2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Sampling based approximate spectral clustering ensemble for partitioning datasets

Yaser Moazzen
Dept. of Electronics and Communications Engineering,
Istanbul Technical University
Ayazaga Kampusu, Ayazaga, Istanbul, Turkey

Kadim Tasdemir
Dept. of Computer Engineering,
Antalya International University
Universite Caddesi 2, Dosemealti, Antalya, Turkey
Email: kadim.tasdemir@antalya.edu.tr

Abstract—**Spectral clustering is able to extract clusters with various characteristics without a parametric model, however it is infeasible for large datasets due to its high computational cost and memory requirement. Approximate spectral clustering (ASC) addresses this challenge by a representative-based partitioning approach which first finds a set of data representatives either by sampling or quantization, then applies spectral clustering on them. To achieve an optimal partitioning with ASC, several sampling or quantization methods together with advanced similarity criteria have been recently proposed. While quantization is more accurate than sampling in expense of heavy computation, and geodesic based hybrid similarity criteria are often more informative than others, there is no unique solution optimum for all datasets. Alternatively, we propose to use ensemble learning to produce a consensus partitioning constructed from different set of representatives and similarity criteria. The proposed ensemble (SASCE) not only produces a relatively more accurate partitioning but also eliminates the need to determine the best pair (the optimum set of representatives and the optimum similarity). Thanks to the efficient similarity definition on the representative level, the SASCE can be powerful for clustering small and medium datasets, outperforming traditional clustering approaches and their ensembles.**

Index Terms—**approximate spectral clustering, cluster ensemble, maximum voting, geodesic distances, hybrid similarity**

I. Introduction

Clustering have been of great interest for data analysis due to its unsupervised nature depending mainly on intrinsic data properties which can be utilized by some (dis)similarity criteria. Among many methods, spectral clustering stands out thanks to its manifold learning approach which exploits pairwise similarities by eigendecomposition of a similarity based graph [1], [2], [3], [4], [5]. Despite its success on extracting clusters with various characteristics without parametric models, its high computational cost and memory requirement make its direct use infeasible for partitioning large datasets. Therefore, approximate spectral clustering (ASC) applies spectral clustering on a subset of data representatives selected by sampling or quantization [6], [7], [8], [9], [10], [11], [12], [13]. Several studies analyze the effects of different sampling and quantization methods in ASC: random sampling [6], [14], selective sampling [8], [11], k-means, random projection trees [11], neural networks (self-organizing maps and neural gas) [12], [13], k-means++ [13]. Wang et al. [8] find selective sampling better performing than random sampling, whereas quantization often outperforms sampling in expense of computational complexity [12], [15]. Besides a representative set selection, ASC brings new similarity definitions on the representative level (such as local data distribution, data topology, geodesic distance and their fusion) [12], [13]. The geodesic based hybrid similarity criteria together with neural gas quantization are outperforming in general, yet, one needs to determine how to select the representative set and similarity criterion for an optimum partitioning in the given dataset [13]. Alternatively, one can fuse all these partitionings obtained by different sets of representatives and similarity criteria into one consensus result by ensemble learning.

Ensemble learning merges partitionings of different input subsets or features, obtained by distinct methods or the same method with several parameter settings, to eliminate the need to determine the optimal selection for the greatest possible classification (or clustering) accuracy. It can be performed by various approaches including majority voting, evidence accumulation, hyper graph operations, metaclustering, or mixture models [16], [17], [18], [19], [20], [21], [22]. Among them, majority voting is commonly preferred due to being the most naive and easy-to-implement approach by counting the labels of the same input sample under various settings. The spectral clustering ensemble in [20], which uses random sampling together with maximum voting and metaclustering algorithm, combines partitionings obtained with different kernel parameter values in similarity definition, for segmentation of relatively small remote sensing images. Despite the ensemble approach, different parameter windows, specific to the datasets, are used to achieve high performance. In addition, when random sampling is used for spectral clustering of large datasets, out-of-sample labeling may become problematic [11], [23]. In contrast, graph based ensemble approaches are often superior to majority voting in expense of heavy computational cost, making them infeasible for large datasets. Despite the fact that the graph based ensemble learning can be run for representative based partitioning of large datasets (such as for approximate spectral clustering) it limits the ensemble on the representative level leading to the use of only one representative set obtained by one sampling/quantization method [24].

