IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 2, 2020, accepted February 14, 2020, date of publication February 28, 2020, date of current version March 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977195

Co-Regularized Discriminative Spectral Clustering With Adaptive Similarity Measure in Dual-Kernel Space

AUGUSTINE MONNEY[1,2], (Member, IEEE), YONGZHAO ZHAN[1], HONGJIE JIA[1], AND BEN-BRIGHT BENUWA[3]

[1]Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
[2]IT Services Directorate, University of Education at Winneba, Winneba 03323, Ghana
[3]School of Computer Science, Data Link Institute, Tema 2481, Ghana

Corresponding author: Yongzhao Zhan (yzzhan@ujs.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672268 and Grant 61906077.

ABSTRACT Spectral clustering is a very popular graph-based clustering technique that partitions data groups based on the input data similarity matrix. Many past studies based on spectral clustering, however, do not consider the global discriminative structure of the dataset. Also, the benefits of using more than one kernel have not been fully exploited with respect to spectral clustering, although it has been established by these past studies that using more than one kernel in clustering can result in a more accurate clustering than those obtained with a single kernel. Multi-kernel approaches, however, tend to be more time consuming compared to single kernel methods. To compensate these drawbacks, we integrate a global discriminative term into the clustering with an adaptive neighbor framework. This is done to preserve both the global geometric information and global discriminative information in a dual kernel space, in an attempt to optimize clustering performance. Via co-regularization, we utilize more than one kernel space to take advantage of the benefits of multiple kernels. We, however, use two heterogeneous kernels to help us reduce clustering time, since the ability to quickly process data is as equally important as its accuracy in this era of information explosion. Since these different kernel spaces admit the same underlying clustering of the data, we approach the problem looking for clustering consistent across the two kernel views. Hence we are able to detect the non-linear intrinsic geometrical information of the dataset. We perform clustering using the obtained indicator matrix from our modified Laplacian utilizing k-means. Our Experimental outcomes show that our approach gives satisfactory results in terms of accuracy and NMI, with time-to-cluster savings in comparison to other state-of-the-art clustering methods using both synthetic and public datasets.

INDEX TERMS Adaptive neighbors, co-regularize, multi-kernel, similarity measure, spectral clustering.

I. INTRODUCTION

Clustering is a very useful procedure in the field of artificial intelligence. Based on clustering results, many analytic approaches could be carried out. As a result, several studies have been done on various clustering approaches including hierarchical clustering methods, central grouping methods, and graph clustering methods with them achieving great success [1]–[3]. Compared with conventional clustering algorithms, spectral clustering (SC) has obvious advantages. It can converge to a global optimum and performs well for sample space of arbitrary shapes and especially suitable for non-convex dataset [4]. On very challenging clustering tasks in real-world applications such as image and video segmentation, spectral clustering is known to perform very well and hence a preferred approach for numerous researchers [5].

It is worth noting that the advancement of information technology has led many applications to possess rich data structure and relations [6]. It is vital to utilize the benefits of these multiple types of information to improve the clustering performance. Multiple kernels can be considered as different views of the same data [7] and are thus considered

The associate editor coordinating the review of this manuscript and approving it for publication was Paul Yoo.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

as such in our discussions. We combine two kernel views via co-regularization. Co-regularization employs two main assumptions for its success [8], [9]. Firstly, the true target functions in each view should agree on the labels for the unlabeled data. Secondly, the views should be independent given the class label. The first assumption makes it possible to reduce the space of likely target hypotheses by searching only over the compatible functions. The second assumption makes it unlikely for compatible classifiers to agree on wrong labels. Thus, a data point in both views is most likely to be assigned to the correct cluster.

Inspired by the above discussions, we propose in this paper a unique spectral clustering algorithm henceforth referred to as Co-Regularized Discriminative Spectral Clustering with Adaptive Similarity Measure in Dual-Kernel Space (CoRDiSC-ASMDKS). We build on the Clustering with Adaptive Neighbors (CAN) model [10]. The CAN model learns data similarity matrix by assigning the adaptive and optimal neighbours for each data point based on the local distances and then imposes a rank constraint on the Laplacian matrix of the data similarity matrix. The efficacy of spectral clustering critically depends on the construction of the graph Laplacian and the ensuing eigenvectors that reflect the cluster structure in the data. We therefore construct an objective function that consists of the discriminative graph Laplacians from all the kernel views of the data and regularize on the eigenvectors of the Laplacians such that the cluster structures resulting from each Laplacian look consistent across all the views.

The proposed approach integrates both global geometrical structure and global discrimination structure in a dual kernel co-regularized framework, to perform spectral clustering. Just as in [10], we assume that the similarity between two points is the probability that the two points are neighbors in the kernel space since [10], [11] shows that probabilistic neighborhood is very effective in measuring similarity in data feature learning.

The proposed approach has the following summary as its main contributions:

1) It introduces discriminability into the normalized Laplacian of each kernel view that ensures that the features to the spectral clustering is more discriminative.
2) It uses a co-regularization approach to combine objectives of the individual kernel spaces with their disagreements to obtain a joint minimization problem that is solved to obtain a class indicator matrix used in k-means for clustering.
3) It improves time for clustering samples compared to other multi-kernel clustering approaches and is satisfactorily robust to noisy data.

The rest of the paper is presented as follows. Section II gives a brief of related works. The proposed approach is elaborated in section III. Experimental results are presented in Section IV and Section V completes the paper.

II. RELATED WORKS

1) SPECTRAL CLUSTERING

SC is based on the algebraic graph theory, which treats data clustering problem as a graph partitioning problem [12]–[15]. It constructs an undirected weighted graph with each node corresponding to a data point, and the weight of the edge connecting the two nodes being the similarity value between the two points [16], [17]. Then, using a certain graph cut method, it divides the graph into connected components, which are called clusters.

Ng *et al.* [18] introduced a theoretical work grounded on matrix perturbation theory that brought forward the conditions under which to expect a good performance of the SC algorithm [19]. Their method found the optimal value of parameter σ to improve spectral clustering. In [11], Du and Shen proposed a self-tuning spectral clustering algorithm, which improved the SC algorithm by locally scaling the parameter in similarity measure. Luxburg [20] summarized comprehensively the main literature related to spectral clustering.

SC has a few challenges including but not limited to its inability to handle big data without using approximation methods such as the Nyström algorithm [21], [22], the power iteration method [23], or linear algebra-based methods [24]–[26]. It however has enormous advantages including its ability to perform very well on very challenging clustering tasks in real-world applications such as image and video segmentation.

2) KERNEL, MULTI-KERNEL AND ENSEMBLE SPECTRAL CLUSTERING

Kernel based methods works by mapping data into high dimensional feature space implicitly defined by the choice of the kernel function. Alzate and Suykens [27] introduced a technique named kernel spectral clustering (KSC), which is based on solving a constrained optimization problem in a primal-dual setting. From [27], casting SC in a learning framework allows to meticulously select tuning parameters such as the natural number of clusters which are present in the data and also, an accurate prediction of the cluster memberships for unseen points. This can be done by projecting the test data in the embedding Eigen space learned during training. Ye and Sakurai [28] based their work on the fact that SC makes use of the spectrum of some normalized similarity matrix that is derived from the data to reveal the cluster structure, and the fact that data is normally very complex, heterogeneous and high dimensional. They measured the similarity of data points not in their original space, but in kernel space to precisely reflect the underlying data structure. This results in better clustering with the most appropriate kernel function chosen. In general, learning graph in kernel space can enhance clustering accuracy due to the incorporation of nonlinearity.

In recent times, many approaches that make use of multiple kernels have also been proposed. This is as a result of the fact that in many applications, there could be multiple possibly beneficial features and thereby multiple affinity matrices. To ensure better clustering results, multiple affinity matrices should be aggregated or fused. Some of these recent methods includes Low-rank Kernel Learning for Graph-based Clustering [29], Clustering with Similarity Preserving [30] and Robust Graph Learning from Noisy Data [31].

Besides multiple kernels, other authors have also considered the ensemble clustering technique, which aims to utilize multiple clusterers to obtain a stronger clusterer. Some of these techniques includes the Ultra-Scalable Spectral Clustering and Ensemble Clustering [32], Enhanced Ensemble Clustering via Fast Propagation of Cluster-wise Similarities [33], Locally Weighted Ensemble Clustering [34], Robust Ensemble Clustering Using Probability Trajectories [35], and a clustering ensemble framework based on elite selection of weighted clusters [36].

3) DISCRIMINATIVE CLUSTER ANALYSIS

It is observed that many algorithms consider the global manifold structure of datasets [12], [37], [38], but fail to consider the discriminative structure which reveals the intrinsic structure of the data distribution. We know that both manifold information and discriminant information are of great importance for clustering and hence we expect to preserve the discriminant information of a dataset in the learning process. Discriminative cluster analysis (DCA) [39] uses discriminative features for clustering rather than generative ones. In [40], [41], both the local manifold structure and the global discriminant information are preserved simultaneously through manifold discriminant learning. In [42], the proposed local discriminative and global integration clustering algorithm (LDMGI) combines the local discriminative models and manifold structure for clustering. Nie *et al* in [43] introduce a new Laplacian matrix into a spectral embedded clustering frame work to capture local and global discriminative information for clustering. In the work of Yang *et al.* [44], the global discriminative regularization term is introduced, which provides a more discriminative information that enhances clustering performance. These algorithms use the global discriminative information, and make their performance to improve.

