Spectral Active Clustering via Purification of the k-Nearest Neighbor Graph

Caiming Xiong cxiong@buffalo.edu
David M. Johnson davidjoh@buffalo.edu
Jason J. Corso jcorso@buffalo.edu

ABSTRACT

Spectral clustering is widely used in data mining, machine learning and pattern recognition. There have been some recent developments in adding pairwise constraints as *side information* to enforce top-down structure into the clustering results. However, most of these algorithms are "passive" in the sense that the side information is provided beforehand. In this paper, we present a spectral active clustering method that actively select pairwise constraints based on a novel notion of node uncertainty rather than pair uncertainty. In our approach, the constraints are used to drive a purification process on the k-nearest neighbor graph—edges are removed from the graph based on the constraints—that ultimately leads to an improved, constraint-satisfied clustering. We have evaluated our framework on three datasets (UCI, gene and image sets) in the context of baseline and state of the art methods and find the proposed algorithm to be superiorly effective.

KEYWORDS

Spectral clustering, Active clustering, kNN Graph, Purification

1. INTRODUCTION

Data clustering is a fundamental problem in data mining and computer vision, such as image segmentation, object classification, gene analysisand social network analysis.Spectral clustering [Von Luxburg, 2007] is one of the most widely used clustering methods developed in recent years, for a variety of reasons including its minimal assumptions on the distribution of the data and strong mathematical underpinnings. In its original unsupervised form, spectral clustering groups samples based on pairwise similarity, which is expressed through a graph Laplacian matrix. However, in this formulation there is no way to ensure that the resulting clusters correspond to the semantic or other user-specified notions of categories in the data.

To address this problem, methods have been proposed that allow the algorithm to integrate pairwise constraints on the data as *side information*[Basu et al., 2004b, Li and Liu, 2009, Lu and Carreira-Perpinán, 2008]. These constraints may be either must-link (the two points belong in the same cluster) or cannot-link (the two points belong in different clusters). These papers have shown that the use of pairwise constraints can significantly improve the correspondence between clusters and semantic labels, when the constraints are selected *well*. Davidson et al. [Davidson et al., 2006] demonstrated that poorly chosen constraints can lead to worse performance than no constraints at all. However, selecting and identifying good constraint sets remains an open problem.

This issue of constraint selection in semi-supervised clustering is compounded by the nature of the state of

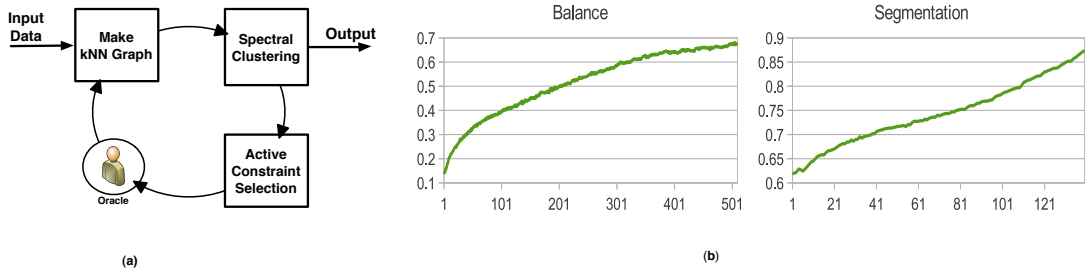

Figure 1: (a)Flowchart for spectral active clustering, which iteratively seeks new constraints based on the current clustering to refine the k-NN graph; (b)Two examples showing the effectiveness k-NN graph purification ($k = 20$). See Section 5. for details on the data sets. The graphs plot the cluster accuracy against number of nodes purified. Cluster accuracy is computed using the well-known V-Measure[Rosenberg and Hirschberg, 2007] algorithm.

the art methods: most are *passive* and require that their entire constraint set be selected before the clustering operation begins. Thus, these methods cannot know what effect a given constraint will have on the algorithm.

