Volume 2, Issue 10, October 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Data fusion in K-means Laplacian Clustering

**T.Kokilavani**
*Research Scholar*
*Department of Computer Science*
*Gobi Arts & Science College.*

**Dr.V.Thiagarasu**
*Associate Professor of Computer Science*
*Gobi Arts & Science College (Autonomous)*
*Gobichettipalaaym-638453.*

***Abstract: In clustering technique, the hard clustering membership values and overlapping concept could not be identified along with non-convex problem. The proposed algorithm uses soft clustering to combine both Laplacians and multiple kernels for clustering analysis. The algorithm is formulated on a Rayleigh quotient objective function. The bi-level optimization is an alternating minimization procedure; it is used to convert the hard clustering to soft clustering. The kernels and Laplacians co-efficient can be optimized automatically by using the methods semi-infinite programming and quadratic constraint quadratic programming .The kernel Laplacians algorithm uses to control the overlapping.***

***Keywords: Clustering, k-means, Data fusion, Data clustering, Soft clustering, hard clustering.***

## I. INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In pattern recognition, data analysis is concerned with predictive modeling: given some training data, to predict the behavior of the unseen test data. This task is also referred to as learning. Often, a clear distinction is made between learning problems that are (i) supervised (classification) or (ii) unsupervised (clustering), the first involving only labeled data (training patterns with known category labels) while the latter involving only unlabeled data. The goal data clustering is to discover the natural grouping(s) of a set of patterns, points, or objects. It defines cluster analysis as "a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics."

Data clustering has been used for the following three main purposes. (i)Underlying structure: to gain insight into data, generate hypotheses, detect anomalies, and identify salient features. (ii)Natural classification: to identify the degree of similarity among forms or organisms (phylogenetic relationship). (iii)Compression: as a method for organizing the data and summarizing it through cluster prototypes. The idea of incorporating several data sources in analysis may be beneficial by reducing the noise, as well as by improving statistical significance and leveraging the interactions and correlations between data sources to obtain more refined and higher-level information, which is known as data fusion.

In bioinformatics [21], considerable effort has been devoted to genomic data fusion, which is an emerging topic pertaining to a lot of applications. At present, terabytes of data are generated by high-throughput techniques at an increasing rate. In data fusion, these tera bytes are further multiplied by the number of data sources or the number of species.

Many algorithms are either similarity-based model-based. In general, the former (e.g., normalized cut) requires no assumption on data densities but simply a similarity function, and usually partitions data exclusively into clusters. In contrast, model-based methods apply mixture models to fit data distributions and assign data to clusters (i.e. mixture components) probabilistically. This soft clustering is often desired, as it encodes uncertainties on data-to-cluster assignments. However, their density assumptions can sometimes be restrictive, e.g. clusters have to be Gaussian-like in Gaussian mixture models. In overlapping clustering as the problem of mapping [13] each data point to a small set of labels that represent cluster membership. The number of labels does not have to be the same for all data points. The objective is to find a mapping so that the similarity between any pair of points in the dataset agrees as much as possible with the similarity of their corresponding sets of labels. In the formulation of complete graph as input, but every edge is associated with a weight, which is a number in [0; 1]. Weights represent similarity between data points and the extent to which data points should be assigned to the same cluster. For defining distances between sets of labels, it considers two measures: a set-intersection indicator function and the Jaccard coefficient. The maximum number of cluster labels allowed, either globally or per data point.

## II. PROBLEM FORMULATION

The objective function is difficult to optimize analytically because data are unlabeled; moreover, the discrete cluster memberships make the problem NP hard. In this multi-assignment step objects are assigned to poles plays an important role in the construction of "overlapping-clusters". The advantage of assigning an object to several clusters is well-known;

 

the assignment method is often based on an arbitrary threshold applied on a membership matrix obtained with a clustering method.

The approach is focused in the overlapping clustering techniques trying to a well-known classical partition technique named K-means using a genetic algorithm approach. K-means is a clustering algorithm that uses a fixed number (K) of clusters and looks for the best division of the dataset (through a predefined metric or distance) in this number of groups.

To finding the problems the K-means cannot be directly applied because it does not allow overlapping. In contrast, it is common for communities to share members. An alternative solution could be soft clustering which allows every one element to belong to several clusters giving a probability of membership, therefore same kind of overlapping for an element can be considered.