In this paper, we propose a sampling based approximate spectral clustering ensemble (SASCE). The SASCE first obtains data representatives using selective sampling to optimize

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

the balance between high computational cost of quantization and low performance of random sampling. Since the sampling has low computational cost, one can obtain many sets of representatives to have alternative partitionings of the dataset. The SASCE then cluster each representative set by spectral clustering using advanced geodesic similarity definitions which utilize different information types [13], instead of the traditional similarity defined by distance based Gauss kernel with different parameter values. The cluster labels of the representatives are assigned to their corresponding data samples. This produces various partitionings of the dataset obtained by different representative sets and similarity definitions. The SASCE merges them into a consensus partitioning by majority voting. By combining different representative sets with advanced similarity definitions harnessing data manifold, the SASCE achieves accuracies superior to other spectral clustering ensembles for large datasets. In addition, thanks to manifold based similarities, the SASCE also provides high accuracies for the medium or small datasets where spectral clustering is feasible.

The paper is outlined as follows. Section II presents the proposed SASCE together with a brief explanation of the approximate spectral clustering, selective sampling, and similarity criteria. Section III first shows the clustering performance of the SASCE based on accuracy and adjusted Rand index on large datasets for which spectral clustering is infeasible due to its computational cost and memory requirement. Then Section III provides SASCE performance for five datasets (from UCI Machine Learning Repository [25]) and a thorough discussion by comparing it to other ensembles based on traditional clustering methods. Section IV concludes the paper.

II. SASCE: Sampling based ASC Ensemble

Being an ensemble of partitionings obtained by ASC, the SASCE has three steps, two of which constitute the ASC (Section II-A): i) determining data representatives by selective sampling (Section II-A1), ii) the spectral clustering of the representatives using advanced similarity definitions (Section II-A2), and assigning the extracted cluster labels to the corresponding data samples, iii) ensemble of data partitionings by majority voting.

At each step of the SASCE, alternative decisions could have been made. However, selective sampling provides a balanced performance between heavy computation cost of quantization methods (such as k-means, neural gas) and relatively low performance of random sampling. Therefore it makes it efficient to obtain different sets of data representatives which can then be clustered to extract different labelling of their corresponding data samples. The advanced geodesic based similarities includes data manifold information which can address the challenges in out-of-sampling labelling to a great extent. The naive ensemble of majority voting provides a simple yet cost effective merging of data partitionings based on different sets of representatives, whereas advanced ensemble methods using graph theoretic approaches are feasible for merging partitionings on the representative level [24]. Depending on the majority voting ensemble on the data level, the SASCE algorithm can be summarized as follows:

1) Obtain n_r sets of data representatives by selective sampling (Section II-A1).
2) Get n_s partitionings of each set of data representatives with spectral clustering using n_s similarity definitions in Section II-A2.
3) Find n_r × n_s cluster labels of data samples using the labels of their corresponding representatives
4) Ensemble n_r × n_s data partitionings by majority voting to obtain a consensus labelling.

A. ASC: Approximate Spectral Clustering

By applying spectral clustering on data representatives obtained by a sampling or quantization method, ASC not only utilizes spectral clustering advantages such as successful extraction of clusters with distinct characteristics without parametric models, but also enables manifold based similarity definitions (such as data topology, local density) on the representative level.

Spectral clustering depends on eigendecomposition of a graph Laplacian matrix L constructed with respect to some optimization criteria [26], [2], [27]. The SASCE employs the method in [2] since there is no clear advantage among spectral clustering methods as long as a normalized graph Laplacian is used [28], [29]. By constructing a weighted undirected graph $G = (V, S)$, where the nodes V are the representatives and the edges S are their pairwise similarities, the normalized Laplacian, L_{norm}, is defined as

$$L_{norm} = D^{-1/2} S D^{-1/2} \quad (1)$$

Here D is diagonal degree matrix D of the similarity matrix S with $d_i = \sum_j s(i,j)$ showing the total similarity of each node. Then, the k eigenvectors $\{e_1, e_2,\ldots,e_k\}$of Lnorm, associated with its k highest eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ are found to construct $E = [e_1, e_2,\ldots,e_k]$ and its normalized variant U (by normalizing the rows of E to have norm 1). Finally, the rows of U are clustered by k-means (or any simple clustering) based on the fact that this eigendecomposition ideally maps submanifolds (clusters) of the dataset in a well separated manner.

