A technical evidence on Spectral Ensemble Clustering via Weighted K-means

DALLI SAI SURESH 1, BODDU JOY STEVENSON 2, SIVAH AKASH 3
1Student,Dept of CSE, GITAM University, Visakhapatnam,Andhra Pradesh,India.
2Student,Dept of CSE, GITAM University, Visakhapatnam,Andhra Pradesh,India.
3Student,Dept of CIVIL, GITAM University, Visakhapatnam,Andhra Pradesh,India.

Abstract— As a promising path for heterogeneous information investigation, agreement bunching has pulled in expanding consideration in late decades. Among different brilliant arrangements, the co-affiliation grid based strategies shape a milestone, which reclassifies accord grouping as a chart parcel issue. By the by, the moderately high time and space complexities block it from wide genuine applications. We hence propose Spectral Ensemble Clustering (SEC) to use the upsides of co-affiliation grid in data coordination however run all the more productively. We reveal the hypothetical proportionality among SEC and weighted K-implies bunching, which significantly diminishes the algorithmic unpredictability. We additionally determine the inert accord capacity of SEC, which to our best learning is the first to connect co-affiliation grid based strategies to the techniques with unequivocal worldwide target capacities. Further, we demonstrate in principle that SEC holds the vigor, generalizability and union properties. We at last stretch out SEC to address the difficulty emerging from deficient essential parcels, in view of which a column division plot for huge information bunching is proposed. Trials on different genuine informational indexes in both gathering and multi-see grouping situations show the prevalence of SEC over some cutting edge techniques. Specifically, SEC is by all accounts a promising possibility for enormous information grouping.

Index Terms—Consensus Clustering, Ensemble Clustering, Spectral Clustering, Co-association Matrix, Weighted K-means

1.Introduction

Ensemble clustering, also known as consensus clustering, is emerging as a promising solution for multi-source and/or heterogeneous data clustering and is attracting increasing academic attention. It aims to find a single partition that most agrees with multiple existing basic partitions [22]. Consensus clustering is of recognized benefit in generating robust partitions, finding bizarre clusters, handling noise and outliers, and integrating solutions from multiple sources [21]. Consensus clustering can be roughly divided into two categories: those with implicit or explicit objectives. Methods that utilize implicit objectives do not set objective functions, but instead directly adopt some heuristics to find approximate solutions. Representative methods include coassociation matrix-based methods [27, 26, 10, 14], graphbased algorithms [22, 8], relabeling and voting methods [1, 9, 19], locally adaptive cluster-based methods [7], and genetic algorithm-based methods [32]. Methods with explicit objectives employ explicit objective functions for consensus clustering. For instance, [23] used K-means to find the solution based on quadratic entropy, which was then generalized in [28] as the paradigm of K-means-based consensus clustering. Other solutions for different objective functions include non-negative matrix factorization [12], EM algorithm [24], simulated annealing [15], combination regularization [30], and hill-climbing method [5]. Many other algorithms for consensus clustering can be found in the survey [13, 25, 20]. Of these consensus clustering methods, the co-association matrix-based method is a landmark. First, the information represented by basic partitions is summarized into a co-association matrix, which measures how many times a pair of instances appears in the same cluster; then a graph partition method can be used to obtain the final consensus clustering. The main contribution of the co-association method is the redefinition of the consensus clustering problem as a classical graph partition problem, so that agglomerative hierarchical clustering, spectral clustering, or other algorithms can directly run on the co-association matrix without much modification. However, the co-association matrixbased method also suffers from some limitations. For instance, the high time and space complexity prevents handling large-scale data clustering and no explicit objective function is used to supervise clustering. In light of this, we propose Spectral Ensemble Clustering (SEC), which employs spectral clustering on the coassociation matrix to find the consensus partition. SEC equivalently results in weighted K-means clustering, which decreases the time complexity from O(n 3) to roughly O(n) and decreases the space complexity from O(n 2) to roughly O(n) as well. We then derive the intrinsic consensus ob- jective of SEC and provide a robustness and generalization analysis. Further we extend SEC to handle incomplete basic partitions. Experimental results on various real-world data sets demonstrate that SEC delivers efficient and high quality clustering compared to some state-of-the-art consensus clustering methods. SEC is also highly robust to incomplete basic partitions with many missing values. Finally, SEC is used to explore big data clustering of Weibo data.

2.EXISTING SYSTEM

As a promising way for heterogeneous data analytics, consensus clustering has attracted increasing attention in recent decades. Among various excellent solutions, the co-association matrix based methods form a landmark, which redefines consensus clustering as a graph partition problem. Nevertheless, the relatively high time and space complexities preclude it from wide real-life applications.

