', E$ where

$$\hat{E} = \begin{bmatrix} E + \delta C_{rc} & \delta C_{rr} \\ \delta C_{cc} & 0 \end{bmatrix}$$

In this formula, the new graph $\hat{G}$ is still a bipartite graph, one can apply traditional spectral co-clustering to obtain the result. For both methods, the main computational cost is to calculate the eigenvectors of certain matrix. Consider Lanczos method to compute the eigenvectors. The complexity of both algorithms are $O(KN(|V_r| + |V_c|)^2)$ where k is the number of eigenvectors desired, N is the number of Lanczos iteration steps.

$(|V_r| + |V_c|)^2$ is the upper bound of the nonzero entries of the matrix M and $\hat{E}$. In the next section, a spectral approach is proposed to directly model the constraints into the formula with a more efficient implementation.

Constrained Co-clustering as Trace Minimization (The SCP Approach):

In this section, introduced the Spectral Constraint Prototype (SCP) algorithm that directly models the objective as trace minimization problem. First of all, given a bipartite graph $(G = V_r, V_{c,} E)$define the co-clustering partition matrix X as

$$X = \begin{bmatrix} X_r \\ X_c \end{bmatrix}$$

Where X_r is the partition on row vertex set V_r, and X_c is the partition on column vertex set V_c. The entry $[X_r]_{ij}$ if and only if the row vertex V_i belongs to cluster j. The normalized cut [16] on the bipartite graph is to minimize the following function:

$$\min_X tr(X^T LX)$$

where

$$L = D - M$$

$$D = \begin{bmatrix} D_r & 0 \\ 0 & D_c \end{bmatrix}$$

And D_r and D_c are diagonal matrices such that $[D_r]_{ij} = \sum_j E_{ij}$ and $[D_c]_{ii} = \sum_j E_{ji}$ The matrix L is called the Laplacian of the graph [16] that has several advantages such as symmetric and positive semidefinite. From the above modeled the co-clustering constraints as a trace norm minimization problem.

EXPERIMENTAL RESULTS

The proposed initialization technique for K-Means is experimented using two UCI Machine learning repository data sets: Lung Cancer Dataset and Lymphography Dataset.

Clustering Accuracy:

Clustering accuracy is calculated for Standard K-Means (random initialization technique), DPDA K-Means (deriving initial cluster centers from data partitioning along the data axis), K-Means using Constrained Spectral Co-clustering (CSC) and the proposed K-Means using SCP in lung cancer dataset and lymphography dataset. Figure 4.1 shows the comparison of the accuracy of clustering results for the proposed method with the standard K-Means, DPDA-K-Means and K-Means using CSC. From the figure, it can be observed that in both the dataset, the accuracy of clustering results of the proposed K-Means using SCP is better than the other methods.

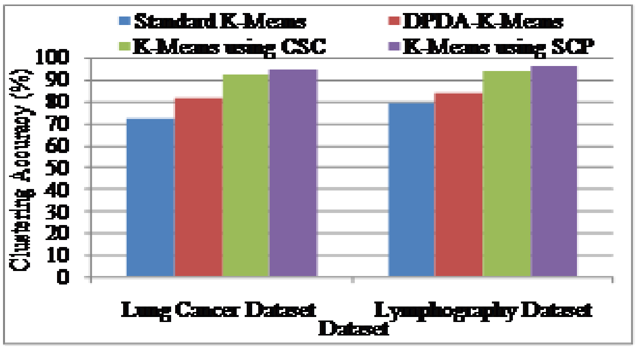


Figure 4.1: Comparison of Clustering Accuracy

Execution Time:

The execution time is calculated based on the running time of the clustering approaches on the two dataset. Figure 4.2 shows the execution time taken by the Standard K-Means, DPDA-K-Means, K-Means using CSC and the proposed K-Means using SCP.