Two important ASC decisions are sampling/quantization method and similarity definition for S. While the pair of a neural gas based quantization and geodesic hybrid similarity definition is shown superior for several datasets [13], quantization requires high computational cost and alternative similarity definitions may outperform for some other datasets. To address these challenges, the SASCE uses selective sampling and harness available similarity definitions, which are briefly explained below.

1) Selective sampling for ASC: By addressing several challenges in sampling such as tendency to over-sample and insufficient sample size, selective sampling provides an ASC performance superior to other sampling methods and similar to k-means quantization [9]. It has three steps. Firstly, h distinguished objects $p_1, p_2, \ldots, p_h$ are selected from a dissimilarity matrix (DNN) of the dataset (N is size of dataset). The first index p_1 is randomly selected from the index set {1, 2, · , N } and a search array A is generated:

$$A = (a_1, a_2, .., a_N) = (d_{1,1}, d_{1,2}, .., d_{1,N}) \qquad (2)$$

The remaining h−1 distinguished objects are iteratively selected to maximize ai, i.e., pi = arg max$j$ aj, and then search array A is updated using

$$\text{A} = (\min((a_1,\ d_{pi-1,N}),\ \ldots\ ,\ \min((a_N,\ d_{pi-1,N})) \qquad (3)$$

to get all distinguished h objects with a max – min farthest point strategy. Secondly, each data sample v_i is assigned to the nearest distinguished object q using

$$q = \arg\min_{j}(d_{p,j}\ , i) \qquad (4)$$

to obtain the receptive fields (R_q) of the h objects. Thirdly, n = $\sum_q n_q$ samples are randomly selected from $R_q s$, where R_q is proportional to the number of samples in R_q.

2) Similarity criteria for ASC: In ASC, the similarity criterion for S is traditionally determined by a Gaussian kernel based on the (Euclidean) distances, $d_{Euc}\ (v_i, v_j)$:

$$S_{Euc}(i,j) = exp\ (-\frac{d^2_{Euc}(vi,vj\ )}{2\sigma_i\sigma_j}) \qquad (5)$$

Here $\sigma_i\sigma_j$ are decaying parameters which can be set as a global optimum using empirical studies [2] or as local distances to the k^{th} nearest neighbor of v_i, v_j [30]. (Note that, for ASC, instead of using data samples, the pairwise similarities between the representatives are calculated.).

Alternatively, different information types such as topology, density were utilized for effective definition of pairwise similarities on the level of representatives [12], [13]. For example, [12] defines similarity as based on a weighted Delaunay triangulation (CONN) exploiting detailed local density together with data topology. CONN(i,j) is defined as the number of data samples inside the subregions $V_{ij} \cup V_{ji}$ of the Voronoi polygons V_i and V_j, where V_i is the set of data samples v for which w_i is the closest representative:

$$CONN(i,j) = |V_{ij} \cup V_{ji}|\ with \qquad (6)$$

$$V_{ij} = \{v \in V_i : \|v - w_j\| \leq \|v - w_k\|\ \forall k \neq i\ \}. \qquad (7)$$

Being a parameterless similarity depending on data characteristics, CONN produces more accurate partitionings than those obtained by distance based approaches [31], [12]. By integrating the distance information into CONN, a hybrid similarity criterion S_{hyb} can be obtained:

$$s_{hyb}(i,j) = s_{Euc}(i,j) \times \exp(\frac{CONN(i,j)}{max_{i,j}CONN(i,j)}) \qquad (8)$$

This enhances s_{Euc} between [1, e] depending on local density distribution. If *CONN (i, j)* = 0 then two representatives w_i, w_j do not have any data sample for which they are the best-matching pairs (i.e. w_i, w_j are not neighbors with respect to the data manifold), producing $s_{hyb}\ (i, j) = s_{Euc}(i,j)$.

