Volume 5, No. 4, April 2014 (Special Issue)
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

Membership Optimization for Clustering Process in Machine Intelligence-A Practical Approach

Asrar Ahmad[1], V.M. Thakare[2], M.A. Atique[3]
SGBAU, Amravati
asrar60@gmail.com[1], vilthakare@yahoo.co.in[2], mohd.atique@gmail.com[3]

Abstract: Clustering is an unsupervised method to divide data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. Most of the approaches perform hard clustering, i.e., they assign each item to a single cluster. This works well when clustering compact and well-separated groups of data, but in many real-world situations, clusters overlap. Thus, for items that belong to two or more clusters, it may be more appropriate to assign them with gradual memberships to avoid coarse-grained assignments of data. The objective of k-means clustering is formulated as a Rayleigh quotient function of the between-cluster scatter and the cluster membership matrix and further combined with nonlinear dimensionality reduction in Hilbert space, where heterogeneous data sources can be easily combined as kernel matrices. The objective to optimizing the kernel combination and the cluster memberships on unlabeled data is non-convex. To solve it, minimization method to optimize the cluster memberships and the kernel coefficients iteratively to convergence is to be applied. Instead of a single fixed kernel, multiple kernels may be used. Recent developments in multiple kernel learning have shown that the construction of a kernel from a number of basis kernels allows for more flexible encoding of domain knowledge from different sources or cues. We here extend the multiple kernel-learning paradigm to fuzzy clustering. The proposed algorithm simultaneously finds the best degrees of membership and the optimal kernel weights for a nonnegative combination of a set of kernels. We also embed the feature weight computation into the clustering procedure. The incorporation of multiple kernels and the automatic adjustment of kernel weights render MKFC more immune to unreliable features or kernels. It also makes combining kernels more practical, since appropriate weights are assigned automatically.

I. INTRODUCTION

Clustering is an unsupervised method to divide data into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. In this paper, numbers of different optimization methods for clustering process are reviewed. All these methods are compared with their complexities and computational time. Most of these approaches perform hard clustering, i.e., they assign each item to a single cluster. This works well when clustering compact and well-separated groups of data, but in many real-world situations, clusters overlap. Thus, for items that belong to two or more clusters, it may be more appropriate to assign them with gradual memberships to avoid coarse-grained assignments of data. This class of clustering methods is called soft- or fuzzy-clustering.In this paper, Optimized data fusion for kernel *k*-means clustering (OKKC) [1], clustering sentence level text using a novel fuzzy relational clustering (FRECCA) [2] and multiple kernel fuzzy clustering (MKFC) [3] algorithms are proposed for membership optimization.

Clustering response is a primitive exploratory approach in data analysis with little or no prior knowledge. In cluster analysis, a group of objects is divided into a number of more or less homogeneous subgroups on the basis of a subjectively chosen measure of similarity, such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups. The main challenge for most of clustering algorithms is their necessity to know the number of clusters for which to look. Some researchers have tried to estimate or determine it automatically. This issue of obtaining the clusters that better fit a dataset, as well as their evaluation, has been the subject of almost all research efforts in this field. Since the clustering is an unsupervised problem, in most cases, the user has no prior knowledge about the actual number of clusters. Obviously, splitting the dataset into smaller or larger clusters will eventuate in merging some separate clusters or breaking down some compact ones. The problem of finding an optimal number of clusters is usually called cluster validity. Once the clusters are obtained by a clustering method, the validity function can help us to verify whether they accurately present the structure of the dataset or not.

