Journal of Machine Learning Research 1 (2020) 1-48 Submitted 4/00; Published 10/00

Query-augmented Active Learning for Large-scale Clustering

Yujia Deng YUJIAD2@ILLINOIS.EDU
Department of Statistics, University of Illinois, Urbana-Champaign,

Yubai Yuan YUBAIY2@ILLINOIS.EDU
Department of Statistics, University of California, Irvine

Haoda Fu FU_HAODA@LILLY.COM
Eli Lilly and Company

Annie Qu AQU2@UCI.EDU
Department of Statistics, University of California, Irvine

Editor:

Abstract

In this paper, we propose an active metric learning method for clustering large-scale data. The key idea of the new active clustering is to choose the informative instance pairs actively in terms of estimating the underlying metrics by incorporating unlabeled instance pairs, which leads to a more accurate and efficient clustering process. Specifically, we formulate estimation and sampling processes based on a unified framework to select a metric favoring the clustering task to incorporate both within-cluster instance constraints and the implicit structure-level constraints. In addition, the proposed method is designed to increase the robustness on uncertainty from human-machine interaction at each step of active learning algorithm. Therefore we iteratively update a metric to gradually refine a continuous cluster structure in the latent space. Numerical studies and theoretical properties all indicate that the proposed method is especially advantageous when the signal-to-noise ratio between relative features and irrelevant features is low, and the dimension of total features is high.

Keywords: Semi-supervised clustering; Active learning; Metric Learning; Selective penalty

1. Introduction

The idea of incorporating experts' domain knowledge or user's feedback has been established in several clustering methods Basu et al. (2004, 2006); Davidson et al. (2006). Specifically, a user can specify a prior in that two instances must belong to the same cluster or different clusters. Alternatively, instead of directly clustering the instances in the original feature space, metric learning Hoi et al. (2010); Niu et al. (2011); Xing et al. (2003) seeks an appropriate distance metric from labeled data in that similar instances have a closer metric distance compared to dissimilar objects, to improve the performance of clustering.

However, the above learning process could be inefficient when only a limited number of constraints. Therefore, the idea of active learning has been adopted in clustering to query the most informative unlabeled instance pairs sequentially in order to achieve efficient and effective clustering. This includes sampling the instances either on boundary Basu et al. (2004); Grira et al. (2005); Mallapragada et al. (2008) or the ones with higher uncertainty Xiong et al. (2014); Huang and Mitchell (2006); Biswas and Jacobs (2014).

The existing active clustering methods have one major challenge in that when the metric space changes, the neighborhood information or the clustering results from the previous step also change

and thus the uncertainty criteria of unlabeled pairs may be ineffective or misleading. Another limitation is that existing active clustering approaches do not pursue dimension reduction. However, identifying and selecting significant features which are more relevant to user's clustering principles are crucial to enhance the similarity within a cluster. In the context of active clustering, pursuing dimension reduction also leads to selecting more interpretable clustering criteria from the user. In addition, most of active clustering approaches do not fully incorporate the history training results in the final model. Nonetheless, the history training information can essentially improve the model performance as shown in the sequential algorithms such as boosting Quinlan et al. (1996).

In this paper, we propose a novel active metric learning for large-scale clustering. We select a metric enhancing the clustering performance by incorporating both within-cluster pairwise constraints and the implicit structure-level constraints. In addition, due to the intrinsic nature of active learning and clustering as a greedy algorithm, we design a robust scheme through iteratively updating a metric to refine a continuous cluster structure in the latent space sequentially, which can lead to a more accurate and interpretable clustering outcome. Moreover, we propose a new active query strategy to select unlabeled pairs which have the highest impact on the distribution of instance pairs over the entire dataset. This provides a more accurate measurement for obtaining the informative unlabeled pairs compared to the existing criteria.

