2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Active Constrained Clustering via Non-Iterative Uncertainty Sampling

Panagiotis Stanitsas
University of Minnesota
stani078@umn.edu

Anoop Cherian
Australian National University
anoop.cherian@anu.edu.au

Vassilios Morellas
University of Minnesota
morellas@cs.umn.edu

Nikolaos Papanikolopoulos
University of Minnesota
npapas@cs.umn.edu

***Abstract*— Active Constraint Learning (ACL) is continuously gaining popularity in the area of constrained clustering due to its ability to achieve performance gains via incorporating minimal feedback from a human annotator for selected instances. For constrained clustering algorithms, such instances are integrated in the form of Must-Link (ML) and Cannot-Link (CL) constraints. Existing iterative uncertainty reduction schemes, introduce high computational burden particularly when they process larger datasets that are usually present in computer vision and visual learning applications. For scenarios that multiple agents (i.e., robots) require user feedback for performing recognition tasks, minimizing the interaction between the user and the agents, without compromising performance, is an essential task.**

In this study, a non-iterative ACL scheme with proven performance benefits is presented. We select to demonstrate the effectiveness of our methodology by building on the well known *K-Means* algorithm for clustering; one can easily extend it to alternative clustering schemes. The proposed methodology introduces the use of the Silhouette values, conventionally used for measuring clustering performance, in order to rank the degree of information content of the various samples. In addition, an efficient greedy selection scheme was devised for selecting the most informative samples for human annotation. To the best of our knowledge, this is the first active constrained clustering methodology with the ability to process computer vision datasets that this study targets. Performance results are shown on various computer vision benchmarks and support the merits of adopting the proposed scheme.

***Index Terms*— Visual Learning, Active Constrained Clustering, Image Clustering Uncertainty Management**

I. Introduction

In the contemporary times of big data analytics, generating quality ground truth data annotations for the purpose of training various machine learning algorithms has become a major challenge. As human annotators are scarce and expensive, targeting manual resources to subsets of datasets that are the most informative, is essential. Such active selection of subsets has by now become a major sub-field in machine learning. The objective of active schemes is to enable the creation of correct and coherent clusters that capture the underlying structure of the data. Selection and ordering of samples for processing is critical to the overall performance of the algorithms since a poorly selected set of samples can have adverse effects ([5], [7] and [17]).

While iterative cluster refinement is commonly utilized for active feedback in large datasets, batch mode active selection is known to be useful in several situations. Such examples are concerned with multi-agent tasks for which the number of times that communication is established between an agent and a human annotator needs to be minimized. To this end, we formulate a novel objective that balances between the uncertainty in cluster memberships and the spatial diversity of data points in the batch selection; this goes beyond a simple ranking of samples based on uncertainty. An efficient greedy selection approach was also introduced towards optimizing the derived objective.

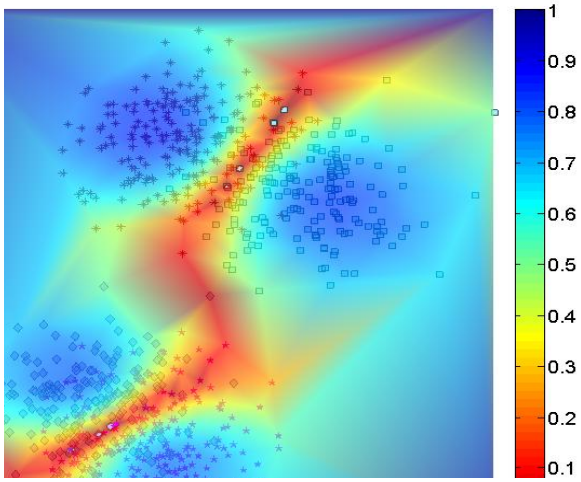

Fig. 1: Silhouette values distribution on a 4-cluster Gaussian synthetic dataset. Red color indicates points of high uncertainty and therefore high informativeness.

In contrast to the standard iterative active learning schemes for ACL, this study introduces a *non-iterative* active sampling methodology, which exploits a notion of information content existing, but not yet discovered, in the data. The non-iterative nature of the proposed methodology lends itself to the processing of datasets which was not computationally feasible with iterative ACL schemes. In addition, ensemble ACL techniques face equivalent computational burdens with their iterative counterparts. The aforementioned computational load of iterative and ensemble ACL schemes is so substantial that as of now, to the best of our knowledge, performance baselines cannot be found at the level of data complexity (dimensionality and number of samples) this effort targets.