Geodesic distance based similarity definitions were also considered in ASC [13]. Geodesic approaches require a neighborhood graph to determine the representatives neighbor in the data manifold. A naive way to obtain this graph is the use of (mutual) k nearest neighbors ($k - nn$) of the representatives w_i, w_j. Their geodesic distance is the length of the shortest path between w_i and w_j:

$$d_{geoknn(w_i,w_j)} = \sum_{lm \in SP_{knn}(w_i,w_j)} d_{Euc}(l,m) \qquad (9)$$

with $SP_{knn}(w_i, w_j)$ is the set of edges in the shortest path between w_i and w_j calculated with the Euclidean distance d_{Euc} and k-nn graph. This definition requires a parameter k which needs to be optimally set for each representative. Instead of k-nn, CONN can be used as a manifold based neighborhood graph exploiting local characteristics for optimal number of neighbors for each representative [32]. The geodesic distance d_{geoadj} with CONN and Euclidean distances d_{Euc} is defined:

$$d_{geoadj(w_i,w_j)} = \sum_{lm \in SP_{adj}(w_i,w_j)} d_{Euc}(l,m) \qquad (10)$$

Here $SPadj(w_i, w_j)$ is the set of edges in the shortest path based on d_{Euc} and CONN. Alternative to the Euclidean distances d_{Euc}, local density based dissimilarity d_{CONN} can be utilized for geodesic distance calculation:

$$d_{geoconn(w_i,w_j)} = \sum_{lm \in SP_{conn}(w_i,w_j)} d_{CONN}(l,m) \qquad (11)$$

$$d_{CONN}(w_i, w_j) = \begin{cases} e^{-\frac{CONN(i,j)}{max_{y,z}CONN(y,z)}} & if\ CONN(i,j)>0 \\ \infty & otherwise \end{cases}$$

$SP_{conn}(w_i, w_j)$ is now the set of edges in the shortest path between w_i and w_j with respect to d_{CONN} and CONN. In addition, a hybrid approach $d_{geohyb}(w_i, w_j)$ can harness both distance and density as:

$$d_{geohyb(w_i,w_j)} = \sum_{lm \in SP_{hyb}(w_i,w_j)} d_{Euc}(l,m) d_{CONN}(l,m) \qquad (12)$$

The geodesic distance based similarities are then obtained by replacing d_{Euc} in (5) with the corresponding distance criterion. They are successful for a wide variety of datasets with different characteristics, where the geodesic hybrid similarity s_{geohyb} is often superior [13]. However, there is no unique similarity achieving the best partitioning for any given dataset. One solution is the empirical selection of the optimal similarity criterion for each application, with respect to some lustering validity indices (or classification accuracy using a test datasetwith ground truth labels). Alternatively, we use ensemble of the partitionings obtained by different criteria to utilize different information types without any criterion selection.

III. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed SASCE using five large datasets from different remote-sensing applications[33], [34], [35]. Table I provides the properties of these datasets. One of these datasets, the Boston remote sensing dataset [33], has 41 features derived from a remotely sensed area with 216000 samples (a 360 × 600 pixel image) to capture eight classes. Three dataset (KARD: Kardjali, PLOV: Plovdiv, VARN: Varna) are multitemporal RapidEye images acquired for land cover analysis to determine lands in good agriculture condition in the frame of the Common Agricul-tural Policy of the European Union [34]. These regions have distinct land cover characteristics: KARD is a mountainous region with significant forest coverage and small agricultural parcels whereas PLOV and VARN are mainly covered with agricultural parcels. Their data have 20D features (5-band images acquired in four months from April to July 2009) with 4 million pixels for KARD and approximately 8 million pixels for PLOV and VARN. The fifth dataset, Bengisu, is a WorldView-2 image (2918 × 4775-pixel) used for land cover identification to extract hazelnut fields [35]. This 8D dataset has four classes (hazelnuts, woodlands, agricultural lands, and others). For SASCE of these five large datasets, we obtain 20 different sets of 1600 data representatives, and cluster them using seven criteria in Section II-A2. This produces 140 partitionings for each dataset, which are then merged by majority voting to obtain the final clustering.