DRAWBACKS

1. This suffers from some non-ignorable drawbacks, particularly when facing data sets of different characteristics.
2. Some basic partitions are unable to do

3.PROPOSED SYSTEM

We projected the Spectral Ensemble Clustering (SEC) algorithm. By identifying the equivalent relationship between SEC and weighted K-means, we decreased the time and space complexities of SEC dramatically. The intrinsic consensus objective function of SEC was also revealed, which bridges the co-association matrix based methods with the methods with explicit global objective functions. We then investigated the robustness, generalizability and convergence properties of SEC to showcase its superiority in theory, and extended it to handle incomplete basic partitions.

ADVANTAGES

1. We finally extend SEC to meet the challenge arising from incomplete basic partitions, based on which a row-segmentation scheme for big data clustering is proposed.

2. This solves some non-ignorable drawbacks, particularly when facing data sets of different characteristics.

4. LITERATURE SURVEY

4.1 SPECTRAL CLUSTERING APPLICATIONS

From the past decade, Spectral Clustering algorithm is evolved as more powerful algorithm in the field of data mining. In the recent research, the Spectral Clustering is extensively used in text mining, information retrieval and image segmentation domains. The obtained results of spectral clustering are very notable.

(i) Image segmentation

Authors in the paper [10] proposed a spectral clustering method applicable for massive images using a mixure of block wise transform and stochastic ensemble consent. In digital image refinement, distribution is important for image description and classification. According to their mechanism the clusters formed for images established on the attributes called intensity of the pixel, color, texture, location, and mix of these. “Major functionality of the spectral clustering is the eigen decomposition of pair wise affinity matrix, which is very complex for high dimensional datasets. The basic idea of this mechanism used by the author is to execute an over-partition of the image with respect to the pixel level using spectral clustering, and hence combine the partitions by using a mix of stochastic ensemble consensus and an iterative approach of spectral clustering at individual segment level. To determine the pixel classifications they used stochastic ensemble in cooperation with both global and local image quality. The current step also removes block wise processing artifacts. Tung et al.[86] also presented the empirical outcomes on a collection of uniform scene images of the normalized cut, the self-tuning spectral clustering. They conclude that "the illustrated mechanism produce partition outcomes that are identical to or superior to the rest of the two techniques”.

(ii) Educational Data Mining

Authors in [11] proposed a methodology to describe Educational Data Mining (EDM) tasks as well, such as making an in-tutor prediction on the KDD Cup 2010 dataset. The immensely inter-disciplinary terrain of Educational Data Mining (EDM) has emerged from a fusion of numerous different areas, some of which covers Machine Learning, Cognitive Science, and Psychology. The major task in EDM is to build computational models and tools to mine data that explored in an educational setting. With rapidly growing data archive from various educational contexts (paper tests, e-learning, Intelligent Tutoring Systems etc.), best tradition in EDM can potentially answer significant research issues about student learning. The task of EDM is justifying instrumental in combining the knowledge derived from the data to combine with theories from cognitive psychology to formulate the best learning settings and methodologies. The results have been very encouraging. In the same vein, this methodology was also tried on the Performance Factor Analysis (PFA) task, the only difference being that the predictor used to train on each cluster was a logistic regression model. Preliminary work has indicated an improvement in the prediction accuracy. The objective of this work was to introduce to the domain of Educational Data Mining the great utility of using spectral clustering methods.

Authors in [12] used spectral clustering to enhance the performance of a new ensemble method proposed in an earlier work. While the objective was to introduce the use of spectral clustering, a very significant result of the work is proving the efficiency of Dynamic Assessment as compared to static assessment. These results show that an Intelligent can not only save time that is wasted on assessment for instruction, but it can also be a better predictor of their performance in post-tests.

(iii) Entity resolution

Authors in [13] proposed an efficient spectral neighborhood (SPAN) algorithm based on spectral clustering. In numerous telecom and web utilization, the demand of entity resolution is getting bigger and bigger. The main aim of Entity resolution is to check the correspondence between the objects in the related source and the identical entity in the real world. The similar problem rises often in the field of data integration when there lacks a specific attribute across several data authority to serve as a real world entity. Blocking is a necessary technique for developing the computational performance of the algorithms for entity resolution. To solve the entity resolution issue, an efficient spectral neighborhood (SPAN) algorithm constitute on spectral clustering is then proposed. SPAN is an unsupervised and unconstrained algorithm and it is suitable in several applications where the number of blocks is unexplored beforehand. SPAN uses the vector space model in the way of characterize every record by a vector of qgrams.