Active selection of informative ML and CL constraints was enabled by the integration of the Silhouette values [13], conventionally used for clustering performance evaluation, and by building on the notion of informativeness of pairwise constraints [5]. An uncertainty sampling scheme is proposed that incorporates in an initial step the results of an unrefined,

yet not committing, clustering task that assigns to each sample a value in the range [-1,1]; larger values indicate a confident (i.e., less informative) assignment. Figure 1 illustrates how the deployment of the silhouette values characterizes the informativeness of various samples on a 4-cluster Gaussian synthetic dataset. Points at the overlapping regions of the clusters receive a low silhouette value (i.e., high informativeness). It is therefore apparent that high informative value of samples reflects the degree of confusion that clustering algorithms face in correctly clustering the respective data points. The strength of the Silhouette value in an ACL scheme will be discussed at Section III-A.

The selection of constraints that utilize the Silhouette values, leads to the creation of a constraint matrix via a 4-step process. In the first step, an unrefined clustering is performed prior to the assignment (second step) of Silhouette values; the *K-Means* algorithm was considered. Subsequently, a greedy selection scheme decides about the samples to be included in the final constraint matrix based on the computed Silhouette values as presented in Section IV-B. In the last step, the *COP-KMeans* algorithm [16] uses the constraint matrix, so produced in the previous steps, resulting in a more refined, partitioning of the data. The two selected clustering algorithms (*K*-Means and *C*OP-KMeans) could be easily substituted by other clustering tools of choice, nevertheless, we select to present our results based on those two. Experimental results presented below, support the merit of the devised approach and prove the utility of the proposed ACL scheme.

II. Related Work

The literature in the domain of active learning is undergoing a continuous growth, following the increasing interest in the computational benefits of active methods. For a thorough review of active learning methods the reader can consider [12] and [14]. Even though the broader active learning literature experiences a rapid growth, the domain of ACL for constrained clustering algorithms is following at a slower pace. As Davidson et al. [5] has shown, constraints should be cautiously integrated in constrained clustering algorithms. Basu et al. [2] paved the way for active constraint selection by deriving an active version of the constrained K-Means [3]. In particular, in [2] an active seeding approach was devised, utilizing samples at maximal distances between them (furthest-first query selection) built on [3]. The work of Bilenko et al. [3], has served as the evaluation workhorse of numerous studies along the lines of active constraint selection [2], [7], [8], [10], [20].

Xiong et al. [20] as well as Huang and Lam [8] utilized active selection techniques, embedded in iterative uncertainty reduction schemes that capitalized on the notion of neighborhoods in the feature space. Mallapragada et al. [10] derived the min-max query selection criterion which was suggested as an alternative to the furthest-first query selection of Basu et al. [2]. Greene and Cunningham [7] proposed a non-iterative ACL scheme for which, the basis, was an ensemble of base clusterings. Thresholds, placed on a co-association matrix whose entries denote the fraction of clusterings in the ensemble for which two points are grouped together, revealed the constraints of the highest value. Biswas and Jacobs [4] developed an elegant methodology that capitalized on Minimum Spanning Forest clustering towards grouping images with their devised scheme termed Active Hierarchical Agglomerative Clustering with Constraints (Active HACC). The uncertainty measure adopted in [4] was the magnitude of change in the Jaccard Coefficient (commonly used for measuring clustering performance) resulting from constraint integrations. The derived $\mathcal{O}(n^4 lg(n))$ complexity per constraint selection for an n-sample problem makes its deployment prohibitive when processing 1K samples or more. Even in an accelerated version of Active-HACC, the complexity still remained high ($\mathcal{O}(n^3 lg(n))$ per constraint selection) while the sophisticated implementation makes the use of this scheme difficult.

In addition to the previous techniques, the area of spectral clustering has also been altered in order to accommodate feedback from a human annotator. Xu et al. [21] made the first attempt towards Active Spectral Clustering (ASC) with the ACCESS (Active Constrained Clustering by Examining Spectral eigenvectorS) algorithm. The eigenvectors of a similarity matrix are used in order to identify points at the boundaries of clusters and user feedback is used to either strengthen or weaken similarities of such points. Shamir et al. [15] as well as Wauthier et al. [19] have completed work in the domain of ASC utilizing only partial knowledge of the similarity matrix. Wang and Davidson [18] iteratively identified constraints that resulted in maximal error reduction, using spectral information.

