International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 4 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195410

A Review article on Semi- Supervised Clustering Framework for High Dimensional Data

M. Pavithra[1], Dr. R. M. S. Parvathi[2]
[1]Assistant Professor, Department of C.S.E, Jansons Institute of Technology, Coimbatore, India
[2]Professor & Dean- PG Studies, Sri Ramakrishna Institute of Technology, Coimbatore, India

ABSTRACT

Cluster analysis methods seek to partition a data set into homogeneous subgroups. It is useful in a wide variety of applications, including document processing and modern genetics. Conventional clustering methods are unsupervised, meaning that there is no outcome variable nor is anything known about the relationship between the observations in the data set. In many situations, however, information about the clusters is available in addition to the values of the features [2]. For example, the cluster labels of some observations may be known, or certain observations may be known to belong to the same cluster. In other cases, one may wish to identify clusters that are associated with a particular outcome variable. This review describes several clustering algorithms (known as "semi-supervised clustering" methods) that can be applied in these situations [3]. The majority of these methods are modifications of the popular k-means clustering method, and several of them will be described in detail. A brief description of some other semi-supervised clustering algorithms is also provided. Cluster formation has three types as supervised clustering, unsupervised clustering and semi supervised. This paper reviews traditional and state-of-the-art methods of clustering [1]. Clustering algorithms are based on active learning, with ensemble clustering-means algorithm, data streams with flock, fuzzy clustering for shape annotations, Incremental semi supervised clustering, Weakly supervised clustering, with minimum labeled data, self-organizing based on neural networks. Incremental semi-supervised clustering ensemble framework (ISSCE) which makes utilization of the advantage of the arbitrary subspace method, the limitation spread approach, the proposed incremental ensemble member choice process, and the normalized cut algorithm to perform high dimensional information clustering [4]. Semi-supervised clustering employs limited supervision in the form of labeled instances or pairwise instance constraints to aid unsupervised clustering and often significantly improves the clustering performance. Despite the vast amount of expert knowledge spent on this problem, most existing work is not designed for handling high-dimensional sparse data.

Keywords : Cluster Ensemble, Semi-Supervised Clustering, Clustering Analysis, High-Dimensional Data.

I. INTRODUCTION

The bunch troupe methodologies are more points of interest and more consideration because of its valuable applications in the regions of example acknowledgment, data mining, bioinformatics, and more one [1]. At the point when contrasted and customary single grouping calculations, bunch gathering methodologies can coordinate various grouping arrangements got from various information sources into a bound together arrangement, and give a more hearty, steady and precise last result. In any case, conventional cluster ensemble approaches have a few statutes of impediments: First they don't

consider how to make utilization of earlier information given by specialists, which are spoken to by Pair savvy limitations. Match shrewd requirements are regularly characterized as the must-connect limitations and they can't interface imperatives. The must-interface limitation implies that two component vectors ought to be doled out to a similar group, while they can't connect requirements implies that two element vectors can't be appointed to a similar cluster. First most of the cluster ensemble methods cannot procure acceptable results on high dimensional datasets. Third not all the ensemble members add to the last result. So as to address the 1 and 2 restrictions, we first propose the random subspace based semi-supervised clustering ensemble framework (RSSCE), joins the irregular subspace method, the imperative proliferation approach [2], and the normalized cut algorithm [3] into the cluster ensemble framework to perform high dimensional information grouping. At that point, the incremental semi-supervised clustering ensemble framework (ISSCE) is intended to expel the copy ensemble members. At the point when contrasted and customary with traditional semi-supervised clustering algorithm, ISSCE is elements by the incremental ensemble member selection (IEMS) handle in view of an as of late proposed worldwide target work and a nearby target work, which decision ensemble individuals dynamically. The nearby target capacity is ascertained in view of an as of late planned closeness work which chooses how comparative two arrangements of properties are in the subspaces. Besides, the computational cost and the space utilization of ISSCE are dissected hypothetically. Labeled data can classify easily, but unlabeled data classification is very challenging task [4].