4.2 RECENT IMPROVEMENTS IN SPECTRAL ENSEMBLE CLUSTERING ALGORITHM

From the past few years, spectral clustering techniques have attained reputation as a mechanism to implement data clustering one of the largest primitive functions of machine learning. All the spectral techniques face few relevant improvements, being the capability to cluster scalar data, and generally give exceptional experimental performance. Furthermore, they are well-studied and backed hypothetically. In our literature survey we identified the major advances in the spectral clustering algorithm. Here we are notifying some of the improvements.

(i) Improvement in time complexity

In paper [14] recommend and review a new spectral clustering algorithm with improvement in time complexity. This algorithm works well with linear time complexity in the order of given input information points. Hence for the massive data sets also this algorithm works accurately in linear time. This fast and improved algorithm is implemented based on the combination of the two important algorithms spectral clustering algorithms and Nystrom methods which are widely used mechanisms in machine learning. Usually these algorithms applied to attain best quality low rank similarities of massive matrices. The suggested algorithm employs the Nystrom similarity on the graph Laplacian to implement clustering. We commenced the hypothesis study of the administration of the algorithm and prove that the oversight limit it attains and we also mention the algorithm performance circumstances corresponding to spectral clustering with the initial graph Laplacian.

(ii) Time and Space Efficient Spectral Clustering via Column Sampling

In the paper [15] authors discussed the performance of the Spectral clustering only certain eigenvectors plays the major role. To investigate those crucial eigenvectors is a typical problem. A simple approach to resolve this issue is with the use of low-rank matrix similarities. One of the simple approach which is most adequate for this problem is the Nystrom technique. In this technique, first it partition m x n columns from the initial n × n matrix, after that regulates a low-rank approximation of the complete matrix by using the correspondence between the sampled columns and the rest of the n - m columns. It is clear that in the performance of the Nystrom technique only some portion of complete matrix is analyzed and stored, hence it automatically reduce the time and space complexities greatly. Fowlkes et al. satisfyingly enforced this to spectral clustering for image partition. Along with the Nystrom technique has also been regularly applied for the issues such as Gaussian processes and manifold training.

(iii) Spectral Clustering on a Budget in paper [16] authors introduced the spectral clustering that performs on the bottom of the budget restraint. They worked under a constraint direction in which they are concentrating only on some specific entries to generate clusters from the affinity matrix even though; the complete matrix is available in our hand. Authors introduced two algorithms for this issue. These two algorithms are explained theoretically as well as practically. The first algorithm is a elementary and it uses randomized procedure to produce the satisfied results. The hypothesis clearly points out that once data is clustered in simple manner then the performance improves automatically. Specifically, for a given simple n × n affinity matrix the clustering is performed within the budget of O˜(n) (i.e., linear time with respect to input data points). The next algorithm is flexible, and has improved experimental performance. On the other side, second algorithm involves higher computational complexity, and the results are not matched with hypothesis.

(iv)Active Spectral Clustering

Authors in the paper [17] introduced a unique training algorithm for Spectral clustering which iterative approach. It measures the affinities in an incremental procedure through intermediate results produced from each and every iteration. For particular applications, measuring the affinity is very complex and ambiguous. They implement this algorithm to preserve execution assessment of the pure affinities and also to measure the accuracy. Based on this information, the algorithm upgrades some specific measures which are approximately unreliable and whose upgradation would useful for removing the uncertainty in clusters. They study these methods on several datasets, including a real world example where affinities are more expensive and ambiguous. From the outcome it is very clear that these methods improve the performance of the clustering relative to other available methods.

(v) Parallel Spectral Clustering proposed in the paper [18] in the case of larger datasets traditional spectral clustering experience several difficulties as both memory usage and time complexity increases corresponding to the increase in the data size. To apply clustering on massive datasets, however, the parallel spectral clustering affected by the scalability issue.