2. Notation and Background

Given n data points in a p-dimensional feature space, $\boldsymbol{x}_i \in \mathbb{R}^p, i = 1, ..., n$, we assume there are K clusters and denote the cluster label of $\boldsymbol{x}_i$'s as $\boldsymbol{l} = (\ell_1, ..., \ell_n)$, where $\ell_i = 1, ..., K$, denotes the index of the cluster that $\boldsymbol{x}_i$ belongs to. We also denote the similarity matrix as $Y \in \{0, 1\}^{n \times n}$, where $y_{ij} = 1$, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ fall into the same cluster, and 0 otherwise. The goal of clustering is to estimate either Y or $\boldsymbol{l}$, since $y_{ij} = \mathbb{1}(\ell_i = \ell_j)$. In *unsupervised* clustering, no elements of Y are known beforehand, while in *semi-supervised* clustering, some elements of Y are queried from the oracle as pairwise constraints. These pairwise constraints are referred to as *similar* and *dissimilar* pairs whose index sets are denoted as $\mathcal{S} = \{(i, j)|y_{ij} = 1\}$ and $\mathcal{D} = \{(i, j)|y_{ij} = 0\}$, respectively, while the unlabeled set is denoted as $\mathcal{U} = \{(i, j) \,|\, (i, j) \notin \mathcal{S} \cup \mathcal{D}\}$.

In addition, we consider the case that the clustering structure is determined on a linear subspace $\mathbb{R}^m \subset \mathbb{R}^p, m \leq p$, i.e., there exists a matrix $M \in \mathbb{R}^{m \times p}$ with orthogonal columns such that $P(y_{ij} = 1|\boldsymbol{x}_i, \boldsymbol{x}_j) = P(y_{ij} = 1|M\boldsymbol{x}_i, M\boldsymbol{x}_j)$. This setting induces a Mahalanobis distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top A (\boldsymbol{x}_i - \boldsymbol{x}_j)$, where $A = M^\top M$ is called the *metric matrix*. By correctly identifying M, or A equivalently, one can improve the clustering performance compared with using the original distance metric. In general, the distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A$ is small if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster and is large if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to different clusters. One way Xing et al. (2003) to learn A is through

$$\min_A \sum_{(i,j)\in\mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A^2, \quad \text{s.t.} \sum_{(i,j)\in\mathcal{D}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A \geq 1, \quad \text{and} \quad A \succeq 0, \tag{1}$$

where $A \succeq 0$ denotes A is positive semi-definite. The above training process aims to minimize the distance between similar pairs while separating the dissimilar pairs to avoid trivial zero solutions.

3. Methodology

3.1 Metric learning with augmented pairwise constraints

One problem of (1) and existing metric learning methods Wagstaff et al. (2001); Lu (2007); Grira et al. (2005) is that only the violation on the queried pairwise constraints is penalized. However, knowing one pairwise relation also provides additional prior information on other pairs implicitly through the underlying cluster structure. This motivates us to generalize the queried pairwise constraints $\mathcal{S} \cup \mathcal{D}$ to all y_{ij}'s through inferring labels on unlabeled instance pairs, and train the metric matrix A using both the queried pairwise constraints and the augmented ones.

Specifically, we first solve for a fuzzy membership matrix $H \in \mathbb{R}^{n\times p}$ by

$$\widehat{H} = \underset{H}{\operatorname{argmin}} \sum_{(i,j)\in S\cup D} (y_{ij} - \boldsymbol{h}_i^T \boldsymbol{h}_j)^2, \quad \text{s.t.} \quad h_{ij} \geq 0, \quad \sum_{k=1}^{K} h_{ik} = 1, \text{ for all } i, \tag{2}$$

where $\boldsymbol{h}_i^T$ is the i-th row of H. In contrast to the hard-thresholding labels ℓ_i, h_{ik} is continuous between $[0, 1]$ and represents the probability that sample i belongs to the cluster k. We only infer $\boldsymbol{h}_i$ with at least one elements of $\{y_{i\cdot}\}$ observed, otherwise we let the elements of $\boldsymbol{h}_i$ be all $1/K$. Note the solution purely uses the constraint information without involving the distance between data points since the distance metric is inaccurate during training, which may lead to erroneous inference.