Recently, [30] proposed a discriminative graph learning method which can preserve the pairwise similarities between samples in an adaptive manner. This was due to the fact that prior kernel-based graph learning mechanisms was not similarity-preserving, hence led to sub-optimal performance. Their method required the learned graph to be close to a kernel matrix, which serves as a measure of similarity in raw data. Our method differs from [30] by learning a discriminative consensus result over a collection of kernels. We take advantage of the fact that different kernel spaces admit the same underlying clustering of data, and hence, learn a view that is consistent across the kernels using co-regularization to detect the nonlinear intrinsic geometrical information of the dataset. We use a modified Laplacian which is discriminative to learn a consistent view that is used to obtain an indicator matrix to be utilized in k-means for clustering.

4) PROBABILISTIC NEIGHBORS WITH KERNEL DISTANCE MEASURING

According to the Reproducing Kernel Hilbert Space (RKHS) theory, we can calculate our Mercer Kernel on a given set of n data points $\{x_i\}_{i=1}^n$, with $X_i \in R^d$ with a function $K : X \times X \rightarrow R$. This can be expressed as;

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \tag{1}$$

with $\varphi : X \rightarrow F$ performing a mapping from the original space X to a high dimensional feature space F. In this way, each coordinate matches a feature of the data items. The distance d_{ij}^{φ} between data points x_i and x_j in the feature space of a Kernel (φ), is measured using the Euclidian distance, but according to the Mercer Kernel theorem, this can be calculated from the kernel values as illustrated in equation (2).

$$d_{ij}^{\varphi} = \sqrt{K(x_i, x_j) + K(x_j, x_j) - 2K(x_i, x_j)} \tag{2}$$

We ensure that our data is centered so that $\sum_{i=1}^{n}(\frac{x_i}{n}) = 0$. Assuming, $\varphi(x_i)$ and $\varphi(x_j)$ are neighbors, then the probability p_{ij} can be said to be $0 \leq p_{ij} \leq 1$, and also the probabilities for all the data points to be connected to $\varphi(x_i)$ satisfy $\sum_{j=1}^{n} P_{ij} = 1$. The probability of data points being neighbors is thus inversely proportional to their distance of separation. Therefore P_{ij} is large if $\left\|\varphi(x_i) - \varphi(x_j)\right\|^2$ is small. The probabilities $P_{ij}(j = 1, \ldots, n)$ of all data points to be connected to $\varphi(x_i)$ can be determined as;

$$\min_{p_i} \sum_{j=1}^{n} \left\|\varphi(x_i) - \varphi(x_j)\right\|_2^2 p_{ij},$$
$$s.t.\ 0 \leq p_{ij} \leq 1, \quad p_i^T \mathbf{1} = 1 \tag{3}$$

where $p_i \in R^{n \times 1}$ is a vector with the j-th element as P_{ij}

Adding a regularizing parameter λP_{ij}^2 to the problem (3), we avoid a trivial solution.

$$\min_{p_i} \sum_{j=1}^{n} \left\|\varphi(x_i) - \varphi(x_j)\right\|_2^2 p_{ij} + \lambda P_{ij}^2,$$
$$s.t.\ 0 \leq p_{ij} \leq 1, \quad p_i^T \mathbf{1} = 1 \tag{4}$$

where λ is the regularization parameter and p_{ij} is the probability that $\varphi(x_i)$ is a neighbor of $\varphi(x_j)$. Solving problem (4), we can assign neighbors to each data point $\varphi(x_i)$.

From the work of Nie *et al.* [10], we can add an additional constraint $rank(L_p) = n - k$ to the problem (4) to help attain optimal neighbor assignment, hence the problem is equivalent to;

$$\min_{P,F} \sum_{i,j=1}^{n} \left\|\varphi(x_i) - \varphi(x_j)\right\|_2^2 p_{ij} + \lambda P_{ij}^2 + 2\gamma Tr(F^T L_p F)$$
$$s.t.\ 0 \leq p_{ij} \leq 1, \quad p_i^T \mathbf{1} = 1,\ F \in R_{n\times k},\ F^T F = I_K \tag{5}$$

where $L_P = I_K - D^{\frac{1}{2}} P D^{\frac{1}{2}}$ is the normalized Laplacian matrix with P as the similarity matrix of the data set, I_K is an identity matrix of size K, and D as the degree matrix with the i-th diagonal entry defined as $\sum_{j=1}(P_{ij} + P_{ji})$ and $F \in R^{n\times k}$ is defined as the weighted indicator matrix.

III. PROPOSED METHOD

The offered algorithm for our technique, Co-regularized Discriminative Spectral Clustering with Adaptive Similarity Measure in Dual Kernel Space (CoRDiSC-ASMDKS) is comprehensively outlined in this segment. Given a data set made up of n data points $\{x_1, x_2, \ldots, x_n\}$, we strive to group the n data points into K clusters $\{C_j\}_{j=1}^K$. To explore the non-linear feature space of the datasets, the data points are projected into high dimensional spaces to learn adaptively the optimal neighbors of each data point in these spaces. In our experiments, the linear and Gaussian radial basis function kernel (RBF) are used. A similarity matrix and a discriminative term is learnt and added to the determined Laplacian in these kernel spaces. The disagreement between the clusterings is measured and combined with the discriminative spectral clustering with adaptive similarity measure objectives of the individual kernel spaces, to obtain a joint minimization problem. Using alternate minimization with respect to each view, the joint minimization problem is solved, and k-means is used to perform the final clustering.

A. CO-REGULARIZED DISCRIMINATIVE SPECTRAL CLUSTERING

Given data with two kernel representations. Let $\varphi^{(v)}(X) = \varphi^{(v)}(X_1), \varphi^{(v)}(X_2), \ldots, \varphi^{(v)}(X_n)$ and $\varphi^{(u)}(X) = \varphi^{(u)}(X_1), \varphi^{(u)}(X_2), \ldots, \varphi^{(u)}(X_1)$ denote samples in kernel representations v and u respectively, and $P^{(v)}$ and $P^{(u)}$ also denote the similarity of $\varphi(X)$ in kernel representations v and u respectively. We introduce a discrimination term ($\mathbf{\Theta}$) as deduced in the next subsection into problem (5) for each kernel representation.

1) DISCRIMINATIVE TERM (Θ)

For a data set with n data points $x_i \in R^d$ having a similarity matrix P, with p_{ij} representing the relationship between x_i and x_j, our target is to group in K clusters $C_{jj} = 1^K$ the data points. To obtain very good clustering results, discriminative information should be considered, hence inspired by the work done by Wang *et al.* [45], discriminability is introduced into our normalized Laplacian to ensure that the spectral clustering is discriminative. We represent the j_{th} column of the indicator matrix F_j as

$$F_j = (0, \ldots, 0, \overbrace{1, \ldots, 1}^{n_j}, 0, \ldots, 0)^T / n_j^{\frac{1}{2}}$$
$$s.t.\ F_{ij} = \begin{cases} 1/\sqrt{n_j}, & \text{if } x_i \in C_j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

with n_j being the size of the cluster C_j. Similar to [46], we formulate our discrimination term as

$$\max_F Tr(F^T(I_n - (I_n + (\frac{1}{\lambda}G)^{-1})F)$$
$$s.t.\ F^TF = I_K \tag{7}$$

where $Tr(\cdot)$ signifies the trace of a matrix, I_n represents an identity matrix of size n, λ which is greater than zero is a regularization parameter, and a symmetrical positive semi-definite Gram matrix is represented as $G = X^TX$.