To overcome these limitations, we propose an spectral active clustering method that actively (iteratively) selects new pair-constraints based on the current clustering (illustrated in Figure 1)(a). As new constraints are selected, we ask an oracle, which generally is human, if the selected constraint is must-link or cannot-link. Then the k-Nearest Neighbor (k-NN) graph is refined to enforce these constraints and the spectral clustering is reevaluated. We note earlier work in the active selection of constraints for semi-supervised clustering exists, but previous active clustering methods either limit the number of clusters to two [Wang and Davidson, 2010, Xu et al., 2005], are computationally expensive [Hoi and Jin, 2008], or use simpler clustering regimes [Basu et al., 2004a, Klein et al., 2002, Mallapragada et al., 2008] (e.g., k-means) that require stringent assumptions on the cluster distribution (see our review in Section 2. for details).

In contrast, our active clustering approach can handle multiple clusters situation within the spectral clustering framework and is comparatively computationally simple. We approach the spectral active clustering problem by iteratively *purifying* the k-NN graph—a process that finds uncertain nodes, select constraings based on these sampled nodes, queries the oracle for new constraints and then refines the k-NN graph to satisfy the new constraints. Theoretical justification for these ideas is discussed in Section 3..

We compare our method and active selection criteria with baseline and state of the art approaches on UCI machine learning datasets [Asuncion and Newman, 2007], two gene datasets [Cho et al., 1998, Iyer et al., 1999] and part of the Caltech image dataset [Fei-Fei et al., 2006]. The results show that given the same number of pairs queried which is selected by our active selection criteria, our method can obtain much better accuracy than the baseline methods.

2. RELATED WORK

Active constraint selection for clustering has been attracting growing interest in the machine learning community due to the difficulty of selecting good constraints *a priori* coupled with the computational burden of having too many constraints and the cost of getting new constraints.

Active k-means clustering. Most preexisting active clustering algorithms are based on k-means and hierarchical clustering. Basu et al. [Basu et al., 2004a] proposed active k-means clustering using the farthest-first strategy. Klein et al. [Klein et al., 2002] developed a cluster-level active querying technique for hierarchical clustering, which works on data sets that exhibit local proximity structure. Mallapragada et al.[Mallapragada et al., 2008] presented another active k-means method based on a special case of the min-max approach, using similarity between points in a pair as a confidence value for must-link constraints. Though active, these methods carry the limitations of the underlying k-means algorithm.

Active spectral clustering. Xu et al.[Xu et al., 2005] propose an active constrained spectral clustering algorithm that examines the eigenvectors to identify the boundary points (of two clusters) and sparse points; then, it queries the oracle for constraints based on the these points. It has shown limited applicability because it

requires many queries to the oracle and assumes that errors in the clustering result only occur on the boundary points (which is only the case if the clusters are already nearly separated). Wang and Davidson[Wang and Davidson, 2010] present another spectral active clustering technique that identifies informative pairs according to the entropy of the pair example, requiring the evaluation of n^2 pairs at each iteration. Both of these prior spectral active clustering approaches have limited their work to two classes, and direct generalizations to multicluster cases are not known. In contrast, our proposed method is suitable for multicluster problems and uses an efficient linear-time method for actively selecting new constraints, as we demonstrate in this paper.

3. THEORETICAL UNDERPINNING AND MOTIVATION

Our ideas are motivated by two independent threads of literature. First, recent results in active learning [Settles, 2010], have demonstrated the potential in actively requesting information on the most uncertain point out of n available nodes. However, applying these techniques directly to the clustering case is problematic, because the algorithm must select a pair rather than a single node, requiring the evaluation of n^2 candidates.