We utilize $\widehat{H}$ to infer additional pairwise constraints. The idea is that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ tend to be similar if $\widehat{\boldsymbol{h}}_i$ and $\widehat{\boldsymbol{h}}_j$ are concordant in the sense that $\widehat{\boldsymbol{h}}_i^\top \widehat{\boldsymbol{h}}_j$ is close to 1; and dissimilar if $\widehat{\boldsymbol{h}}_i^\top \widehat{\boldsymbol{h}}_j$ is close to 0. In the completely random case, we have $\widehat{h}_i = \widehat{h}_j = (1/K, ..., 1/K)$ and $\widehat{\boldsymbol{h}}_i^\top \widehat{\boldsymbol{h}}_j = 1/K$. Therefore, we use $1/K$ as a threshold for the concordance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and define the augmented constraints as $\widetilde{\mathcal{S}} = \{(i,j)|\widehat{\boldsymbol{h}}_i^\top \widehat{\boldsymbol{h}}_j > 1/K\}$ and $\widetilde{\mathcal{D}} = \{(i,j)|\widehat{\boldsymbol{h}}_i^\top \widehat{\boldsymbol{h}}_j < 1/K\}$. Then we use the augmented pairwise constraints to train metric matrix through

$$\begin{aligned} \min_{A} \quad & Loss(A) \triangleq \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A^2 + \frac{1}{|\widetilde{\mathcal{S}}|} \sum_{(i,j)\in\widetilde{\mathcal{S}}} w_{ij} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A^2, \\ s.t. \quad & \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A + \frac{1}{|\widetilde{\mathcal{D}}|} \sum_{(i,j)\in\widetilde{D}} w_{ij} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A \geq 1, \quad A \succeq 0, \end{aligned} \tag{3}$$

where $|\cdot|$ denotes the set cardinality, and $w_{ij} = \frac{K}{K-1} \max\{\widehat{\boldsymbol{h}}_i^T \widehat{\boldsymbol{h}}_j - \frac{1}{K}, 0\} - K \min\{\widehat{\boldsymbol{h}}_i^T \widehat{\boldsymbol{h}}_j - \frac{1}{K}, 0\}$. Compared with (1), the second terms in the loss function and the constraint of (3) are the augmented similar constraints and dissimilar constraints, respectively, with $w_{ij} \in [0, 1]$ quantifying the certainty of the inference. In particular, we impose less weight on the augmented constraints that are similar to random guess.

3.2 Metric aggregation through selective penalty

Although the metric matrix is trained sequentially in most of the existing active clustering methods, the history training result is not incorporated in the final model. However, the metric matrix learned during the previous steps provides extra information for identifying the features related to the user-specified clustering principles.

We propose to aggregate the metric matrices learned in each step to extract the underlying true features by imposing an adaptive penalty. We denote the minimizer of (3) at tth step ($t = 1, ..., T-1$) as A^t and the rank of eigenvalues of A^t in ascending order as $\boldsymbol{r}^t$, $t = 1, ..., T-1$. Intuitively, we aim to extract a clustering-oriented subspace by penalizing the unrelated features. Note that imposing penalty on all features makes little effects since clustering is invariant to the scale change of metric. Instead, we enlarge the relative scale between the true features and the unrelated features by penalizing selectively. To determine which features are unrelated, we aggregate the training results of the previous $T-1$ steps and impose penalty adaptively according to the eigenvalues of $A^t, t = 1, ..., T-1$. In general, the features with smaller eigenvalues on average are less relevant in clustering and thus are supposed to have less weight.

We let $\bar{\boldsymbol{r}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \boldsymbol{r}^t$ be the average rank of the features. To shrink the irrelevant features, we penalize the top q features with the smallest entries in terms of $\bar{\boldsymbol{r}}$ and denote the index set of the penalized features as $\mathcal{G}^T$, where q is the predetermined number. For Tth step, we train the metric

matrix by adding selective penalty on A through

$$\begin{aligned} \min_{A} \quad & Loss(A) + \gamma \sum_{p \in \mathcal{G}^T} |\sigma_p(A)|, \\ s.t. \quad & \frac{1}{|\mathcal{D}^T|} \sum_{(i,j)\in\mathcal{D}^T} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A + \frac{1}{|\widetilde{\mathcal{D}}^T|} \sum_{i,j} w_{ij} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A \geq 1, \quad A \succeq 0, \end{aligned} \tag{4}$$

where λ and γ are two tuning parameters and $\sigma_p(A)$ denotes the pth eigenvalue of A.

