Information-theoretic Classification Accuracy: A Criterion that Guides Data-driven Combination of Ambiguous Outcome Labels in Multi-class Classification

Shandong Mathematical Society
Annual Academic Conference

Chihao Zhang

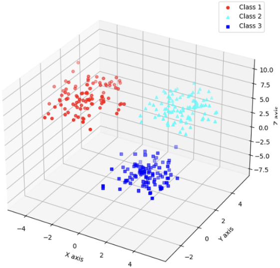
Janary 4, 2025

Acamdemy of Mathmatics and Systems Science, CAS

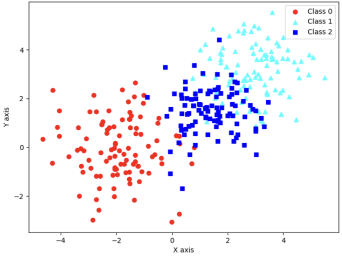
Background

- Outcome labeling ambiguity and subjectiveness are ubiquitous
  - Common in biomedical applications, e.g., disease diagnosis/prognosis
  - Data are inherently noisy
  - Labels may be mislabeled or labeled inconsistently by different graders [KGR+18]
- Ambiguous outcome labels would inevitably deteriorate prediction accuracy

Motivation example I



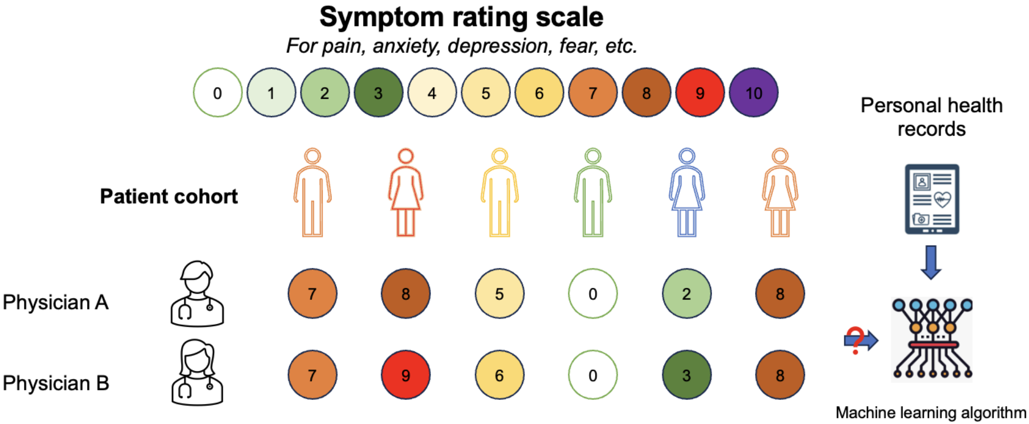
Full covariates



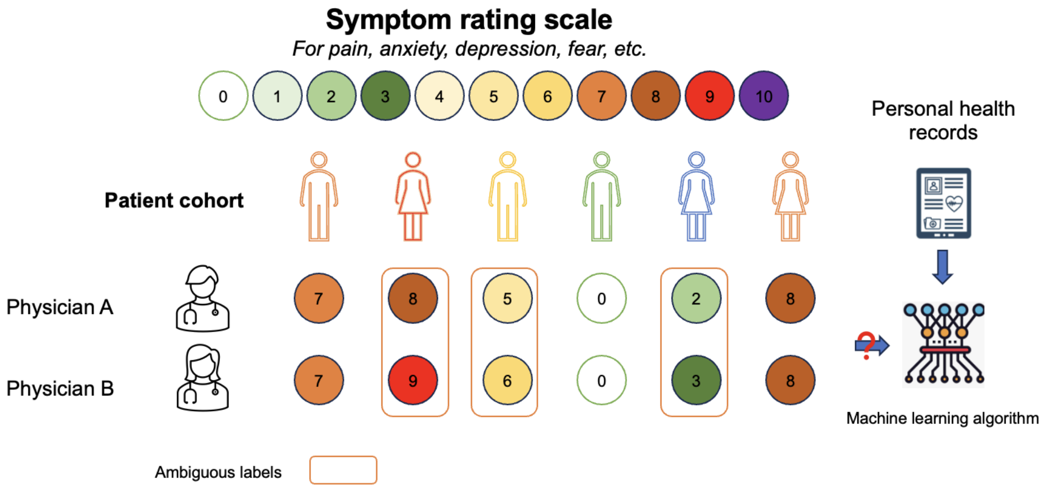
Partial covariates

- **Case:** Train a classifier on partial/low-quality data annotated with full/high-quality data.
- Problem: Uncertainty about whether the available information can sufficiently predict classes.