Ye *et al* shows in [46] that $tr(F^TF) = K$ and for a given similarity matrix P, the equivalent of problem (7) can be given as

$$\mathbf{\Theta} = \min_F Tr(F^T\hat{G}F)$$
$$s.t\ F^TF = I_K \tag{8}$$

where $\hat{G} = (I_n + \frac{1}{\lambda}G)^{-1}$ is treated as a Gram matrix

2) DISCRIMINATIVE SPECTRAL CLUSTERING WITH ADAPTIVE SIMILARITY MEASURE IN KERNEL SPACE

A discrimination term ($\mathbf{\Theta}$) is introduced into problem (5) for each kernel representation. Thus for example, in the v-kernel representation, the following equation is obtained;

$$\min_{P^{(v)},F^{(v)}} \sum_{i,j=1}^{n}(\left\|\varphi^{(v)}(x_i) - \varphi^{(v)}(x_j)\right\|_2^2 p_{ij}{}^{(v)} + \lambda^{(v)}P_{ij}^{(v)2}$$
$$+ 2\gamma^{(v)}\left(\alpha_1\, Tr(F^{(v)^T}L_p^{(v)}F^{(v)}) + (1-\alpha_1)\Theta^{(\mathbf{v})}\right)$$
$$s.t.\ 0 \le p_{ij}{}^{(v)} \le 1, \quad p_i^{(v)^T}\mathbf{1} = 1,\ F^{(v)} \in R^{n\times k},$$
$$F^{(v)^T}F^{(v)} = I_K \tag{9}$$

where $\alpha_1 \in [0, 1]$

Expanding problem (9) we obtain

$$\min_{P^{(v)},F^{(v)}} \sum_{i,j=1}^{n}(\left\|\varphi^{(v)}(x_i) - \varphi^{(v)}(x_j)\right\|_2^2 p_{ij}{}^{(v)} + \lambda^{(v)}P_{ij}^{(v)2}$$
$$+ 2\gamma^{(v)}Tr\big(\alpha_1(F^{(v)^T}L_p^{(v)}F^{(v)})$$
$$+ (1-\alpha_1)(F^{(v)^T}\hat{G}^{(v)}F^{(v)})\big)$$
$$s.t.\ 0 \le p_{ij}{}^{(v)} \le 1, p_i^{(v)^T}\mathbf{1} = 1, F^{(v)} \in R^{n\times k},$$
$$F^{(v)^T}F^{(v)} = I_K \tag{10}$$

This is equivalent to

$$\min_{P^{(v)},F^{(v)}} \sum_{i,j=1}^{n}(\left\|\varphi^{(v)}(x_i) - \varphi^{(v)}(x_j)\right\|_2^2 p_{ij}{}^{(v)} + \lambda^{(v)}P_{ij}^{(v)2}$$
$$+ 2\gamma^{(v)}Tr\big(F^{(v)^T}[(\alpha_1\, L_p^{(v)}) + (1-\alpha_1)\hat{G}^{(v)}]F^{(v)}\big)$$
$$s.t.\ 0 \le p_{ij}{}^{(v)} \le 1, \quad p_i^{(v)^T}\mathbf{1} = 1,\ F^{(v)} \in R^{n\times k},$$
$$F^{(v)^T}F^{(v)} = I_K \tag{11}$$

where $F^{(v)} \in R^{(n\times K)}$ is defined as the weighted indicator matrix, $\hat{G^{(v)}}$ is treated as a Gram matrix, $L_P{}^{(v)}$ is the normalized Laplacian matrix with $P^{(v)}$ as the similarity matrix of the data set and $\alpha_1 \in [0, 1]$.

We solve problem (11) as our optimization problem for each kernel representation of our data.

a: UPDATE P AND λ WITH F FIXED

When F is fixed, the problem (11) can be reformed as

$$\begin{aligned}\min_{P^{(v)},F^{(v)}} \sum_{i,j=1}^{n}(&\left\|\varphi^{(v)}(x_i)-\varphi^{(v)}(x_j)\right\|_2^2 p_{ij}{}^{(v)}+\lambda^{(v)}P_{ij}{}^{(v)2}\\ &+2\gamma^{(v)}Tr\big(F^{(v)^T}[(\alpha_1\, L_p^{(v)})+(1-\alpha_1)\hat{G}^{(v)}]F^{(v)}\big)\\ s.t.\ \forall i\ 0\leq p_{ij}{}^{(v)}&\leq 1,\quad p_i{}^{(v)^T}\mathbf{1}=1 \end{aligned}\tag{12}$$

If we assigned an arbitrary function value $f_i^{(v)} \in R^{(K\times 1)}$ to each node, it can be verified that

$$\begin{aligned}&\min_{P^{(v)}}\sum_{i,j=1}^{n}\left\|f_i^{(v)}-f_j^{(v)}\right\|_2^2 p_{ij}{}^{(v)}\\ &\quad=2\gamma^{(v)}Tr\big(F^{(v)^T}[(\alpha_1 L_p^{(v)})+(1-\alpha_1)\hat{G}^{(v)}]F^{(v)}\big)\end{aligned}\tag{13}$$

where $F^{(v)} \in R^{n\times k}$ with the $i-th$ row formed by $f_i^{(v)}$

We can therefore obtain

$$\begin{aligned}\min_{P^{(v)},F^{(v)}} \sum_{i,j=1}^{n}(&\left\|\varphi^{(v)}(x_i)-\varphi^{(v)}(x_j)\right\|_2^2 p_{ij}{}^{(v)}\\ &+\lambda^{(v)}P_{ij}^{(v)^2})+\sum_{i,j=1}^{n}\left\|f_i^{(v)}-f_j^{(v)}\right\|_2^2 p_{ij}{}^{(v)}\\ s.t.\ \forall i\ 0\leq p_{ij}{}^{(v)}&\leq 1,\quad p_i^{(v)^T}\mathbf{1}=1\end{aligned}\tag{14}$$

We solve problem (14) individually for each i. If we denote $d_{ij}^{\varphi x(v)} = \|\varphi^{(v)}(x_i)-\varphi^{(v)}(x_j)\|_2^2$, $d_{ij}^{f(v)} = \|(f_i)^{(v)}-(f_j)^{(v)}\|_2^2$ and $d_i^{\varphi(v)} \in R^{n\times 1}$ as a vector with the j-th element as $d_{ij}{}^{(v)} = d_{ij}^{\varphi x(v)}+d_{ij}^{f(v)}$, problem (14) can be re-written into vector form as:

$$\begin{aligned}&\min_{p_i{}^{(v)}}\left\|p_i{}^{(v)}+\frac{d_i{}^{\varphi(v)}}{2\lambda_i{}^{(v)}}\right\|_2^2\\ &s.t.\ \forall i\ 0\leq p_{ij}{}^{(v)}\leq 1,\quad p_i^{(v)^T}\mathbf{1}=1\end{aligned}\tag{15}$$

For each i, the Lagrangian function of problem (15) is;

$$\begin{aligned}\tau(p_i{}^{(v)},\mu_i{}^{(v)},\sigma_i{}^{(v)})=&\frac{1}{2}\left\|p_i{}^{(v)}+\frac{d_i^{\varphi(v)}}{2\lambda_i{}^{(v)}}\right\|_2^2\\ &-\mu_i{}^{(v)}(p_i^{(v)^T}\mathbf{1}-1)-\sigma_i^{(v)^T}p_i{}^{(v)}\end{aligned}\tag{16}$$

where $\mu_i{}^{(v)}$ and $\sigma_i{}^{(v)} \geq 0$ are the Lagrangian multipliers. Solving problem (16) similarly as in [10], we obtain $\lambda^{(v)}$ and $p_{ij}{}^{(v)}$ as;

$$\lambda^{(v)}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}(d_{i,k+1}^{\varphi^{(v)}})-\frac{1}{2}\sum_{j=1}^{k}d_{ij}^{\varphi(v)}\right)\tag{17}$$

$$p_{ij}{}^{(v)}=\frac{(d_{i,k+1}^{\varphi^{(v)}})^2-(d_{ij}^{\varphi^{(v)}})^2}{kd_{i,k+1}^{\varphi^{(v)}}{}^2-\sum_{j=1}^{k}(d_{ij}^{\varphi^{(v)}})^2}\tag{18}$$

where k is the k-nearest number

b: UPDATE F WITH P FIXED

When P is fixed, the objective function in problem (11) can be reformed as:

$$\begin{aligned}&\min_{F^{(v)}}Tr\big(F^{(v)^T}[(\alpha_1 L_p^{(v)})+(1-\alpha_1)\hat{G}^{(v)}]F^{(v)})\big)\\ &s.t.\ F^{(v)}\in R^{n\times k},\quad F^{(v)^T}F^{(v)}=I_K\end{aligned}\tag{19}$$

Here, the constant $F^{(v)}$ is relaxed as in traditional clustering approaches. From the Ky Fan theorem [47], the solution to problem (19) can be derived as the eigenvectors of the matrix $(\alpha_1\ L_p^{(v)}) + (1-\alpha_1)\hat{G}^{(v)}$ corresponding to the smallest K eigenvalues. Hence, in the convergence of $P^{(v)}$, an optimal solution $F^{(v)}$ to problem (19) is attained.

Likewise for view u, its solution is the eigenvectors of the matrix $(\alpha_2\ L_p^{(u)})+(1-\alpha_2)\hat{G}^{(u)}$ corresponding to the smallest K eigenvalues. Therefore as in problem (19), the optimal solution $F^{(u)}$ is attained in the convergence of $P^{(u)}$.

