Active Clustering with Unknown Evaluation Metric

Illinois Statistics Office

1 Problem Formulation

Given N samples $X_i \in \mathbb{R}^m,\ i = 1, \cdots, N$ and a feature map $f : \mathbb{R}^m \to \mathbb{R}^p$ which maps X_i to a p-dimensionl feature space. For the clustering task, we hod as a function g to map the features to K clusters: $g : \mathbb{R}^p \to \mathbb{R}, f(X) \longmapsto k, k = 1, \cdots, K$. Additionally, we define a evaluation function f to estimate the performance of the clustering method. Usually for clustering, this metric can be adjusted rand index [1], mutual information based scores [2] and so on. In our case, however, this evaluation metric is unknown and relies on the feedback of the human (or the oracle). Denote the function space of g as $\mathcal{G}$, then the problem can be described as

$$\max_{g \in \mathcal{G}} h(g \circ f(\mathbf{X})). \tag{1}$$

However, since we cannot exhaust the choices of $g \in \mathcal{G}$, we usually constrains our solution on a parameterized space with parameter $\boldsymbol{\theta} \in \mathbb{R}^q$, so that the optimization can be done on a q-dimensional Euclidean space. Denote the corresponding function as $g_{\boldsymbol{\theta}}$, then the problem (1) can be approximated by

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^q} h(g_{\boldsymbol{\theta}} \circ f(\mathbf{X})) \tag{2}$$

2 Framework and Discussion

First of all, to optimize $\boldsymbol{\theta}$, we need to get a reasonable estimation of h, which relies on the human interaction. We call the procedure of selecting samples and asking human for feedback as a "query". For each query, a pair of parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are chosen and the corresponding clustering method $g_{\boldsymbol{\theta_1}}$ and $g_{\boldsymbol{\theta_2}}$ are applied. The human then chooses $D < N$ samples and determine which clustering method is better, i.e., to inference if

$$h(g_{\boldsymbol{\theta_1}} \circ f(\boldsymbol{X})) > h(g_{\boldsymbol{\theta_2}} \circ f(\boldsymbol{X})) \tag{3}$$

based on the performance of the D chosen examples. The human interaction will continue until the budget runs out or the human fails to distinguish the difference of the performance of $g_{\boldsymbol{\theta}_1}$ and $g_{\boldsymbol{\theta}_2}$.

In an active learning framework, the key of solution to (2) is to design an consistent and efficient query strategy to maximize $h(g_{\boldsymbol{\theta}} \circ f(\mathrm{X}))$ with limited attempts of $\boldsymbol{\theta}$.

A sub-problem is to explore a sampling method of D samples so they will represent the behavior of the two clustering method reasonably well.

For the sub-problem, an intuitive way is to select the samples that on the boundaries of the clusters like [3]. After applying the clustering method $g_{\boldsymbol{\theta}_1}$ and $\boldsymbol{\theta_2}$, we can get two boundary setsB_1 and B_2 according to the two clustering results respectively. Then we can select the D samples from $B_1 \cup B_2$. However, the boundary based sampling method may result in sampling bias [4].

For the main problem, our setting is quite different from that in the conventional active learning context. To be specific, in active supervised learning, the evaluation metric h is always known. And instead of comparing the behavior of $g_{\boldsymbol{\theta}_1}$ and $g_{\boldsymbol{\theta}_2}$ to get an estimation of h, the performance of the model can be computed and compared directly. The human interaction here is to label new samples, which are used to train optimal $\boldsymbol{\theta}$. And based on optimized $\boldsymbol{\theta}$, people choose which unlabeled sample to be presented to the oracle. The loop continues until the budget is exhausted. In a more complex case [5][6], the human interaction is not limited to labeling the samples but also the features extracted by f.

On the other hand, for active unsupervised learning task, the metric h is unknown, but it is under some constraints and does not require human to determine. For example, [3] add "must-link" and "cannot-link" to the loss function. This is equivalent to consider $\ell_{\boldsymbol{\theta}} = h \circ g_{\boldsymbol{\theta}}$ and optimize it as a whole.

3 Literature review

- [7] propose a scheme to improve the efficiency of clustering by performing the clustering procedure on the selected sample data instead of whole dataset. The original data is the proximity data that are similarity values obtained by comparing pairs of entities. The selection of data for clustering is based on the expected value of information which measures the gain in classification accuracy by incorporating addition data.

- [6] propose a framework of active supervise learning, whose goal is to learn the output's distribution conditioned on the input features. The full conditional distribution of output is modeled through conditional random field. During each iteration in training, the algorithm ask user to label a feature selected by some information entropy criterion to reduce the model uncertainty. Specifically, this is not feature selection since we are not determining which features will be part of the model but determining the features for which supervisory feedback will be most helpful to the model.