II. BACKGROUND

The objective of *k*-means clustering is formulated as a Rayleigh quotient function of the between-cluster scatter and the cluster membership matrix and further combined with nonlinear dimensionality reduction in Hilbert space, where heterogeneous data sources can be easily combined as kernel matrices. The objective to optimizing the kernel combination and the cluster memberships on unlabeled data is non-convex. To solve it, an alternating minimization method is applied to optimize the cluster memberships and the kernel coefficients iteratively to convergence. When the cluster membership is given, the kernel coefficients as kernel Fisher discriminants (KFD) using least-squares support vector machine (LS-SVM) is to be optimized. The objectives of KFD and k-means are combined in a unified model; thus the two components optimize toward the same objective; therefore, the OKKC algorithm [1] solving this objective converges locally. The OKKC method extends the idea of Multiple Kernel Learning to an unsupervised problem. In comparison with hard clustering methods, in which a pattern belongs to a single cluster, fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. However, because most sentence similarity measures do not represent sentences in a common metric space, conventional fuzzy clustering approaches based on prototypes or mixtures of

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

Gaussians are generally not applicable to sentence clustering.

Clustering text at the document level is well established in the Information Retrieval (IR) literature, where documents are typically represented as data points in a high-dimensional vector space in which each dimension corresponds to a unique keyword, leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents (e.g., tf-idf values of the keywords). This type of data, which we refer to as "attribute data," is amenable to clustering by a large range of algorithms. Since data points lie in a metric space, we can readily appl y prototype-based algorithms such as k-Means, Isodata, Fuzzy c-Means (FCM) and the closely related mixture model approach, all of which represent clusters in terms of parameters such as means and co-variances, and therefore assume a common metric input space. Since pair-wise similarities or dissimilarities between data points can readily be calculated from the attribute data using similarity measures such as cosine similarity, we can also apply relational clustering algorithms such as Spectral Clustering and Affinity Propagation. To distinguish it from attribute data, we refer to such data as "relational data." A broad range of hierarchical clustering algorithms can also be applied.

A novel fuzzy relational clustering algorithm (FRECCA) [2] is inspired by the mixture model approach; model the data as a combination of components. However, unlike conventional mixture models, which operate in a Euclidean space and use a likelihood function parameterized by the means and co-variances of Gaussian components, The FRECCA algorithm abandon use of any explicit density model (e.g., Gaussian) for representing clusters. Instead, it use a graph representation in which nodes represent objects, and weighted edges represent the similarity between objects. Cluster membership values for each node represent the degree to which t he object represented by that node belongs to each of the respective clusters, and mixing coefficients represent the probability of an object having been generated from that component. By applying the PageRank algorithm to each cluster, and interpreting the PageRank score of an object within some cluster as likelihood, it can then use the Expectation-Maximization (EM) framework to determine the model parameters (i.e., cluster membership values and mixing coefficients). The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair-wise similarities.

Fuzzy c-means (FCM) is one of the most promising fuzzy clustering methods. In most cases, it is moreflexible than the corresponding hard-clustering algorithms. Unfortunately, as with other clustering methods that are based on the L2 -norm distance in the observation space, it has been shown that while it is effective for spherical clusters, it does not perform well for more general clusters. Thus, kernel-based clustering has been pro-posed to perform clustering in a typically higher dimensional feature space spanned by embedding maps and corresponding kernel functions. The FCM algorithm has also been extended to the kernel FCM algorithm, which yields better performance. However, for such kernel-based methods, a crucial step is the combination or selection of the best kernels among an extensive range of possibilities. This step is often heavily influenced by prior knowledge about the data and by the patterns that we expect to discover. Unfortunately, it is unclear which kernels are more suitable for a particular task.

The problem is aggravated for many real-world clustering applications, in which there are multiple potentially useful cues. For such applications, to apply kernel-based clustering, it is often necessary to aggregate features from different sources into a single aggregated feature. However, these features are often not equally relevant to clustering; some are irrelevant, and some are less important than others.

A multiple kernel fuzzy c-means (MKFC) [3] algorithm that extends the fuzzy c-means algorithm with a multiple kernel-learning setting finds the best degrees of membership and the optimal kernel weights for a nonnegative combination of a set of kernels. By incorporating multiple kernels and automatically adjusting the kernel weights, MKFC is more immune to ineffective kernels and irrelevant features. This makes the choice of kernels less crucial.