After acquiring $\widehat{A}$ through (4), we solve for the cluster membership by performing pairwise constrained Kmeans (PCKmeans) Basu et al. (2004).

3.3 Active query with minimum expected pairwise uncertainty

We propose a new query strategy by selecting the instance pair that has the largest impact on the uncertainty of all the instance pairs. To start with, we introduce the neighborhood structure of instances Xiong et al. (2014) which increases the number of constraints one can obtain through one query. A neighborhood N_m contains the instances that are confirmed to belong to the same cluster, i.e. $(i,j) \in \mathcal{S}$ for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in N_m$ and $(i,j) \in \mathcal{D}$ for any $\boldsymbol{x}_i \in N_m$, $\boldsymbol{x}_j \in N_{m'}, m \neq m'$. We denote the collection of neighborhoods as $\mathcal{N} = \{N_m\}_{m=1}^L$, where L is the number of neighborhood at the current step, $L \leq K$. Then by identifying the neighborhood of an instance, we can generate its pairwise relationship with all other instances in the existing neighborhoods.

During the sequential query procedure, the unlabeled pair is queried based on certain uncertainty criteria. We propose to utilize the expected decrease of uncertainty measured by cross-instance entropy. Specifically, we let $R \in \mathbb{R}^{n\times L}$ be the neighborhood membership matrix, where $r_{im} = P(\boldsymbol{x}_i \in N_m)$. If the data is sampled independently, then the probability that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same neighborhood can be computed as $p_{ij} = \sum_{m=1}^L P(\boldsymbol{x}_i \in N_m, \boldsymbol{x}_j \in N_m) = \sum_{m=1}^L r_{im} r_{jm}$. We measure the uncertainty of the entire dataset by summation of the entropy of each pair,

$$Q(R) = -\sum_{i,j} \left\{ p_{ij} \log_2 p_{ij} + (1 - p_{ij}) \log_2 (1 - p_{ij}) \right\}.$$

With more pairwise constraints queried, $Q(R)$ decreases monotonically. In particular, $Q(R)$ drops to 0 when the membership of each instance is known. Consequently, the most informative instance at the tth step minimizes the expected uncertainty of the $(t+1)$th step conditioning on the tth step, i.e., $E(Q(R^{(t+1)})|R^{(t)})$, which can be estimated by

$$u^t(\boldsymbol{x}_i) = \sum_{m=1}^{L^t} r_{im}^t Q(\widetilde{R}_{-im}^{(t+1)}), \tag{5}$$

where $\widetilde{R}_{-im}^{(t+1)} \in \mathbb{R}^{n\times L^{t+1}}$ is defined elementwise by $\widetilde{r}_{ij}^{(t+1)} = r_{kj}^t$, if $k \neq i$; 0, if $k = i, j \neq m$; and 1, if $k = i, j = m$. In other words, $\widetilde{R}_{-i,m}^{(t+1)}$ denotes the neighborhood membership matrix assuming knowing $\boldsymbol{x}_i$ belongs to the mth neighborhood, and $u^t(\boldsymbol{x}_i)$ estimates the expected uncertainty after obtaining the neighborhood membership of $\boldsymbol{x}_i$. Then we select the instance whose neighborhood membership is unknown to minimize (5), $\boldsymbol{x}^* = \text{argmin}_{\boldsymbol{x}_i \notin \mathcal{N}^t} u^t(\boldsymbol{x}_i)$. The complete algorithm combining Section 3.1, 3.2 and 3.3 is summarized in Algorithm 1.