The above schemes use an iterative strategy in general, i.e., each iteration querying the user for a data class membership, later repeating the constrained clustering step. Unfortunately, such schemes might not work for large datasets; for which batch mode active selection might be the appropriate. Thus, we further propose a non-iterative Silhouete metric based active sampling scheme in which the batch of data points to be queried are selected by optimizing a cost function that balances the uncertainty of points regarding their cluster memberships and the diversity of their spatial arrangements. The latter property helps avoid points that are uncertain and are also in the proximity of other uncertain points, whose labels can be obtained via label propagation. However, our objective is combinatorial and thus difficult to maximize. We propose an approximate but efficient greedy algorithm to solve this objective.

III. Uncertainty Measure for Active Selection

ML and CL constraints are embedded onto the proposed scheme in the form of a matrix. Constraint matrices are symmetric, square matrices of size equal to the number of samples and they capture pairwise relationships between selected points. For example, for l selected points $\binom{l}{2}$, pairwise constraints are derived based on the agreement or disagreement provided by the user labels. While random techniques tend to not take into account the difficulty in

clustering particular points, the proposed ACL scheme attempts to principally rank and subsequently select informative samples (i.e., high uncertainty or low Silhouette values), substantially aiding in the clustering process. Davidson et al. [5] and Wagstaff et al. [17] denoted the importance of using appropriate constraint matrices in the context of constrained clustering. In their work, they explicitly identified the importance in and underlined the lack of a procedure to construct constraint matrices that enhance the process of maximizing the information content embedded in the constraint matrix.

In that context, pairwise constraints for points that lie simultaneously near the boundaries of multiple clusters, thus exhibiting a higher degree of difficulty during clustering (i.e., high Silhouette value), should be stronger candidates in the selection process. In particular, CL constraints for such points can steer the algorithm to finding optimal solutions and subsequently providing a clustering assignment of higher accuracy.

A. Uncertainty Sampling Metric

Constructing a constraint matrix of high informativeness was made possible, in this work, by the embedding of the Silhouette values [13] which from now on will be referred to as Uncertainty Sampling Measure (USM). The concept of Silhouettes (not to be confused with the silhouettes of objects) was proposed by Rousseew [13] as a performance measure in cluster analysis. However, in this context, we elevate the concept of Silhouettes to that of a USM. The USM computation is based on a measure of closeness between data points, as well as on the assignment of points to clusters and is defined as follows: suppose after an initial clustering, let the data point x_i be assigned to cluster π_z, then

$$S_i = \frac{\beta_i - \alpha_i}{\max(\alpha_i,\ \beta_i)} \tag{1}$$

where, $\alpha_i = \frac{1}{|\pi_z|} \sum_{j \in \pi_z} || x_i - x_j ||^2$ is the average distance of point i with the rest of the points that belong in cluster π_z and $\beta_i = \min_{z' \neq z} \frac{1}{|\pi_{z'}|} \sum_{l \in \pi_{z'}} || x_i - x_l ||^2$ is the minimum average distance between point i and all clusters that i is not a member; closeness of data points is based on Euclidean distance, although other metrics can be used. USM values are associated with all points in the dataset. As is clear from Equation (1), S_i ranges in $[-1, 1]$. A value of 1 characterizes a sample of low clustering uncertainty, while a value of -1 corresponds to a sample of high uncertainty and therefore high informativeness.

The USM can at a high level, be viewed as a best versus second best measure as underlined by Equation (1). One example of the proposed strategy on a synthetic, 4-cluster Gaussian dataset is presented in Figure 2. As it is shown, the proposed ACL scheme automatically selects the points (denoted by the red triangles) that contain the highest informative value, which are located at the regions where the clusters overlap. Furthermore, points that are not proximal to cluster centers are also considered as uncertain and thus selected by the proposed scheme.

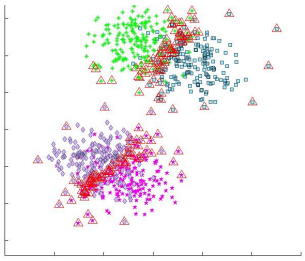

Fig. 2: Active query sampling for the 4-cluster Gaussian dataset using USM values. Points in red denote samples that were selected by the ACL scheme.

IV. Active Constraint Learning

The proposed ACL scheme capitalizes on the aforementioned notion of *informativeness* (Equation (1)) and constructs a sampling scheme around it. Below, we devise a four-step procedure for refining the clustering outcome:

1) Compute an unrefined clustering of the dataset.
2) Compute the resulting USM values.
3) Select the most informative samples to query the user using our greedy selection scheme.
4) Execute *COP-Kmeans* to deliver the final clustering.