Instead of a single fixed kernel, multiple kernels may be used. Recent developments in multiple kernel learning have shown that the construction of a kernel from a number of basis kernels allows for more flexible encoding of domain knowledge from different sources or cues.

The rest of this paper is organized as follows. In Section 3, we discuss previous work done on various methodologies, and in Section 4, we discuss the analysis and discussion. We present the proposed methodology in Section 5, and we present possible outcome and results in Section 6. We conclude this paper in Section 7.

III. PREVIOUS WORK

Over the past decades, many clustering algorithms have been proposed, including k-means clustering, mixture models, spectral clustering, locality-sensitive hashing, and maximum margin clustering. Most of these approaches perform hard clustering, i.e., they assign each item to a single cluster. This works well when clustering compact and well-separated groups of data, but in many real-world situations, clusters overlap.

To learn the optimal combination of multiple information sources as similarity matrices (kernel matrices), Lange and Buhmann's algorithm is proposed. Lange and Buhmann's algorithm uses non negative matrix factorization to maximize posteriori estimates of data point assignments to partitions. To combine the similarity matrices, a cross-entropy objective is minimized to seek good factorization and the weights assigned on similarity matrices are optimized.

The OKKC algorithm [1] is related to the Nonlinear Adaptive Metric Learning (NAML) algorithm proposed for clustering. Although NAML is also based on multiple kernel extension of k-means clustering, the mathematical objective and the solution are different from OKKC. In NAML, the metric of k-means is constructed based on the Mahalanobis distance. NAML optimizes the objective iteratively at three levels: the cluster assignments, the kernel coefficients, and the projection in the Representer Theorem. The k-means objective in OKKC approach is constructed in Euclidean space and the algorithm optimizes the cluster assignments and kernel coefficients in a bi-level procedure.

The first successful fuzzy relational clustering model is generally considered to be Hathaway et al.'s Relational

Fuzzy c-Means (RFCM) algorithm. However, RFCM is a variant of Fuzzy c-Means, and is implicitly based on the notion of prototype. Thus, while RFCM operates on relational data input, it still requires that the relation expressed by this data be Euclidean (i.e., it assumes that there exists a set of data points in some space such that the squared Euclidean distance between points in this space match those in the dissimilarity relation). Non-Euclidean relations can be transformed into Euclidean ones by a transformation that adds a positive number to all off-diagonal elements of the dissimilarity matrix, but the problem is to determine an appropriate value for such that this Euclidean condition is met without leading to excessive loss of cluster information.

The k-Medoid family of algorithms are based on the observation that in k-Means (and most other prototype-based algorithms), the only step that involves calculating Euclidean distances is the minimization step, in which cluster means and co-variances are updated. By restricting prototypes to being data points, k-Medoid algorithms avoid the need to calculate distances, since all calculations can be performed on the basis of pair wise relations. This idea forms the basis of the Partitioning Around Medoids (PAM) algorithm, which performs hard (or crisp) clustering. Fuzzy versions of k-Medoids have also been proposed. Like k-Means, methods based on k-Medoids are highly sensitive to the initial (random) selection of centroids, and in practice it is often necessary to run the algorithm several times from different initializations. To overcome these problems, Frey & Dueck proposed Affinity Propagation, a technique which simultaneously considers all data points as potential centroids. Frey and Dueck have shown how Affinity Propagation can be applied to the problem of extracting representative sentences from text. A fuzzy variant of Affinity Propagation was recently proposed in Gewiniger.

A novel fuzzy relational clustering algorithm [2] is inspired by the mixture model approach, it model the data as a combination of components. The result is a fuzzy relational clustering algorithm which is generic in nature, and can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities.