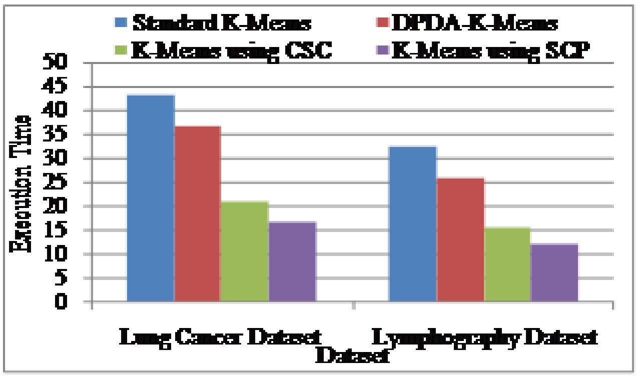


Figure 4.2: Comparison of Execution Time

It can be observed that the time required for execution using the proposed K-Means using SCP is very low, whereas, more time is needed by other clustering techniques for execution.

CONCLUSION

Numerous applications depend upon the clustering techniques. The most commonly used clustering technique is K-Means clustering. But the initialization K-Means often make the clustering results reaching the local optima. So to overcome this disadvantage a novel initialization technique is proposed. The novel initialization technique consists of two steps namely Spectral Co-clustering and Incorporating Constraints as vertices using Spectral Constraint Prototype. The experiments for this proposed initialization technique is conducted on two UCI Machine learning repository data sets. The data sets used are lung cancer dataset and lymphography dataset. From the results, it is revealed that the clustering accuracy of the proposed initialization technique using Spectral Constraint Prototype is very high when compared against the Standard K-Means, DPDA K-Means and K-Means using CSC. Furthermore, the experimental section also reveals that the proposed initialization technique takes very lesser time for execution than other techniques.

REFERENCES

[1] Shi Yong and Zhang Ge, “Research on an improved algorithm for cluster analysis”, International Conference on Consumer Electronics, Communications and Networks (CECNet), Pp. 598 – 601, 2011.

[2] B. Chen, P.C. Tai, R. Harrison and Yi Pan, “Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis”, IEEE Computational Systems Bioinformatics Conference, Pp. 105 – 108, 2005.

[3] P.S. Bradley and U.M. Fayyad, “Refining Initial Points for K-Means Clustering,” ACM, Proceedings of the 15th International Conference on Machine Learning, pp. 91-99, 1998.

[4] Yan Zhu, Jian Yu and Caiyan Jia, “Initializing K-means Clustering Using Affinity Propagation”, Ninth International Conference on Hybrid Intelligent Systems (HIS '09), Vol. 1, Pp. 338 – 343, 2009.

[5] Kohei Arai and Ali Ridho Barakbah, “Hierarchical K-means: an algorithm for centroids initialization for K-means”, Saga University, Vol. 36, No.1, 25-31, 2007.

[6] Madhu Yedla, Srinivasa Rao Pathakota and T. M. Srinivasa, “Enhancing K-Means Clustering Algorithm with Improved Initial Center”, Vol. 1, 121-125, 2010.

[7] Manjunath Aradhya, Francesco Masulli, and Stefano Rovetta “Biclustering of Microarray Data based on Modular Singular Value Decomposition”, Proceedings of CIBB 2009

[8] F. Yuan, Z.H. Meng, H.X. Zhang C.R. and Dong C. R, “A New Algorithm to Get the Initial Centroids,” Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.

[9] B. Borah and D.K. Bhattacharyya, “An Improved Sampling-based DBSCAN for Large Spatial Databases”. In Proceedings of the International Conference on Intelligent Sensing and Information, Pp. 92, 2004.

[10] Brian S. Everitt, “Cluster analysis”. Third Edition, 1993.

[11] M. Halkidi, Y. Batistakis and M. Vazirgiannis, “Clustering

Validity Checking Methods: Part II". In Proceedings of the ACM SIGMOD International Conference on Management of Data, Volume 31, Issue 3, pages19 – 27, September 2002.

[12] Jieming Wu, Wenhu Yu, "Optimization and Improvement Based on K-Means Cluster Algorithm", Second International Symposium on Knowledge Acquisition and Modeling (KAM '09), Vol. 3, Pp. 335 – 339, 2009.

[13] Yuval Kluger, Ronen Basri, Joseph T. Chang and Mark Gerstein, "Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions", Genome Research - GENOME RES, Vol. 13, No. 4, Pp. 703-716, 2003.

[14] Xiaoxiao Shi, Wei Fan and Philip S. Yu, "Efficient Semi-supervised Spectral Co-clustering with Constraints", IEEE International Conference on Data Mining, Pp. 1043-1048, 2010.

[15] I.S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning", Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001.

[16] J. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 8, Pp. 888–905. 2000.