4. Numerical studies

Simulation data The simulated data consists of two parts, $\boldsymbol{x} = \left((\boldsymbol{x}^1)^\top, (\boldsymbol{x}^2)^\top\right)^\top$, where $\boldsymbol{x}^1 \in \mathbb{R}^{p_1}$ are true features which determine the cluster memberships, and $\boldsymbol{x}^2 \in \mathbb{R}^{p_2}$ are the irrelevant features. Specifically, $\boldsymbol{x}^1$ is sampled from a Gaussian mixture model, i.e., $\boldsymbol{x}^1 | z^1 \sim \mathcal{N}(\boldsymbol{\mu}_{z^1}^1, \boldsymbol{I}_{p_1})$, where z^1 is

Algorithm 1 Query-augmented active clustering with metric aggregation

Input: Data $\{\boldsymbol{x}_i\}$, budget T, number of clusters K.
Output: Cluster label $\boldsymbol{l}$.
Initialization: Single neighborhood $\mathcal{N} = \{N_1\}, N_1 = \{\boldsymbol{x}_1\}$, where $\boldsymbol{x}_1$ is randomly selected. Let $\mathcal{S} = \mathcal{D} = \varnothing$ and $t = 0$.
while $t \leq T$, **repeat**:

1. (*Active query*) Select the most informative instance $\boldsymbol{x}^*$ to minimize (5). Sort $N_i \in \mathcal{N}$ in decreasing order of $p(\boldsymbol{x}^* \in N_i)$, query $\boldsymbol{x}^*$ against an instance $x_1 \in N_1$, update $\mathcal{S}$ or $\mathcal{D}$ according to the feedback, $t \leftarrow t + 1$.

2. (*Metric Learning*) Train metric A^t with augmented queries (3).

3. Repeat step 1 and 2 for the rest of the neighborhood until $t > T$ or a similar link between $\boldsymbol{x}^*$ and $\boldsymbol{x}_i$ is found. Let $N_i = N \cup \{\boldsymbol{x}^*\}$. If no similar link is found, treat $\boldsymbol{x}^*$ as a new neighborhood. Let $N^* = \{\boldsymbol{x}^*\}$, and $\mathcal{N} = \mathcal{N} \cup N^*$.

(*Metric aggregation*) Compute $\mathcal{G}^T$ based on $\{A^t\}_{t=1}^T$, solve for $\widehat{A}$ with selective penalty (4).
(*Semi-supervised clustering*) Cluster the instances with the learned metric $\widehat{A}$ using PCKmeans.

uniformly distributed over $1, \cdots, p_1$, and the elements of $\boldsymbol{\mu}^1_{z^1}$ are all zero except the z^1th element equals c. Here c denotes the distance between the center of the clusters and the origin. The larger c is, the easier the data can be clustered. We let the cluster label $\ell = z^1$ and the number of clusters $K = p_1$, so the cluster memberships are fully determined by the first p_1 features. For the irrelevant features, we let $\boldsymbol{x}^2 | z^2 \sim \mathcal{N}(\boldsymbol{\mu}^2_{z^2}, \boldsymbol{I}_{p_2})$, where z^2 is uniform over $1, \cdots, p_2$, and the elements of $\boldsymbol{\mu}^1_{z^1}$ are all zero except the z^1th element equals c. An illustration of the simulation data with $p_1 = p_2 = 3$ is shown in Figure 1. In this experiment, we select $p_1 = 6$, $p_2 = 3$, $c = 5$ and $K = 6$. We

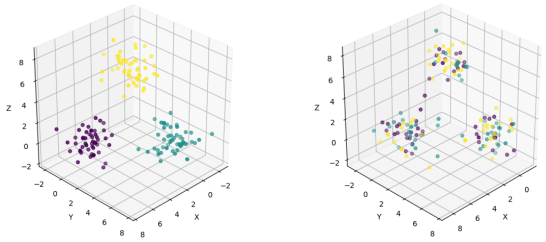


Figure 1: Simulated data with $p1 = p2 = 3$ and $K = 3$, showing the first three features (*left*) and the last three features (*right*). Each color shows the cluster label of data points, determined by the first three features.