Clustering methods that can be applied to partially labeled data or data with other types of outcome measures are known as semi-supervised clustering methods (or sometimes as supervised clustering methods). They are examples of semi-supervised learning methods, which are methods that use both labeled and unlabeled data. This review will briefly describe several semi-supervised clustering methods that can be applied to different types of partially labeled data sets. The review will focus primarily on variations of k-means clustering, since most existing semi-supervised clustering methods are modified versions of k-means clustering. However, a brief description of some semi-supervised hierarchical clustering methods will also be provided [1].

Existing methods for semi-supervised clustering can be generally grouped into three categories. First, the constraint based methods aim to guide the clustering process with pairwise instance constraints [5] or initialize cluster centroids by labeled instances [4]. Second, the distance-based methods employ metric learning techniques to get an adaptive distance measure used in the clustering process based on the given pairwise instance constraints [2]. Finally, the hybrid method proposed. [6] unifies the first two methods under a general probabilistic framework. However, most existing semi-supervised methods are not designed for handling high-dimensional data. It is well known that the traditional Euclidean notion of density is not meaningful in high-dimensional data sets [7]. Since most semi-supervised clustering techniques are based on proximity or density, they often have difficulties in dealing with high-dimensional data. Therefore, it is necessary to integrate feature reduction into the process of semi-supervised clustering. The key challenge is how we can incorporate supervision into dimensionality reduction such that the reduced data can still capture the available class information.

II. LITERATURE SURVEY

In this section, we provide a review of related works on using user provided information to improve data clustering. We first discuss some algorithms in which

prior knowledge is in the form of labeled data. Next, we describe other algorithms for which pair wise constraints are required to be known a priori. SS-constrained-Kmeans and SS-seeded-Kmeans [3] are the two well-known algorithms in semi-supervised clustering with labels. The SS-constrained-Kmeans seeds the k-means algorithm with the given labeled data and keeps that labeling unchanged through-out the algorithm. Moreover, it is appropriate when the initial seed labeling is noise-free, or if the user does not want the labels of the seed data to change.

Kmeans algorithm changes the given labeling of the seed data during the course of the algorithm. Also, it is applicable in the presence of noisy seeds, since it does not enforce the seed labels to remain unchanged during the clustering iterations and can therefore abandon noisy seed labels after the initialization step. Semi-supervised clustering with labels has been successfully applied to the problem of document clustering. It [5] proposed incorporating background knowledge into document clustering by enriching the text features using WordNet.1 In Jones et al. [4], some words per class and a class hierarchy were sought from the user in order to generate labels and build an initial text classifier for the class. A similar technique was the user is made to select interesting words from automatically selected representative words for each class of documents. These user identified words were then used to re-train the text classifier. Active learning approaches have also found applications in semi-supervised clustering. [1] Has proposed to convert a user recommended feature into a mini-document which is then used to train an SVM classifier. This approach has been extended. Which adjusts SVM weights of the key features to a predefined value in binary classification tasks? Recently, it [6] presented a probabilistic generative model to incorporate A number of previous works adopt feature selection approaches to choose an optimal gene subset in the task of cancer classification. For example, Mundra and Rajapakse [7] integrated the minimum-redundancy maximum relevancy filter into the support vector machine recursive feature elimination approach to select an optimal gene subset and improve the accuracy of cancer classification.

It [3] proposed a top-r feature selection approach to perform gene selection with respect to classification accuracy from microarray data. It[1] proposed a feature selection approach in combination with a radial basis function based neural network to perform gene selection and improve the performance of cancer classification. The multi-criterion fusion based recursive feature elimination (MCF-RFE) algorithm to select an optimal gene subset from gene expression data sets. A feature selection approach based on an efficient margin based sample weighting algorithm to improve the performance of gene selection investigated how to use model-based entropy to perform feature selection from gene expression data. Mao and Tang [5] adopted a recursive measure to find an optimal gene subset.

III. RELATED WORK

The related literature on semi-supervised clustering can be grouped into three categories: constraint-based methods, distance-based methods, and a combination of constraint based and distance-based methods. For constraint-based methods, the cop-kmeans algorithm [2] guides the cluster allocation process by a constraint motivated heuristic objective function. However, this algorithm strictly enforces the clustering process such that any violation of the given pairwise constraints is forbidden, which limits its use, especially in a noisy environment. In contrast, our version of semi-supervised clustering algorithm allows some relaxation of the pairwise constraints. Also, It [4] proposed a seeded K-means which tries to get better initial cluster centroids from the labeled instances in addition to constraining the clustering process, while their supervised cluster initialization is

based on the labeled instances instead of pairwise constraints. For distance-based methods, Cohn et al. [1] used gradient descent for weighted Jensen-Shannon divergence in the context of EM clustering.