## III. LITERATURE SURVEY

Most efficient clustering algorithms are first development in the low dimensional space so the cluster the data in the lower dimensional space. Typically, these algorithms also work under a separation requirement, which is measured by the minimum distance between the means of any two mixture components. The k-means clustering algorithm is used for binning the biometric modalities in our approach. Using the k-means, the binning of feature vector is performed to set of input vectors is divided into N bins performs a s hard-partition of the entire data-space. It also reduces the penetration rate and consequently the false acceptance rate. In the finger print segmentation task, the background blocks and foreground blocks of most fingerprints (fair quality) fall into two clusters with high density, and the number of clusters can be directly set to the statistical. Separately of the two clusters helps the k-means algorithm to achieve good clustering performance.

In the Fingerprint biometrics k-means is performed to cluster the blocks into two clusters, that is, background and foreground cluster. The outputs of k-means are two clusters without specifying background or foreground, preliminary classification process to determine which cluster indicates the foreground blocks.

## IV. PROPOSED SYSTEM

The proposed system is a novel algorithm to combine multiple kernels and Laplacians for clustering analysis. The new algorithm is formulated on a Rayleigh quotient objective function and is solved as a bi-level alternating minimization procedure. Using the algorithm, the coefficients of kernels and Laplacians can be optimized automatically.

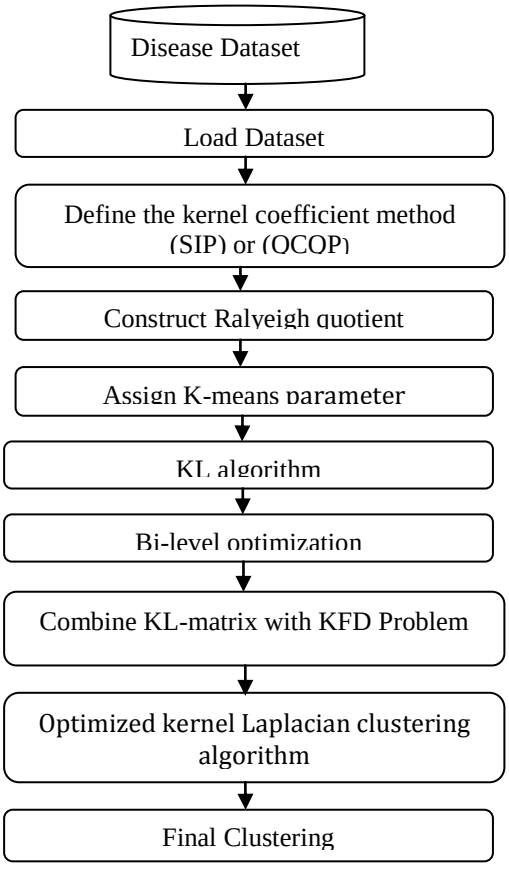


Figure 1. The process flow of proposed clustering.

 

In the Fig(1), The disease dataset as input to be find the final clustering result along with the KL algorithm and bi-level optimization procedure technique.

The kernel co-efficient values are generated to make a supervised data (i.e., trained data). Using k-means how many group of clusters to be formed with the data set. Ralyeigh quotient function to find out the Hermitian matrix and non-zero vector for finding the number of data points belonging to cluster.

In KL algorithm to solve the membership values and degree of assigned on Laplacians can be solved via the following KFD problem, to optimize the bi-level procedure determining the pair wise similarity to describing hyper plane high-dimensional feature space in projection matrix.

The bi-level optimization procedure algorithm can be easily extended to incorporate different criteria in clustering and KFD. It is also possible to deal with overlapping cluster membership, known as "soft clustering". In many applications such as bioinformatics, a gene or protein may be simultaneously related to several biomedical concepts so it is necessary to have a "soft clustering" algorithm to combine multiple data sources.

*A.Optimized Kernel Laplacian Clustering*

The coefficients assigned to the Laplacians and the kernels are optimized in parallel. To terminate the iterations when the cluster membership matrix is stops changing. The tolerance value $\epsilon$ is a constant value as the stopping rule of KL, and implementation it is set to 0.05. In the implementation, the final cluster assignment is obtained using the KL algorithm on $A(\gamma)$.

## V. OVERLAPPING CLUSTERING

Soft clustering algorithms which can identify overlapping clustering mixture modeling and then make a hard assignment of each data point to one or more clusters using a threshold on cluster membership probability.

The main idea is to formulate overlapping clustering as an optimization problem in which each data point is mapped to a small set of labels, representing membership to different clusters. The objective is to find a mapping so that the distances between data points agree as much as possible with distances taken over their label sets.