Second, theoretical convergence analyses of spectral clustering have shown that the structure of the k-NN graph has significant influence on the ultimate clustering results. Maier et al.[Maier et al., 2009] first showed that the spectral clustering result is not independent of the graph structure, and Ting et al. [Ting et al., 2010] developed a framework for analyzing the graph with shrinking neighborhoods and demonstrated that changing the method by which nearest neighbors are determined could yield superior clustering results. Moreover, we have found empirically that the accuracy of the clustering result increases as the k-NN graph is purified, a concept that will be defined in Section 3. below; our results further expound this evidence (Section 5.). the rest of paper.

Given the graph G with the node set $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ that each $\mathbf{x}_i$ overloaded to be both a node in the graph and a sample $\mathbf{x}_i \in \mathbb{R}^m$; and the edge set E that ($e_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected). $\mathbf{W} = \{w_{ij}\}$ is the associated pairwise similarity matrix, which is non-negative and symmetric, From $\mathbf{W}$, denote the Laplacian and normalized Laplacian matrix of G as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ and $\overline{\mathbf{L}} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$, respectively, where $\mathbf{D}$ is degree matrix of graph G: $D_{ij} = \sum_j w_{ij}$ if $i = j$, else 0.

Spectral graph theory focuses on understanding the properties of the eigenvalues and eigenvectors of the two Laplacians, $\mathbf{L}$ and $\overline{\mathbf{L}}$. A typical objective function emphasizes the smoothness of the eigenvectors,

$$\Omega(f) = \sum_{ij} w_{ij}(\frac{f_i}{\sqrt{D_{ii}}} - \frac{f_j}{\sqrt{D_{jj}}})^2 = tr(f^T\overline{\mathbf{L}}f) \tag{1}$$

This measurement penalizes large changes between two connected nodes in the graph.It follows that the structure of the graph G, specifically the k-NN structure from which the sparse $\mathbf{W}$ is constructed, determines the ultimate clustering output, which may differ significantly from the ground truth. Must-link and cannot-link constraints may be used in semi-supervised clustering methods to impose the ground truth structure on the k-NN graph, but most existing methods accomplish this by selecting constraints randomly and beforehand. They are thus less likely to choose constraints that will adequately impact the clustering.

In this paper, our method adopts an active approach to selecting constraints that chooses new constraints based on current clustering outputs and the k-NN graph. Let us define a new property for the k-NN graph: *Purity*. Assume for now we are given a k-NN graph and ground truth knowledge of the cluster structure. Denote l_i to be the clustering membership index of node i in the ground truth clustering (assuming one exists for this discussion). Purity is the average fraction of the k neighbors for each node that lie within the same ground truth cluster as that node:

$$\text{Purity}(G) = \frac{1}{n}\sum_{i=1}^{n}\frac{\#\{l_j = l_i, j \in N_i\}}{\#\{N_i\}} \ , \tag{2}$$

where n is total number of nodes; N_i is the set of neighbors of the node i in the k-NN graph; and $\#$ is the set cardinality operator.

Ideally, all the connected nodes in the k-NN graph are in the same ground truth cluster, yielding a Purity of 1 and a trivially separable connected component graph. But, in practice, because of imperfect distance metrics

and complex underlying data distributions, the generated k-NN graph is nearly always impure, with many edges connecting nodes of different ground truth clusters. Due to these "bad edges," the spectral clustering result will likely differ from the ground truth assignment. Therefore, by removing so-called bad edges we are be able to improve the spectral clustering result (i.e., bring it closer to the ground truth clustering assignments).

We report our findings in Figure 1(b). We find that k-NN graph purification significantly improves the accuracy of the clustering results; this improvement occurs even with randomly selected constraints but more so for our proposed selection strategies.We further demonstrate this idea with an experiment—see Section 5. for details on the data used.

