



Information-theoretic Classification Accuracy (ITCA)

How to Combine Ambiguous Outcome Labels in Multi-class Classification?

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

<http://jsb.ucla.edu>

Data quality challenge in classification: outcome labeling

- Outcome labeling ambiguity and subjectiveness
 - common in biomedical applications, e.g., disease diagnosis/prognosis
 - e.g., inconsistent labels by different graders
- Ambiguous outcome labels → deteriorated prediction accuracy



Motivating application

- Traumatic brain injury (TBI) patients: rehabilitation or longer hospital stay?
- Predict patients' rehabilitation outcomes (each has $K_0 = 7$ levels) from admission features
- The prediction accuracy is low



Motivating application

- Traumatic brain injury (TBI) patients: rehabilitation or longer hospital stay?
- Predict patients' rehabilitation outcomes (each has $K_0 = 7$ levels) from admission features
- The prediction accuracy is low

Combine adjacent, ambiguous outcome levels?



1. **Classification in the presence of labeling noise** (Frénay and Verleysen, 2013)
 - (1) using **robust losses** or **ensemble learning**
(Freund, 2001; Beigman and Klebanov, 2009; ...)
 - (2) **removing data points** that are likely mislabeled
(Zhang et al., 2006; Thongkam et al., 2008; ...)
 - (3) modeling labeling noise using **data generative models**
(Swartz et al., 2004; Kim and Ghahramani, 2008; ...)



2. Set-valued prediction

(1) conformal prediction

(Vovk et al., 2005; Balasubramanian et al., 2014; ...)

(2) set-based utility maximization

(Corani and Zaffalon, 2008; Del Coz et al., 2009; Zaffalon et al., 2012; Mortier et al., 2021)



2. Set-valued prediction

(1) conformal prediction

(Vovk et al., 2005; Balasubramanian et al., 2014; ...)

(2) set-based utility maximization

(Corani and Zaffalon, 2008; Del Coz et al., 2009; Zaffalon et al., 2012; Mortier et al., 2021)

Existing methods do not guide **global class combination**



Trade-off between classification accuracy and resolution

Classification accuracy can be boosted at the cost of losing resolution

- Combining all outcome labels into one, we obtain a 100% accurate classifier



Trade-off between classification accuracy and resolution

Classification accuracy can be boosted at the cost of losing resolution

- Combining all outcome labels into one, we obtain a 100% accurate classifier

A principled method is called to balance the trade-off:

- How to characterize the “resolution”?
- How to properly balance the accuracy and resolution?



Trade-off between classification accuracy and resolution

Classification accuracy can be boosted at the cost of losing resolution

- Combining all outcome labels into one, we obtain a 100% accurate classifier

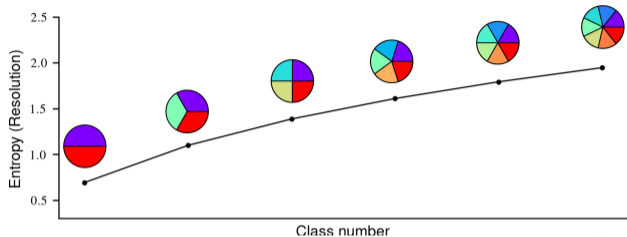
A principled method is called to balance the trade-off:

- How to characterize the “resolution”?
- How to properly balance the accuracy and resolution?

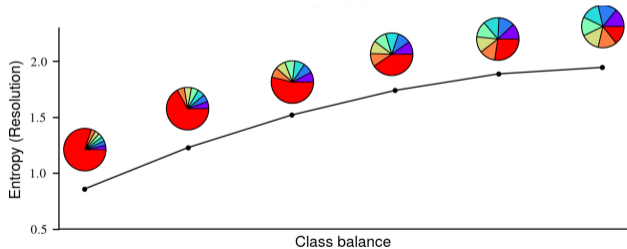
We proposed the information-theoretic classification accuracy (ITCA)



Entropy of outcome label distribution characterizes the “resolution”



For balanced classes:
the larger the class number,
the higher the resolution

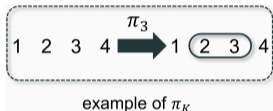


Given the number of classes:
the more balanced,
the higher the resolution



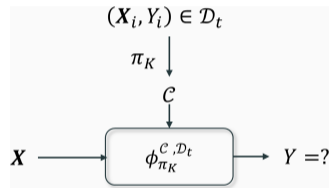
Notations for class combination

- $\pi_K: [K_0] \rightarrow [K]$ where $K < K_0$



$$\pi_3^{-1}(1) = \{1\}, \pi_3^{-1}(2) = \{2, 3\}, \pi_3^{-1}(3) = \{4\}$$

- Given the training data \mathcal{D}_t , a classification algorithm \mathcal{C} , and a class combination π_K , denote the trained classifier by $\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}$



Out-of-sample accuracy (ACC)

Given class combination π_K , training data \mathcal{D}_t , and classification algorithm \mathcal{C}
 \implies **classifier** $\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}$, whose ACC is evaluated on validation data \mathcal{D}_v

$$\begin{aligned} \text{ACC}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) &:= \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i)) \\ &= \sum_{k=1}^K \underbrace{p_{\pi_K}^{\mathcal{D}_v}(k)}_{\substack{\text{proportion of the} \\ \text{combined class } k}} \cdot \underbrace{\frac{\sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = k, \pi_K(Y_i) = k)}{1 \vee \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\pi_K(Y_i) = k)}}_{\substack{\text{conditional accuracy of } \phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t} \\ \text{for the combined class } k}}, \end{aligned}$$

where $p_{\pi_K}^{\mathcal{D}_v}(k) := \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\pi_K(Y_i) = k)$