The problem is aggravated for many real-world clustering applications, in which there are multiple potentially useful cues. For such applications, to apply kernel-based clustering, it is often necessary to aggregate features from different sources into a single aggregated feature. However, these features are often not equally relevant to clustering; some are irrelevant, and some are less important than others. As most clustering methods do not embed a feature selection capability, such feature imbalances often necessitate an additional process of feature selection, or feature fusion, before clustering. Instead of a single fixed kernel, multiple kernels may be used. Recent developments in multiple kernel learning have shown that the construction of a kernel from a number of basis kernels allows for more flexible encoding of domain knowledge from different sources or cues. However, as observed by Zhao *et al.*, previous multiple kernel-learning approaches have focused on supervised and semi-supervised learning. A notable exception is their work on multiple kernel maximum margins clustering, which is designed for hard clustering.

IV. ANALYSIS AND DISCUSSION

A. *Optimized Data Fusion for Kernel k-means Clustering:*

a. *Objective Of K-Means Clustering:*

In k-means clustering, a number of k prototypes are used to characterize the data, and the partitions $\{C_j\}$ j=1….k are determined by minimizing the distortion as

$$\min \sum_{j=1}^{k} \sum_{\vec{x}_i \in \mathcal{C}_j} \|\vec{x}_i - \vec{\mu}_j\|^2,$$

Where $\vec{x_i}$ is the ith data sample, $\vec{\mu_j}$ is the prototype (mean) of the jth partition Cj, k is the number of partitions (usually predefined). It is known that above equation is equivalent to the trace maximization of the between-cluster scatter S_b

$$\max_{a_{ij}} \operatorname{trace} S_b,$$

Where a_{ij} is the hard cluster assignment

$$a_{ij} \in \{0,1\}, \textstyle\sum_{j=1}^{k} a_{ij} = 1$$

And

$$S_b = \sum_{j=1}^{k} n_j(\vec{\mu}_j - \vec{\mu}_0)(\vec{\mu}_j - \vec{\mu}_0)^T,$$

Where $\vec{\mu_o}$ is the global mean, $n_j = \sum_1^N a_{ij}$ is the number of samples in C_j. Without loss of generality, we assume that the data X □ $IR^{M\times N}$ has been centered such that the global mean is $\vec{\mu_o} = 0$. To express $\vec{\mu_j}$ in terms of X, we denote a discrete cluster membership matrix A □ $IR^{N\times K}$ as

$$A_{ij} = \begin{cases} \frac{1}{\sqrt{n_j}}, & \text{if } \vec{x}_i \in C_j, \\ 0, & \text{if } \vec{x}_i \notin C_j, \end{cases}$$

Then $A^TA = I_k$ and the objective of k-means can be equivalently written as

$$\begin{aligned} &\max_{A} \quad \operatorname{trace}(A^T X^T X A), \\ &\text{s.t.} \quad A^T A = I_k, \quad A_{ij} \in \left\{0, \frac{1}{\sqrt{n_j}}\right\}. \end{aligned}$$

The discrete constraint in above equation makes the problem NP-hard to solve.

The k-means objective extended to Hilbert space F and multiple data sets are incorporated, given by

$$\begin{aligned} Q1: \quad &\max_{A,\vec{\theta}} \mathcal{J}_{Q1} = \operatorname{trace}(A^T \mathbf{\Omega} A), \\ &\text{s.t.} \quad A^T A = I_k, \quad A_{ij} \in \left\{0, \frac{1}{\sqrt{n_j}}\right\}, \\ &\mathbf{\Omega} = \sum_{r=1}^{p} \theta_r G_r, \\ &\theta_r \geq 0, \ r = 1, \ldots, p \\ &\sum_{r=1}^{p} \theta_r^{\delta} = 1. \end{aligned}$$

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

b. Bi-level optimization of k-means on multiple kernels:

The objective in above equation is difficult to optimize analytically because the data are unlabeled; moreover, the discrete cluster memberships make the problem NP hard. Our strategy is to optimize the two parameters iteratively.

