# Uncertainty Reduction For Active Image Clustering via a Hybrid Global-Local Uncertainty Model

**Caiming Xiong, David M. Johnson** and **Jason J. Corso**
State University of New York at Buffalo
Department of Computer Science and Engineering
338 Davis Hall, Buffalo, NY, 14260-2500
{cxiong,davidjoh,jcorso}@buffalo.edu

## Abstract

We propose a novel combined global/local model for active semi-supervised spectral clustering based on the principle of uncertainty reduction. We iteratively compute the derivative of the eigenvectors produced by spectral decomposition with respect to each item/image, and combine this with local label entropy provided by the current clustering results in order to estimate the uncertainty reduction potential of each item in the dataset. We then generate pairwise queries with respect to the best candidate item and retrieve the needed constraints from the user. We evaluate our method using three different image datasets—faces, leaves and dogs, and consistently demonstrate performance superior to the current state-of-the-art.

## Introduction

Semi-supervised clustering plays a crucial role in artificial intelligence for its ability to enforce top-down semantic structure while clustering data that is often noisy or incomplete (Basu, Bilenko, and Mooney 2004; Li and Liu 2009; Chen and Zhang 2011). It has the potential to be a powerful tool in many problems, including facial recognition and plant categorization (Biswas and Jacobs 2012).

In visual surveillance, for example, there is significant demand for automated grouping of visual elements, whether they are faces, plants or actions in video. However, obtaining large amounts of training data for this problem is problematic—expecting typical humans to label a large set of strangers' faces or plant species is not realistic. However, a human worker probably *can* reliably determine whether two faces belong to the same person or two plants to the same species, making it quite feasible to obtain pairwise constraints for this problem by adopting a low-cost crowdsourcing tool such as Mechanical Turk.

However, even when using relatively inexpensive human labor, any attempt to apply semi-supervised clustering methods to large-scale problems must still consider the cost of obtaining large numbers of pairwise constraints. To overcome these problems, researchers have begun exploring *active* constraint selection methods which allow clustering algorithms to intelligently select constraints based on the structure of the data and/or clustering results (Basu, Banerjee, and Mooney 2004; Xu, Desjardins, and Wagstaff 2005; Wang and Davidson 2010; Xiong, Johnson, and Corso 2012; Biswas and Jacobs 2012).

In this paper, we propose a novel hybrid global/local uncertainty model that we use to perform efficient and effective item-based constraint selection in an online iterative manner. In each iteration of the algorithm, we find the item that will yield the greatest predicted **reduction** in the **uncertainty** of the clustering, and generate pairwise questions. The proposed framework is as follows (more details in next section):

1 Randomly choose a single item, assign it to the first *certain set* and initialize the pairwise constraint set as empty. A *certain set* is those items from the same cluster based on the user query responses.
2 **Constrained Clustering:** cluster all items into $n_c$ clusters using the current pairwise constraint set.
3 **Informative Item Selection:** choose the most informative item based on our **Global/Local** uncertainty model.
4 **Pairwise Constraint Queries:** use pairwise user queries to assign the selected item to a certain set and generate pairwise constraints.
5 Repeat steps 2-4 until the human is satisfied with the clustering result or the query budget is reached.

We run our method on face, leaf (Kumar et al. 2009), and dog (Khosla et al. 2011) image datasets and find it consistently outperforms existing techniques.

## Hybrid Global-Local Active Clustering

First denote the data set $X = \{x_1, x_2, \cdots, x_n\}$, and the corresponding similarity matrix $\mathbf{W} = \{w_{ij}\}$. Also denote the set of certain item sets $\mathcal{Z} = \{Z_1, \cdots, Z_m\}$, where $Z_i$ is a set such that $Z_i \subset X$ and $Z_i \cap Z_j = \emptyset \quad \forall j \neq i$, and an item set $\mathcal{U} = \bigcup_i Z_i$ containing *all* current certain items.

### Constrained clustering with pairwise constraints

In order to incorporate pairwise constraints into spectral clustering, we adopt a simple and effective method called spectral learning (Kamvar et al. 2003). Whenever we obtain new pairwise constraints the algorithm directly modifies the current similarity matrix $W^t$, producing a new matrix $W^{t+1}$ that reflects the information in those constraints. We then proceed with the standard spectral clustering procedure.