Motivating example II



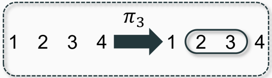
Motivating example II



An ad hoc solution: combining ambiguous outcome labels

Boost accuracy by combining ambiguous outcome labels

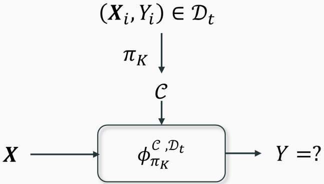
- **Class combination** π_K: $[K_0] \to [K]$ where $K < K_0$



example of π_K

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

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



Problems:

- Loosing prediction resolution
- Ad hoc, lacking a principled method

Trade-off between classification accuracy and resolution

Classification accuracy can be boosted at the cost of loosing prediction resolution

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

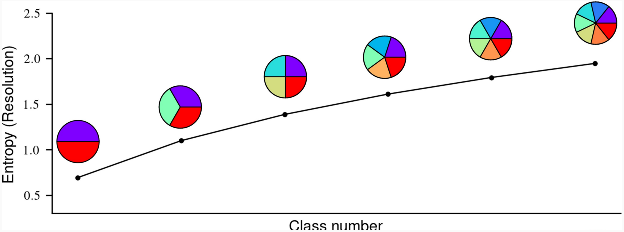
A **principled** method is called to balance the trade-off:

- How to characterize the “resolution”?

- How to properly balance the accuracy and resolution?

We proposed a criterion to guide class combination from an information-theoretic perspective

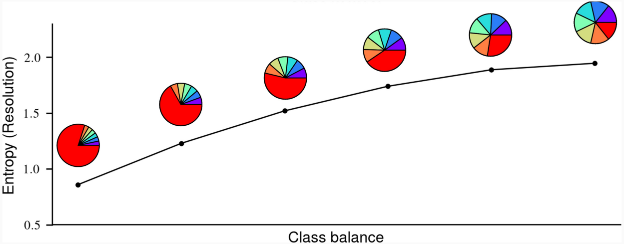
Observation: entropy of outcome label distribution characterizes the resolution



For balanced classes:

the larger the class number,

the higher the resolution



Given the number of classes:

the more balanced,

the higher the resolution

Information-theoretic classification accuracy (ITCA)

Definition of ITCA

Given class combination π_K, training data $\mathcal{D}_t$, evaluation data $\mathcal{D}_e$, and classification algorithm $\mathcal{C} \Longrightarrow$ classifier $\phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t}$

$\hat{p}_{k_0} := \mathbb{I}(Y_i = k_0)/n$ indicates the proportion of k_0-th original class in $\mathcal{D}_t \cup \mathcal{D}_e$

$$
\begin{aligned}
&\mathrm{ITCA}(\pi_K; \mathcal{D}_t, \mathcal{D}_e, \mathcal{C}) \\
&:= \sum_{k=1}^{K} \underbrace{\left[-\left(\sum_{k_0 \in \pi_K^{-1}(k)} \hat{p}_{k_0} \right) \log \left(\sum_{k_0 \in \pi_K^{-1}(k)} \hat{p}_{k_0} \right) \right]}_{\substack{\text{contribution of the combined class } k \\ \text{to the entropy of } \pi_K(Y)}} \cdot \underbrace{\frac{\sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{D}_e} \mathbb{I}(\phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t}(\boldsymbol{X}_i) = k,\ \pi_K(Y_i) = k)}{1 \bigvee \sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{D}_e} \mathbb{I}(\pi_K(Y_i) = k)}}_{\substack{\text{conditional accuracy of } \phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t} \\ \text{in the combined class } k}},
\end{aligned}
$$

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

Exhaustive search is prohibitive even K_0 is moderate

Table 1: The number of allowed class combinations π_K's given K_0

Label	K_0					
Type	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-k}'s that maximizes the ITCA
- **Breadth-first search**: track all the combination that can improve ITCA at each round