(vi)A Text Image Segmentation Method Based on Spectral Clustering

Images generally contain rich messages from textual information, such as street name, construction identification, public transport stops and a variety of signal boards. The textual information assists the understanding the essential content of the images. If computers can automatically recognize the textual information from an image, it will be highly valuable to improve the existing technology in image and video retrieval from high-level semantics [19]. For instance, road signs and construction identification in a natural environment can be captured into images by cameras and the textual information will be detected, segmented, and recognized automatically by machines. These messages then can be synchronized as human voice to be used as instructions for visually impaired person. In addition to the example, textual information extraction plays a major role in images retrieval based on contents, cars auto-drive, vehicle plate recognition and automatics. In general, automatic textual extraction consists of text detection, localization, binarization and recognition etc. In a natural scene texts could have different backgrounds and characters in the text message can also have variety of forms. And, existing OCR (Optical Character Recognition) engine can only deal with printed characters against clean backgrounds and cannot handle characters embedded in shaded, textured or complex backgrounds. So that characters are separated from the text in the detected region accurately is very necessary.

4.3 MAJOR CHALLENGES IN CLUSTERING

The process of clustering has many common and specific issues, which are not represented in the literatures, are summarized below. This summarization helps us to select optimal clustering methods.

(i) Data representation: clustering algorithm performance is considered as one of the important factor of data representation, the clusters are compact and simple if the representation is good as for K-means finds them. But unfortunately there is no such good representation. Domain knowledge is used to guide the choice of representation. **(ii) Number of Clusters:** In data clustering the most difficult problems is the finding of clusters automatically. on running

a clustering algorithm for variable values of K the best value of K is traced out by criterion function. MML, minimum message length criteria is used b authors in conjunction with Gaussian mixture model in order to estimate the values of K. They started the approach with large number of clusters, and slowly merged the clusters which made a decrease in the MML criterion. Another commonly used approach is Gap Statistics. Here the main assumption is that when dividing data into an optimal number of clusters. The obtained partition is more similar to the accidental perturbation. For the number of clusters a non parametric was introduced by Dirichlet Process (DP). It is used in probabilistic models in the derivation of distribution of posterior which are most likely computed clusters. A number of clusters of Bayesian prior of non parametric are introduced with a key idea. It is very difficult to decide the exact value of K for more meaningful clusters.

4.4 Summary:

While new clustering algorithms continue to be developed, some issues still have to be resolved. Some problems and research directions as pointed in the literature have to be addressed:There is a need to achieve tighter integration between clustering algorithms and application requirements. Each application has its own requirements: some of them just need a global partition of the data while others need to have the best partition with great precision. Generally, in mining applications, the goal is not to provide all the clusters of the search results but a summarized list of the different topics of the query. Users can after easily figure out what they are exactly searching for by selecting the target topic. Showing images from the target category in which the user is truly interested is much more effective and efficient than returning all the clusters or all the mixed images. There is a need for clustering algorithms that lead to computationally efficient solutions for large-scale data. Not all clustering algorithms can deal with large scale issues. There is a need for stable and robust clustering algorithms that lead to stable solutions even in the presence of noisy data. There is a need to use any available a priori information concerning the nature of the dataset and the goal/domain of the application in order to decide which data
representation is the most suitable and which clustering method is the most appropriate. There is a need to have generic clustering that can be applied to any type of data. There is a need for benchmark data with available ground truths and diverse data sets from various domains to evaluate any kind of clustering algorithm because current benchmarks are limited to a small dataset that can be applied only for a limited

choice of clustering methods. As said above, the growing amount of data leads to diverse data (both structured and unstructured). Raw images, text, video are considered as unstructured data because they do not follow a specific format, in contrast to structured data where there is a semantic relationship between objects. Generally, clustering approaches are applied without taking into account the structure of the data. It is precisely for these reasons, that new algorithms are being developed. In the literature several authors presents an overview of clustering techniques and highlights some emerging, and useful, trends in data clustering, some of which are presented below.

Another problem in clustering is the scalability, which is stated as Large-scale clustering: large size datasets are being handled by the clustering algorithms. Some of them are based on efficient nearest neighbor's search and use trees as in literature or random projections as in. When clustering algorithms are summarized then large data set are converted to small data set. Dataset as with the BIRCH algorithm in contrast to sampling based methods like CURE algorithm which creates a sub-sample, when clustering is performed on a small data set then it is transferred to large data set.

It is more difficult when a cluster is formed by the mixture of heterogeneous components in the multi way. A classical clustering method leads to poor performances. Co-clustering treats this problem, and has been successfully applied to document clustering (at the same time both the word and documents are clustered together, this leads to multi way clustering here a set of objects are being clustered simultaneously by the heterogeneous components.