## Pairwise query generation

After clustering and item selection, we must use pairwise queries to assign the selected informative item $x_i$ to the correct certain set. We do this by finding the single item $\mathbf{x}_l$ in each certain set which is closest to $x_i$, and then querying (in order of ascending Euclidean distance) the true relationship between $x_i$ and each $x_l$, stopping when we obtain a must-link constraint. We then add $\mathbf{x}_i$ to the certain item set containing that $x_l$, or create a new certain set $Z_{m+1}$ and add $\mathbf{x}_i$ to it if no must-link constraints are found.

Since the relation between the new item and all certain sets in $\mathcal{Z}$ is known, we can now generate new pairwise constraints between the selected item $\mathbf{x}_i$ and all items in $\mathcal{U}$ without submitting any further queries to the human.

## Global/local uncertainty model

In this section, we propose a global-local model for finding the most informative item. We first assume that obtaining the most informative item and querying pairs to make the chosen item "certain" can decrease the uncertainty of the data set as a whole and thus improve the clustering result.

To estimate the uncertainty-reducing impact of the selected item, we adopt a first-order approximation using both a global measure and a local one

$$x_i = \underset{x_i}{\operatorname{argmax}}\, \mathcal{G}(x_i; L^t, X, \mathcal{U}) \cdot \mathcal{L}(x_i; C^t, X) \qquad (1)$$

where $\mathcal{G}(x_i; L^t, X, \mathcal{U})$ is global model used to estimate the slope of the uncertainty reduction based on derivatives of eigenvectors of the Laplacian $L^t$ at $x_i$; $\mathcal{L}(x_i; C^t, X)$ is the local model to estimate the step scale factor.

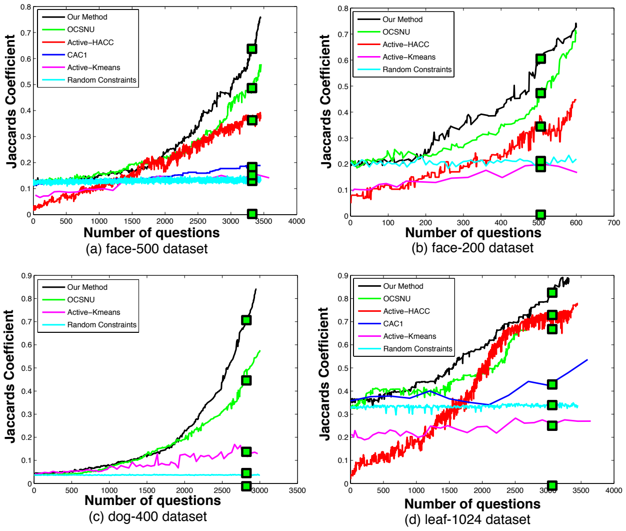
**Global Model for estimating the slope of the uncertainty reduction** The spectral learning algorithm updates the similarity matrix by $W^{t+1} = W^t + \delta W^t$ at each iteration. The clustering result is trivial to compute based on the $n_c$ eigenvectors using k-means. Thus, we approximate the derivative of uncertainty reduction by using the first order change of the eigenvectors inspired by matrix perturbation theory (Stewart and Sun 1990).

Assume $L^t \approx \sum_{j=1}^{n_c} \lambda_j v_j v_j^T$. In each iteration, for any chosen point $x_i$, the relation between $x_i$ and $x_k \in X_m$ will be known, where $X_m = \{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\}$ sampled from each certain set $Z_i \in \mathcal{Z}$. Thus $w_{ik}^t$ in $W^t$ will be updated for clustering. Therefore we define our global model as:

$$\begin{aligned} \mathcal{G}(x_i; L^t, X, \mathcal{U}) &= \sum_{x_k \in X_m} \left| \sum_{j=1}^{n_c} \frac{dv_j}{dw_{ik}^t} \right| \qquad (2) \\ &= \sum_{x_k \in X_m} \left| \sum_{j=1}^{n_c} \sum_{p \neq j} \frac{v_j^T [\partial L^t / \partial w_{ik}^t] v_p}{\lambda_j - \lambda_p} v_p \right| . \end{aligned}$$

It sums up the influence on the first order change of eigenvectors by altering elements $w_{ik}^t$ in similarity matrix $W^t$ as the estimate of the influence of item $x_i$.