Alternative criteria that may guide class combination

Adjusted accuracy (AAC)

$$\mathrm{AAC} := \frac{1}{|\mathcal{D}_e|} \sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{D}_e} \frac{\mathbb{1}\left(\phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t^r}(\boldsymbol{X}_i) = \pi_K(Y_i)\right)}{\sum_{k_0 \in \pi_K^{-1}(\pi_K(Y_i))} \hat{p}_{k_0}}$$

Combined Kullback-Leibler divergence (CKL)

$$\mathrm{CKL} := D_{\mathrm{KL}}\left(\widehat{F}_{\pi_K,\mathcal{D}_e} \mid\mid \widehat{F}_{\pi_{K_0},\mathcal{D}_e}\right) + D_{\mathrm{KL}}\left(\widehat{F}_{\phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t},\mathcal{D}_e} \mid\mid \widehat{F}_{\pi_K,\mathcal{D}_e}\right)$$

Prediction entropy (PE)

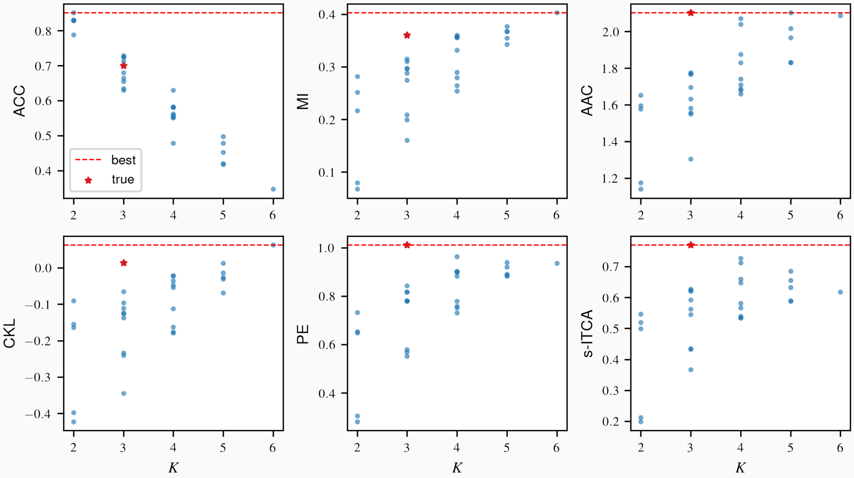
$$\mathrm{PE} := \sum_{k=1}^{K} - \frac{\sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{D}_e} \mathbb{1}\left(\phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t}(\boldsymbol{X}_i) = \pi_K(Y_i) = k\right)}{|\mathcal{D}_e|} \cdot \log\left(\frac{\sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{D}_e} \mathbb{1}\left(\phi_{\pi_K}^{\mathcal{C},\mathcal{D}_t}(\boldsymbol{X}_i) = \pi_K(Y_i) = k\right)}{|\mathcal{D}_e|}\right)$$

Commonly used criteria

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

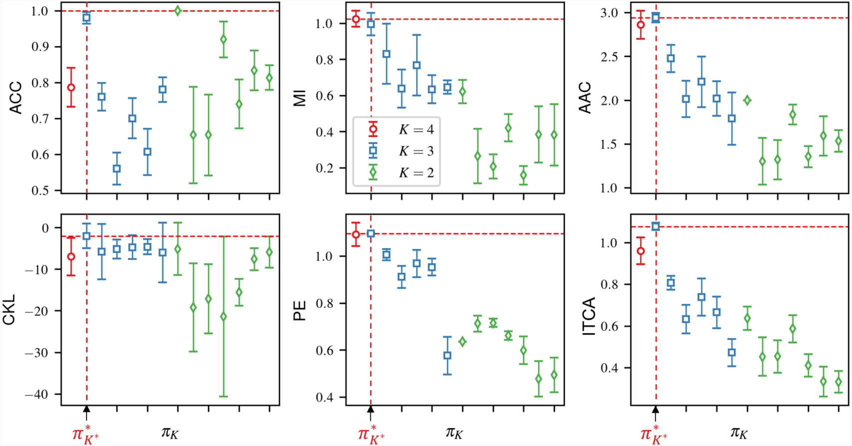
Simulation studies

ITCA finds the true class combination with a clear gap (simulated data)