It [6] combined the Newton Raphson method and iterative projection together to learn a Mahalanobis distance for K-means clustering. It [4] proposed a more efficient algorithm for learning the distance metric with side information, which utilized Canonical Correlation Analysis (CCA) to approximate LDA. In general, the metric learning used in the distance based method, which is equivalent to learning an adaptive weight for each dimension, is either based on iterative algorithms, such as gradient descent and Newton's method, or involves some matrix operations. However, the distance based method has high computational cost when applied to the high-dimensional data. Indeed, data represented in matrix is often singular when the sparsity of the data is high. This makes some matrix operations, such as inversion, computationally intractable. For hybrid methods, [5] introduced a general probabilistic framework which unifies the constraint-based and distance-based method into the Hidden Markov Random Field (HMRF).

The proposed HMRF-EM algorithm can interweave the constrained clustering and distance learning interactively in the process of semi-supervised clustering. Also, the related literature on feature reduction includes Principle Component Analysis (PCA) [2] which tries to find a low rank approximation to represent the high-dimensional data, and Fisher's Linear Discriminant Analysis (LDA) [1] which tries to find one or more directions along which different classes can be best separated while the variance of each class is minimized given the label for each instance. The PCA method works in an unsupervised manner where the class information is not available, which makes the reduced dataset incapable of capturing the original class information.

IV. SEMI-SUPERVISED CLUSTERING METHODS

We will now briefly outline several semi-supervised clustering methods. These methods will be organized according to the nature of the known outcome data [1]. First, we will consider the simplest case, namely the case where the data is partially labeled. In other words, the cluster assignments are known for some subset of the observations. We will then consider the case where some sort of relationship between the features is known, and finally the case where one seeks to identify clusters associated with a particular outcome variable [2].

A. PARTIALLY LABELED DATA

In some situations, the cluster assignments may be known for some subset of the data. The objective is to classify the unlabeled observations in the data to the appropriate clusters using the known cluster assignments for this subset of the data [3].

In a certain sense, this problem is equivalent to a supervised classification problem, where the objective is to develop a model to assign observations in a data set to one of a finite set of classes based on a training set where the true class labels are known [4]. However, traditional supervised classification methods may be inefficient when only a small subset of the data is labeled. For example, if one wishes to classify web pages into a discrete number of groups, one can easily collect millions of unlabeled observations, but classifying any given observation requires human intervention (and hence is likely to be slow) [5].

Similarly, if one wishes to develop a method to classify e-mails as "spam" or "not spam," then one can easily collect numerous unlabeled observations, but the proportion of labeled observations will be much smaller. For these types of problems, conventional supervised classification methods may be inefficient since they typically do not use unlabeled data to

build the classification algorithm. Thus, the vast majority of the available data will not be used [6].

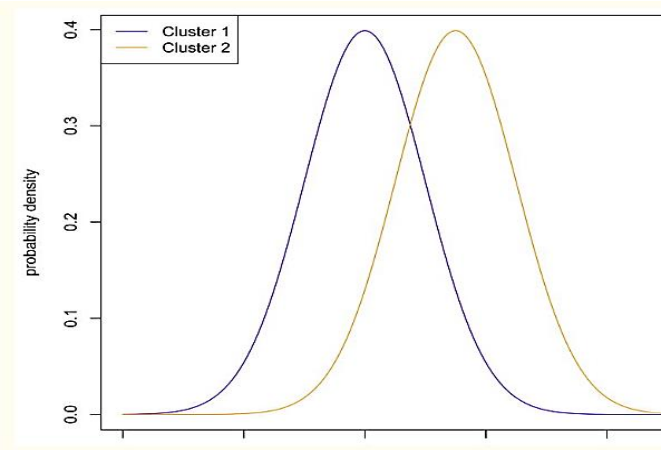
$$C_i = \arg\min_k \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