3) CO-REGULARIZATION

In our procedure, we want to take advantage of the benefits of using more than one kernel. As stated earlier, two kernel views; (u) and (v), are used. The indicator matrix F in each of these views is solved for as in problem (19). In the objective function offered by the proposed approach, it is desired that, new representation (in terms of rows of $F^{(\cdot)}$'s) is comparable across all the kernels views. We therefore strengthen the pair-wise similarities of samples under this representation. Thus, implementing the spectral clustering hypotheses (which are based on the $F^{(\cdot)}$'s) to be the same across all the kernels. The disagreement between clusterings in the two views is measured as;

$$D(F^{(v)},F^{(u)})=\left\|\frac{S_{F^{(v)}}}{\|S_{F^{(v)}}\|_2^2}-\frac{S_{F^{(u)}}}{\|S_{F^{(u)}}\|_2^2}\right\|_2^2\tag{20}$$

where $S_{F^{(v)}}$ and $S_{F^{(u)}}$ are the similarity matrix for $F^{(v)}$ and $F^{(u)}$ respectively. Normalizing the similarity matrices of the views makes them comparable across all views. Since $S_{F^{(v)}} = F^{(v)}F^{(v)^T}$ and $S_{F^{(u)}} = F^{(u)}F^{(u)^T}$, then, the graphs with the learned similarity $S_F^{(v)}$ and $S_F^{(u)}$ will have exactly K connected components. Thus, $\|S_{F^{(v)}}\|_2^2 = K$ and $\|S_{F^{(u)}}\|_2^2 = K$. Substituting this into Equation (20) and ignoring the constant additive and scaling terms that depend on the number of clusters, we get

$$D(F^{(v)},F^{(u)})=-tr\left(F^{(v)}F^{(v)^T}F^{(u)}F^{(u)^T}\right)\tag{21}$$

We choose to let $\alpha_1\ L_p^{(v)} + (1-\alpha_1)\hat{G}^{(v)}$ and $\alpha_2\ L_p^{(u)} + (1-\alpha_2)\hat{G}^{(u)}$ in the objectives of individual kernel spaces in equation (19) to be represented as $Z^{(v)}$ and $Z^{(u)}$. Hence, combining problem (21) with the discriminative spectral clustering with adaptive similarity measure objectives of individual kernel spaces, the following joint minimization problem is obtained.

$$\min_{\substack{F^{(v)}\in R^{n\times K}\\ F^{(u)}\in R^{n\times K}}} tr\big(F^{(v)^T} Z^{(v)} F^{(v)}\big) + tr\big(F^{(u)^T} Z^{(u)} F^{(u)}\big) - \delta tr\big(F^{(v)} F^{(v)^T} F^{(u)} F^{(u)^T}\big)$$
$$s.t\ F^{(v)} \in R^{n\times k},\quad F^{(v)^T} F^{(v)} = I_K,\ F^{(u)^T} F^{(u)} = I_K \quad (22)$$

where δ is a trade-off of the spectral clustering objectives and the disagreement term.

Using alternating minimization with respect to $F^{(v)}$ and $F^{(u)}$, problem (22) is solved. Hence, for a given $F^{(u)}$, the optimization problem in $F^{(v)}$ becomes:

$$\min_{F^{(v)}\in R^{n\times K}} tr\big(F^{(v)^T} (Z^{(v)} - \delta F^{(u)} F^{(u)^T}) F^{(v)}\big)$$
$$s.t\ F^{(v)} \in R^{n\times k},\quad F^{(v)^T} F^{(v)} = I_K \quad (23)$$

The alternate minimization process is repeated until convergence. We monitor the convergence by the change in the value of the objective between successive iterations, and stop when the difference falls below a minimum threshold of $\epsilon = 10^{-4}$. The solution $F^{(v)}$ is given by the smallest-k eigenvectors of this modified Laplacian $Z^{(v)} - \delta F^{(u)} F(u)^T$. The final step in our approach is to use either of $F^{(v)}$ or $F^{(u)}$ in k-means to perform the final clustering. We summarise our procedure as in Algorithm 1.

Algorithm 1 CoRDiSC-ASMDKS Method

```
Input: Data matrix $X \in R^{(d\times n)}$; Parameters $\alpha, k, c$;
Output: Cluster indexes of $x_1, x_2, \ldots, x_n$
for kernel $v, u$ do
    Initialize $P$ by the optimal solution to the problem (4).
    while not converge do
        for each $i$ do
            update the $i$-th row of $P$ by solving problem (15), where $d_i^{\varphi} \in R^{(n\times 1)}$ is a vector with the j-th element as $d_{ij} = d_{ij}^{\varphi x} + d_{ij}^{f}$
        end for
    end while
    Update $F$, which is formed by the $K$ eigenvectors of $Z = (\alpha L_p + (1-\alpha)\hat{G})$ corresponding to the $K$ smallest eigenvalues
end for
Determine $D(F^{(v)}, F^{(u)})$
Solve problem (22) using alternate minimization until convergence
Use $F^{(v)}$ or $F^{(u)}$ in k-means to obtain the final clustering
```

4) COMPLEXITY ANALYSIS

The computational cost involved in solving the main aspects of our proposed CoRDiSC-ASMDKS technique is shown in this sub-section. Our method, CoRDiSC-ASMDKS has a general complexity of

$$O(n^2 logn + nk) + O(n^2) + O(2n^3)$$

This includes initializing P in equation (4), updating the i-th row of P in equation (15), and computing the disagreements between the two kernel views. It includes additionally the general measure for complexity analysis of computing eigenvectors from a dense matrix. However in our case, we solved the eigenproblem by applying sparse eigensolvers [49]. We used ARPACK, which is a variants of Lanczos/Arnoldi factorization with a complexity of

$$\Big(O(h^3) + O(nh) + O(nk) + O(h-c)\Big) * A$$

where A is the number of restarted Arnoldi, $h > c$ is the Arnoldi length used to compute the first c eigenvectors of our modified Laplacian matrix.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed method by comparing it with other state-of-the-art spectral clustering methods.

A. SETUP

All algorithms were implemented on Matlab R2016a (revision 9.0.0.341360) 64-bit, running on a windows7 Intel Core^{TM} i3-4170 CPU @3.7GHz 3.70GHz processor with an 8GB installed memory. Using already implemented tools of Matlab, ARPACK and the Kernel Methods Toolbox (KMBOX), we adapted the open source Matlab codes as presented by [18], [48] and [50]. We use the Gaussian radial basis function (RBF) and Linear kernels in our experiments. We use the same initialization, pre-setting the neighborhood parameter value $k = 10$. We select the parameters α and δ as in subsection IV-E. We repeat our experiments ten times and record the best values.

B. DATA SELECTION

To evaluate the performance of our approach, experiments are conducted on synthetic and publicly available data sets. We first perform experiments on three simple 2D synthetic data; 2S + Circle, 8-Gaussian and the 4S + Noise, to ascertain the usefulness of our method. We then implement our algorithm on other eight publicly available data sets with various degrees of challenges from the UCI Machine Learning Repository [51], the MNIST [52] and Trec database repositories to further evaluate the performance of our algorithm. These data sets are from different fields. We use three face databases; JAFFE, UMIST, Yale and ORL, a handwritten digits database; MNIST, a toy image database; COIL20, and BA which is a binary alpha digits data set. These were taken under different configurations, so some of them are corrupted severely. We use also, one biological dataset; LUNG. The last data sets is a text corpora from TREC 2.

Table 1 and 2 summarizes the characteristics of the synthetic and public datasets.

TABLE 1. Characteristics of the synthetic datasets.

Dataset	Distribution	No. of categories	Size
2S+Circle	Random	3	400
8-Gaussian	Gaussian	8	300
4S+Noise	Random	5	380

TABLE 2. Characteristics of the public datasets.

Dataset	No. of samples	No. of features	No. of categories
UMIST	575	400	20
ORL	400	1024	40
JAFFE	213	676	10
MNIST	6000	748	10
Lung	203	3312	5
COIL20	1440	1024	20
BA	1404	320	36
TR41	878	7454	10
YALE	165	1024	15

C. COMPARATIVE METHODS

Discussing effectively the general performance of our proposed algorithm requires choosing a good set of comparative methods. Our proposed approach was compared with both graph-based and kernel clustering methods. The kernel methods includes both single and multiple kernel approaches.

For our synthetic data, many legacy methods as well as state-of-the-arts approaches were used to establish the worthiness of our proposed approach. These methods include K-means, Normalized Cut, Ratio Cut, Self-Tuning Spectral Clustering (ST-SC) [53], Local density adaptive similarity measurement for spectral clustering (DA-SC) [54], spectral clustering based on k-nearest neighbour (kNN-SC) [55], Spectral clustering with adaptive similarity measure (ASM-SC) [56] and spectral clustering with adaptive similarity measure in kernel space (ASMK-SC) [28].

With that established, the proposed approach is then tested on publicly available datasets. For these publicly available dataset, we compare our approach specifically with the Robust Kernel K-means (RKKM) [57], Self-Tuning Spectral Clustering (ST-SC) [53], Clustering with Adaptive Neighbor (CAN) [10], Spectral clustering with adaptive similarity measure (ASM-SC) [56], Spectral clustering with adaptive similarity measure in Kernel space (ASMK-SC) [28], Robust Graph learning from Noisy data (RGC) [31], Low-rank Kernel Learning for Graph-based Clustering (LRKL) [29] and Clustering with Similarity Preserving (SPC and mSPC) [30].