Each one of the steps will be discussed at more length next.

A. Unrefined Clustering

The deployment of an unrefined clustering is the first step of the proposed ACL scheme. Although, the choice of such an algorithm is not committing, experimentation with the *K-Means* algorithm and its kernelized version, *Kernel K-Means* was performed. The rationale behind the choice of working in a kernel space is driven by the inherent limitation of Euclidean metric to faithfully encode the closeness of points that generally lie on non-linear manifolds.

Once the unrefined clustering step is performed in a kernel space, USM values are obtained for all the transformed samples in the kernel space. It is interesting to notice the new form that coefficients α_i and β_i of Equation (1) take in the kernel space as presented in Equations (2) and (3), where $K(x_i, x_j) = \langle \phi(x_i),\ \phi(x_i) \rangle$. For this study 4 kernels were considered for purposes of demonstration as presented in Table I; linear, polynomial, sigmoid and Radial Basis Function. One should note that no extensive efforts were expended to find the optimal performance kernel.

$$\alpha_i = \frac{1}{|\ \pi_z\ |} \sum_{j \in \pi_z} \Big(K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j) \Big) \tag{2}$$

$$\beta_i = \min_{z' \neq z} \frac{1}{|\pi_{z'}|} \sum_{l \in \pi_{z'}} \Big(K(x_i, x_i) + K(x_l, x_l) - 2K(x_i, x_l) \Big) \tag{3}$$

TABLE I: Tested kernel functions.

Linear Kernel	$K_L(x_i, x_j) = x_i^T x_j + c$
Polynomial Kernel	$K_P(x_i, x_j) = (\alpha x_i^T x_j + c)^d$
Sigmoid Kernel	$K_S(x_i, x_j) = tanh(\alpha x_i^T x_j + c)$
RBF Kernel	$K_R(x_i, x_j) = exp(-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2})$

B. Greedy Query selection

In this section the process of optimally selecting the most informative samples is presented. An optimization problem, as formulated in Equation (4), attempts to find those samples that balance informativeness and spatial dispersion along the boundaries between clusters. This is reflected in the two terms of Equation (4). The first term accounts for the informativeness of the sample while the second disperses the queried samples along the boundaries. The combinatorial nature of the formulation, yields an NP hard problem whose solution is obtained via a greedy approximation. The greedy component of the scheme sequentially augments the list of selected samples such that maximum gain can be achieved at every step. The connection between the current formulation and optimization over submodular functions is apparent and under ongoing investigation.

$$\begin{aligned} \underset{Q \subset \{1,\ldots,n\}}{\text{maximize}} \quad & \sum_{i \in Q} \tilde{S}_i \; + \sum_{(i \neq j) \in Q} \tilde{d}(x_i, x_j) \\ \textbf{subject to} \quad & |Q| = N_Q \end{aligned} \tag{4}$$

where Q is the near optimal set of queries that the algorithm seeks, N_Q is the number of queries the algorithm is allowed to perform, $\tilde{S}_i = \frac{1 - S_i}{2 \sum_{l=1}^{n} S_i}$ and $\tilde{d}(x_i, x_j)$ is the squared Euclidean distance normalized with respect to the sum of distances for all active selection candidates per round as presented in line 8 of Algorithm 1. The reason for normalizing distances is so that uncertainty and distance are equally considered when deciding about the next point to augment the set of actively selected points, restricting the second term to rapidly grow and musk the impact of USM in the selection process.

Algorithm 1 provides the pseudocode for the developed greedy scheme. $\lfloor \frac{N_Q}{k} \rfloor$ points are selected for cluster i by sequentially augmenting the list Q_i with candidates that achieve maximal gains at every step. First, USM values are transformed from the original range of $[-1, 1]$ to the new range of $[0, 1]$ thus becoming appropriate for this maximization formulation as shown in line 5 of the pseudocode. However, this transformation results in the assignment of greater USM values to high informative samples in contrast to the initial range interpretation. Line 6 is responsible for initializing the query list Q_i of cluster i, with its first two samples which are derived such that they contain the largest amount of information while being far apart. Finally, the while loop in lines 7-11 recursively augments Q_i by the sample in cluster i that achieves the largest incremental gain for Equation (4) when added to the set.