$$\bar{x}_{kj} = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_{ij}$$

B. SEMI-SUPERVISED HIERARCHICAL CLUSTERING

The majority of existing semi-supervised clustering methods are based on k-means clustering or other forms of partitional clustering. Comparatively few semi-supervised hierarchical clustering methods have been proposed [2]. This is partly due to the fact that the problem must be formulated differently for hierarchical clustering. As noted earlier, most semi-supervised partitional clustering methods utilize either partially labeled data or known constraints (e.g. "must-link" or "cannot-link" constraints) on the observations. It is more difficult to define such constraints for hierarchical clustering, since hierarchical clustering links all observations in a data set at some level of the clustering hierarchy [3]. Thus, a "must-link" constraint will always be satisfied at some level of the hierarchy and likewise a "cannot-link" constraint will always be violated.

Hence, semi-supervised hierarchical clustering methods have considered different types of constraints. For example, Miyamoto and Terami require observations linked by a "must-link" constraint to be clustered together at the lowest possible level of the hierarchy [1]. They further require that observations separated by a "cannot-link" constraint must not be part of the same clustering hierarchy. Thus, rather than identifying a single clustering hierarchy, the method of Miyamoto and Terami returns several clustering hierarchies [5]. A separate hierarchy is produced for each observation that is part of a "cannot-link" constraint. Several related methods have been proposed to perform hierarchical clustering subject to such constraints [4].



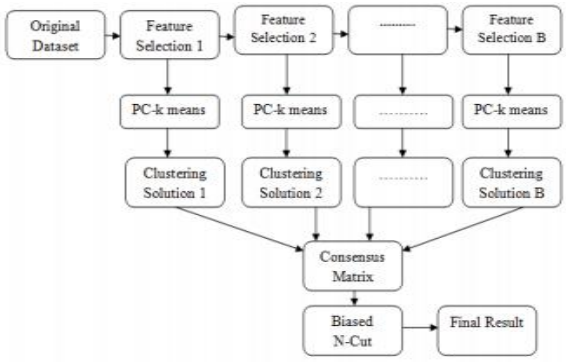
C. ENSEMBLE SEMI-SUPERVISED CLUSTERING

In our work so far, we have assumed constraints to be noise-free. We have also assumed the weights on the constraints to be uniform (PCKMeans) or changed the weights based on the "difficulty of satisfying the constraints" (unified model). An interesting problem in the PCC model would be the choice of the constraint weights in the general case of noisy constraints [2]. Given a set of noisy constraints, we can create an ensemble of semi-supervised clusters, each of which put different weights on the constraints and possibly get different clustering's [3]. We propose a scheme for creating an ensemble of PCC clusters and combining their results using boosting. Each PCC clustered can be considered as a weak learner taking pairwise data points as input, and giving an binary output decision of "same-cluster" or "different-cluster" [1]. The must-link and cannot-link constraints can be considered as the training data for each weak learner.

Given a set of input constraints, the PCC clustered initially sets all constraints to have uniform weight and performs clustering. After clustering is

completed, the clustered categorizes each pair of points as "same-cluster" or "different-cluster", based on whether the pair ended up in the same cluster. In the first stage, to generate a set of new data sets and to remove noisy genes FS-SSCE adopted for feature selection [5]. It is known that, the feature selection approaches are divided into two most important types: supervised feature selection approaches and unsupervised feature selection approaches [4].

To implement the pair wise constrained clustering framework known as PC-K means to estimate the labels Y of the cancer samples. we propose to view clustering solutions as new attributes of the original data set, and adopt feature selection approaches, such as feature selection based on mutual information maximization, mutual information feature selection, max-relevance min redundancy, joint mutual information, double input symmetrical relevance, conditional infomax feature extraction, interaction capping and conditional redundancy, to perform clustering solution selection [6].



D. APPROXIMATION ALGORITHMS FOR SEMI-SUPERVISED CLUSTERING

Another interesting research direction is considering how semi-supervision affects approximation algorithms for some clustering methods, e.g., KMedian. The KMedian problem, which was explained in briefly, is similar to the facility location problem [2]. In the facility location problem, we are given a set of demand points and a set of candidate facility sites with costs of building facilities at each of them. Each demand point is then assigned to its closest facility, incurring a service cost equal to the distance to its assigned facility [4]. The goal is to select a subset of sites where facilities should be built, so that the sum of facility costs and the service costs for the demand points is minimized [3].