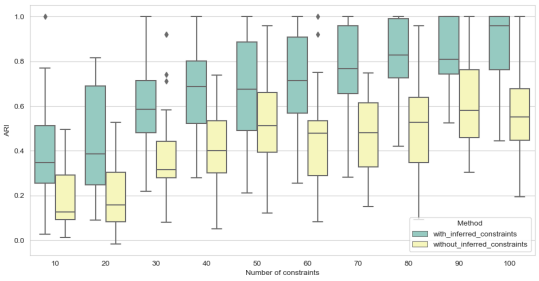


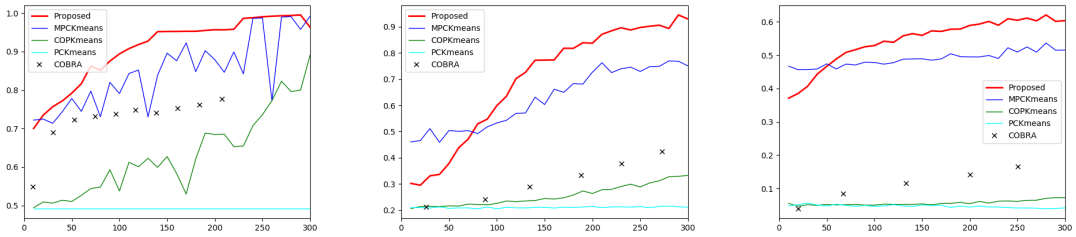
Figure 2: The ARI comparisons of the simulation setting with $p_1 = 6, p_2 = 3$ and $c = 5$ using the proposed method with (*green*), and without (*yellow*) augmented constraints, based on 30 replications for each number of constraints.

compare cases with or without augmented constraints $\widetilde{\mathcal{S}}$ and $\widetilde{\mathcal{D}}$. Their performances are evaluated using the adjusted random index (ARI) , where a higher ARI indicates more consistent clustering result with the truth. Figure 2 demonstrate that incorporating the augmented constraints improves the clustering performance under variant numbers of queried constraints. The advantage is more obvious when the number of constraints increases since the similarity matrix Y is less sparse and the augmented constraints are more accurate.

Real data We apply the proposed method on three real datasets with high dimensional features. The first dataset is the breast cancer diagnostic data which contains 569 samples with 30 features

and 2 clusters. The second dataset is MEU-Mobile dataset which records 71 keystroke features of 9 users with 459 samples in total. The third dataset is the urban land cover dataset which contains 675 multi-scale remote sensing images. Figure 3 shows the proposed method performs the best compared to other semi-supervised clustering approaches based on the average ARI with different number of constraints, including COP-Kmeans Wagstaff et al. (2001), pairwise constrained Kmeans (PCKmeans) Basu et al. (2004), metric pairwise constrained Kmeans (MPCKmeans) Bilenko et al. (2004) and constraint-based repeated aggregation (COBRA) Van Craenendonck et al. (2018).

Furthermore, we investigate the interpretability of the selected features of urban land cover data. We compare the weights of each feature in a decreasing order as shown in Figure 4. The top three features extracted by the proposed method correspond to the normalized difference vegetation index (NDVI) on three different resolution scales, respectively. With all the features, the Kmeans results an ARI of 0.03, however, with the extracted three features, the ARI increases to 0.29 without imposing any pairwise constraints. This further implies that the proposed method is able to identify the underlying feature space which is highly interpretable.



(a) breast cancer (b) MEU-Mobile (c) Urban land cover

Figure 3: Performance comparison on real data. Competing methods integrate NPU strategy.

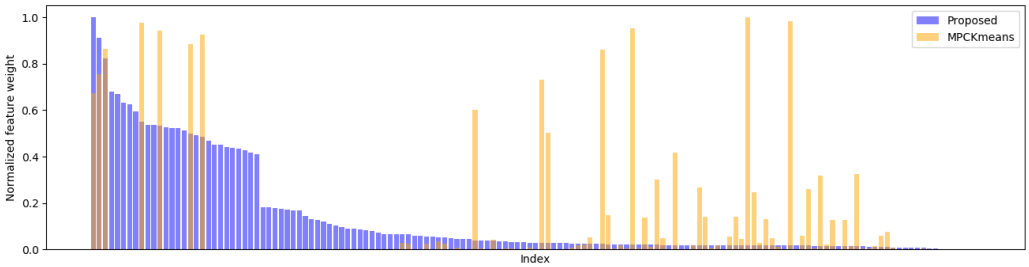


Figure 4: Estimated weights of 147 features against the feature index of the urban land cover dataset, showing the group structure using the proposed method compared with MPCKmeans.