D. EVALUATION METHODS

In this paper, the provided label of each sample is matched with the label computed by the various clustering methods to give the clustering result. Accuracy(ACC) and Normalized Mutual Information(NMI) are used for measuring the clustering performance [58]. The accuracy is defined as follows:

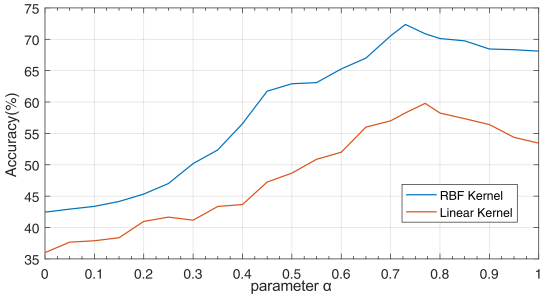


FIGURE 1. Effect of α for kernels.

$$ACC = \frac{\sum_{i=1}^{n} \delta(l_i, map(c_i))}{n},$$

where l_i is the label of data and c_i is the lable result gotten by clustering. If $a = b$, $\delta(a, b)$ equal 1,else, $\delta(a, b)$ equal 0. $map(c_i)$ is the permutation mapping function that best map each cluster label c_i to the equivalent label from dataset. Let L be the true label provided by the dataset and L' be the label gotten from clustering algorithm. The mutual information between L and L' is defined as follows:

$$MI(L,L') = \sum_{Li \in L} \sum_{L'_i \in L} p(l_i, l'j) log_2 \frac{p(l_i, l'_j)}{p(l_i)p(l'_j)},$$

where $p(l_i)$ and $p(l'$ are the marginal probability distribution functions of L and L'. $p(l_i, l'_j)$ is the joint probability distribution function of L and L'. However, in our experiments, we use the NMI for our performance comparison and is defined as follows:

$$NMI(L,L') = \frac{MI(L,L')}{max(H(L), H(L')},$$

where $(H(L)$ and $H(L')$ are the entropies of $p(L)$ and $p(L')$. The NMI takes values in [0, 1]. If NMI equals 1, the two clusters labels are identical; otherwise, they are independent. In our experiments however, we express our NMI as a percentage for easy appreciation.

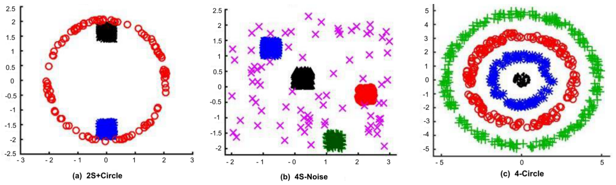
E. PARAMETER SELECTION

1) SELECTION OF α

We demonstrate how the parameter α are selected for our approach using the Jaffe dataset in Figure 1. The parameter α, has a discriminatory balancing effect on the dataset which help us to attain optimal performance of each kernel optimization equation of our proposed method. We set α carefully through experiments to achieve optimal performance. This is done by choosing a suitable α first by evaluating it in a sample of the whole dataset. In theory, α should take a range of 0 to 1. We discovered that on the datasets we used, the best α parameter was in the range of 0.59 to 0.77 depending on the specific dataset. We vary α from 0 to 1 to determine the best α parameter for each kernel. Thus we record the best performance for each kernel acting alone.

TABLE 3. Performance in terms of accuracy (acc) and NMI on synthetic datasets.

Dataset	Performance metric (%)	K-means	DA-SC	ST-SC	kNN-SC	Ratio Cut	Normalize Cut	ASM-SC	ASMK-SC	CoRDiSC-ASMDKS
2S+Circle	NMI	61.32	63.47	67.5	67.26	68.25	68.19	67.13	68.47	**70.52**
	ACC	86.34	87.81	89.8	88.7	89.31	89.16	89.65	90.20	**92.18**
8-Gaussian	NMI	90.24	94.83	96.35	96.37	96.63	96.41	96.82	97.63	**98.96**
	ACC	88.78	92.06	94.89	94.97	95.68	95.57	95.95	96.52	**97.87**
4S+Noise	NMI	70.65	81.42	82.47	81.16	81.57	81.66	81.90	82.66	**86.14**
	ACC	84.13	89.72	89.84	88.26	88.92	88.63	89.67	90.77	**93.62**



(a) 2S+Circle (b) 4S-Noise (c) 4-Circle

FIGURE 2. Synthetic datasets.

As per the results shown, when we set α_1 which is the α for the RBF kernel to 0.73, and α_2 which is the α for the Linear kernel to 0.77, we obtain the best accuracy. It is worth noting that the curves have an upward trend in the values $\alpha = 0$ to 0.73, $\alpha = 0$ to 0.77 and begin to decline afterwards. When $\alpha = 1$, it implies that our model has completely neglected the effect of discrimination. We can conclude from the graph that 0.73 and 0.77 are fine choices for α_1 and α_2 for the Jaffe data set. Also, it may not be a good idea to set a very low α value, otherwise there can be an imbalance between the discriminatory term (Θ) and the normalized Laplacian matrix L_P with P as the similarity matrix of the data set which may lead to a low performance of the model.

2) SELECTION OF δ

In order to select the optimum co-regularization parameter (δ), we experimented with different values of δ in the range 0 to 0.1, and observed how it affected performance. We observed that on most of the data sets, δ showed best performance in the range of $\delta = 0.009$ to 0.052. The best performance of the parameter δ for each dataset was chosen for use in our experiments. For instance, the parameter δ on the Jaffe data set has a best performance at $\delta = 0.014$.

We observed generally that accuracy increases as δ increased from 0 until it reaches its highest point. It then starts decreasing with local ups and downs until δ reaches 0.1.

F. EXPERIMENTS ON SYNTHETIC DATASETS

As shown in Figure 2, three synthetic datasets are used in our experiments.

The NMI and accuracy results of the various methods in comparison to CoRDiSC-ASMDKS on synthetic datasets are illustrated in table 3. It is evident from the experimental results in table 3 that the proposed CoRDiSC-ASMDKS method outdoes all the baseline approaches on the synthetic datasets. CoRDiSC-ASMDKS, for example, recorded an NMI value of 70.52% on the 2S + Circle dataset, surpassing ASMK-SC by 2.05%, ASM-SC by 3.39%, Normalized cut by 2.33%, Ratio Cut by 2.27%, ST-SC by 3.02%, kNN-SC by 3.26%, DA-SC by 7.05% and K-means by 9.20%. The results in table 3 also show that most of the methods performed better than K-means on the 4-Circle, 4-Corner, and 2-Spiral datasets. It is also noted that CoRDiSC-ASMDKS recorded the highest accuracy of 92.18%, 97.87% and 93.62% for the 2S + Circle, 8-Gaussian and 4S + Noise datasets respectively, surpassing ASMK-SC which is the second-best method by 1.98%, 1.35% and 2.85% respectively.

Using box diagrams to illustrate Accuracy(%) and NMI(%) on the 8-Guassian dataset; which contains eight clusters of data obeying the eight Gaussian distributions, and also 2S + Circle dataset made up of three clusters of data randomly distributed in two squares and one circle, we again show the pre-eminence of our approach with the other comparing methods in Figures 3(*a* and *b*), and 4(*a* and *b*).

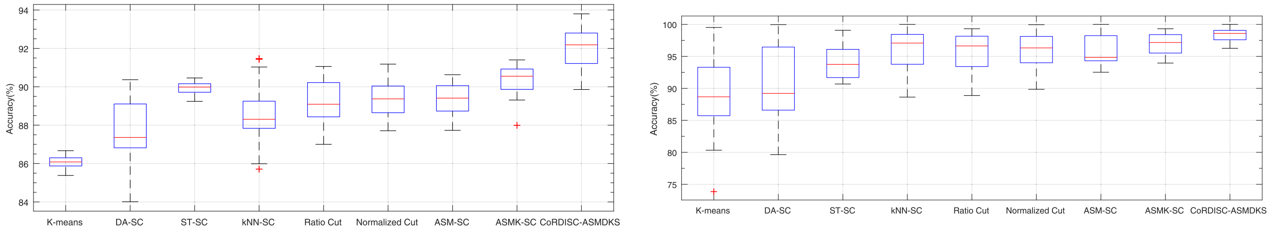
CoRDiSC-ASMDKS shows the highest value Accuracy% and NMI% amongst all the comparing methods from the results. The poorest performance is recorded by K-means as a result of the fact that all the other methods improve upon its performance.

We demonstrate how varying the parameter k affects performance using selected synthetic datasets. Since kNN-SC, ASM-SC and ASMK-SC have the same parameter k, we compare the proposed method with these three methods in the experiments. Figures 5(*a*-*d*) presents the clustering results in terms of accuracy and NMI, varying k on the 2S + Circle and 8-Guassian datasets.

We find that on the 2S + Circle data, our proposed CoRDiSC-ASMDKS method obtains the best accuracy at $k = 9$, whereas ASM-SC and ASMK-SC obtains the best accuracy at $k = 10$. We also find that, kNN-SC is able to cluster correctly when $k = 6$, however, it is very unstable as k varies. This illustrates the fact that different methods performs differently with different values of k.