Out-of-sample accuracy (ACC)

Given class combination π_K , training data \mathcal{D}_t , and classification algorithm \mathcal{C}
 \implies **classifier** $\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}$, whose ACC is evaluated on validation data \mathcal{D}_v

$$\begin{aligned} \text{ACC}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) &:= \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i)) \\ &= \sum_{k=1}^K \underbrace{p_{\pi_K}^{\mathcal{D}_v}(k)}_{\text{proportion of the combined class } k} \cdot \underbrace{\frac{\sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = k, \pi_K(Y_i) = k)}{1 \vee \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\pi_K(Y_i) = k)}}_{\text{conditional accuracy of } \phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t} \text{ for the combined class } k}, \end{aligned}$$

where $p_{\pi_K}^{\mathcal{D}_v}(k) := \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\pi_K(Y_i) = k)$

ACC is dominated by major classes



Information-theoretic classification accuracy (ITCA)

$$\text{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C})$$
$$:= \sum_{k=1}^K \underbrace{\left[-p_{\pi_K}^{\mathcal{D}_v}(k) \cdot \log(p_{\pi_K}^{\mathcal{D}_v}(k)) \right]}_{\text{contribution of the combined class } k \text{ to the entropy of } \pi_K(Y)} \cdot \underbrace{\frac{\sum_{(\mathbf{X}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}_i) = k, \pi_K(Y_i) = k)}{1 \vee \sum_{(\mathbf{X}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\pi_K(Y_i) = k)}}_{\text{conditional accuracy of } \phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t} \text{ for the combined class } k}},$$



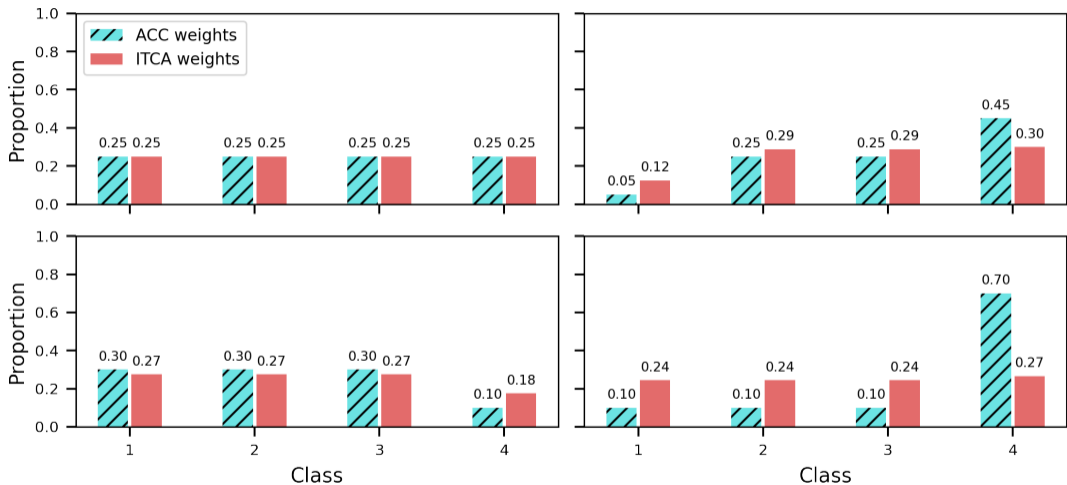
Information-theoretic classification accuracy (ITCA)

$$\text{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) := \sum_{k=1}^K \underbrace{\left[-p_{\pi_K}^{\mathcal{D}_v}(k) \cdot \log(p_{\pi_K}^{\mathcal{D}_v}(k)) \right]}_{\text{contribution of the combined class } k \text{ to the entropy of } \pi_K(Y)} \cdot \underbrace{\frac{\sum_{(\mathbf{X}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}_i) = k, \pi_K(Y_i) = k)}{1 \vee \sum_{(\mathbf{X}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\pi_K(Y_i) = k)}}_{\text{conditional accuracy of } \phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t} \text{ for the combined class } k},$$

- ITCA is entropy-weighted out-of-sample prediction accuracy
- ITCA is also a class-accuracy-weighted entropy



Comparison of class weights in ACC and ITCA



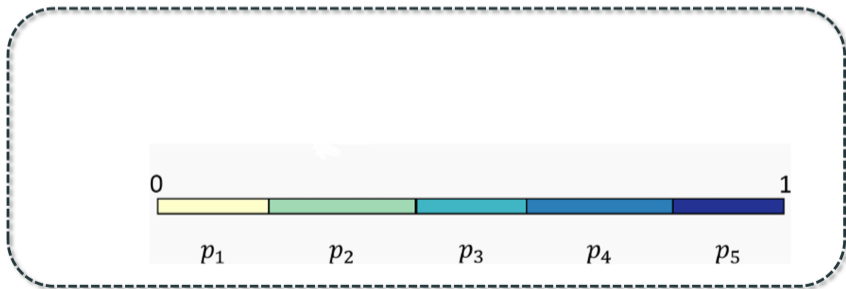
ITCA overweighs minor classes



Alternative definition of ITCA

$$\text{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) = \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} -\log p_{\pi_K}^{\mathcal{D}_v}(\pi_K(Y_i)) \cdot \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i)) ,$$