5. Discussion

In this paper, we propose to impute the membership matrix based on the information given pairwise constraints. The augmented pairwise relationships provide extra information to the clustering, which may be typically ignored in the existing methods. In addition, we present a novel method to integrate the history training information during the active clustering procedure. We add penalty selectively to the potentially unrelated features, and utilize the trained metric matrix to recover the underlying feature space. The proposed method is able to recover the true features and outperforms the existing methods under the high-dimension setting with a limited number of pairwise constraints. The proposed framework is also suitable for online training. Both constraint augmentation and metric aggregation can be adapted into an incremental way without retraining each time new constraints are added to improve the computation efficiency, which are worth future investigation.

References

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active Semi-Supervision for Pairwise Constrained Clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 333–344. Society for Industrial and Applied Mathematics, April 2004. ISBN 978-0-89871-568-2 978-1-61197-274-0. doi: 10.1137/1.9781611972740.31. URL https://epubs.siam.org/doi/10.1137/1.9781611972740.31.

Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J. Mooney. Probabilistic Semi-Supervised Clustering with Constraints. In Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, pages 73–102. The MIT Press, September 2006. ISBN 978-0-262-03358-9. doi: 10.7551/mitpress/9780262033589.003.0005. URL http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262033589.001.0001/upso-9780262033589-chapter-5.

Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Twenty-first international conference on Machine learning - ICML '04*, page 11, Banff, Alberta, Canada, 2004. ACM Press. doi: 10.1145/1015330.1015360. URL http://portal.acm.org/citation.cfm?doid=1015330.1015360.

Arijit Biswas and David Jacobs. Active Image Clustering with Pairwise Constraints from Humans. *International Journal of Computer Vision*, 108(1-2):133–147, May 2014. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-013-0680-6. URL http://link.springer.com/10.1007/s11263-013-0680-6.

Ian Davidson, Kiri L Wagstaff, and Sugato Basu. Measuring constraint-set utility for partitional clustering algorithms. In *European conference on principles of data mining and knowledge discovery*, pages 115–126. Springer, 2006.

Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Active semi-supervised fuzzy clustering for image database categorization. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval - MIR '05*, page 9, Hilton, Singapore, 2005. ACM Press. ISBN 978-1-59593-244-0. doi: 10.1145/1101826.1101831. URL http://portal.acm.org/citation.cfm?doid=1101826.1101831.

Steven C.h. Hoi, Wei Liu, and Shih-Fu Chang. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(3):1–26, August 2010. ISSN 15516857. doi: 10.1145/1823746.1823752. URL http://portal.acm.org/citation.cfm?doid=1823746.1823752.

Yifen Huang and Tom M. Mitchell. Text clustering with extended user feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 413, Seattle, Washington, USA, 2006. ACM Press. ISBN 978-1-59593-369-0. doi: 10.1145/1148170.1148242. URL http://portal.acm.org/citation.cfm?doid=1148170.1148242.

Zhengdong Lu. Semi-supervised clustering with pairwise constraints: A discriminative approach. In *Artificial Intelligence and Statistics*, pages 299–306, 2007.

P. K. Mallapragada, R. Jin, and A. K. Jain. Active query selection for semi-supervised clustering. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, December 2008. doi: 10.1109/ICPR.2008.4761792.

Gang Niu, Bo Dai, Makoto Yamada, and Masashi Sugiyama. SERAPH: Semi-supervised Metric Learning Paradigm with Hyper Sparsity. *arXiv:1105.0167 [cs, stat]*, May 2011. URL http://arxiv.org/abs/1105.0167. arXiv: 1105.0167.

J Ross Quinlan et al. Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730, 1996.

Toon Van Craenendonck, Sebastijan Dumancic, and Hendrik Blockeel. COBRA: A Fast and Simple Method for Active Clustering with Pairwise Constraints. *arXiv:1801.09955 [cs, stat]*, January 2018. URL `http://arxiv.org/abs/1801.09955`. arXiv: 1801.09955.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *ICML*, 2001.

Eric P. Xing, Michael I. Jordan, Stuart J. Russell, and Andrew Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

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, January 2014. ISSN 1041-4347. doi: 10.1109/TKDE.2013.22.