5.SPECTRAL ENSEMBLE CLUSTERING

Let X = fx_1; : : : ; x_ng$^{>}$ 2 $R^{n\ d}$ represent the data matrix containing n instances in d dimensions. $_i$ is a crisp basic partition of X with K_i clusters generated by some traditional clustering algorithm, and $_i$(x) 2 f1; 2; ; K_ig represents the cluster label of instance x. Given r basic partitions of X in = f $_1$; $_2$; ; $_r$g, a co-association matrix $\mathbf{S}_{n\ n}$ is

Algorithm 1 Spectral Ensemble Clustering (SEC)

```
Input: = f 1; 2;    ; rg: r basic partitions.
K: the number of clusters.
Output:     : the consensus partition.
1: Build the binary matrix B = [b(x)] by Eq. 2;
2: Calculate the weight for each instance x by
3: Call weighted K-means on B' = [b(x)=w_b(x)] with the weight w_b(x) and return the partition ;
```

6. EXPERIMENTAL RESULTS

In this section, we evaluate SEC on abundant real-world data sets of different domains, and compare it with several state-of-the-art algorithms across both ensemble clustering and multi-view clustering areas. In the first scenario, each data set is provided with a single view and basic partitions are produced by some random sampling schemes. In the second scenario, however, each data set is provided with multiple views and each view generates either one or mul-tiple basic partitions by random sampling. Finally, a case study on large-scale Weibo data shows the ability of SEC for big data clustering.

6.1 Scenario I: Ensemble Clustering

6.1.1 Experimental Setup

Data. Various real-world data sets with true cluster labels are used for evaluating the experiments in the scenario of ensemble clustering. Table 2 summarizes some impor-tant characteristics of these data sets obtained from UCI[1], CLUTO[2], and LIBSVM[3] repositories, respectively.

Tool. SEC is coded in MAT̄LAB. The kmeans function in MATLAB with either squared Euclidean distance (for UCI and LIBSVM data sets) or cosine similarity (for CLUTO data sets) is run 100 times to obtain basic partitions by varying p the cluster number in [K; n], where K is the true cluster number and n is the data size. For two relatively large data sets letter and mnist, the cluster numbers of basic parti-tions vary in [2; 2K] for meaningful partitions. The baseline methods include consensus clustering with category utility function (CCC, a special case of KCC [12]), graph-based con-sensus clustering methods (GCC, including CSPA, HGPA and MCLA) [1], co-association matrix with agglomerative hierarchical clustering (HCC with group-average, single-linkage and complete-linkage) [5], and probability trajectory based graph partitioning (PTGP) [22]. These baselines are selected for the following reasons: GCC has great impacts in the area of consensus clustering; CCC shares common grounds with SEC by employing a K-means-like algorithm; both HCC and PTGP are co-association matrix based meth-ods, and the former is a very famous one and the latter is newly proposed. All the methods are coded in MATLAB and set with default settings. The cluster number for SEC and all baselines is set to the true one for fair comparison. All basic partitions are equally weighted (i.e., =1). Each algorithm runs 50 times for average results and deviations.

Validation. We employ external measures to assess clus-ter validity. It is reported that the normalized Rand index (R_n for short) is theoretically sound and shows excellent properties in practice [48], which therefore is adopted in our study. R_n is defined as follows:

6.2 Validation of Effectiveness

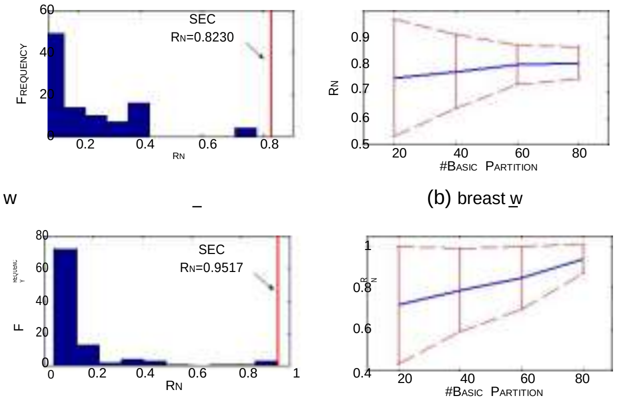
Here, we compare the performance of SEC with that of baseline methods in consensus clustering. Table 3 (Left side) shows the clustering results, with the best results highlighted in bold red and the second best in italic blue.