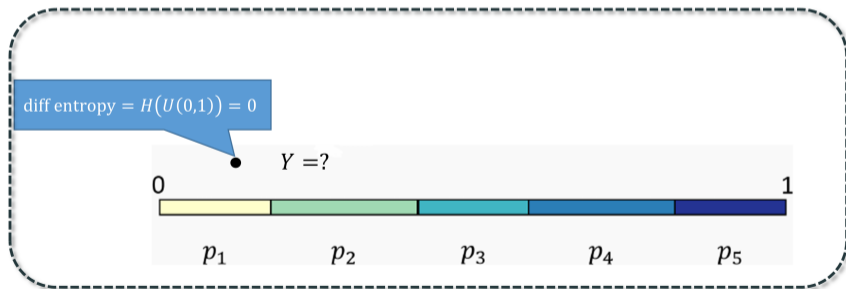
Represent π_K 's K combined classes as K non-overlapping intervals in $[0, 1]$



Alternative definition of ITCA

$$\text{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) = \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} -\log p_{\pi_K}^{\mathcal{D}_v}(\pi_K(Y_i)) \cdot \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i)) ,$$

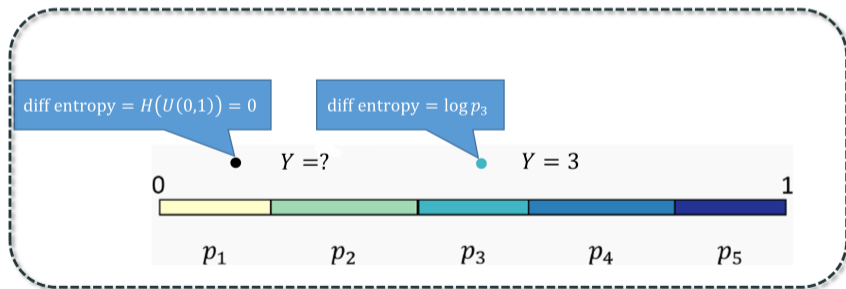
Represent π_K 's K combined classes as K non-overlapping intervals in $[0, 1]$



Alternative definition of ITCA

$$\text{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) = \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} -\log p_{\pi_K}^{\mathcal{D}_v}(\pi_K(Y_i)) \cdot \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i)),$$

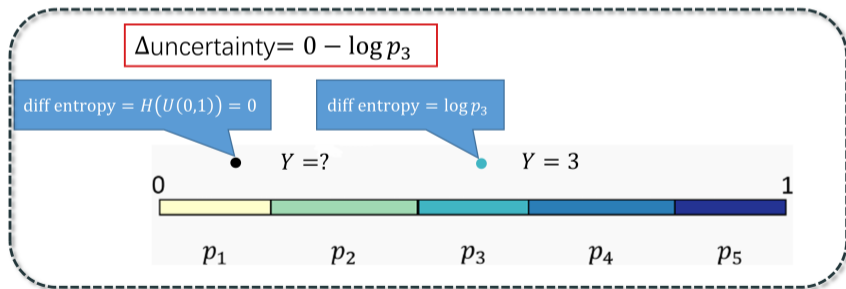
Represent π_K 's K combined classes as K non-overlapping intervals in $[0, 1]$



Alternative definition of ITCA

$$\text{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_v, \mathcal{C}) = \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} -\log p_{\pi_K}^{\mathcal{D}_v}(\pi_K(Y_i)) \cdot \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i)),$$

Represent π_K 's K combined classes as K non-overlapping intervals in $[0, 1]$



Five alternative criteria that may guide class combination (also our proposal)

Adjusted accuracy (AAC)

$$\text{AAC} := \frac{1}{|\mathcal{D}_v|} \sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \frac{\mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i))}{p_{\pi_K}^{\mathcal{D}_v}(\pi_K(Y_i))}$$

Combined Kullback-Leibler divergence (CKL)

$$\text{CKL} := D_{\text{KL}}(\hat{F}_{\pi_K, \mathcal{D}_v} \parallel \hat{F}_{\pi_{K_0}, \mathcal{D}_v}) + D_{\text{KL}}(\hat{F}_{\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}, \mathcal{D}_v} \parallel \hat{F}_{\pi_K, \mathcal{D}_v})$$

Prediction entropy (PE)

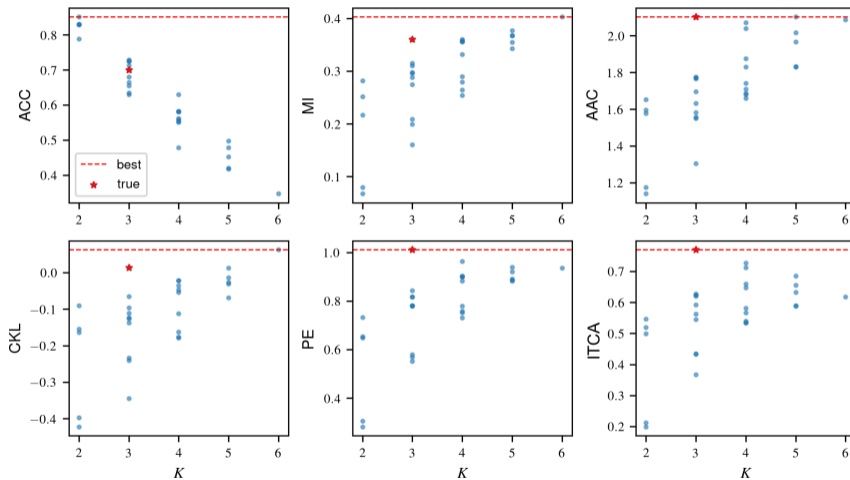
$$\text{PE} := \sum_{k=1}^K - \frac{\sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i) = k)}{|\mathcal{D}_v|} \cdot \log \left(\frac{\sum_{(\mathbf{x}_i, Y_i) \in \mathcal{D}_v} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{x}_i) = \pi_K(Y_i) = k)}{|\mathcal{D}_v|} \right)$$