Algorithm 1: Greedy Query Selection

Data: data set $X_{n \times d}$, labels $L_{1 \times n}$, number of queries N_Q, number of clusters k, USM values $S_{1 \times n}$.
Result: Near optimal set of queries Q.

1 Let $(Q_1, Q_2, \ldots, Q_k)$ be the collection of queries for every cluster;
2 **for** $i \leftarrow 1$ **to** k **do**
3 $\quad S^{(i)}$ = set of USM values of points in cluster i;
4 $\quad X^{(i)}$ = set of samples in cluster i;
5 $\quad \tilde{S}^{(i)} = \frac{1 - S^{(i)}}{2 \sum S^{(i)}}$ (Transformed USM values);
6 $\quad Q_i = \{Q_i \cup u \cup v\} \;\mid\; \max_{(u,v)} (\tilde{S}_u^{(i)} \tilde{S}_v^{(i)} \;||\; x_u - x_v \;||^2)$
$\quad\quad p = 2$ (Counter of selected samples for cluster i);
7 $\quad$ **while** $p < \lfloor \frac{N_Q}{k} \rfloor$ **do**
8 $\quad\quad \max_{c \notin Q_i} \left(\tilde{S}_c^{(i)} + \frac{\sum_r ||x_c - x_r||^2}{\sum_c \sum_r ||x_c - x_r||^2} \right)$;
9 $\quad\quad$ where $r \in Q_i$;
10 $\quad\quad Q_i = \{Q_i \cup c\}$;
11 $\quad\quad p++$
12 **return** $(Q_1, Q_2, \ldots, Q_k)$

Operations at lines 3 and 4 of the pseudocode have complexity of $\mathcal{O}(n)$, while line 6 results in $\mathcal{O}(n)$ for a precomputed pairwise distance matrix. Finally, the block of lines 7-11 has complexity of $\mathcal{O}(n^2 \frac{N_Q}{k})$. The overall complexity of the proposed greedy selection structure is $\mathcal{O}(n^2 N_Q)$. This is orders of magnitude lower than available alternatives as also presented in the following section.

V. Experiments

The performance of the selected uncertainty sampling scheme was demonstrated through a series of experiments on synthetic as well as benchmark computer vision datasets. As mentioned in earlier sections to the best of our knowledge, alternative active constrained clustering schemes for making comparisons at this scale are not available. Figure 3 supports this claim by establishing a complexity comparison with [4]. The necessary operations are orders of magnitude larger when compared to the proposed selection scheme. The blue bars correspond to our selection scheme, while cyan and yellow bars correspond to Fast-Active-HACC and Active-HACC respectively. It should also be noted that the y-axes in Figure 3 appears in logarithmic scale. Even though [4] requests pairwise constraints rather than ordinal feedback this comparison focusing on the selection scheme can still provided useful conclusions on the efficiency of our scheme. The main focus of this section is to present the performance gains when the proposed ACL scheme is adopted against random constraint selection.

The experiments performed, used a variable number of allowable queries. Thirty (30) iterations were executed for each of the available cases and the respective results were averaged to extract the mean performance. The performance of the ACL scheme was tested in both feature as well as a

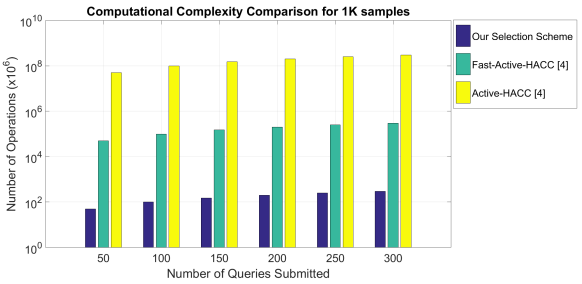

Fig. 3: Comparing the required operations for providing active feedback on a 1K sample dataset.

kernel space. Four different kernels (Table I) were tested on each dataset and the results for the one that performed best are presented in the following sections. The effort on tuning the parameters of the selected kernels was minor since this deviates from the scope of this study.

Although the proposed ACL scheme promotes a more consistent and repeated selection of samples, deviation from this behavior is attributed to the specific clustering procedure deployed during the first step of the process; this also justifies the 30 iterations performed to obtain a mean performance. All comparisons were evaluated using the Adjusted Rand index (AR index) which measures similarities between different clusterings with higher values corresponding to better performance results.