**Local Uncertainty Model for Approximating the Step Scale Factor** To evaluate the step size, under the assumption that highly uncertain items will likely make the uncertainty reduction go longer along the slope, we propose to estimate the local uncertainty of the item using entropy. Specifically we evaluate entropy among local cluster labels produced the current clustering results.



(a) face-500 dataset
(b) face-200 dataset
(c) dog-400 dataset
(d) leaf-1024 dataset

Figure 1: Performance of various active clustering methods with increasing numbers of queries in image sets. *Best viewed in color.*

First, consider the current clustering result $C^t = \{c_1, c_2, \ldots, c_{n_c}\}$, where $c_i$ is a cluster and $n_c$ is the number of clusters. We can then define a simple local non-parametric model based on similarity matrix $W$ for determining the probability of $x_i$ belonging to cluster $c_j$: $P(c_j|x_i) = \sum_{x_l \in c_j} w_{il} / \sum_{x_l \in \mathcal{N}(x_i)} w_{il}$, where $\mathcal{N}(x_i)$ is the $\mathbf{k}$ nearest neighbor points to $x_i$. Then the uncertainty of item $x_i$ can be defined, based on entropy, as $\mathcal{L}(x; C^t, X) = -\sum_{j=1}^{n_c} P(c_j|x_i) \log P(c_j|x_i)$.

Then according to Equation 1, we obtain our global-local strategy to select the item which will yield the greatest uncertainty reduction at each iteration.

# Experiments

**Dataset and Protocol:** We evaluate our proposed framework and informativeness measures on three image datasets: leaf images, dog images and face images datasets. We set $\mathbf{k} = 20$.We use Jaccard's Coefficient (Pang-Ning, Steinbach, and Kumar 2006) as cluster evaluation metrics that is used in other active image clustering paper.

**Comparison methods:** Random Constraints, Active-HACC (Biswas and Jacobs 2012), OCSNU (Xiong, Johnson, and Corso 2012), CAC1 (Biswas and Jacobs 2011) and Active-KMeans (Basu, Banerjee, and Mooney 2004).

**Results:** Figure 1 shows the Jaccard's Coefficient of different active clustering algorithms with varying numbers of pairwise constraints queried. Our algorithm always performs best on all image sets. And comparing with other method, our method performs better in face-500 and dog-400 dataset than in face-200 and leaf-1042, which is expected due to the intrinsic dataset complexity.

# References

Basu, S.; Banerjee, A.; and Mooney, R. 2004. Active semi-supervision for pairwise constrained clustering. In *ICDM*.

Basu, S.; Bilenko, M.; and Mooney, R. 2004. A probabilistic framework for semi-supervised clustering. In *SIGKDD*. ACM.

Biswas, A., and Jacobs, D. 2011. Large scale image clustering with active pairwise constraints. In *ICML 2011 Workshop on Combining Learning Strategies to Reduce Label Cost*.

Biswas, A., and Jacobs, D. 2012. Active image clustering: Seeking constraints from humans to complement algorithms. In *CVPR*. IEEE.

Chen, L., and Zhang, C. 2011. Semi-supervised variable weighting for clustering. In *SDM*.

Kamvar, K.; Sepandar, S.; Klein, K.; Dan, D.; Manning, M.; and Christopher, C. 2003. Spectral learning. In *IJCAI*. Stanford InfoLab.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*.

Kumar, N.; Berg, A.; Belhumeur, P.; and Nayar, S. 2009. Attribute and simile classifiers for face verification. In *ICCV*, 365–372. IEEE.

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

Pang-Ning, T.; Steinbach, M.; and Kumar, V. 2006. Introduction to data mining. *WP Co.*

Stewart, G., and Sun, J. 1990. *Matrix perturbation theory*, volume 175. Academic press New York.

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

Xiong, C.; Johnson, D.; and Corso, J. J. 2012. Online active constraint selection for semi-supervised clustering. In *ECAI 2012 AIL Workshop*.

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

We gratefully acknowledge support from NSF CAREER IIS-0845282 and DARPA CSSG (D11AP00245, D12AP00235).