Commonly used criteria

- **Accuracy (ACC)**
Classification
- **Mutual Information (MI)** Clustering



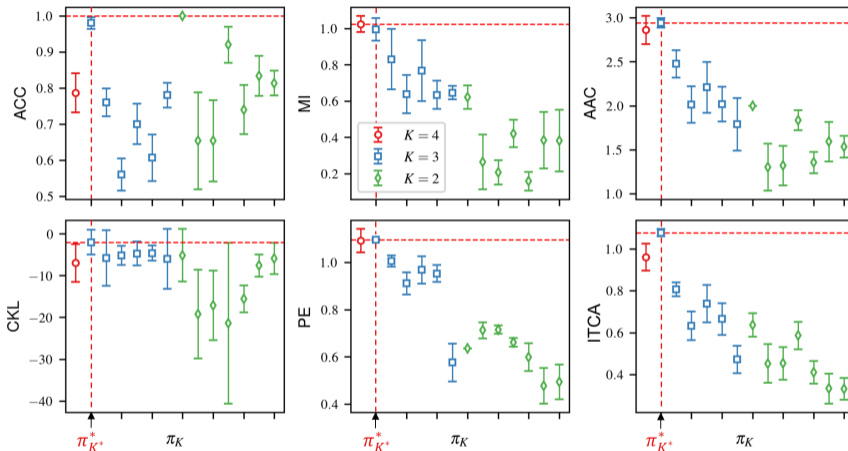
ITCA finds the true class combination (simulated data)



Simulated data with $K_0 = 6$ observed classes; $K^* = 3$ true classes; $\mathcal{C} = \text{LDA}$



ITCA finds the true class combination (the Iris data)



$K^* = 3$ classes (*setosa*, *versicolor*, and *virginica*); the *setosa* class is linearly separable from the other two classes; $K_0 = 4$ (the *setosa* class is randomly split into two equal-sized classes)



ITCA finds the true combination in most cases

Criterion	<u># successes</u>	Average	Max	<u># successes</u>	Average	Max
	<u># datasets</u>	Hamming	Hamming	<u># datasets</u>	Hamming	Hamming
		LDA			RF	
ACC	6/127	2.54	6	7/127	2.53	6
MI	7/127	2.51	6	11/127	2.33	6
AAC	15/127	2.02	6	15/127	1.98	6
CKL	3/127	3.68	6	5/127	2.87	5
PE	101/127	0.47	4	94/127	0.46	3
ITCA	120/127	0.12	3	120/127	0.08	2

The performance of six criteria on the 127 simulated datasets with $K_0 = 8$
LDA = linear discriminant analysis; RF = random forest



Exhaustive search is prohibitive even K_0 is moderate

The number of allowed class combinations π_K 's given K_0

Label Type	K_0					
	2	4	6	8	12	16
Nominal	1	14	202	4139	4213596	$\sim 10^{10}$
Ordinal	1	7	31	127	2047	32767

Two heuristic search strategies

- **Greedy search**: starting from π_{K_0} , in the k -th round, find the best combination among the allowed π_{K_0-k} 's that maximizes the ITCA
- **Breadth-first search (BFS)**: track all the combination that can improve ITCA at each round



Effectiveness of the greedy and BFS search strategies

Strategy	$\frac{\# \text{ successes}}{\# \text{ datasets}}$	Average	Max	Average # class
		Hamming	Hamming	combinations examined
Exhaustive	120/127	0.13	3	127.00
Greedy search	119/127	0.12	3	22.64
BFS	119/127	0.10	2	53.98
Greedy (pruned)	119/127	0.10	2	12.01
BFS (pruned)	119/127	0.10	3	27.41

Performance of ITCA using five search strategies and LDA on the 127 simulated datasets with $K_0 = 8$



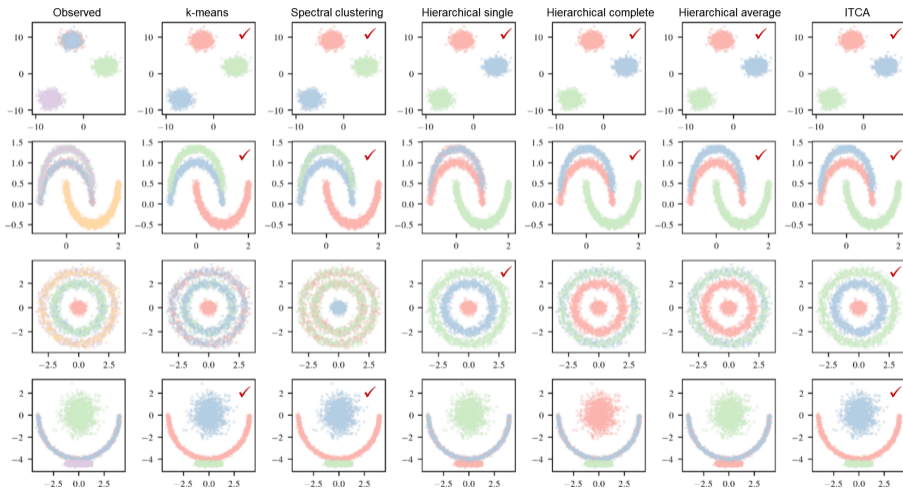
Using clustering algorithms to guide class combination?