The KMedian problem is similar to facility location, but with a few differences — in KMedian there are no facility costs and there is a bound on the number of facilities that can be opened [5]. The KMedian objective is to select a set of facilities so as to minimize the sum of the service costs for the demand points. We propose a semi-supervised extension to KMedian to handle constraints on the demand points. The constrained KMedian problem would be additionally given an input set of must-link and cannot-link constraints on the demand points (i.e., two demand points should be or should not be assigned to the same facility), and the goal would be to minimize an objective function that is the sum of the service costs for the demand points and the cost of violating the constraints [6].

V. CONCLUSION & FUTURE WORK

From above these contents we can conclude that there are various methods we can use to form cluster in semi supervised clustering. Each method has its own some benefits and limitations. For constant dataset all methods are ok ,but for updated data incremental semi supervised clustering would be more useful, because in this the data is continuously entered in system, continuously update data, and form new clusters as per their contents, and sometimes changes clusters as per user demands [2]. This data is labeled or unlabeled or in shape so incremental can work on all these type of data than other methods. So incremental semi supervised clustering is can be used method of clustering approach. Which create correct cluster on given

mixed type datasets? The proposed algorithm is based on semi-supervised hierarchical clustering frame in which the clusters are formed gradually from a small amount of labeled examples as seeds by assigning unlabeled examples to the existed clusters according to their distances [3]. In the hierarchical clustering procedure, dimensionality reduction is incorporated, and the number of dimensions is reduced gradually as the final clusters are formed [1]. The criterion of dimensionality reduction is dependent on both the labeled data in the current clusters and the unlabeled data that have not been assigned to the current clusters [5]. Through the iterative clustering – dimensionality reduction – clustering procedure, the harmony between clustering and dimensionality reduction is reached, and these two tasks are integrated into a harmonious system.

The experimental results also demonstrate the effectiveness of our method. However, how to automatically determine suitable values for the parameters in our methods, and how to improve the computational effectiveness for large scale data sets, are need to be further studied in the future [4]. In this paper, we want to study other aspects of semi-supervised clustering, like: (1) the effect of noisy, probabilistic or incomplete supervision in clustering; (2) model selection techniques for automatic selection of number of clusters in semi-supervised clustering; (3) ensemble semi-supervised clustering. In future, we want to study the effect of semi-supervision on other clustering algorithms, especially in the discriminative clustering and online clustering framework. We also want to study the effectiveness of our semi-supervised clustering algorithms on other domains, e.g., web search engines (clustering of search results), astronomy (clustering of Mars spectral images) and bioinformatics (clustering of gene microarray data) [6].

VI. REFERENCES

[1]. S. Shalini, R.Raja, "An Improved Semi-Supervised Clustering Algorithm Based on Active Learning ", International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1, March 2014.

[2]. Ashraf Mohammed Iqba, Abidal rahman Moh'd, and Zahoor Ali Khan,"Semi-supervised Clustering Ensemble by Voting", University, Halifax, Canada, 2010.

[3]. Aloysius George, " Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm", The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013.

[4]. Handl J, Knowles J, "On semi-supervised clustering via multi objective optimization", Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation (GECCO 2006); pp. 1465–1472, 2016.

[5]. S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in Proc. SIAM Int. Conf. Data Mining, pp. 1–8, 2014.

[6]. Tang W, Xiong H, Zhong S, Wu J, " Enhancing semi-supervised clustering: a feature projection perspective", Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 707–716, 2015.

[7]. Miyamoto S, Terami, "A. Semi-supervised agglomerative hierarchical clustering algorithms with pairwise constraints", Proceedings of the 2010 IEEE International Conference on Fuzzy Systems (FUZZ 2010), pp. 1–6, 2010.

Cite this article as : M. Pavithra, Dr. R. M. S. Parvathi, "A Review article on Semi- Supervised Clustering Framework for High Dimensional Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 102-108, July-August 2019. Available at doi : https://doi.org/10.32628/CSEIT195410 Journal URL : http://ijsrcseit.com/CSEIT195410