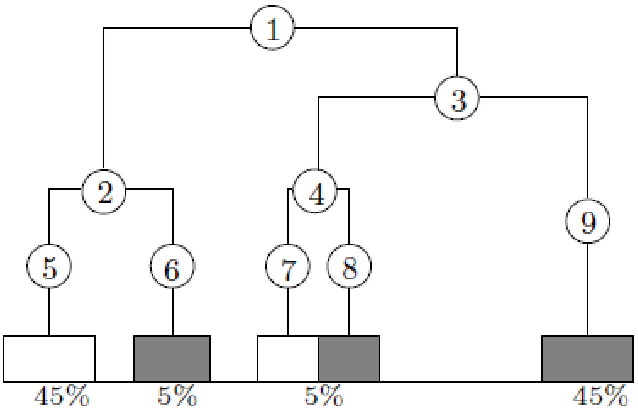

Figure 1: Hierarchical clustering during active learning, from [6]

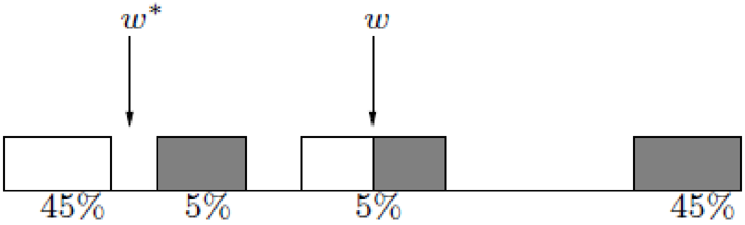

Figure 2: Shortcoming of the selection on the boundary and the consequent sampling bias

- [4] propose a hierarchical clustering based active learning algorithm. It divides the whole process into three steps: 1. hierarchical clustering 2. clustering adaptive sampling and labeling 3. supervised learning on the resulting fully labeled data. During step 2, it computes the weight based on the purity of the (labeled) subtrees and Wald's average linkage clustering. The advantage of the hierarchical tree based method is that it will avoid the "sampling bias" to some extent. The query part is asking the human to label/ classify the samples. (**The crieterion for pruning the hierarchical tree needs further review.(admissible set)**)

- [3] combine the fuzzy clustering and active supervised learning. The boundary of the clusters are defined as the data points with the lowest probability that belongs to the cluster. The human is then asked to add constraints of "must-link" and "cannot-link" constraints of the "most valuable pairs" (MVP) based on the extended boundary to refine the clustering as shown in the figure below.

Notice: 1) The inputs in their experiment are extracted features (histograms) of dimension 100 instead of the original images. 2) They use Mahalanobis distance to compute the dissimilarity between the samples.

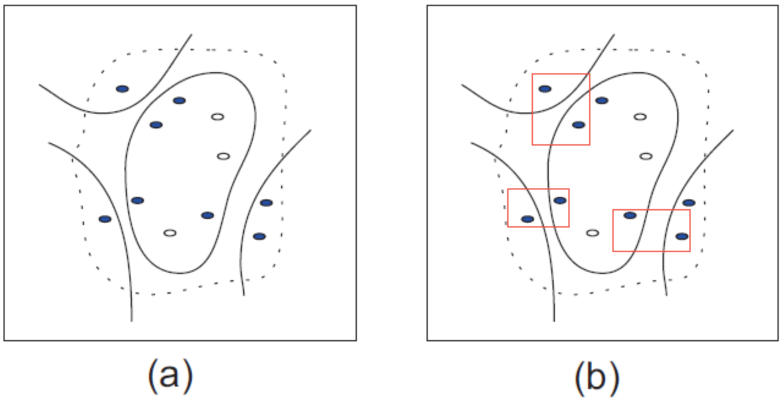

Figure 5: Informative pairwise constraints can be selected on the extended boundary (dashed line) of the less-defined cluster

Figure 3: Figure from [3]

- [5] propose a framework of active learning to simultaneously using the user's feedback about the selected instances and features to improve the learning. During each iteration the model is retrained based on the instances and features selected from candidates, which are obtained based on information criterion. The training procedure stops when the budget of user supervision is reached. The main advantage is that the human-chosen features significantly accelerate learning compared to traditional active learning.

References

[1] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

[2] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary?

In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA, 2009. ACM.

[3] 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.

[4] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 208–215, Helsinki, Finland, 2008. ACM Press.

[5] Hema Raghavan, Omid Madani, and Rosie Jones. Active learning with feedback on features and instances. *Journal of Machine Learning Research*, 7(Aug):1655–1686, 2006.

[6] Gregory Druck, Burr Settles, and Andrew McCallum. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP '09*, volume 1, page 81. Association for Computational Linguistics, 2009.

[7] Thomas Hofmann and Joachim M. Buhmann. Active data clustering. In *Advances in Neural Information Processing Systems*, pages 528–534, 1998.