We also cluster five datasets (Iris, Breast Cancer Wisconsin-BCWS, Yeast, Statlog and Pen Digit) from UCI Machine Learning Repository [25]. The UCI datasets (Iris, Breast Cancer Wisconsin-BCWS-, Yeast, Statlog, and Pen Digits) have a variety of characteristics such as the number of data points, the number of clusters, and the number of features (Table I). For these datasets, we set the number of data representatives as 10% of the number of data samples to have sufficient number of representatives for knowledge extraction. We have 20 sets of representatives which are clustered by the seven similarity criteria, resulting in 140 partitionings for ensemble. For performance evaluation we use two measures: clustering accuracy and adjusted Rand index (ARI) [36]. The clustering accuracy is calculated as the percentage of the correctly clustered data points based on the ground truth labels. ARI is a measure of agreement between labels obtained by the clustering process and the other labels defined by external criteria for the same data. Depending on the relation between data points of the same cluster together with the correct separation of data points into different clusters, ARI sensitively indicates the relation between each datum and its target label to provide a good measure for multi-class problems [36]. Table II shows the accuracies obtained by the proposed SASCE and the mean accuracies for ASC with the corre- sponding similarity criterion, for five large datasets. Note that different criterion achieves the best accuracy (s_{geohyb} for Boston, s_{CONN} for KARD and Bengisu, s_{geoadj} for PLOV, and $s_{geoconn}$ for VARN) while s_{geohyb} and s_{geoadj} achieve an average accuracy of upto 90%. To achieve the best performance by the proposed SASCE would be of great importance since it eliminates the need for selecting the optimum criterion. In addition, the SASCE improves the best individual clustering performance: (Boston: from 92,53% to 93,15%; KARD: from 95,14% to 95,72%; PLOV: from 91,52% to 92,62%; VARN:from 91,76% to 95,35%; and Bengisu: from 79,77% to 82,02%). A similar performance improvement can also be shown based on the ARI values at Table III. Despite the possibility of having a different similarity criterion as the optimum one with respect to ARI assessment, the ARI values favor the consensus partitioning of the SASCE as the best one.

TABLE I
THE DATASETS USED IN THE STUDY. N : NUMBER OF DATA POINTS; c: NUMBER OF CLASSES; f: NUMBER OF FEATURES

Dataset	N	c	f
Iris	150	3	4
BCWS	699	2	9
Yeast	1484	10	8
Statlog	6435	6	4
Pen	10992	10	16
Boston	216000	8	41
KARD	4000000	4	20
PLOV	8000000	4	20
VARN	8000000	4	20
Bengisu	1393345	4	8

Table IV shows the accuracies for the datasets from UCI Machine Learning Repository. Since ARI values indicate a similar evaluation for large datasets, we omit them in this assessment. The SASCE provides the best accuracies for these commonly available datasets as well (Iris: from 86,7% to 88,03%; BCWS: from 96,04% to 96,85%; Yeast: from 39, 55% to 54, 45%, Statlog: from 70, 29% to 77, 05% and Pen Digit: from 73, 05% to 80, 95%).

To indicate the advantage of SASCE in harnessing different information types, we compare the SASCE to two other ensembles based on majority voting. First, we obtain spectral clustering ensemble (SE) which uses the traditional Euclidean based similarity with different decaying parameter σ values (20 σ values: 10 uniformly distributed values between 0and 1 and 10 values between 1 and 10). This ensemble exploits advantages of spectral clustering with one similarity definition. Second, we obtain a hierarchical clustering ensemble (HACE) which merges partitionings obtained with six different linkage methods (single, complete, average, centroid, median and Ward's measure). Despite considering different criteria for within-cluster and between-clusters (dis)similarity, hierarchical clustering also depends on Euclidean distance based similarities between data points. Table V compares the ensemble accuracies of the SASCE, SE and HACE for five UCI datasets. The SASCE provides the best performance for four of them, by exploiting manifold characteristics through various types of information.