Firstly, it is obvious that SEC shows clear advantages over other consensus clustering baselines, with 10 best and 9 second best results out of the total 19 data sets; in particular, the margins for the three data sets: wine, la12 and mm are very impressive. To fully compare the performance of different algorithms, we propose a measurement score as follows:

score(A_i) = ${}^{P}j$ $\frac{Rn(Ai;Dj)}{}$, where $R_n(A_i; D_j)$ denotes

$\max_i R_n(A_i;D_j)$

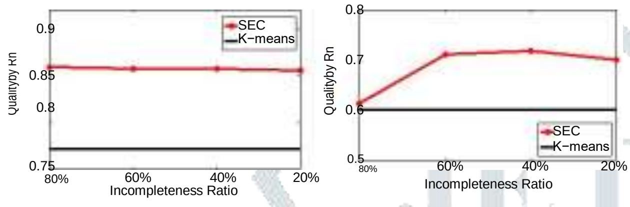
the R_n value of the A_i algorithm on the D_j data set. This score evaluates certain algorithm by the best performance achieved by the state-of-the-art methods. From this score, we can see that SEC exceeds other consensus clustering methods by a large margin.



(a) breast w (b) breast w

(c) cranmed (d) cranmed

Fig. 2. Impact of quality and quantity of basic partitions.



(a) mm (b) reviews

Fig. 3. Performance of SEC with different incompleteness ratios.

Let us take a close look at HCC, which as SEC also leverages co-association matrix for consensus clustering. It is obvious that SEC outperforms HCC with group-average (HCC GA) completely, in 13 out of 19 data sets, although HCC GA is already the second best among the baselines. The implication is two-fold: First, the superior performances of SEC and HCC GA indicate that the co-association matrix indeed does well in integrating information for consensus clustering; Second, a spectral clustering is much better than a hierarchical clustering in making the most of a co-association matrix. The reason for the second point is complicated, but the lack of explicit global objective function in HCC variants might be one of them; that is, unlike CCC or SEC, HCC variants have no utility function to supervise the process of consensus learning, and therefore could perform much less stably than SEC. This is supported by the extremely poor performances of HCC GA on cacmcisi and mm

in Table 3, with negative R_n values even poorer than that of random labeling. Similar observations can be found for the newly proposed algorithm PTGP on mm, which employs the mini-cluster based core co-association matrix but also lacks of utility functions for consensus learning.

We finally turn to CCC, which shares with SEC the K-means clustering in consensus clustering but assigns equal weights to instances. From Table 3, the performance of CCC seems much poorer than that of SEC, especially on breast w and cacmcisi. This indicates that equally weighting of data instances might not be appropriate for consensus learning. In contrast, starting from the spectral clustering view of a co-association matrix, SEC enforces the weights of the instances in large clusters in a quite natural way, and finally leads to superior performances.

6.2.1 Validation of Efficiency

Table 3 (Right side) shows the average execution time of various consensus clustering methods with 50 repetitions.

TABLE 4
Experimental Data Sets for Scenario II

View	Digit	3-Sources	Multilingual	4-Areas
1	Pixel (240)	BBC (3560)	English (9749)	Conference (20)
2	Fourier (74)	Guardian (3631)	German (9109)	Term (13214)
3	-	Reuters (3068)	French (7774)	-
#Instances	2000	169	600	4236
#Classes	10	6	6	4

Since HCC variants have similar execution time, we here only report the results of HCC GA due to limited space.

It is obvious that the K-means-like methods, such as SEC and CCC, get clear edges to competitors, and HCC runs the slowest for adopting hierarchical clustering. This indeed demonstrates the value of SEC in transforming spectral clustering of co-association matrix into weighted K-means clustering. On one hand, we make use of co-association matrix to integrate the information of basic partitions nicely. On the other hand, we avoid generating and handling co-association matrix directly but make use of weighted K-means clustering on the binary matrix to gain high effi-ciency. Although PTGP runs faster than HCC, it needs much more memory and fails to deliver results for two large data sets letter and mnist.

6.2.2 Validation of Robustness

Fig. 2(a) and Fig. 2(c) demonstrate the robustness of SEC by taking breast w and cranmed as example. We choose these two data sets due to their relatively well-structured clusters

— it is often difficult to observe the theoretical properties of an algorithm given very poor performances. We can see that for each data set, the majority of basic partitions are of very low quality. For example, the quality of over 60 basic partitions on cranmed is below 0.1 in terms of R_n. Nevertheless, SEC performs excellently (with $R_n > 0{:}95$) by leveraging the diversity among poor basic partitions. Similar phenomena also occur on some other data sets like breast w, which indicates the power of SEC in fusing diverse information from even poor basic partitions.