a) ***Optimizing the Kernel Coefficients as Simplified KFD:*** Given a single data set and labels of two classes, to find the linear discriminant in F we need to maximize

$$\max_{\vec{w}} \frac{\vec{w}^T S_b^{\Phi} \vec{w}}{\vec{w}^T \left(S_w^{\Phi} + \rho I\right) \vec{w}},$$

b) ***The Role of Cluster Assignment:*** It is worth clarifying the transformations of cluster assignment in the proposed algorithm.

c) ***Solving the Simplified KFD as LS-SVM Using Multiple Kernels:*** The pseudo code to solve the LS-SVM MKL is presented in Algorithm1:

$$\text{Obtain the initial guess } \vec{\beta}^{(0)} = [\vec{\beta}_1^{(0)}, \ldots, \vec{\beta}_k^{(0)}]$$

$$\tau = 0$$

$$\textbf{while } (\Delta_u > \varepsilon)$$

$$\textbf{do} \begin{cases} step1: \text{Fix } \vec{\beta}, \text{ solve } \vec{\theta}^{(\tau)} \text{ then obtain } u^{(\tau)} \\ step2: \text{Compute the kernel combination } \Omega^{(\tau)} \\ step3: \text{Solve the single LS-SVM for the optimal } \vec{\beta}^{(\tau)} \\ step4: \text{Compute } f_1(\vec{\beta}^{(\tau)}), \ldots, f_{p+1}(\vec{\beta}^{(\tau)}) \\ step5: \Delta = \left| 1 - \frac{\sum_{j=1}^{p+1} \theta_j^{(\tau)} f_j(\vec{\beta}^{(\tau)})}{u^{(\tau)}} \right| \\ step6: \tau := \tau + 1 \end{cases}$$

$$\text{comment: } \tau \text{ is the indicator of the current loop}$$

$$\textbf{return } (\vec{\theta}^{(\tau)}, \vec{\beta}^{(\tau)})$$

Algorithm1. SIP-LS-SVM-MKL (G_1…….G_p, F)

c. Optimized data fusion for kernel k-means clustering:

The main characteristic is that the cluster assignments and the coefficients of kernels are optimized iteratively and adaptively until convergence. The coefficients assigned to multiple kernel matrices leverage the effect of different kernels in data integration to optimize the objective of clustering.

The optimized kernel k-means clustering (OKKC) algorithm is presented in Algorithm2:

$$\text{comment: Obtain the } \Omega^{(0)} \text{ by the initial guess of } \theta_1^{(0)}, \ldots, \theta_p^{(0)}$$

$$\tilde{A}^{(0)} \leftarrow \text{PCA}(\Omega^{(0)}\Omega^{(0)}, k)$$

$$A(0) \leftarrow \text{K-MEANS } (\tilde{A})$$

$$\gamma = 0$$

$$\textbf{while } (\Delta A > \epsilon)$$

$$\textbf{do} \begin{cases} step1: F^{(\gamma)} \leftarrow A^{(\gamma)} \\ step2: \Omega^{(\gamma+1)} \\ \qquad \leftarrow \textbf{SIP-LS-SVM-MKL}(G_1, G_2, \ldots, G_p, F^{(\gamma)}) \\ step3: \tilde{A}^{(+1)} \leftarrow \textbf{PCA}(\Omega^{(\gamma+1)}\Omega^{(\gamma+1)}, k) \\ step4: A^{(\gamma+1)} \leftarrow \text{K-MEANS}(\tilde{A}^{(\gamma+1)}) \\ \qquad \textbf{OR} \\ \qquad A^{(\gamma+1)} \leftarrow \textbf{QR}(\tilde{A}^{(\gamma+1)}) \\ step5: \Delta A = \|A^{(\gamma+1)} - A^{(\gamma)}\|^2 / \|A^{(\gamma+1)}\|^2 \\ step6: \gamma := \gamma + 1 \end{cases}$$

$$\textbf{return } (A^{(\gamma)}, \theta_1^{(\gamma)}, \ldots, \theta_p^{(\gamma)})$$