- **K -means-based class combination:** compute the k_0 -th class center $(\sum_{i=1}^n \mathbb{I}(Y_i = k_0) \mathbf{X}_i) / (\sum_{i=1}^n \mathbb{I}(Y_i = k_0))$; use the K -means clustering to cluster the K_0 class centers into K^* clusters
- **Spectral-clustering-based class combination:** compute the K^* -dimensional spectral embeddings of $\mathbf{X}_1, \dots, \mathbf{X}_n$; apply the K -means-based class combination approach
- **Hierarchical-clustering-based class combination:** compute the K_0 class centers; apply the hierarchical clustering to the centers

For all clustering-based class combination approaches, K^* must be **pre-specified**



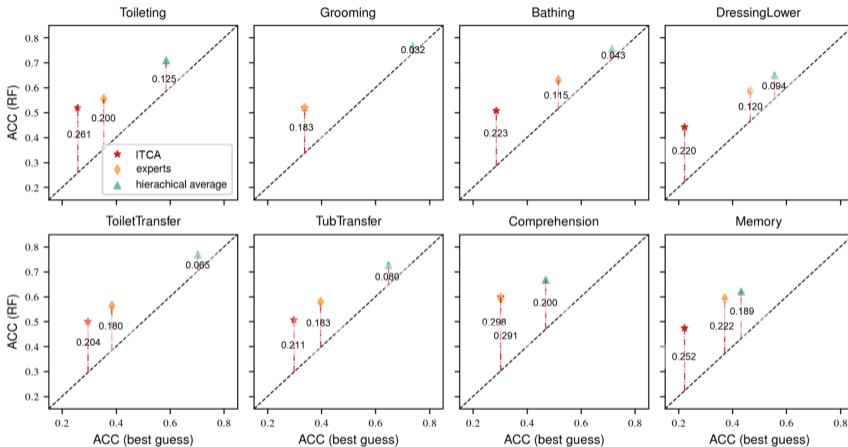
ITCA outperforms clustering-based class combination approaches



Only ITCA (\mathcal{C} = Gaussian kernel SVM) finds the true combination in all cases



Application 1: prognosis of rehabilitation outcomes of TBI patients



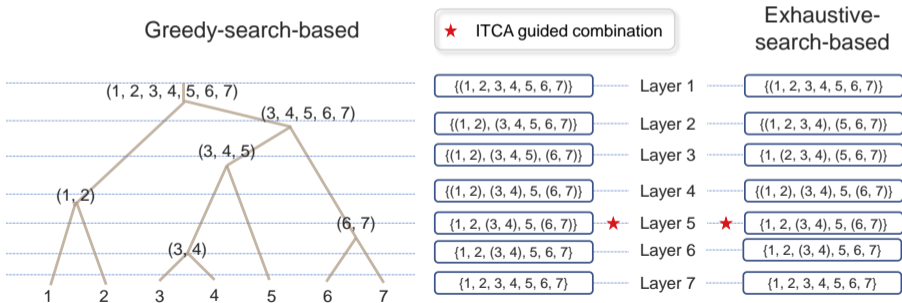
ITCA consistently leads to **more balanced** levels and a **more significant improvement** from the **best guess** (assigning every patient to the level that has the most patients)



ITCA allows multi-layer prediction

For each $K = 1, \dots, K_0$, choose the combination π_K that maximizes the ITCA

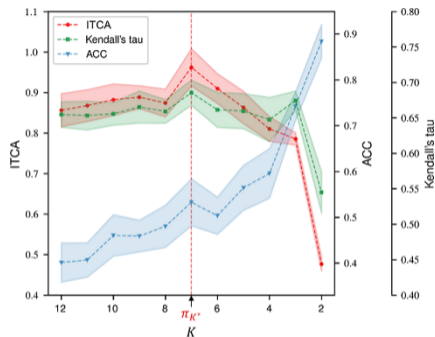
- **Nested-search-based:** classes in each layer are combined from the classes in the layer below
- **Exhaustive-search-based:** no nested constraint



Application 2: prediction of glioblastoma cancer patients' survival time

Glioblastoma cancer is one of the most aggressive cancer types

- **Task:** Predict patients' survival time
- **Approach 1:** survival analysis (Cox regression)
- **Approach 2:** discretize survival time (classification)
 - **Challenge:** How to define survival time intervals?
 - **Solution:** Discretize survival time into small intervals and combine them with ITCA



ITCA ($\mathcal{C} = \text{NN}$) vs. ACC vs. Kendall's tau



Application 2: prediction of glioblastoma cancer patients' survival time

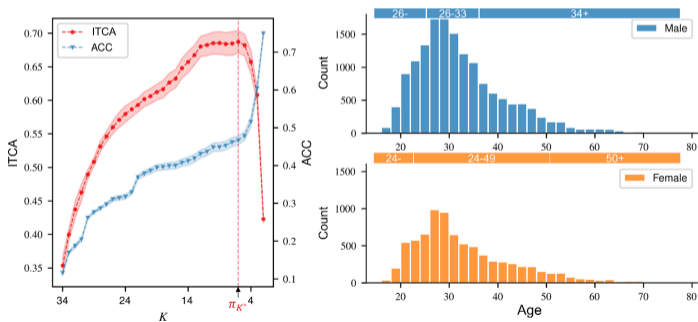
- We use a 3 layered neural network (NN) or logistic regression (LR) with a modified cross entropy loss function for censored data
- $K_0 = 12$
- ITCA finds $K = 7$ for LR and NN (with different π_K 's)