TABLE II

ACCURACIES OF SASCE FOR LARGE DATASETS. THE MEAN ACCURACY (AND STANDARD DEVIATION) FOR EACH SIMILARITY CRITERION IS ALSO GIVEN. THE BEST PERFORMANCE FOR EACH DATASET IS INDICATED IN **BOLD** WHEREAS THE BEST SIMILARITY CRITERION IS IN *italics*.

Dataset	s_{Euc}	s_{CONN}	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}	SASCE
Boston	92,10 (1,3)	68,76 (2,4)	69,25 (2,4)	86,83 (3,0)	92,52 (1,7)	85,21 (1,6)	*92,53* (2,0)	**93,15**
KARD	90,13 (2,4)	*95,14* (1,7)	94,33 (1,5)	76,27 (5,8)	92,99 (1,6)	88,54 (2,4)	93,42 (1,8)	**95,72**
PLOV	90,36 (2,1)	88,52 (1,2)	89,74 (1,9)	65,99 (1,9)	*91,52* (1,4)	88,20 (2,0)	91,02 (1,7)	**92,62**
VARN	91,24 (1,0)	91,45 (1,7)	91,70 (0,8)	83,06 (6,5)	91,73 (0,9)	*91,76* (1,1)	91,46 (1,0)	**95,35**
Bengisu	77,89 (1,07)	*79,77* (0,6)	78,63 (1,1)	55,37 (2,0)	79,65 (1,1)	76,23 (1,4)	79,65 (1,0)	**82,02**
Average	88,34	84,73	84,73	73,50	89,68	85,99	89,62	91,77

TABLE III

ARI VALUES OF SASCE FOR LARGE DATASETS.THE MEAN ARI VALUE (AND ITS STANDARD DEVIATION) FOR EACH SIMILARITY CRITERION IS ALSO GIVEN. THE BEST PERFORMANCE FOR EACH DATASET IS INDICATED IN **BOLD** WHEREAS THE BEST SIMILARITY CRITERION IS IN *italics*.

Dataset	s_{Euc}	s_{CONN}	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}	SASCE
Boston	0,90 (0,01)	0,70 (0,02)	0,70 (0,01)	0,84 (0,02)	*0,91* (0,00)	0,83 (0,00)	*0,91* (0,00)	**0,95**
KARD	0,90 (0,01)	*0,94* (0,00)	0,91 (0,01)	0,83 (0,01)	0,90 (0,02)	0,88 (0,00)	0,92 (0,01)	**0,96**
PLOV	0,88 (0,02)	0,85 (0,01)	0,89 (0,00)	0,66 (0,01)	0,88 (0,00)	0,86 (0,00)	*0,90* (0,00)	**0,93**
VARN	0,90 (0,00)	*0,91* (0,01)	*0,91* (0,03)	0,80 (0,05)	*0,91* (0,01)	*0,91* (0,01)	*0,91* (0,02)	**0,96**
Bengisu	*0,81* (0,01)	0,80 (0,01)	0,79 (0,00)	0,64 (0,00)	0,80 (0,00)	0,78 (0,02)	0,80 (0,01)	**0,85**
Average	0,88	0,84	0,84	0,75	0,88	0,85	0,89	0,93

TABLE IV

ACCURACIES OF SASCE FOR DATASETS FROM UCI MACHINE LEARNING REPOSITORY. THE MEAN ACCURACY (AND STANDARD DEVIATION) FOR EACH SIMILARITY CRITERION IS ALSO GIVEN. THE BEST PERFORMANCE FOR EACH DATASET IS INDICATED IN **BOLD** WHEREAS THE BEST SIMILARITY CRITERION IS IN *italics*.