G. EXPERIMENTS ON PUBLIC DATASETS

In the experiments on public datasets, we present the best NMI and Accuracy performance of the proposed method in comparison to the comparative methods. We organize our results in two tables. The first table (table 4), presents a comparison of the proposed approach with graph and single kernel based methods. The second table (table 5) presents



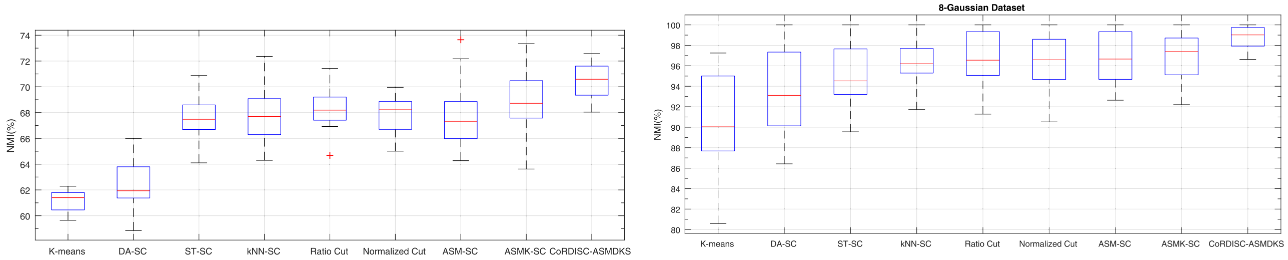
(a) 2S+Circle Dataset

(b) 8-Guassian Dataset

FIGURE 3. Box plot of accuracies for selected synthetic datasets.

TABLE 4. Performance in terms of accuracy (ACC) and NMI for graph and single kernel based methods on public datasets.

Dataset	Performance metric (%)	RKKM	ST-SC	CAN	SPC	ASM-SC	ASMK-SC	CoRDiSC-ASMDKS
UMIST	NMI	63.05	64.31	67.09	85.31	84.07	85.13	**87.01**
	ACC	41.81	42.34	62.55	70.62	68.37	70.22	**73.95**
ORL	NMI	74.23	79.41	76.59	**86.06**	82.36	82.83	84.13
	ACC	54.96	57.68	61.50	**75.75**	66.88	67.09	72.97
JAFFE	NMI	83.47	88.03	97.31	**98.62**	93.40	95.37	98.39
	ACC	75.61	75.14	98.12	98.03	91.22	93.42	**98.54**
MNIST	NMI	51.52	52.91	60.74	**68.71**	63.18	66.86	69.87
	ACC	56.93	57.68	61.81	65.32	62.48	64.37	**66.29**
Lung	NMI	49.58	54.34	60.17	65.63	63.39	65.81	**69.33**
	ACC	64.95	70.23	74.26	82.98	83.17	85.79	**88.08**
COIL20	NMI	74.63	78.25	91.55	89.57	90.87	92.23	**94.47**
	ACC	61.64	59.89	84.58	83.88	80.71	86.60	**89.25**
BA	NMI	57.82	50.76	49.32	58.46	57.27	59.77	**62.89**
	ACC	42.17	31.07	36.82	48.72	48.49	49.05	**50.62**
YALE	NMI	52.29	52.92	57.67	61.32	61.45	61.62	**64.41**
	ACC	48.09	49.42	58.79	60.53	59.28	61.56	**65.68**
TR41	NMI	60.77	61.33	51.13	**71.22**	60.38	62.54	68.94
	ACC	56.76	63.52	62.87	**72.89**	61.40	64.49	72.08



(a) 2S+Circle

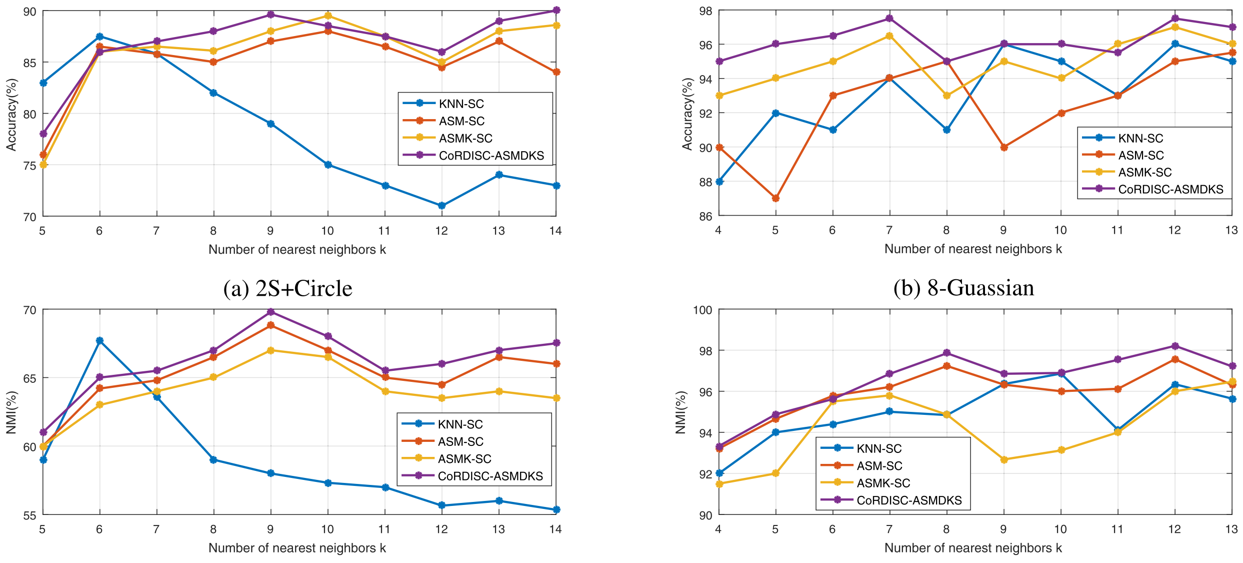
(b) 8-Guassian

FIGURE 4. Box plot of NMI's for selected synthetic datasets.

a performance comparison with multiple kernel based methods.

Examining the results, it is observed that comparing the results of approaches involving kernel methods to non-kernel methods, the kernel methods generally recorded improvements over the clustering results of non-kernel methods, more significantly on the public datasets. This can be attributed to Mercer Kernels giving a more general way of representing complex data through which clusters can accurately be identified. Table 4 shows that the proposed approach performs much better than most of the single kernel methods indicating that our procedure may have benefited from the advantages of using more than one kernel in clustering, and the discrimination that took place at each kernel level. We see SPC performing well in terms of accuracy and NMI on the ORL and TR41 datasets but unable to perform same on the other datasets.

From table 5, we see that the performance of the proposed approach, closely matches the other comparative methods. It obtained the best performance in terms of accuracy and NMI on the LUNG and COIL20 datasets. In terms of accuracy for example on the COIL20 dataset, it surpassed RGC by



(a) 2S+Circle

(b) 8-Guassian

(c) 2S+Circle

(d) 8-Guassian

FIGURE 5. Accuracy (%) and NMI (%) varying the parameter *k* on selected synthetic datasets.

TABLE 5. Performance in terms of NMI and accuracy (ACC) for multiple kernel based methods on public datasets.

Dataset	Performance metric (%)	CoRDiSC-ASMDKS	RGC	LRKL	mSPC
UMIST	NMI	87.01	**87.03**	86.64	86.42
	ACC	73.95	73.76	72.17	**74.25**
ORL	NMI	84.13	84.35	85.10	**85.93**
	ACC	72.97	73.00	73.50	**75.43**
JAFFE	NMI	98.39	98.13	**98.73**	97.36
	ACC	98.54	98.59	**98.60**	98.14
MNIST	NMI	69.87	69.83	69.40	**71.93**
	ACC	66.29	**66.31**	65.81	66.22
Lung	NMI	**69.33**	68.12	66.54	68.46
	ACC	**88.08**	87.91	86.86	87.34
COIL20	NMI	**94.47**	91.58	93.23	92.30
	ACC	**89.25**	85.42	86.70	85.11
BA	NMI	62.89	**64.89**	63.20	61.81
	ACC	50.62	**51.00**	50.50	49.67
YALE	NMI	64.41	**65.29**	64.57	61.36
	ACC	65.68	64.85	**66.06**	63.03
TR41	NMI	68.94	67.35	62.85	**70.50**
	ACC	72.08	70.16	63.48	**80.41**

3.83%, LRKL by 2.55% and mSPC by 4.14%. On the LUNG dataset, it outperformed RGC by 0.17%, LRKL by 1.22% and mSPC by 0.74% in terms of accuracy. For the other datasets, its performance though not the best was quite significant. In terms of accuracy, it recorded 73.95%, 66.29%, 50.62%, 68.68% and 72.08% to become the second best approach for the UMIST, MNIST, BA, YALE, and TR4 datasets. In terms of NMI, it also recorded the second best approach for the UMIST, JAFFE, MNIST and TR4 datasets respectively. It was however, the third best approach for the BA and YALE datasets with reference to NMI performance. Accuracy for the proposed approach was third best for the JAFFE dataset.