Model	ITCA	Kendall's tau	p-value
NN (K_0 survival time intervals)	0.8565 ± 0.0410	0.6547 ± 0.0181	$2.11e-14$
LR (K_0 survival time intervals)	0.6354 ± 0.0620	0.6024 ± 0.0244	$1.64e-11$
NN (ITCA-guided combined intervals)	0.9623 ± 0.0464	0.6855 ± 0.0178	$1.27e-15$
LR (ITCA-guided combined intervals)	0.8196 ± 0.0222	0.6236 ± 0.0240	$5.34e-10$
Cox regression (risk scores)	-	0.6303 ± 0.0542	$2.04e-13$



Application 3: prediction of user demographics using cell phone behavioral data

- Predicting the demographics (gender and age) of users using behavioral data is an essential task in advertising
- We first divided male and female users into 17 age groups
- ITCA with with XGBoost classification algorithm



TalkingData mobile user demographics data, 23556 users 818 features after processing

Application 4: detection of similar cell types inferred from scRNA-seq

- scRNA-seq data are commonly used to identify cell types
- However, the annotated cell types are often subjective
- ITCA provides a data-driven approach to detect similar cell types due to over-clustering
- Apply ITCA with LDA algorithm to Hydra data (25052 cells, 40 PCs, $K_0 = 38$ cell types) (Siebert et al., Science, 2019)



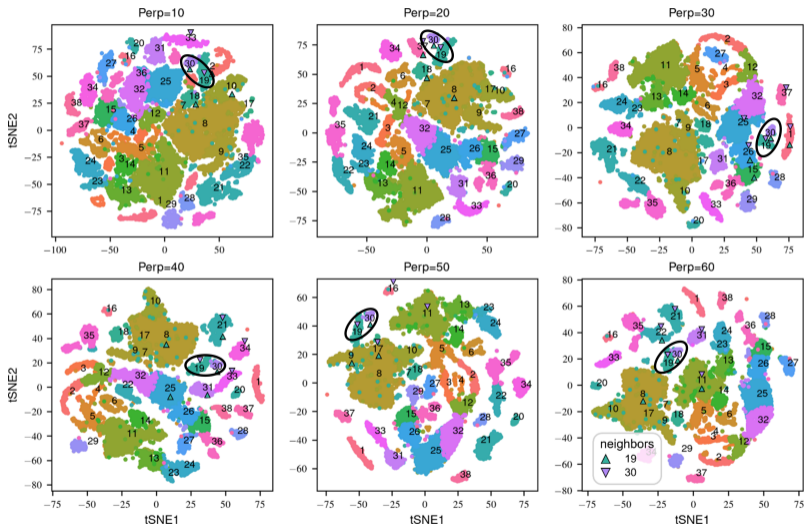
ITCA suggests that cell types 19 and 30 are similar

- Cell type 19 (endodermal epithelial cells in tentacles)
- Cell type 30 (endodermal epithelial cells in tentacle nematocytes—suspected phagocytosis doublets)
- Cell types 19 and 30 are consistent neighbors in t-SNE embedding

Perplexity	Neighbors of cell type 19	Neighbors of cell type 30
10	2, 10, 18, 30	19 , 25, 33
20	8, 18, 30	1, 19 , 37
30	1, 15, 25, 26, 30	1, 19 , 25, 37
40	8, 21, 25, 30 , 31	19 , 21, 33, 34
50	9, 17, 30	11, 16, 17, 19
60	8, 11, 22, 30	11, 19 , 21, 22, 31



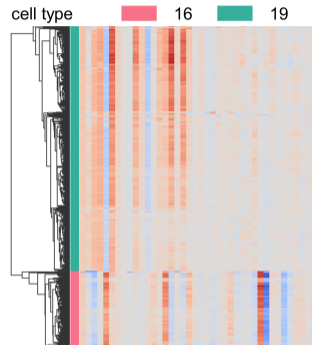
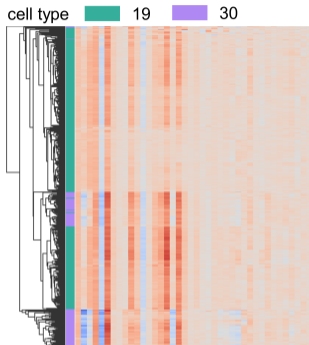
t-SNE embedding of Hydra data w.r.t different perplexities



Cell types 19 and 30 are consistent neighbors

Cell types 19 and 30 share similar gene expression patterns

- The gene expression patterns of cell types 19 and 30 are barely distinguishable
- As a control, cell types 16 and 19 are well separated