Dataset	s_{Euc}	s_{CONN}	s_{hyb}	s_{geoknn}	s_{geoadj}	$s_{geoconn}$	s_{geohyb}	SASCE
Iris	63,47 (9,2)	68,96 (14,7)	69,01 (14,3)	56,39 (7,4)	80,94 (12,4)	*86,7* (7,8)	80,87 (12,4)	**88,03**
BCWS	*96,04* (0,6)	88,8 (11,6)	92,64 (2,0)	90,00 (6,5)	65,38 (0,4)	65,38 (0,4)	65,38 (0,4)	**96,85**
Yeast	36,60 (0,9)	*39,55* (2,8)	35,65 (3,8)	34,52 (2,1)	37,89 (3,0)	31,88 (1,7)	37,84 (2,8)	**54,45**
Statlog	*70,29* (2,1)	61,05 (17,5)	67,38 (2,6)	39,24 (9,5)	64,23 (14,3)	56,26 (9,9)	62,22 (14,5)	**77,05**
Pen Digit	69,90 (1,4)	59,91 (15,7)	*73,05* (13,7)	58,70 (6,2)	56,49 (8,9)	58,20 (9,5)	56,48 (8,6)	**80,95**

TABLE V

ACCURACIES OF DIFFERENT ENSEMBLES FOR DATASETS FROM UCI MACHINE LEARNING REPOSITORY. SASCE: THE PROPOSED SAMPLING BASED APPROXIMATE SPECTRAL CLUSTERING ENSEMBLE; SE: SPECTRAL CLUSTERING ENSEMBLE (DIFFERENT σ VALUES); HACE: HIERARCHICAL CLUSTERING ENSEMBLE (DIFFERENT LINKAGES INCLUDING AVERAGE, CENTROID, COMPLETE, SINGLE, MEDIAN)

Dataset	SASCE	SE	HACE
Iris	88,03	88,67	89,33
BCWS	96,85	96,28	94,56
Yeast	54,45	50,27	33,42
Statlog	77,05	65,77	59,04
Pen Digit	80,95	77,27	58,84

IV. CONCLUSION

Approximate spectral clustering not only makes spectral clustering feasible for large datasets but also enables alternative similarity definitions based on density distribution and data topology on the representative level. Instead of empirical selection of the optimal similarity criterion, we proposed the SASCE which is an ASC ensemble utilizing available information types based on majority voting. The SASCE not only eliminates the need for empirical selection of the best similarity criterion, but also outperforms the best individual clustering accuracy. In addition, the SASCE employs selective sampling for extraction of data representatives, instead of quantization methods that are computationally heavy. The SASCE hence provides accurate partitionings in a relatively fast manner.

ACKNOWLEDGMENT

This study was funded by TUBITAK Career Grant No. 112E195. Tas¸demir is also supported by FP7 Marie Curie Career Integration Grant (IAM4MARS).

REFERENCES

[1] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[2] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: analysis and an algorithm,” in *T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in NIPS 14. MIT Press*, 2002.

[3] M. Meila and J. Shi, “A random walks view of spectral segmentation,” in *8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2001.

[4] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[5] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, “A survey of kernel and spectral methods for clustering,” *Pattern recognition*, vol. 41, no. 1, pp. 176–190, 2008.

[6] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, “Spectral grouping using the nystrom method,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 214–225, 2004.

[7] J. C. Bezdek, R. J. Hathaway, J. M. Huband, C. Leckie, and K. Ramamohanarao, “Approximate clustering in very large relational data,” *Int’l Journal of Intelligent Systems*, vol. 21, no. 8, pp. 817–841, 2006.

[8] L. Wang, J. C. Bezdek, C. Leckie, and R. Kotagiri, “Selective sampling for approximate clustering of very large data sets,” *International Journal of Intelligent Systems*, vol. 23, no. 3, pp. 313–331, 2008.

[9] L. Wang, C. Leckie, K. Ramamohanarao, and J. Bezdek, “Approximate spectral clustering,” in *Advances in Knowledge Discovery and Data Mining*. Springer, 2009, pp. 134–146.

[10] D. Yan, L. Huang, and M. I. Jordan, “Fast approximate spectral clustering,” in *Proceedings of the 15th ACM SIGKDD Int’l Conference on Knowledge Discovery and Data Mining*, 2009, pp. 907–916.

[11] L. Wang, C. Leckie, R. Kotagiri, and J. Bezdek, “Approximate pairwise clustering for large data sets via sampling plus extension,” *Pattern Recognition*, vol. 44, no. 2, pp. 222–235, 2011.

[12] K. Tas¸demir, “Vector quantization based approximate spectral clustering of large datasets,” *Pattern Recognition*, vol. 45, no. 8, pp. 3034–3044, 2012.