Algorithm2. OKKC (G_1, G_2… G_p, k)

B. Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm:

The fuzzy relational clustering algorithm uses the PageRank score of an object within a cluster as a measure of its centrality to that cluster. These PageRank values are then treated as likelihoods. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients.

a. ***Initialization:*** Assume here that cluster membership values are initialized randomly, and normalized such that cluster membership for an object sums to unity over all clusters. Mixing coefficients are initialized such that priors for all clusters are equal.

b. ***Expectation step:*** The E-step calculates the PageRank value for each object in each cluster. PageRank values for each cluster are calculated with the affinity matrix weights W_{ij} obtained by scaling the similarities by their cluster membership values;

$$w_{ij}^m = s_{ij} \times p_i^m \times p_j^m,$$

c. ***Maximization step:*** Since there is no parameterized likelihood function, the maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step.

C. Multiple Kernels Fuzzy Clustering Algorithm:

A brief comparison of optimized kernel k-means clustering (OKKC), novel fuzzy relational clustering (FRECCA) and multiple kernel k-means clustering (MKFC) algorithms are as shown in Table 1.

Table 1: Comparison between OKKC, FRECCA and MKFC Algorithm

Clustering methods	Advantages	Disadvantages
optimized kernel k-means clustering (OKKC)	1.Its simple optimization procedure and low computational complexity. 2.OKKC is comparable to the best candidates in comparison, which improves the performance. 3.The optimization procedure of OKKC is bi-level, which is simpler than the tri-level architecture of the NAML algorithm.	1.For clustering of pen digit and journal data, large amount of memory is required. 2.Performance of OKKC depends on nature of data, performance on disease data degrades significantly.
fuzzy relational clustering (FRECCA)	1.The FRECCA algorithm is able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode on a challenging data set. 2.The FRECCA algorithm is capable of identifying overlapping clusters of semantically related sentences.	1.The major disadvantage of the FRECCA algorithm is its time complexity. Since Page Rank must be applied to each cluster in each EM cycle and this can lead to long convergence times if the problem involves a large number of objects and/or clusters.
multiple kernel k-means clustering (MKFC)	1.MKFC performs feature selection automatically and provides better clustering results. 2.MKFC requires less iteration than the average of the KFCs. This indicates that MKFC converges more quickly. 5.MKFC is easy to implement.	1.MKFC was slightly slower than MKKM. 2.MKFC is not ranked 1 for each individual dataset; hence kernel combination does not yield the best performance in every single case.

V. PROPOSED METHODOLOGY

While fuzzy c-means is a popular soft-clustering method, its effectiveness is largely limited to spherical clusters.

By applying kernel tricks, the kernel fuzzy c-means algorithm attempts to address this problem by mapping data with nonlinear relationships to appropriate feature spaces. Kernel combination, or selection, is crucial for effective kernel clustering. Unfortunately, for most applications, it is uneasy to find the right combination.

The proposed method based on the three stages as objective function, optimizing membership and optimizing weight.

A. Objective function:

To discover nonlinear relationships among data, kernel methods use embedding mappings that map features of the data to new feature spaces.

B. Optimizing membership:

The goal of proposed algorithm is to simultaneously find combination weights w, memberships U, and cluster centers V, which minimize the objective function.

C. Optimizing weight:

It can be seen that when the weights w and cluster centers V are fixed, the optimal memberships U can be obtained. Now, let us assume that the memberships are fixed. We seek to derive the optimal centers and weights to combine the kernels.