Simulated data with $K_0 = 6$ observed classes; $K^* = 5$ true classes; $\mathcal{C} = \text{LDA}$

ITCA finds the true class combination with a clear gap (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 at the most cases

Criterion	# successes / # datasets (LDA)	Average Hamming (LDA)	Max Hamming (LDA)	# successes / # datasets (RF)	Average Hamming (RF)	Max Hamming (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

Table 2: The performance of six criteria on the 127 simulated datasets with $K_0 = 8$

Effectiveness of the greedy and BFS search strategies

Strategy	# successes / # datasets	Average Hamming	Max Hamming	Average # class 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

Table 3: Performance of ITCA using five search strategies and LDA on the 127 simulated datasets with $K_0 = 8$. ITCA failed in seven cases where $K* = 2$ and I will give a theoretical explanation later.

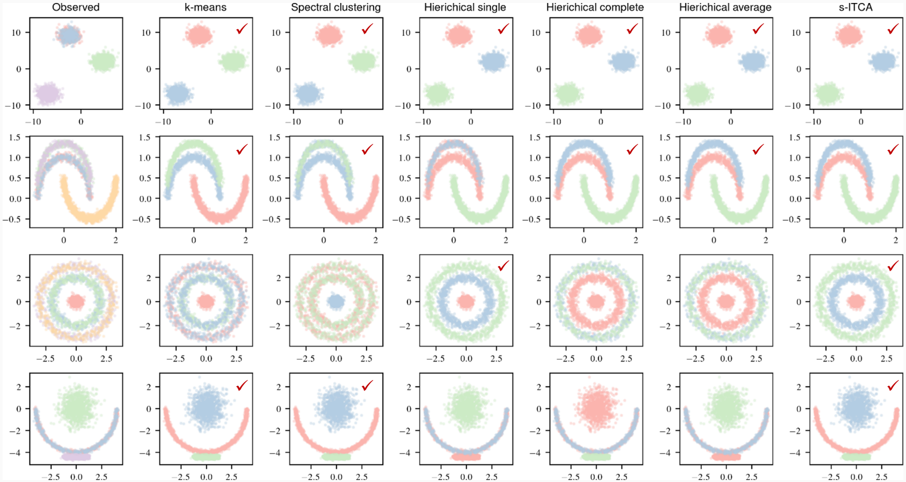
Using clustering algorithms to guide class combination

While ITCA provides a powerful data-driven approach for combining ambiguous classes, one may consider using a clustering algorithm

- **K-means-based class combination**: compute the k_0-th class center $\left(\sum_{i=1}^{n} \mathbb{I}(Y_i = k_0)\boldsymbol{X}_i\right) / \left(\sum_{i=1}^{n} \mathbb{I}(Y_i = k_0)\right)$; 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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$; 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 predefined

ITCA outperforms clustering-based class combination approaches



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

Some theoretic remarks

Population-level ITCA (p-ITCA)

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

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

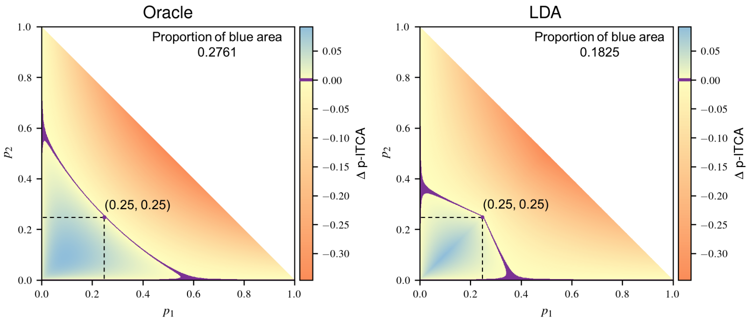
Definition (oracle classifier)
Given K_0 observed classes, let $S \subseteq [K_0]$ be a set of classes that share the same distribution. A classifier $\phi^*_{\pi_{K_0}}$ is an oracle classifier if that for any $(\boldsymbol{X}_i, Y_i)$ where $Y_i \in S$, $\phi^*_{\pi_{K_0}}$ predicts the label $s \in S$ by $\text{Multi}(1, [|S|], [p_s / \sum_{s \in S} p_s])$