To solve the main optimization problem, suggest a bi-level optimization. The iterative step of our algorithm requires solving non-convex optimization sub problems, solving using a multiple kernels and Laplacians clustering.

*A.Non-Overlapping Clustering*

In non-overlapping clustering, the k cluster labels, denoted by $L = \{1,\ldots, n\}$, and the task is to assign cluster labels for each object in O . The clustering is defined by a labeling function $l : O \rightarrow L$. The objective is to assign labels to objects so that, to the largest possible extent, similar objects get assigned the same label. The correlation-clustering problem provides a precise formulation for such an objective The intuition underlying the above problem is that if two objects u and v are assigned to the same cluster we should pay the amount of their dissimilarity $1 - s(u, v)$, while if they are assigned to different clusters we should pay the amount of their similarity s(u, v). In the expresses of cost function as the number of object pairs that have similarity 0 and are clustered together plus the number of object pairs that have similarity 1 and are clustered in different clusters.

## VI. ALGORITHM DESIGN

*A. Clustering by Kernels and Laplacians with Bi-level Optimization:*

An approach to find out the optimal-weighted convex linear combinations, to extend by incorporating multiple kernels and Laplacians for clustering,

To optimize Laplacians matrix the generalized KL algorithm with Rayleigh quotient objective,

$$L=D^{-1/2}WD^{-1/2}$$

Where, D denotes diagonal matrix and W denotes Affinity matrix.

The trace maximization of the between-cluster[18], the norm vectors of discriminant projections are merely the eigenvectors[20] of the between-class scatter

$$\text{Maximize trace } (A^T X^T X A)$$

The objective of KL integration becomes,

$$\text{Minimize trace } (A^T LA)\text{-}(1\text{-}K) \text{ trace}$$

 

$(A^T G_C A)$

Subject to $A^T A = I_K$

$$0 \le \kappa \le 1.$$

The algorithm formulate kernel coefficient learning as semi-infinite programming (SIP) [15], which is much more accomplished and able to be changed in size of cluster and then the quadratic constraint quadratic programming (QCQP) [14] also proposed.

*B. Mathematical Expression*

To design and develop a mathematical model for Kernel and Laplacians as generalized Ralyeigh quotient for clustering: The objective of the KL integration to combine X (attribute dataset) and *W* (graph affinity matrix) for clustering can be defined as

$$O_{KL} = kJ_{SC} + (1-k)J_{KM}$$

Where $J_{SC}$ and $J_{KM}$ are, respectively, the objectives of SC and KM clustering, $\kappa \in [0, 1]$ is a coefficient adjusting the effect of the two objectives.

The weighted scalar cluster membership matrix is given by,

$$A_{ab} = \begin{cases} \dfrac{1}{\sqrt{n_b}} & \text{if } \vec{x}_a \in C_b \\ 0 & \text{if } \vec{x}_a \notin C_b, \end{cases}$$

Where $n_b$ is the number of data points belonging to cluster $C_{b...}$ $A^T A = IK$, the objective of k-means in [16] can be equivalently written as where *IK* denotes a K X K identity matrix.

TABLE I. KERNEL LAPLACIANS AND BI-LEVEL BI-LEVEL OPTIMIZATION ALGORITHM

| | |
|---|---|
| Step 1: | Construct a objective function of kernels and Laplacians clustering with optimal-weighted convex linear combinations. |
| Step 2: | The optimal coefficients assigned to the Laplacians and the kernels. The kernel parameter is replaced by the coefficients assigned on each individual data sources. |
| Step 3: | To solve the Rayleigh quotient problem with eigenvalue problem to reduces the optimization coefficients with given cluster memberships. |
| Step 4: | To determining the pair wise discriminating hyper plane high-dimensional feature space in projection matrix of KFD. |
| Step 5: | To solve the bi-level optimization of Step 2 to 4 to optimize toward the same objective as a Rayleigh quotient iterative optimization converges to a local optimum. |

The algorithm is used to control the overlapping in clustering the data. The performance is scientifically validated on two real-life data fusion applications. The proposed Optimized Kernel Laplacian Clustering (OKLC) algorithms perform a

 

lot better than other methods. The kernel parameter is replaced by the coefficients assigned on each individual data sources. To solve the Rayleigh quotient problem with eigen value problem to reduces the optimization coefficients with given cluster memberships