[13] K. Tas¸demir, B. Yalcin, and I. Yildirim, “Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures,” *Pattern Recognition*, vol. 48, no. 4, pp. 1459–1471, 2015.

[14] T. Xiang and S. Gong, “Spectral clustering with eigenvector selection,” *Pattern Recognition*, vol. 41, no. 3, pp. 1012–1029, 2008.

[15] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, “Parallel spectral clustering in distributed systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 568–586, 2011.

[16] A. Strehl and J. Ghosh, “Cluster ensembles - a knowledge reuse framework for combining multiple partitions,” *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 583–617, March 2002.

[17] S. Dudoit and J. Fridlyand, “Bagging to improve the accuracy of a clustering procedure,” *Bioinformatics*, vol. 19, no. 9, pp. 1090–1099, 2003.

[18] A. Topchy, A. K. Jain, and W. Punch, “A mixture model for clustering ensembles,” in *SIAM Int’l Conf. Data Mining*, 2004, pp. 379–390.

[19] A. L. N. Fred and A. K. Jain, “Combining multiple clusterings using evidence accumulation,” *IEEE Trans. Pattern Analysis Machine Intelligence (PAMI)*, vol. 27, no. 6, pp. 835–850, 2005.

gence (PAMI), vol. 27, no. 6, pp. 835–850, 2005.

[20] X. Zhang, L. Jiao, F. Liu, L. Bo, and M. Gong, “Spectral clustering ensemble applied to SAR image segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2126–2136, July 2008.

[21] F. Tung, A. Wong, and D. A. Clausi, “Enabling scalable spectral clustering for image segmentation,” *Pattern Recognition*, vol. 43, no. 12, pp. 4069 – 4076, 2010.

[22] J. Jia, X. Xiao, and B. Liu, “Similarity-based spectral clustering ensemble selection,” in *Fuzzy Systems and Knowledge Discovery (FSKD), 9th Int’l Conference on*, May 2012, pp. 1071–1074.

[23] X. Peng, L. Zhang, and Z. Yi, “Scalable sparse subspace clustering,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 430–437.

[24] K. Tasdemir, Y. Moazzen, and I. Yildirim, “An approximate spectral clustering ensemble for high spatial resolution remote-sensing images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 5, pp. 1996–2004, 2015.

[25] A. Asuncion and D. Newman, “UCI machine learning repository.” [Online].Available: http://www.ics.uci.edu/_mlearn/MLRepository.html

[26] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888–905, 2000.

[27] M. Meila and J. Shi, “A random walks view of spectral segmentation,” in *8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2001.

[28] U. von Luxburg, “A tutorial on spectral clustering,” *J. Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[29] D. Verma and M. Meila, “A comparison of spectral clustering algorithms,” Unv of Washington, Tech. Rep. UW TR CSE-03-05-01, 2003.

[30] L. Zelnik-Manor and P. Perona, “Self-tuning spectral clustering,” in *Advances in Neural Information Processing Systems*, 2004.

[31] K. Taşdemir and E. Merényi, “Exploiting data topology in visualization and clustering of self-organizing maps,” *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 549–562, 2009.

[32] K. Tas¸demir, Y. Moazzen, and I. Yildirim, “Geodesic based similarities for approximate spectral clustering,” in *22nd Int’l Conference on Pattern Recognition, Stockholm, Sweden, 24-28 August*, 2014.

[33] G. A. Carpenter, S. Martens, and O. J. Ogas, “Self-organizing information fusion and hierarchical knowledge discovery: a new framework using ARTMAP neural networks,” *Neural Networks*, vol. 18, no. 3, pp. 287–295, 2005.

[34] K. Tas¸demir, P. Milenov, and B. Tapsall, “A hybrid method combining som-based clustering and object-based analysis for identifying land in good agricultural condition,” *Computers and Electronics in Agriculture*, vol. 83, pp. 92 – 101, 2012.

[35] S. Reis and K. Tas¸demir, “Identification of hazelnut fields using spectral and gabor textural features,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 5, pp. 652–661, September 2011.

[36] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, pp. 193–218, 1985.