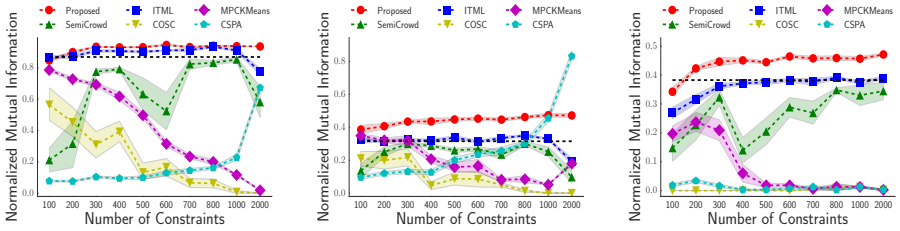
Supplementary Material: Clustering from Multiple Uncertain Experts

1 Results on *Dermatology, Glass, Heart*

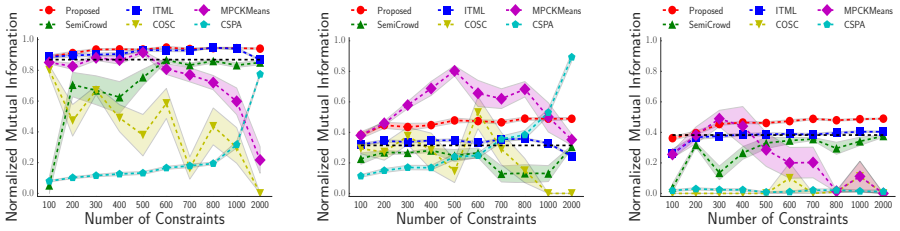
1.1 Case 1: Unequal Accuracies



(a) Dermatology (b) Glass (c) Heart

Figure 1: After setting accuracy parameter $\boldsymbol{\alpha} = \boldsymbol{\beta} = (0.95, 0.85, 0.75, 0.65, 0.55)$ to make different experts have unequal accuracies, each plot shows NMI against the number of constraints for competing approaches on one UCI dataset.

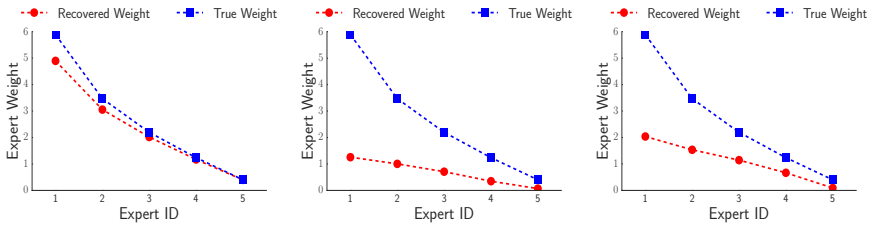
1.2 Case 2: Equal Accuracies



(a) Dermatology (b) Glass (c) Heart

Figure 2: After setting accuracy parameter $\boldsymbol{\alpha} = \boldsymbol{\beta} = (0.9, 0.9, 0.9, 0.9, 0.9)$ to make different experts have equal accuracies of good quality, each plot shows NMI against the number of constraints for competing approaches on one UCI dataset.

1.3 Expert Weights in Case 1



(a) Dermatology (b) Glass (c) Heart

Figure 3: After setting accuracy parameter $\boldsymbol{\alpha} = \boldsymbol{\beta} = (0.95, 0.85, 0.75, 0.65, 0.55)$ to make different experts have unequal accuracies, each plot shows the recovered expert weights (in red) and true expert weights (in blue) for all experts on a UCI dataset.

2 Parameter Settings

All approaches need to specify K, the number of clusters. For UCI benchmark datasets, we set K to be the number of classes. For the COPD dataset, we set $K = 4$ according to a recent study on COPD subtyping [2].

Proposed: We set $\lambda = 10^{-3}$ in all experiments and observe that the results are very stable for a wide range of values of λ (from 10^{-5} to 10^{-1}).

SemiCrowd: We set d_0, d_1, two thresholds used to filter out uncertain sample pairs in the average similarity matrix, are set to be 0 and 0.8 respectively.

ITML: As is suggested by the author [3], we set lower and upper bounds associated with the constraint terms to be the 5^{th} and 95^{th} percentiles of the observed distribution of distances between pairs of points within the dataset.

COSC: We use a Gaussian kernel to construct the similarity matrix and set the scale parameter to be the median of pairwise Euclidean distances [4]. After specifying the constraints, we directly run the author's MATLAB implementation [6].

MPCKMeans: We specify the constraints according to the instructions listed on the author's website and directly run the author's Java implementation [1].

CSPA: After computing the average similarity matrix, we use spectral clustering [5] to obtain the cluster labels.

3 Complexity Analysis

The time complexity of the optimization algorithm is dominated by gradient computation w.r.t. W and $\mathbf{b}$, which can be written as $\mathcal{O}\left((K+M)l^2 + Kld\right)$, where K is the number of clusters, M is the number of experts, d is the number of features in the data matrix, l is the number of samples provided with constraints by the experts, therefore l is upper bounded by $\min(2c, n)$, where c is the number of constraints and n is the number of samples. When the number of constraints grows large enough to cover all the samples in the dataset, the time complexity becomes quadratic w.r.t sample size n.

The space complexity involves the data matrix $X \in \mathbb{R}^{n \times d}$, experts' constraints stored in sparse matrices with cost $\mathcal{O}(cM)$, square matrix $\gamma \in \mathbb{R}^{l \times l}$. Therefore, the space complexity is at the scale $\mathcal{O}(nd + cM + l^2)$. When the number of constraints grows large enough to cover all the samples in the dataset, the space complexity becomes quadratic w.r.t sample size n.

References

[1] M. Bilenko. Mpckmeans implementation, 2004.

[2] P. J. Castaldi, J. Dy, J. Ross, Y. Chang, G. R. Washko, D. Curran-Everett, A. Williams, D. A. Lynch, B. J. Make, J. D. Crapo, et al. Cluster analysis in the copdgene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*, 2014.

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007.

[4] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[5] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *NIPS*, 2:849–856, 2002.

[6] S. Rangapuram. Cosc implementation, 2012.