Algorithm 1 Spectral Active Clustering algorithm

Input: data points $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, k, t_{max}
Initialization:
Build the k-NN graph G^0 as the initial purified k-NN graph, set $t = 0$;
repeat
According to the purified k-NN graph G^t, calculate the similarity matrix W and normalized Laplacian matrix $\overline{\mathbf{L}} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2}$. And do spectral clustering.
Based on the clustering result, select the most informative node and obtain the corresponding pairs/edges according to the active selection methods in section 4.2;
Ask the oracle for constraints on the selected pairs/edges, and purify the k-NN graph to obtain graph G^{t+1} based on the response of the oracle.
$t = t + 1$
until ($t > t_{max}$ || oracle requests stop)
Output the final clustering result.

4. SPECTRAL ACTIVE CLUSTERING VIA k-NN GRAPH PURIFICATION

We now present our innovation for spectral active clustering via k-NN graph purification. Recall the basic flow of the algorithm depicted earlier in Figure 1(a). We first construct the initial k-NN graph from the input data, then compute an initial spectral clustering result. Using this result, we actively select new constraints, query the oracle and use the responses to purify the k-NN graph. We iterate this process until the oracle is satisfied or a fixed number of constraints have been generated. Note that the trivial clustering—i.e., the oracle has provided a full set of constraints that lets us fully purify the graph and thus each cluster is represented as a connected component in the k-NN graph—is not plausible as there are n^2 possible constraints and we implicitly seek as few as possible.

In this work, we make the common assumption that values for both k (the number of neighbors in the k-NN graph) and for k (the number of clusters) have been provided by the user. In our experiments, we use $k = 20$ and set the number of clusters based on each data set.

4.1 The Spectral Clustering Sub-Routine

At the beginning of the spectral active clustering procedure, an initial k-NN graph is formed by the distance of each two data points. During each iteration of the method (see Figure 1(a) for an overview and Algorithm 1 for a more precise description of the method), the selection algorithm will generate new constraints, purify the k-NN graph, and compute a new spectral clustering. At the core of our spectral active clustering method is the underlying spectral clustering method. Here, we use the NJW algorithm proposed by Ng et al.[Ng et al., 2001].

4.2 Active Selection

After obtaining the clustering results based on NJW, our method will then actively search for new constraints to further purify the k-NN graph and thus improve the clustering. As we described earlier, the goal in searching for new constraints is to find the nodes in the graph with the most "bad edges," which can provide the greatest

increase to the graph's purity if those bad edges are removed. However, in practice, we cannot directly evaluate the purity or bad edges because we do not know the ground truth. So, we rely on node uncertainty, which represents a node's confidence with its cluster assignment. A node with high uncertainty is likely to have bad edges in the k-NN graph which can be removed. In this section, we propose two strategies for computing node uncertainty and thus actively selecting new constraints.

4.2.1 Uncertainty based on the k-NN graph

Recall that when the k-NN graph is pure, the clustering result should match the connectivity in the k-NN graph. This implies that, in a pure graph, nodes connected in the k-NN graph will be assigned to the same cluster. On the other hand, in an impure graph the clustering result usually does not match the k-NN graph connectivity, and the graph will thus contain nodes whose neighbors are assigned to different clusters. As a result, the connections between these nodes and their neighbors are likely to contain bad edges. Based on this observation, we define an entropy criterion to measure the uncertainty of a node $\mathbf{x}_j$:

$$H_1(\mathbf{x}_j) = -\sum_i P_j(i) \log P_j(i) \ , \tag{3}$$

where $P_j(i) = \frac{\#\{c_z=i, z\in N_j\}}{\#\{N_j\}}$ is the ratio of neighbors of node j in the graph that are assigned to cluster i during clustering (notation c_z denotes the cluster index of neighbor z), and $0 \log 0 = 0$. If the entropy of a node is high, the relation between that node and its neighbors is disordered. If the entropy is small or 0, then the cluster assignments of the node and its neighbors obey the k-NN graph, and querying the node is unlikely to yield any useful information. So we select the node that has the highest entropy, and thus the most uncertainty:

$$\mathbf{x}_j^* = \underset{\mathbf{x}_j}{\operatorname{argmax}} H_1(x_j) \ . \tag{4}$$

We then obtain the edges:

$$E(x_j^*) = \{(x_i, x_j^*) \| e_{ij^*} = 1, x_i \in G\} \ . \tag{5}$$

The edge set contains edges that connect $\mathbf{x}_j^*$ to its corresponding neighbors in the k-NN graph and query the oracle for new constraints along these edges.