A. Results on the 4-cluster Gaussian dataset

The first dataset processed in this work is illustrated in Figure 2 and consists of 4 Gaussian clusters, each containing 200 points and exhibiting two overlapping regions. The results obtained on this dataset are presented in Figure 4 for 7 different cases of queries allowed. This corresponds to the following percentages of labels provided to the algorithm: (5%, 10%, 15%, 20%, 25%, 30% and 35%). For the case that 15% of the labels were provided to the algorithm, the proposed ACL scheme in the feature space achieved performance gains of 8% when compared with random selection. In addition, by applying a polynomial kernel transformation to the dataset, performance gains that reached 6% were also achieved, when compared with the random selection strategy.

An important observation on the results of this experiment is that the ACL scheme achieves an equivalent performance with the random selection scheme for a fraction of the supervision. One such example is the case when 35% of the labels were used via a random selection scheme yielding an AR index of 83.0%, when only 20% of the labels achieved an equivalent performance via the ACL scheme. The ACL scheme for the transformed space was not able to outperform the ACL scheme when directly applied to the feature space. A more appropriate kernel transformation could have resulted in a more refined clustering outcome, nevertheless identifying optimal transformations goes beyond the scope of this study.

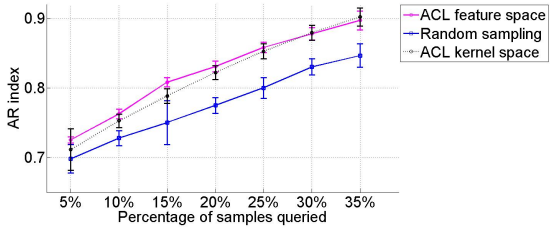

Fig. 4: Performance results on the 4-cluster Gaussian synthetic dataset. For ACL in kernel space a polynomial kernel was utilized.

B. Results on the MIT Scene dataset

The MIT outdoor scene dataset (`http://people.csail.mit.edu/torralba/code/spatialenvelope`) is a collection of 2688, RGB images of size 256x256 that are categorized in the following 8 categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways. Figure 5 illustrates one sample from each of the aforementioned categories. In the process of clustering the MIT dataset, GIST descriptors [11] were extracted which resulted in 512 dimensional vector representations for every sample. Due to the high variability that each clusters exhibits, the MIT scene dataset still remains a challenge for machine learning algorithms.

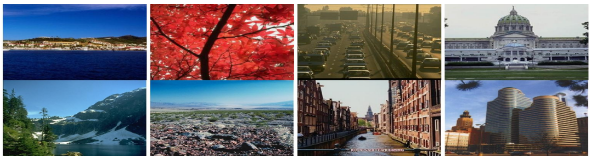

Fig. 5: Samples from the MIT dataset.

The results on the MIT scene dataset demonstrate the strength of the proposed methodology as presented in Figure 6. Performing the unrefined clustering step in a kernel space transformed by a sigmoid kernel, provided the highest AR index values for the whole spectrum of the available queries (6%, 12%, 18%, 24%, 30%, 36%). It can be seen that for the ACL scheme in the kernel space, it achieved 12% increase in the AR index for the case that 24% of the labels were provided to the algorithm. Another observation that supports the practicality of the ACL scheme can be visualized by the smaller number of labels that the proposed scheme requires when compared to random selection. In particular for the MIT dataset, the ACL strategy in the kernel space was able to achieve an AR index of 61.5% for the case that 18% of the labels were utilized. In contrast, the random selection strategy required 40% more labels in order to achieve an equivalent performance.

C. Results on the USPS dataset

The USPS handwritten digit dataset is one of the most popular benchmark datasets in the field of computer vision since its creation; in most cases it is divided in a training and

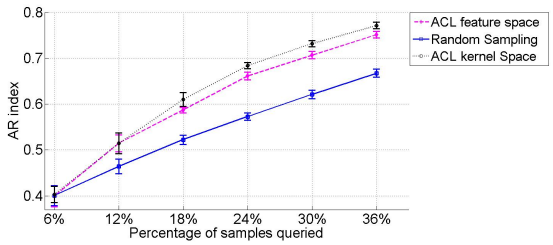

Fig. 6: Performance results on the MIT scene dataset. For ACL in kernel space a sigmoid kernel was utilized.

a test set. One sample from each category of the USPS is presented in Figure 7. For this study the USPS dataset was obtained along with the software package accompanying the work of Li et al. [9] in semi-supervised clustering (`http://www.ee.columbia.edu/~zgli/`). It contains a selection of 9,298 gray-scale images of size 16x16. The feature vector for each sample is a 256-dimensional representation of the image formed by concatenating the rows of the image. Even though several classification approaches have obtained almost perfect results on the USPS dataset, it still remains a demanding task for clustering algorithms.