6.2.3 Validation of Generalizability and Convergence

Next, we check the generalizability and convergence of SEC. Fig. 2(b) and Fig. 2(d) show the results by varying the num-ber of basic partitions from 20 to 80 for breast w and cranmed, respectively. Note that the above process is repeated 20 times for average results. Generally speaking, it is clear that with the increasing number of basic partitions (i.e., r), the performance of SEC goes up and becomes stable gradually. For instance, SEC achieves satisfactory result from breast w with only 20 basic partitions, but it also suffers from high volatility given such a small r; when r goes up, the variance becomes narrow and stabilizes in a small region.

6.2.4 Effectiveness of Incompleteness Treatment

Here, we demonstrate effectiveness of SEC in handling in-complete basic partitions (IBP). The row-segmentation strat-egy is employed to generate IBPs. In detail, data instances are firstly randomly sampled with replacement, with the sampling ratio going up from 20% to 80%, to form over-lapped data subsets and generate IBPs; SEC is then called to ensemble these IBPs and obtain a consensus partition. Note that for each ratio, the above process repeats 100 times

TABLE 5

Clustering Results in Scenario II (by R_n)

Data sets	Digit	3-Sources	Multilingual	4-Areas
ConKM	0.58 0.06	0.16 0.08	0.12 0.04	0.00 0.00
ConNMF	0.49 0.06	0.28 0.09	0.22 0.02	0.03 0.06
ColNMF	0.39 0.03	0.20 0.05	0.22 0.02	0.11 0.14
CRSC	0.64 0.03	0.30 0.04	0.24 0.01	0.00 0.00
MultiNMF	**0.65 0.03**	0.22 0.06	0.22 0.02	0.00 0.00
PCV	0.56 0.00	N/A	N/A	0.01 0.00
SEC	0.44 0.05	**0.55 0.09**	**0.25 0.03**	**0.56 0.09**

Note: N/A means no result due to more than two views data sets.

TABLE 6

Clustering Results in Scenario II with pseudo views (by R_n)

Data sets	Digit	3-Sources	Multilingual	4-Areas
ConKM	0.62 0.09	0.09 0.05	0.15 0.04	0.00 0.00
ConNMF	0.51 0.05	0.25 0.04	0.21 0.00	0.02 0.06
ColNMF	0.43 0.07	0.14 0.09	0.20 0.00	0.04 0.08
CRSC	0.66 0.02	0.32 0.02	0.25 0.04	0.00 0.00
MultiNMF	0.65 0.06	0.23 0.08	0.22 0.01	0.00 0.01
PCV	N/A	N/A	N/A	N/A
SEC	**0.69 0.06**	**0.62 0.09**	**0.29 0.03**	**0.67 0.09**

Note: N/A means no result due to more than two views data sets.

to obtain IBPs, and unsampled instances are omitted in the final consensus learning. It is intuitive that a lower sampling ratio leads to smaller overlaps between IBPs and thus worse clustering performances. Fig. 3 shows the sample results on and reviews, where the horizontal line indicates the K-means clustering result on the original data set and serves as the baseline unchanged with the sampling ratio. As can be seen, SEC keeps providing stable and competitive results as the sampling ratio goes down to 20%, which demonstrates the effectiveness of incompleteness treatment of SEC.

6.3 Scenario II: Multi-view Clustering

6.3.1 Experimental Setup

Data. Four real-world data sets, i.e., UCI Handwritten Digit, 3-Sources, Multilingual and 4-Areas listed in Table 4, are used in the experiments. UCI Handwritten Digit[4] consists of 0-9 handwritten digits obtained from the UCI repository, where each digit has 200 instances with 240 features in pixel view and 76 features in Fourier view. 3-Sources[5] is collected from three online news sources: BBC, Guardian and Reuter, from February to April in 2009. Of these docu-ments, 169 are reported in all three sources (views). Each document is annotated with one of six categories: busi-ness, entertainment, health, politics, sports and technology. Multilingual[6] contains the documents written originally in five different languages over 6 categories. We here use the sample suggested by [35], which has 100 documents for each category with three views in English, German and French, respectively. 4-Areas[7] is derived from 20 conferences in four areas including database, data mining, machine learning and information retrieval. It contains 28,702 authors and 13,214 terms in the abstract. Each author is labeled with one or multiple areas, and the cross-area authors are removed for unambiguous evaluation. The remainder has 4,236 au-thors in both conference and term views.

Tool. We compare SEC with a number of baseline algorithms including ConKM, ConNMF, ColNMF [32], CRSC [34], MultiNMF [35] and PVC [36]. All the competitors are with default settings whenever possible. Gaussian ker-nel is used to build the affinity matrix for CRSC. The trade-off parameter is set to 0.01 for MultiNMF as suggested in Ref. [35]. For SEC, we employ the kmeans function in MATLAB to generate one basic partition for each view, and then call SEC to fuse them with equal weights into a consensus one. Each algorithm is called 50 times for the average results.