High entropy of a node signifies that the node is highly uncertain about the cluster to which it belongs. Hence, high entropy of a node suggests that it is impure. We therefore select the node whose entropy is highest and query the oracle about its edges.

4.3 Purifying the k-NN graph

After a node has been selected and its k-NN edges are queried by the oracle, we obtain two sets of constraints on the edges: the must-link set L_M and the cannot-link set L_C. Note that at any point in the algorithm, the current sets L_M and L_C include constraints from the current iteration as well as all prior iterations. Purification is a two step process. First, we take the current sets L_M and L_C and augment them based on the k-NN graph structure: some unknown edge constraints can be inferred from the known ones. For example, we search for all transitive closure groups of the L_M edges. We explain the constraint augmentation process here.

1. First, use transitive closure of the must-link edges to get connected components consisting of all points connected by must-link constraints. Call these connected components *collapsed points*.
2. Then, augment the must-link set L_M by adding the edges that connect two nodes belonging to the same collapsed point.
3. If there is an edge from cannot-link set L_C that connects two collapsed points, then other edges that connect the two collapsed point in the k-NN graph are assigned as cannot link and added into the set L_C.

Second, we apply the constraints to directly purify the graph. When we complete the constraint augmentation, we apply the constraints to purify the graph by deleting all the edges in the cannot link set L_C from the k-NN graph to increase the purity of the k-NN graph. We also set the similarity value of edges in the must-link L_M to 1. Then we repeat the spectral clustering algorithm based on the new purified k-NN graph.

Table 1: UCI Datasets, GENE Datasets and 5-CLASS IMAGE dataset

Name	$\#Classes$	$\#Instances$	$\#Features$	Name	$\#Classes$	$\#Instances$	$\#Features$
Segmentation	7	210	19	Cho's	5	386	17
Breast	3	683	10	Iyer's	11	517	12
Balance	3	625	4	Pyramidal HOG	5	485	6300

5. EXPERIMENTS

We test the proposed spectral active clustering (SAC) algorithm on UCI machine learning datasets [Asuncion and Newman, 2007], two gene datasets (Cho's[Cho et al., 1998] and Iyer's[Iyer et al., 1999]) and a five-class image dataset that is randomly sampled from Caltech-101 [Fei-Fei et al., 2006] with images represented by codebooked pyramidal dense HOG features [Dalal and Triggs, 2005]. More details are in Table 1.

At the beginning of all experiments, the k-NN graph is built by calculating the distance matrix using Euclidean distance and applying a Gaussian kernel to the distances to generate the similarity weights. We use $k = 20$ in all experiments. To evaluate our proposed spectral active clustering algorithm , we use the following set of methods, including baselines and the state of the art:

- **Random**: the first baseline algorithm is the proposed spectral active clustering (SAC) algorithm, but the pairwise constraints are randomly sampled. This baseline lets us evaluate the proposed active constraint selection systematically within the same clustering paradigm.
- **CCSKL**: Constrained Spectral Clustering[Li and Liu, 2009] with randomly sampled constraints. This is the state of the art multiclass semi-supervised spectral clustering method.
- **CCSKL+P-Random**: the CCSKL method with randomly sampled constraints and our proposed graph purification with those constraints, which lets us evaluate if the proposed purification method can improve other spectral clustering methods.
- **CCSKL+P-Active**: CCSKL using the constraints selected by our active selection method, with the constraint purifying the graph.
- **SAC**: This is our proposed algorithm selecting the most uncertain node based on the k-NN graph.