The heatmaps of the first 40 principal components

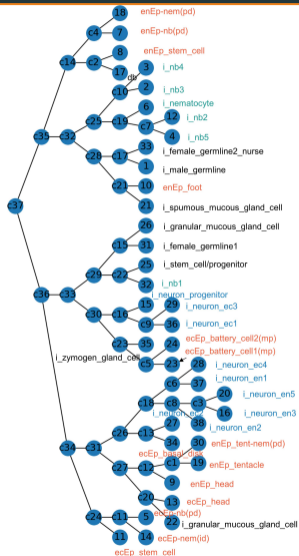
Cell type 19 has 458 cells

Cell type 30 has 134 cells

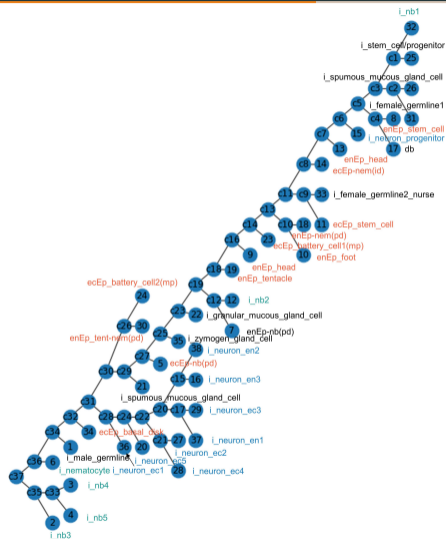
Cell type 16 has 143 cells



ITCA guides the construction of a cell-type hierarchy



ITCA



hierarchical clustering

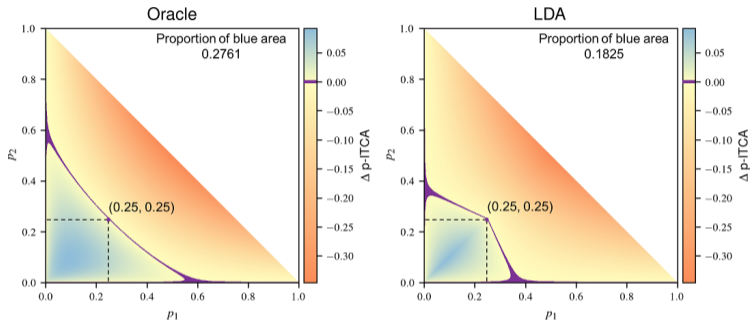


The choice of classification algorithm

- ITCA is adaptive to all classification algorithms
- ITCA is comparable across classification algorithms
- Users can choose the most suitable classification algorithms for different tasks
 - **Prediction**: a strong classification algorithm that maximizes ITCA
 - **Detection of similar classes**: a weak classification algorithm (e.g., LDA)



Theoretical analysis: class-combination regions

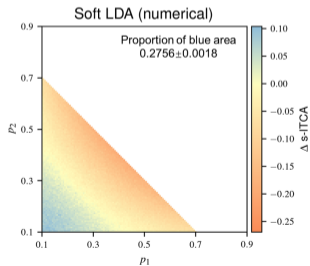
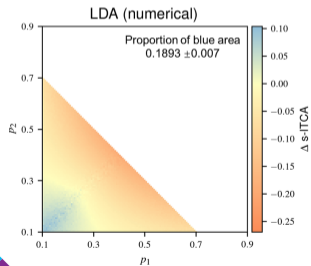
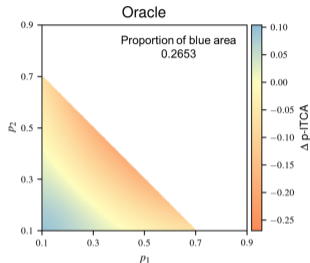
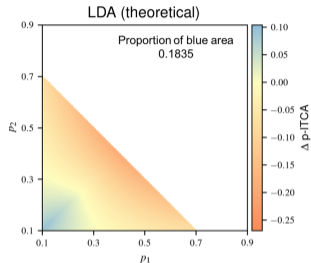


Blue area: the two classes will be combined

- p -ITCA will not combine the same classes when the proportion of the combined class is large



Enhance the ability of LDA for discovering the true class combination



Soft LDA

For probabilistic classification algorithms, soft prediction of class labels can better guide class combination



- ITCA guides the combination of ambiguous outcome labels by balancing classification accuracy and resolution
- Extensive simulation studies verify the effectiveness of ITCA
- Multiple real-world applications demonstrate the application potential of ITCA
- Future work:
incorporate user-specified constraints on class combination
allow one combined class (?)

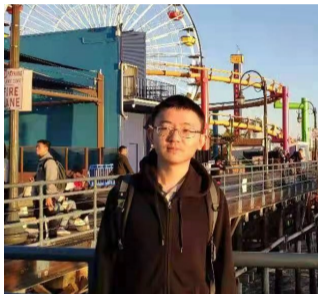


Acknowledgements

Zhang, C., Chen, Y.E.,
Zhang, S., and Li, J.J.

[arXiv:2109.00582](https://arxiv.org/abs/2109.00582)

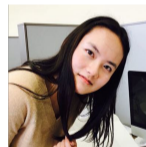
*Journal of Machine
Learning Research*
23(341):1–65.



Chihao Zhang
AMSS, CAS



Prof. Shihua Zhang
AMSS, CAS



Yiling Elaine Chen
UCLA



Appendix

Censored cross entropy (CCE)

The commonly used loss function for NN is the cross entropy (CE):

$$\text{CE} = - \sum_{i=1}^K I(Y_i = k) \log[\phi(X_i)]_k,$$

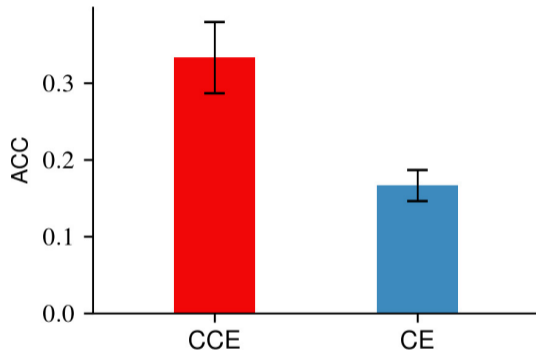
is not suitable for censored data. We propose the censored cross entropy (CCE):

$$\begin{aligned} \text{CCE} = & - \sum_{k=1}^K O_i I(Y_i = k) \log[\phi(X_i)]_k \\ & - (1 - O_i) \sum_{k > Y_i} \frac{p_k}{1 - \sum_{l \leq Y_i} p_l} \log[\phi(X_i)]_k, \end{aligned}$$

where O_i is binary and $O_i = 0$ indicates that the data is right censored.