3. CONCLUSION

As mentioned before, thousands of clustering algorithms have been proposed in the literature in many different scientific disciplines. This makes it extremely difficult to review all the published approaches. Nevertheless, clustering methods differ on the choice of the objective function, probabilistic generative models, and heuristics. We will briefly review some of the major approaches. Clusters can be defined as high density regions in the feature space separated by low-density regions.

REFERENCES

[1]. Sidhu, Nimrat Kaur, and Rajneet Kaur. "Clustering in data mining." International Journal of Computer Trends and Technology (IJCTT) 4, no. 4 (2013): 710-714.

[2]. Aggarwal, Charu C., and Chandan K. Reddy, eds. Data clustering: algorithms and applications. CRC press, 2013.

[3]. Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.

[4]. Ghosh, Joydeep, and Ayan Acharya. "Cluster ensembles." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, no. 4
(2011): 305-315.

[5]. Topchy, A. Jain, and W. Punch, “Combining multiple weak clusterings,” in Proceedings of ICDM, 2003.

[6]. Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k -means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24, no. 7 (2002): 881-892.

[7]. Bradley, Paul S., Usama Fayyad, and Cory Reina. Scaling EM (expectation-maximization) clustering to large databases. Redmond: Technical Report MSR-TR-98-35, Microsoft Research, 1998.

[8]. Liu, Hongfu, Tongliang Liu, Junjie Wu, Dacheng Tao, and Yun Fu. "Spectral ensemble clustering." In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 715-724. ACM, 2015.

[9]. Liu, H., Wu, J., Liu, T., Tao, D., & Fu, Y. (2017). Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 29(5), 1129-1143.

[10]. Tung, Frederick, Alexander Wong, and David A. Clausi. "Enabling scalable spectral clustering for image segmentation." Pattern Recognition 43.12 (2010): 4069-4076.

[11]. Trivedi, S., Pardos, Z., Sárközy, G., & Heffernan, N. (2010, June). Spectral clustering in educational data mining. In Educational Data Mining 2011.

[12]. Trivedi, S., Pardos, Z. A., & Heffernan, N. T. (2011, June). Clustering students to generate an ensemble to improve standard test score predictions. In International Conference on Artificial Intelligence in Education (pp. 377-384). Springer Berlin Heidelberg.

[13]. Shu, L., Chen, A., Xiong, M., & Meng, W. (2011, April). Efficient spectral neighborhood blocking for entity resolution. In Data Engineering (ICDE), 2011 IEEE 27th International Conference on (pp. 1067-1078). IEEE.

[14]. Choromanska, Anna, et al. "Fast spectral clustering via the nyström method." International Conference on Algorithmic Learning Theory. Springer, Berlin, Heidelberg, 2013.

[15]. Li, Mu, Xiao-Chen Lian, James T. Kwok, and Bao-Liang Lu. "Time and space efficient spectral clustering via column sampling." In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 2297-2304. IEEE, 2011.

[16]. Shamir, Ohad, and Naftali Tishby. "Spectral clustering on a budget." In International Conference on Artificial Intelligence and Statistics, pp. 661-669. 2011.

[17]. Wauthier, Fabian L., Nebojsa Jojic, and Michael I. Jordan. "Active spectral clustering via iterative uncertainty reduction." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[18]. Chen, Wen-Yen, et al. "Parallel spectral clustering in distributed systems." IEEE transactions on pattern analysis and machine intelligence 33.3 (2011): 568-586.

[19]. Wu, Rui, Jianhua Huang, Xianglong Tang, and Jiafeng Liu. "A Text Image Segmentation Method Based on Spectral Clustering." Computer and Information Science 1, no. 4 (2008): 9.

Author's Profile:

DALLI SAI SURESH pursing B. Tech in Dept of CSE in 2018, respectively. , GITAM University, Visakhapatnam,Andhra Pradesh,India saisureshdalli6@gmail.com.

BODDU JOY STEVENSON pursing B. Tech in Dept of CSE in 2018, respectively. , GITAM University, Visakhapatnam,Andhra Pradesh,India joystevenson01@gmail.com

SIVAH AKASH pursing B. Tech in Dept of CIVIL in 2018, respectively. , GITAM University, Visakhapatnam,Andhra Pradesh,India akash.universe@yahoo.com