Experiment Setup. To measure the performance, we adopt the well-known two measurement: Rand Index and V-Measure[Rosenberg and Hirschberg, 2007] metric for determining cluster accuracy, which defines entropy-based measures for the completeness and homogeneity of the clustering results, and computes the harmonic mean of the two.

Figure 2 shows our comparative results using the V-Measure accuracy. Our active clustering framework is demonstrably effective at achieving high accuracy with fewer constraints than the state of the art CCSKL method, and uniformly outperforms it. We note the improvement in accuracy of the CCSKL method when we incorporate the proposed graph-purification step—implying that graph-purification, in and of itself, may be important to the broader spectral clustering community.

However, the degree to which active selection methods outperform random selection methods varies greatly among the different datasets that may be a consequence of the structure of certain datasets.

Figure 3 shows our comparative results using the Rand Index accuracy. Once again, our active methods show a clear advantage over both the baseline using random constraint selection and CCSKL, though the extent to which they outperform the baseline is variable. There are a few cases where the V-Measure and Rand Index disagree—for instance, on the performance of the k-NN purification method on the Caltech data. These likely stem from the fact that the Rand Index tends to report inflated results for problems with 3 or more classes, due to the preponderance of cannot-link constraints between the nodes, which will mostly be satisfied regardless of the actual correctness of the clusters. Despite these few conflicts, the two measures generally agree on which method achieves the best performance on each dataset.

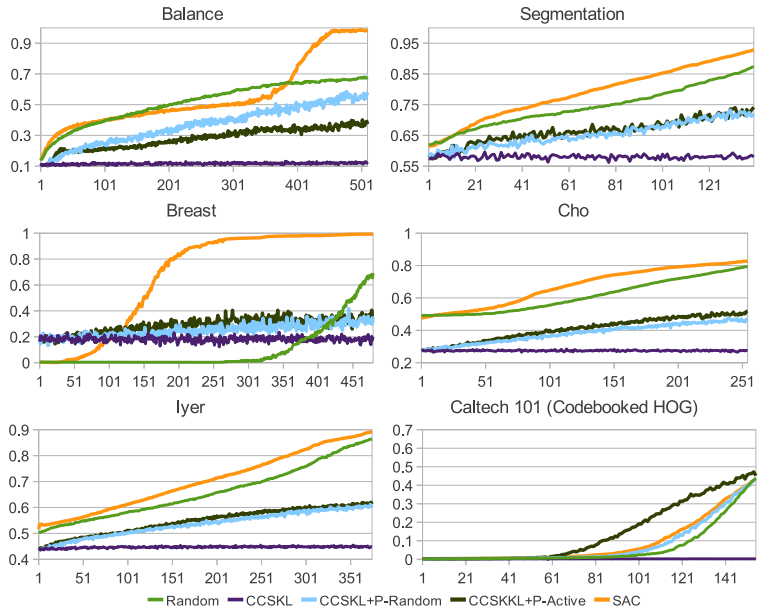

Figure 2: V-Measure accuracy (vertical axes) with increasing iteration number (twenty constraints per iteration) on the datasets. Our active methods generally outperform the baseline approaches, sometimes by a considerable margin. *View in color.*

6. CONCLUSION

In this paper, we have considered the problem of active constraint selection for semi-supervised spectral clustering. Our paper makes two contributions: first, we describe a method for semi-supervised spectral clustering by purifying the k-NN graph; second, we propose one method for actively sampling constraints by transforming the pair-uncertainty problem into a node-uncertainty problem. Our comparative results on several benchmarks demonstrate superior performance to the baseline and state of the art semi-supervised spectral clustering method using the V-measure and the Rand Index measurement.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), DARPA CSSG (HR0011-09-1-0022 and D11AP00245). Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, ARO, NSF or the U.S. Government.