It is important to note however, that no single multi-kernel method performed unilaterally well on all the datasets. However, although our method wasn't the best approach, its performance is significant considering the fact that it used only two kernels. Also, comparing the time it takes to cluster, our approach has lower computational cost. We show this in our next section.

H. TIME-TO-CLUSTER FOR MULTI-KERNEL BASED METHODS ON PUBLIC DATA SETS

For us to appreciate the performance of our method further, we measure the time it takes for each method to cluster selected public datasets. Table 6 shows the obtained results in seconds (s).

It is clear that our approach records the best time to cluster. The time-to-cluster per sample for the proposed algorithm is

TABLE 6. Clustering time in seconds (s) for multi-kernel methods at k = 10.

Dataset	CoRDiSC-ASMDKS	RGC	LRKL	mSPC
UMIST	$0.13183s$	$0.47273s$	$0.44544s$	$0.43300s$
ORL	$0.10721s$	$0.38024s$	$0.33104s$	$0.34164s$
JAFFE	$0.01293s$	$0.04692s$	$0.03996s$	$0.03066s$
MNIST	$0.18248s$	$0.56168s$	$0.45268s$	$0.54734s$
LUNG	$0.17926s$	$0.45041s$	$0.42913s$	$0.47484s$
COIL20	$0.92908s$	$2.80994s$	$2.77073s$	$2.62695s$
YALE	$0.05972s$	$0.17492s$	$0.15164s$	$0.12480s$

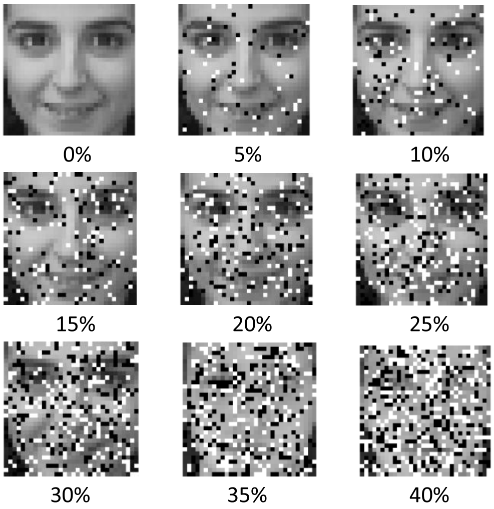


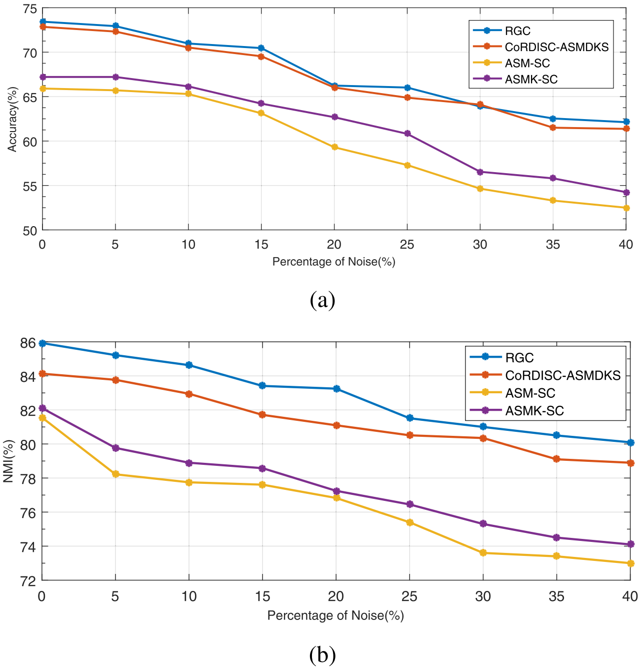
FIGURE 6. Sample image from ORL dataset with varied noise ratio.

more than three times as fast as the multi-kernel comparative methods. On the UMIST dataset, it was 3.67 times faster than RBC, 3.38 times faster than LRKL and 3.28 times faster than mSPC. This results from the fact that our model uses only two kernels, and hence is able to process the clusters faster although its performance reduces slightly. But in this era of big data, ability to process data quickly is quite important and hence the performance of our model in terms of accuracy and NMI can be said to be highly satisfactory.

I. ROBUSTNESS TO NOISY DATA

In this sub-section, we investigate how robust our proposed approach CoRDiSC-ASMDKS is in comparison to the Robust Graph learning from Noisy data (RBC) method [31], spectral clustering with adaptive similarity measure (ASM-SC) [56] and spectral clustering with adaptive similarity measure in kernel space (ASMK-SC) [28] to noisy dataset. It is worth noting that methods such as Robust Graph learning from noisy data (RBC) uses robust graph learning scheme to learn reliable graphs from real-world noisy data by adaptively removing noise and errors in the raw data. Our approach does not aim to remove noise but demonstrate the importance of discriminatory information in noise handling.

We introduce salt & pepper noise in an incremental order of 5%, 10% up to 40% to the ORL datasets and record their average performance. Figure 6 shows sample data with varied noise ratio.



(a)

(b)

FIGURE 7. Robustness of CoRDISC-ASMDKS in comparison to RGC, ASM-SC and ASMK-SC on ORL dataset with varied noise ratio.

This experiment is performed using the tuned optimal parameter settings for the algorithms and we record the best performances at $k = 10$. Figure 7 shows that the proposed approach did not degrade drastically with the introduction of noise and performed better than ASM-SC and ASMK-SC. It is noticed that, with the introduction of noise, its performance degraded in a similar fashion as the RGC technique, although lagged behind RBC. At 30% noise level, it is seen that our methods accuracy performance is a little better than RGC.

The proposed approach is able to find the global discriminatory data structure, identify noisy samples and obtain satisfactory performance against noise. What this implies is that, incorporating discriminatory information in noise removing clustering techniques can further improve their performance.

V. CONCLUSION

In this paper, a novel co-regularized discriminative spectral clustering method with adaptive similarity measure in a dual-kernel space is proposed. This enabled us to exploit some of the benefits of multiple kernels, since multi-Kernels can offer a more general way to represent data by which clusters can be more accurately identified. Taking advantage of the fact that different kernel spaces admit the same underlying clustering of data, we learned a view that is consistent across the kernels using co-regularization to detect the non-linear intrinsic geometrical information of the dataset. Our approach essentially differs from existing spectral clustering methods by learning a discriminative consensus result over a collection of kernels. We use a modified Laplacian which is discriminative to learn a consistent view that is used to obtain an indicator matrix which is utilized in k-means to obtain the clustering.

An extensive experimental study on synthetic and public data sets demonstrates that CoRDiSC-ASMDKS obtains satisfactory clustering quality with improved clustering time compared to other state-of-the-art clustering methods. As future work, we will consider extending our work to multiple kernel co-regularization whiles maintaining good time to cluster, to further improve the accuracy of data clustering.

REFERENCES

[1] Y. Zhao, Y. Yuan, F. Nie, and Q. Wang, "Spectral clustering based on iterative optimization for large-scale and high-dimensional data," *Neurocomputing*, vol. 318, pp. 227–235, Nov. 2018.

[2] M. A. Rahman, K. L.-M. Ang, and K. P. Seng, "Unique neighborhood set parameter independent density-based clustering with outlier detection," *IEEE Access*, vol. 6, pp. 44707–44717, 2018.

[3] J. Hou and A. Zhang, "Enhanced dominant sets clustering by cluster expansion," *IEEE Access*, vol. 6, pp. 8916–8924, 2018.

[4] S. Ding, H. Jia, L. Zhang, and F. Jin, "Research of semi-supervised spectral clustering algorithm based on pairwise constraints," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 211–219, Oct. 2012.

[5] J. Xu, H. Li, P. Liu, and L. Xiao, "A novel hyperspectral image clustering method with context-aware unsupervised discriminative extreme learning machine," *IEEE Access*, vol. 6, pp. 16176–16188, 2018.

[6] X.-D. Wang, R.-C. Chen, F. Yan, Z.-Q. Zeng, and C.-Q. Hong, "Fast adaptive K-means subspace clustering for high-dimensional data," *IEEE Access*, vol. 7, pp. 42639–42651, 2019.

[7] M. Ji, H. Rao, Z. Li, J. Zhu, and N. Wang, "Partial multi-view clustering based on sparse embedding framework," *IEEE Access*, vol. 7, pp. 29332–29343, 2019.

[8] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Proc. CVPR*, Jun. 2011, pp. 1977–1984.

[9] S. Wang, E. Zhu, J. Hu, M. Li, K. Zhao, N. Hu, and X. Liu, "Efficient multiple kernel k-means clustering with late fusion," *IEEE Access*, vol. 7, pp. 61109–61120, 2019.

[10] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2014, pp. 977–986.