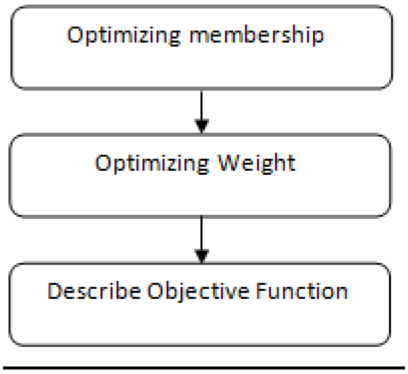

Figure 1. Framework of membership optimization for clustering

VI. POSSIBLE OUTCOMES AND RESULT

A. Performance Measures:

Membership degrees make it possible for us to measure the performance of these algorithms using either hard-clustering measures or soft-clustering measures.

a. ***Hard-Clustering Measures****:* Most clustering measures are designed for the evaluation of the results of hard clustering, in which each data item is assigned to a single class.To use this kind of measure for soft clustering, one must convert the membership degrees to hard assignments. We take the conventional approach for such assignments, i.e., we assign each data item to the cluster with the highest membership degree. Hard-clusterings measures can be roughly categorized into pair-counting-based measures (e.g., Rand index (RI) and adjusted Rand index (ARI), set matching-based measures (e.g., *H* criterion), and information theoretic- based measures (e.g., mutual information and normalized mutual information (NMI).
b. ***Soft-Clustering Measures****:* The casting of soft clustering to hard clustering often fails to faithfully reflect the performance of soft-clustering algorithms. For example, different fuzzy partitions (with potentially widely divergent spatial distributions) may result in the same crisp partition; accordingly, both will have the same hard-clustering measure. This loss of information, caused by the disposal of the fuzzy membership values, makes the hard clustering measures unable to discriminate between overlapped and nonoverlapped clusters. As such, these hard-clustering measures might not be appropriate for the assessment of fuzzy clustering algorithms. To get around these drawbacks, Campello proposed a fuzzy extension of the RI and other related indexes. The extended index is obtained by first rewriting the

formulation of the RI in a fully equivalent form using basic concepts from set theory.

VII. CONCLUSION

This paper exhausted different clustering algorithm for optimizing procedure and provide low computational complexity. The results of these clustering algorithm show that the algorithm is able to achieve superior performance. Most of the approaches perform hard clustering, i.e., they assign each item to a single cluster. This works well when clustering compact and well-separated groups of data, but in many real-world situations, clusters overlap. In FRECCA algorithm is capable of identifying overlapping clusters of semantically related sentences. It is also mentioned that feature selection automatically and provides better clustering results.

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 generic fuzzy clustering algorithm that can be applied to any relational clustering problem and application to several non sentence data sets has shown its performance to be comparable to Spectral Clustering and k-Medoid benchmarks. It can also be used within more general text mining settings such as query-directed text mining.

VIII. FUTURE SCOPE

Graph-based methods are an exciting area of research within the pattern recognition community. All These clustering methods can be extending to perform hierarchical clustering. The concept in natural language documents usually display some type of hierarchical structure, where as these proposed algorithm identifies only flat clusters. The future objective is to extend these ideas to the development of a hierarchical fuzzy relational clustering algorithm. In the future, it is useful in open topics, such as strategies for setting the fuzzification degree or choosing the basis kernels.

IX. REFERENCES

[1]. Shi Yu, Le´on-Charles Tranchevent, Xinhai Liu, Wolfgang Glanzel, Johan A.K. Suykens, Bart De Moor and Yves Moreau, “Optimized Data Fusion for Kernel k-Means Clustering”, IEEE Transaction On Pattern Analysis And Machine Intelligence, Vol. 34, No. 5, PP. 1031-1039, May 2012.

[2]. Andrew Skabar and Khaled Abdalgader, “Clustering Sentence-Level Text Using A Novel Fuzzy Relational Clustering Algorithm,” IEEE Transactions On Knowledge and Data Engineering, Vol. 25, No. 1, PP. 62-75, January 2013.

[3]. Hsin-Chien Huang, Yung-Yu Chuang, and Chu-Song Chen, “Multiple Kernel Fuzzy Clustering”, IEEE Transactions on Fuzzy Systems, Vol. 20, No. 1, PP. 596-604, February 2012.