TABLE II

OPTIMIZED KERNEL LAPLACIANS CLUSTERING

| | |
|---|---|
| Step 1: | To Fix the kernel-Laplacian combined matrix and optimal membership can be found as the dominant K number of eigenvectors of Ω. |
| Step 2: | The membership and optimal angle assigned on Laplacians can be solved via the following KFD problem. |
| Step 3: | The threshold value is set to 1, the optimized solution is sparse, which assigns dominant values to only one or two Laplacians (kernels) and zero values to the others. |
| Step 4: | To distinguish relevant sources from a large number of irrelevant data sources from step 3. |
| Step 5: | Set the sparse co-efficient values(1,2,3) to optimize the sparse, average and non-sparse coefficients on kernels and Laplacians. |

## VII. APPLICATIONS

Developments in overlapping clustering have mainly been driven by the concrete needs of applications. For instance, driven by the need to cluster microarray gene expression data, various methods for overlapping clustering and overlapping bi-clustering have been proposed. To detecting communities in social networks is a problem that has been studied extensively, only few researchers have addressed the problem of detecting overlapping communities.

To finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records. In web applications the document classification; clustering weblog data to discover groups of similar access patterns. It also used to the Information ideal with several thematic for a single document, biological data with several metabolic functions for one gene.

## VIII. RESULT

The algorithm is proposed that used to control the overlapping in clustering the data. The performance is scientifically validated on two real-life data fusion applications. The proposed Optimized Kernel Laplacian Clustering (OKLC) algorithms perform a lot better than other methods. The performance of the proposed bi-level optimization algorithm on the Disease gene data set with bi-level optimization kernel k –means.

To gain a meaning cluster and calculations is compute all the pairwise combination of 29 disease The related gene of each pair disease combinations are selected and cluster into two group then the performance is evaluated using the diseases lables.The gene related to both disease in paired combination are removed before clustering in total there are less than 5 percent gene being remove. Finally the average performance all 406 paired combination is used as the overall clustering performance. Moreover, the coefficients of kernels and Laplacians optimized by OKLC show some correlation with the rank of performance of individual data source. Though in our evaluation the *K* values are predefined, in practical studies, the optimal cluster number can be consistently estimated from the Eigen spectrum of the combined kernel Laplacians matrix.

TABLE III. PERFORMANCE OF GENE DISEASE DATASET

| Algorithm | KL | CSPA | HGPA | MCLA |
|---|---|---|---|---|
| Similarity between two data clustering's | 0.7336 | 0.6703 | 0.6673 | 0.6571 |

 

| | | | | |
|---|---|---|---|---|
| P-Value | - | 8.84 | 4.7 | 6.55 |
| Mutual information | 0.7758 | 0.7173 | 0.7141 | 0.7128 |
| P-value | - | 1.2 | 5.1 | 2.31 |

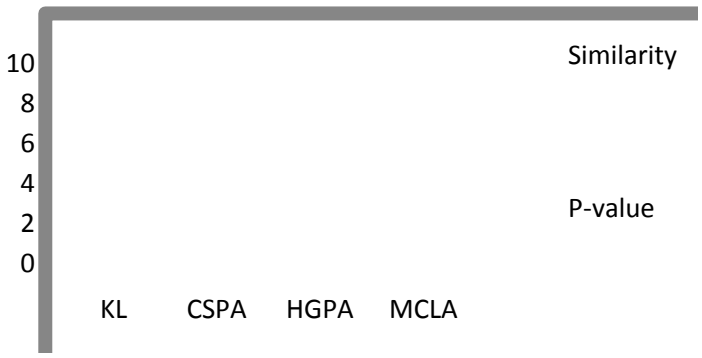


Figure 2.( a) Chart show the performance of Gene Data Disease set

The Fig 2(a) show the performance on the gene disease data set using the algorithm KL control over the overlapping. The algorithm CSPA, HGPA and MCLA are clustering group method propose [22] to optimize mixture Co-efficient.

TABLE IV
PERFORMANCE OF JOURNAL DATASET

 

| Algorithm | Similarity between two data clustering 's | P-Value | Mutual inform ation | P-value |
|---|---|---|---|---|
| KL | 0.5859 | - | 0.7451 | - |
| CSPA | 0.4367 | 5.66 | 0.6362 | 4.23 |
| HGPA | 0.5040 | 8.47 | 0.6872 | 7.42 |
| MCLA | 0.4731 | 2.26 | 0.6519 | 5.26 |