REFERENCES

A. Asuncion and D.J. Newman. UCI machine learning repository. 2007.

S. Basu, A. Banerjee, and R.J. Mooney. Active semi-supervision for pairwise constrained clustering. In *ICDM*, pages 333–344, 2004a.

S. Basu, M. Bilenko, and R.J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004b.

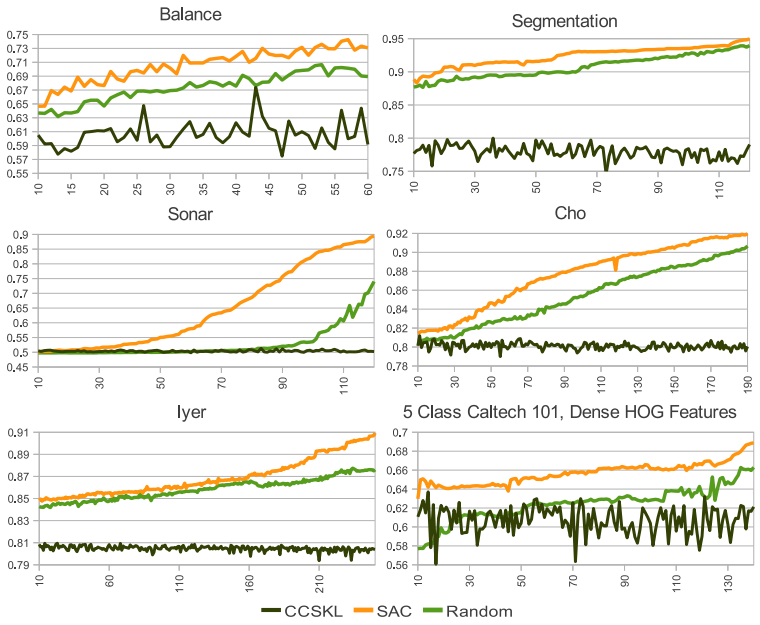

Figure 3: Rand Index accuracy (vertical axes) with increasing iteration number (twenty constraints per iteration) on the datasets. Our active methods generally outperform the baseline approaches. *View in color.*

R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell*, 2(1):65–73, 1998.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005. ISBN 0769523722.

I. Davidson, K. Wagstaff, and S. Basu. Measuring constraint-set utility for partitional clustering algorithms. *PKDD*, pages 115–126, 2006.

L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, pages 594–611, 2006. ISSN 0162-8828.

S.C.H. Hoi and R. Jin. Active kernel learning. In *ICML*, pages 400–407. ACM, 2008.

V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J. Hudson, M.S. Boguski, et al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398):83, 1999.

D. Klein, S.D. Kamvar, and C.D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML*, 2002.

Z. Li and J. Liu. Constrained clustering by spectral kernel learning. In *ICCV*, pages 421–427. IEEE, 2009.

Z. Lu and M.A. Carreira-Perpinán. Constrained spectral clustering through affinity propagation. In *CVPR*, pages 1–8. IEEE, 2008.

M. Maier, U. Von Luxburg, and M. Hein. Influence of graph construction on graph-based clustering measures. *NIPS*, 22: 1025–1032, 2009.

P.K. Mallapragada, R. Jin, and A.K. Jain. Active query selection for semi-supervised clustering. In *ICPR*, pages 1–4. IEEE, 2008.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL'07*, pages 410–420, 2007.

B. Settles. Active learning literature survey. Technical report, 2010. (unpublished report).

D. Ting, L. Huang, and M.I. Jordan. An Analysis of the Convergence of Graph Laplacians. In *ICML*. Citeseer, 2010.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

X. Wang and I. Davidson. Active Spectral Clustering. In *ICDM*, 2010.

Q. Xu, M. Desjardins, and K. Wagstaff. Active constrained clustering by examining spectral eigenvectors. In *Discovery Science*, pages 294–307. Springer, 2005.