CCE improves the accuracy



Performance of neural networks with CCE and CE as the loss functions, respectively.



Population-level ITCA (p-ITCA)

We define the **population-level ITCA (p-ITCA)** of π_K as

$$\text{p-ITCA}(\pi_K; \mathcal{D}_t, \mathcal{C}) := \sum_{k=1}^K [-\mathbb{P}(\pi_K(Y) = k) \log \mathbb{P}(\pi_K(Y) = k)] \cdot \mathbb{P}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}) = \pi_K(Y) | \pi_K(Y) = k)$$


Definition (oracle classifier)

Given K_0 observed classes, let $S \subseteq [K_0]$ be a set of classes that share the same distribution. A classifier $\phi_{\pi_{K_0}}^*$ is an oracle classifier if that for any (\mathbf{X}_i, Y_i) where $Y_i \in S$, $\phi_{\pi_{K_0}}^*$ predicts the label $s \in S$ with equal probability

Definition (class combination boundary curve)

$K_0 > 2$, there exist two classes $S = \{1, 2\}$ that follow the same distribution. The other classes' distributions are different from S . π_{K_0-1} only combines class 1 and 2 into one class

$$\text{p-ITCA}(\pi_{K_0}; \mathcal{D}_t, \mathcal{C}) = \text{p-ITCA}(\pi_{K_0-1}; \mathcal{D}_t, \mathcal{C})$$

 is the class combination boundary curve of S

When should we combine two classes i and j ?

Assumption (property of the classifier)

Considering a class combination π_{K-1} that only combines two class labels i and j , classifiers $\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}$ and $\phi_{\pi_{K-1}}^{\mathcal{C}, \mathcal{D}_t}$ satisfies

$$\sum_{k \in [K] \setminus \{i, j\}} [-\mathbb{P}(\pi_K(Y) = k) \log \mathbb{P}(\pi_K(Y) = k)] \cdot \mathbb{P}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}) = \pi_K(Y) | \pi_K(Y) = k) \geq \sum_{k \in [K] \setminus \{i, j\}} [-\mathbb{P}(\pi_{K-1}(Y) = k) \log \mathbb{P}(\pi_{K-1}(Y) = k)] \cdot \mathbb{P}(\phi_{\pi_{K-1}}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}) = \pi_{K-1}(Y) | \pi_{K-1}(Y) = k)$$

The property holds if ϕ is oracle. It also holds if ϕ is constructed from one-vs-all classifiers



Prune search space by combination criteria

Proposition (class combination criterion)

If Assumption 1 holds, class i and j will be combined by p-ITCA if and only if:

$$\frac{\mathbb{P}(\phi_{\pi_{k-1}}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}) = \pi_{k-1}(Y) | Y \in \{i, j\}) \geq p_i \log p_i \mathbb{P}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}) = Y | Y = i) + p_j \log p_j \mathbb{P}(\phi_{\pi_K}^{\mathcal{C}, \mathcal{D}_t}(\mathbf{X}) = Y | Y = j)}{(p_i + p_j) \log(p_i + p_j)}$$

- RHS ≥ 1 , p-ITCA cannot be improved by combining classes
- The combination criterion help prune the search space
- If $p_i + p_j = 1$ (there are only two classes), we should not combine the two classes



Properties of search strategies with the oracle classification algorithm

Definition (π_K 's induced partition)

Given K_0 observed classes, a class combination π_K 's induced partition is defined as K subsets of $[K_0]$: $\pi_K^{-1}(1), \dots, \pi_K^{-1}(K)$. That is, $\pi_K^{-1}(k) \cap \pi_K^{-1}(k') = \emptyset$ if $1 \leq k \neq k' \leq K_0$, and $\cup_{k=1}^K \pi_K^{-1}(k) = [K_0]$.

Definition (set of split true class combinations \mathcal{A}^*)

Suppose $\pi_{K^*}^*$ is the true class combination. We define

$\mathcal{A}^* := \{\pi_K : \forall k \in [K], \exists k' \in [K^*] \text{ s.t. } \pi_K^{-1}(k) \subset \pi_{K^*}^{*-1}(k')\}$ as the set of split true class combinations such that, in \mathcal{A}^* , each combination π_K 's induced partition is nested under the true class combination $\pi_{K^*}^*$'s induced partition; that is, each combined class defined by π_K is a subset of a combined class defined by $\pi_{K^*}^*$.



The optimality of BFS

Theorem (characteristics of the search strategies)

Suppose there are K_0 observed classes. Denote the class combinations found by the exhaustive search, BFS and greedy search with the oracle classification algorithm by $\pi_{K_{ES}}^{ES}$, $\pi_{K_{BFS}}^{BFS}$ and $\pi_{K_{GS}}^{GS}$, which correspond to K_{ES} , K_{BFS} and K_{GS} combined classes, respectively. Then $\pi_{K_{ES}}^{ES}, \pi_{K_{BFS}}^{BFS}, \pi_{K_{GS}}^{GS} \in \mathcal{A}^*$, the set of split true class combinations, and $\pi_{K_{ES}}^{ES} = \pi_{K_{BFS}}^{BFS}$.