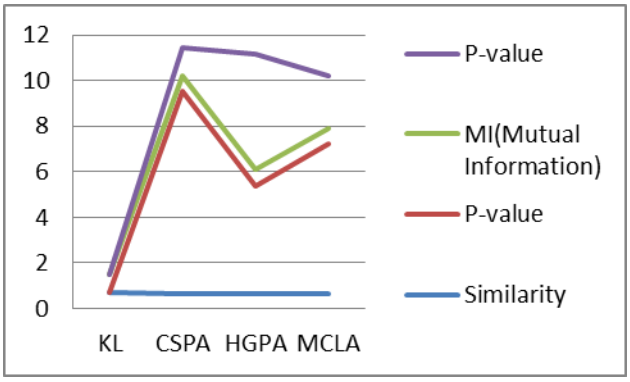


Fig 2 ( b) Chart show the performance of Journal Disease set

The Fig 2(b) show the performance Journal Disease set that not most successful. The KL and bi-level optimize is comparable to the best which indicates that the optimize data fusion improve performance.

## IX. CONCLUSION

 

To discover the optimized data fusion KL algorithm indeed improves the performance on artificial data, digit identification data and disease gene data. In the final step of data integration process the problem of data fusion in the large environment has been identified. The proposed KL algorithm with bi-level optimization process demonstrates the advantage of combining and the maximum advantage of information from different data structures and sources. It is actually useful in biomedical and many other applications, where there is a flow of interest to combine match based information and interaction based relationships in numerical analysis and machine learning.

## X. REFERENCES

[1] S. Yu, T. Falck, A. Daemen, L.C. Tranchevent, J. Suykens, B. De Moor, and Y. Moreau, "L2-Norm Multiple Kernel Learning and Its Application to Biomedical Data Fusion," BMC Bioinformatics, vol. 11, no. 309, pp. 1-53, 2010.
[2] Girolami M. Mercer kernel-based clustering in feature space. IEEE Trans. Neural Netw. 2002;13:780-784.
[3] Mika S, et al. Fisher discriminant analysis with kernels. IEE N.N. Singal. Process. 1999;9:41-48
[4] Suykens JAK, et al. Least Squares Support Vector Machines. Singapore: World Scientific Publishing; 2002
[5] Ye J, et al. Multi-class discriminant kernel learning via convex programming. J. Mach. Learn. Res. 2008;9:719-758.
[6] Yu S, et al. Gene prioritization and clustering by multi-view text mining. BMC Bioinformatics 2010a;11:1-28.
[7] Liu X, et al.Weighted hybrid clustering by combining text mining and bibliometrics on large-scale journal database. J. Am. Soc. Inform. Sci. Technol.2010;61:1105-1119.
[8] Strehl A, Ghosh J . Cluster ensembles: a knowledge Reuse framework for combining multiple partitions. J. Mach. Learn. Res. 2002;3:583-617.
[9] Yu S, et al L2-norm multiple kernel learning and its application to biomedical data fusion. BMC Bioinformatics 2010b;11:1-53.
[10]Von Luxburg U. A tutorial on spectral clustering. Stat. Compute. 2007;17:395-416.
[11] Fred ALN, Jain AK Combining multiple clusterings using evidence accumulation. IEEE Trans. PAMI 2005;27:835-85
[12] R. Hettich and K.O. Kortanek, "Semi-Infinite Programming: Theory,Methods, and Applications," SIAM Rev., vol. 35., no. 3, pp. 380-429, 1993.
[13] W. Tang, Z. Lu, and I.S. Dhillon, "Clustering with Multiple Graphs," Proc.IEEE Ninth Int'l Conf. Data Mining, 2009.
[14] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge Univ. Press,2004.
[15] R. Hettich and K.O. Kortanek, "Semi-Infinite Programming: Theory,Methods, and Applications," SIAM Rev., vol. 35., no. 3, pp. 380-429, 1993.
[16] H.G. Ayad and M.S. Kamel, "Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters," IEEE Trans. Pattern Analysis.
[17] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral Relaxation for KMeans Clustering," Proc.Advances in Nerual Information Processing, vol. 14,pp. 1057-1064, 2001.
[18] S. Theodoridis and K. Koutroumbas, Pattern Recognition, fourth ed.Academic Press, 2009.
[19] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.
[20] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, second ed. JohnWiley & Sons, 2001.
[21] S. Yu, L.-C. Tranchevent, B. De Moor, and Y. Moreau, "Gene Prioritizationand Clustering by Multi-View Text Mining," BMC Bioinformatics, vol. 11,no. 28, pp. 1-48, 2010.
[22] A. Strehl and J. Ghosh, "Clustering Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

 