Fig. 7: Samples from the USPS dataset.

For the USPS dataset 5 cases of allowed queries were examined (6%, 11%, 15%, 19%, 23%) as presented in Figure 8. The ACL scheme in the feature space was able to provide performance gains that reached 6.2% when compared with the random selection scheme for the case that 19% of the labels were provided to the algorithm. For the case that 6% of the labels were utilized, the random selection scheme performed better than the ACL scheme. For most query percentages the proposed scheme required 4% less labels than the random selection scheme in order to achieve the same or higher performance which in this dataset translates in 370 labels. Finally, for the case that a sigmoid kernel transformation was applied to the dataset, the performance curve was not superior to the random selection scheme for 2 out of the 5 number of queries.

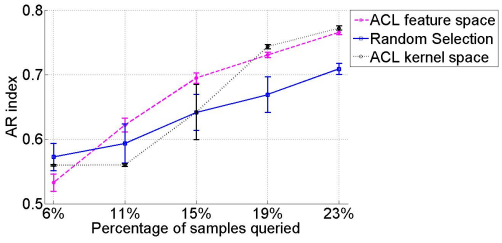

Fig. 8: Performance results on the USPS dataset. For ACL in kernel space a sigmoid kernel was utilized.

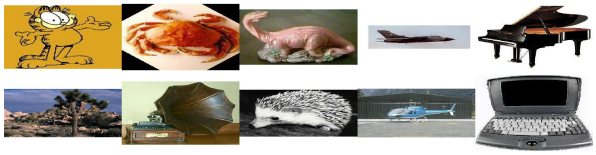

Fig. 9: Samples from 10 categories of the Caltech 101 dataset.

D. Results on the Caltech 101 dataset

The Caltech 101 dataset is a collection of 9,144 images of 102 categories with RGB images of size, approximately, 200x300 pixels [6]. Ten (10) random samples from the Caltech 101 are presented in Figure 9. For this particular dataset dimensionality reduction techniques were applied in order to process the data. GIST descriptors [11] were extracted and via a PCA analysis, the principal components that accounted for 85% of the variance were identified. This reduced the dimensions of the dataset to 60 from the original 512. Five (5) cases of allowed queries were used for the experiments on this dataset (5.5%, 11%, 16.5%, 22%, 27.5%). Even though the obtained AR index on this dataset, did not exceed 30%, this performance cannot be directly compared against other methodologies that used this dataset. For example the clustering performance presented in [1] on Caltech 101, operating on a reduced number of samples (1,959 instead of 9,144) and clusters (50 clusters rather than 102) reached an approximate 55% performance. Even though the performance achieved in this study remains low, it can still be used to draw conclusions regarding the behavior of the proposed ACL scheme against random selection for the complete set of data.

Specifically, as illustrated in Figure 10, the ACL scheme in the feature space outperformed both the ACL method in a transformed by a sigmoid kernel space as well as the random selection scheme. Although the magnitude of the performance gains on this dataset does not match the gains on the previously discussed datasets, it was still able to reach an increase of 2.2% for the case when 27.5% of the labels were provided to the algorithm. One exception lies in the case that algorithm received the minimum percentage of labels (5.5%). This was the only case that the random selection performed better than the two ACL schemes for the same reasons that were discussed on the previous datasets (i.e., the selection of high informative samples does not span the complete spatial support of the clusters in the data representation space).

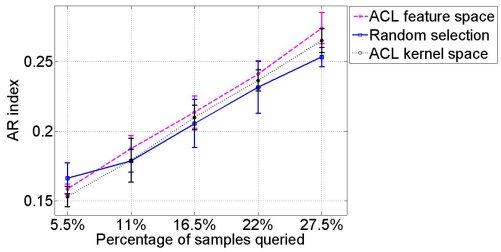

Fig. 10: Performance results on the Caltech 101 dataset. For ACL in kernel space a sigmoid kernel was utilized.