[11] L. Du and Y.-D. Shen, "Unsupervised feature selection with adaptive structure learning," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Sydney, NSW, Australia, 2015, pp. 209–218.

[12] R. Shang, Z. Zhang, L. Jiao, W. Wang, and S. Yang, "Global discriminative-based nonnegative spectral clustering," *Pattern Recognit.*, vol. 55, pp. 172–182, Jul. 2016.

[13] H. Jia, S. Ding, H. Ma, and W. Xing, "Spectral clustering with neighborhood attribute reduction based on information entropy," *J. Comput.*, vol. 9, no. 6, pp. 1316–1324, Jun. 2014.

[14] X. Yang, S. Liang, H. Yu, S. Gao, and Y. Qian, "Pseudo-label neighborhood rough set: Measures and attribute reductions," *Int. J. Approx. Reasoning*, vol. 105, pp. 112–129, Feb. 2019.

[15] Z. JingMao and S. YanXia, "Review on spectral methods for clustering," in *Proc. 34th Chin. Control Conf. (CCC)*, Jul. 2015, pp. 3791–3796.

[16] H. Jia, S. Ding, H. Zhu, F. Wu, and L. Bao, "A feature weighted spectral clustering algorithm based on knowledge entropy," *J. Softw.*, vol. 8, no. 5, pp. 1101–1108, May 2013.

[17] S. Ding, H. Jia, M. Du, and Q. Hu, "*p*-spectral clustering based on neighborhood attribute granulation," in *Proc. Int. Conf. Intell. Inf. Process.*, 2016, pp. 50–58.

[18] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001, pp. 849–856.

[19] S. Mouysset, J. Noailles, D. Ruiz, and C. Tauber, "Spectral clustering: Interpretation and Gaussian parameter," in *Data Analysis, Machine Learning and Knowledge Discovery*. Cham, Switzerland: Switzerland: Springer, 2014, pp. 153–162.

[20] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Aug. 2007.

[21] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nystrom method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[22] G. Liu, H. Chen, X. Sun, N. Quan, L. Wan, and R. Chen, "Low-complexity nonlinear analysis of synchrophasor measurements for events detection and localization," *IEEE Access*, vol. 6, pp. 4982–4993, 2018.

[23] F. Lin and W. W. Cohen, "Power iteration clustering," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, Jun. 2010, pp. 655–662.

[24] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. S. Huang, "Incremental spectral clustering by efficiently updating the eigen-system," *Pattern Recognit.*, vol. 43, no. 1, pp. 113–127, Jan. 2010.

[25] I. Koutis and H. Le, "Spectral modification of graphs for improved spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4859–4868.

[26] Q. Wang, Z. Qin, F. Nie, and X. Li, "Spectral embedded adaptive neighbors clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1265–1271, Apr. 2019.

[27] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with Out-of-Sample extensions through weighted kernel PCA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 335–347, Feb. 2010.

[28] X. Ye and T. Sakurai, "Spectral clustering with adaptive similarity measure in kernel space," *Intell. Data Anal.*, vol. 22, no. 4, pp. 751–765, Jun. 2018.

[29] Z. Kang, L. Wen, W. Chen, and Z. Xu, "Low-rank kernel learning for graph-based clustering," *Knowl.-Based Syst.*, vol. 163, pp. 510–517, Jan. 2019.

[30] Z. Kang, H. Xu, B. Wang, H. Zhu, and Z. Xu, "Clustering with similarity preserving," *Neurocomputing*, vol. 365, pp. 211–218, Nov. 2019.

[31] Z. Kang, H. Pan, S. C. H. Hoi, and Z. Xu, "Robust graph learning from noisy data," *IEEE Trans. Cybern.*, to be published.

[32] D. Huang, C.-D. Wang, J. Wu, J.-H. Lai, and C. K. Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Trans. Knowl. Data Eng.*, to be published.

[33] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[34] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, May 2018.

[35] D. Huang, J.-H. Lai, and C.-D. Wang, "Robust ensemble clustering using probability trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1312–1326, May 2016.

[36] H. Parvin and B. Minaei-Bidgoli, "A clustering ensemble framework based on elite selection of weighted clusters," *Adv. Data Anal. Classification*, vol. 7, no. 2, pp. 181–208, Apr. 2013.

[37] H. Lu, Z. Fu, and X. Shu, "Non-negative and sparse spectral clustering," *Pattern Recognit.*, vol. 47, no. 1, pp. 418–426, Jan. 2014.

[38] G. Akbarizadeh and M. Rahmani, "Efficient combination of texture and color features in a new spectral clustering method for PolSAR image segmentation," *Nat. Acad. Sci. Lett.*, vol. 40, no. 2, pp. 117–120, 2017.

[39] F. De la Torre and T. Kanade, "Discriminative cluster analysis," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 241–248.

[40] P. Li, J. Bu, Y. Yang, R. Ji, C. Chen, and D. Cai, "Discriminative orthogonal nonnegative matrix factorization with flexibility for data representation," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1283–1293, Mar. 2014.

[41] C. Wang, E. Zhu, X. Liu, L. Gao, J. Yin, and N. Hu, "Multiple kernel clustering with global and local structure alignment," *IEEE Access*, vol. 6, pp. 77911–77920, 2018.

[42] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.

[43] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1796–1808, Nov. 2011.

[44] Y. Yang, Y. Yang, H. T. Shen, Y. Zhang, X. Du, and X. Zhou, "Discriminative nonnegative spectral clustering with out-of-sample extension," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1760–1771, Aug. 2013.

[45] F. Wang, C. Ding, and T. Li, "Integrated KL (K-means–Laplacian) clustering: A new clustering approach by combining attribute data and pairwise relations," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 38–48.

[46] J. Ye, Z. Zhao, and M. Wu, "Discriminative K-means for clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1649–1656.

[47] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1057–1064.

[48] P. Perona and W. Freeman, "A factorization approach to grouping," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany, 1998, pp. 655–670.

[49] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 568–586, Mar. 2011.

[50] J. Shi, S. Belongie, T. Leung, and J. Malik, "Image and video segmentation: The normalized cut framework," in *Proc. Int. Conf. Image Process. (ICIP)*, 1998, pp. 943–947.

[51] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[52] C. Y. LeCun, C. Christopher, and J. C. Burges. *The MNIST Dataset*. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[53] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1601–1608.

[54] X. Zhang, J. Li, and H. Yu, "Local density adaptive similarity measurement for spectral clustering," *Pattern Recognit. Lett.*, vol. 32, no. 2, pp. 352–358, Jan. 2011.

[55] M. Lucinska and S. T. Wierzchon, "Spectral clustering based on K-nearest neighbor graph," in *Proc. 11th IFIP TC8 Int. Conf. Comput. Inf. Syst. Ind. Manage. (CISIM)*, Venice, Italy, Sep. 2012, pp. 254–265.

[56] X. Ye and T. Sakurai, "Similarity measure based on adaptive neighbors for spectral clustering," in *Proc. 9th Int. Conf. Mach. Learn. Comput.*, Singapore: ACM, 2017, pp. 405–409.

[57] L. Du, P. Zhou, L. Shi, H. Wang, M. Fan, and W. Wang, "Robust multiple kernel k-means using l21-norm," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, Buenos Aires, Argentina, Jul. 2015, pp. 3476–3482.

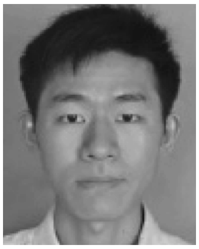
[58] H. Wang and G. Liu, "Two-level-oriented selective clustering ensemble based on hybrid multi-modal metrics," *IEEE Access*, vol. 6, pp. 64159–64168, 2018.

AUGUSTINE MONNEY (Member, IEEE) is currently pursuing the Ph.D. degree with the School of Computer Science and Communication Engineering, Jiangsu University, China. His research interests include clustering, data mining, video semantic analysis, and machine learning.

YONGZHAO ZHAN was born in Sanming, Fujian, China, in 1962. He received the B.S. degree from Fuzhou University, China, in 1984, the M.S. degree from Jiangsu University, China, in 1990, and the Ph.D. degree from Nanjing University, China, in 2000, all in computer science.

He is currently a Professor with the School of Computer Science and Communication Engineering, Jiangsu University. He has authored more than 80 articles. His research interests include big data, multimedia, and the Internet of Vehicles.

Dr. Zhan was a recipient of the Science and Technology Progress Award from the Zhenjiang Government, in 2006, and from the Jiangsu Government, in 2013.



HONGJIE JIA received the Ph.D. degree in computer application technology from the China University of Mining and Technology, Xuzhou, China, in 2017.

He is currently a Lecturer with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include clustering, data mining, and machine learning.

BEN-BRIGHT BENUWA received the Ph.D. degree in computer application technology from Jiangsu University, Zhenjiang, China, in 2018.

He is currently a Lecturer with the School of Computer Science, Data Link Institute, Tema, Ghana, and the University of Education at Winneba, Winneba, Ghana. His research interests include clustering, data mining, video semantic analysis, and machine learning.