VI. Discussion and Future Work

This study introduced an ACL scheme for the analysis of structurally complex datasets in constrained clustering setups. The proposed scheme significantly differs from other active methods since it can operate on datasets at the scale (dimensionality and number of samples) of many popular computer vision benchmarks. This ability is not present in other efforts along the lines of active constrained clustering due to their, in most cases, iterative manner of query selection that involves recursive (or ensembles of) clusterings. A four step process was devised towards actively selecting the most informative samples for querying the user. First, an unrefined clustering step derives a preliminary assignment of the data in clusters that is then used to compute the proposed USM values. This step can be executed both in the original feature space or a kernel space using the kernel version of a selected clustering algorithm as well as the derived version of the USM in the kernel space. A greedy selection algorithm identifies samples that balance informativeness (USM values) and spatial dispersion (closeness metrics) along the boundaries between clusters which are later used to query the user. The acquired labels for the selected samples are used to construct a constraint matrix that summarizes the available pairwise ML and CL relationships. A constrained clustering scheme takes advantage of the derived ML and CL constraints to provide a refined clustering for the data.

Experiments were performed on 4 different datasets; one synthetic datasets of low dimensionality as well as 3 more complex computer vision benchmarks. The obtained results, which reached 12% of performance gains when adopting the proposed ACL scheme, provide guarantees in adopting this methodology. Future directions include the establishment of a more computationally effective method for solving the greedy selection problem in the context of submodular function optimization. Another future direction that is currently under consideration, is to enhance the proposed structure with additional directional information, as defined by the fitting of local tangential hyperplanes. Finally, from an application perspective the intention is to transition this methodology in real world automation tasks. In such cases even a small reduction in the number of necessary samples for a targeted performance can transition the deployment of such a system from practically unacceptable to feasible.

In a subsequent step the proposed scheme will be tested in a real world scenario where multiple agents will utilize the annotation effort of a human towards performing object recognition tasks closing the loop between the theoretical work that has been developed and a real world scenario.

VII. Acknowledgements

This material is based upon work supported by the National Science Foundation through grants #IIP-0934327, #CNS-1039741, #SMA-1028076, #CNS-1338042, #IIS-1427014, #IIP-1439728, #IIP-1432957, #OIA-1551059 and #CNS-1514626.

References

[1] S. Anand, S. Mittal, O. Tuzel, and P. Meer. Semi-supervised kernel mean shift clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1201–1215, June 2014.

[2] S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *International Conference on Data Mining*, volume 4, pages 333–344. SIAM, 2004.

[3] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the International Conference on Machine Learning*, page 11. ACM, 2004.

[4] A. Biswas and D. Jacobs. Active image clustering: Seeking constraints from humans to complement algorithms. In *International Conference on Computer Vision and Pattern Recognition*, pages 2152–2159. IEEE, 2012.

[5] I. Davidson, K. L. Wagstaff, and S. Basu. *Measuring constraint-set utility for partitional clustering algorithms*. Springer, 2006.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[7] D. Greene and P. Cunningham. Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering. In *European Conference on Machine Learning*, pages 140–151. Springer, 2007.

[8] R. Huang and W. Lam. Semi-supervised document clustering via active learning with pairwise constraints. In *International Conference on Data Mining, 2007*, pages 517–522. IEEE, 2007.

[9] Z. Li, J. Liu, and X. Tang. Constrained clustering via spectral regularization. In *International Conference on Computer Vision and Pattern Recognition*, pages 421–428. IEEE, 2009.

[10] P. K. Mallapragada, R. Jin, and A. K. Jain. Active query selection for semi-supervised clustering. In *International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[11] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[12] F. Olsson. *A literature survey of active machine learning in the context of natural language processing*. Swedish Institute of Computer Science, 2009.

[13] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[14] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66, 2010.

[15] O. Shamir and N. Tishby. Spectral clustering on a budget. In *International Conference on Artificial Intelligence and Statistics*, pages 661–669, 2011.

[16] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al. Constrained k-means clustering with background knowledge. In *International Conference of Machine Learning*, volume 1, pages 577–584, 2001.

[17] K. L. Wagstaff, S. Basu, and I. Davidson. When is constrained clustering beneficial, and why? *Ionosphere*, 58(60.1), 2006.

[18] X. Wang and I. Davidson. Active spectral clustering. In *International Conference on Data Mining*, pages 561–568. IEEE, 2010.

[19] F. L. Wauthier, N. Jojic, and M. I. Jordan. Active spectral clustering via iterative uncertainty reduction. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 1339–1347. ACM, 2012.

[20] S. Xiong, J. Azimi, and X. Z. Fern. Active learning of constraints for semi-supervised clustering. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):43–54, 2014.

[21] Q. Xu, K. L. Wagstaff, et al. Active constrained clustering by examining spectral eigenvectors. In *Discovery Science*, pages 294–307. Springer, 2005.