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

$$\text{CC}(\pi_{K_0-1} || \pi_{K_0}; \mathcal{D}_t, \mathcal{C}) := \{(p_1, p_2) \in \Omega : \text{p-ITCA}(\pi_{K_0}; \mathcal{D}_t, \mathcal{C}, p_1, p_2) = \text{p-ITCA}(\pi_{K_0-1}; \mathcal{D}_t, \mathcal{C}, p_1, p_2)\}$$

is the class-combination curve

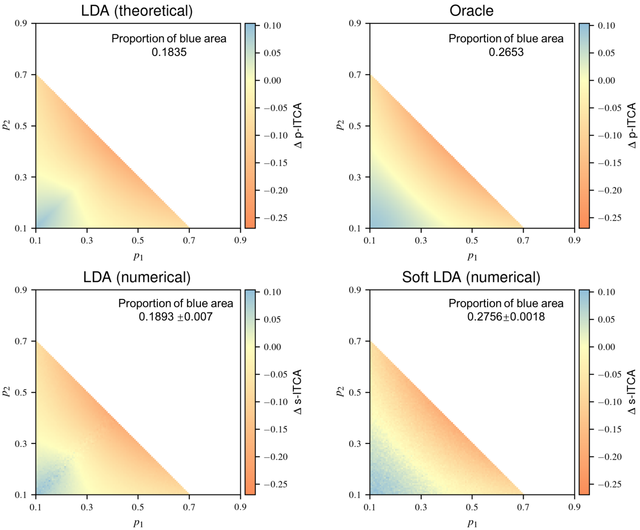
Different classification algorithms induce different CC-curves



Blue area means that p-ITCA increase after combination (orange means decrease), purple indicates the boundary.

- p-ITCA will not combine classes 1 and 2 when the proportions of the combined class is large
- LDA has a much smaller chance to discover the true class combination

Enhance the ability of LDA to discover the true combination



Soft LDA
Soft assigns label to $\boldsymbol{X}$ randomly with a multinomial distribution $\mathrm{Mult}(1, \mathrm{softmax}(\delta))$ where δ is the decision score where *delta* is the decision score $\delta = (\delta_1, \cdots, \delta_K)$

- We can show that Soft LDA is the same as the oracle classification algorithm when $||\boldsymbol{\mu}||/\sigma^2 \to \infty$

The choice of classification algorithm

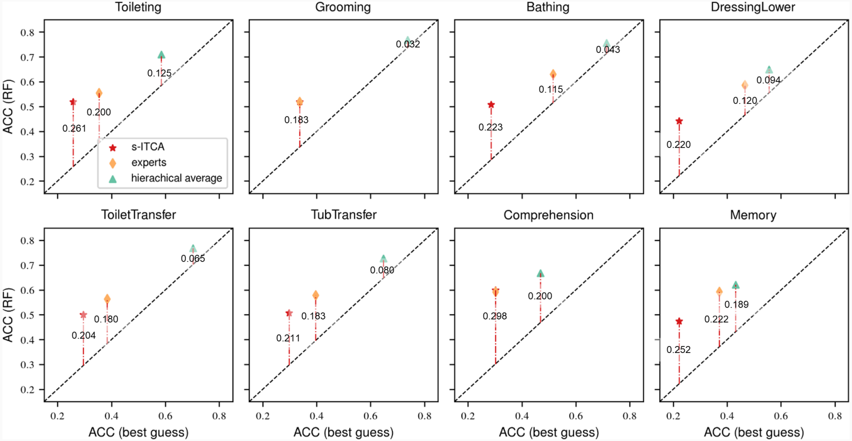
- ITCA is adaptive to all classification algorithms
- ITCA is comparable across different classification algorithms
- Users can choose the most suitable classification algorithms for different tasks
  - **Prediction**: a strong classification algorithm that maximizes ITCA
  - **Detection of similar classes**: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

Applications

ITCA refines prognosis of rehabilitation outcomes of TBI patients

- Rehabilitation outcomes of traumatic brain injury (TBI) patients is costly
- Predict rehabilitation outcomes (17 FIMs, each is a $K_0 = 7$ level outcome) for individual patients from their admission features
- The prediction accuracy of the trained classifier ($\mathcal{C} = \text{RF}$) on the original data is relatively low

Experts' suggestion vs. ITCA guided class combination

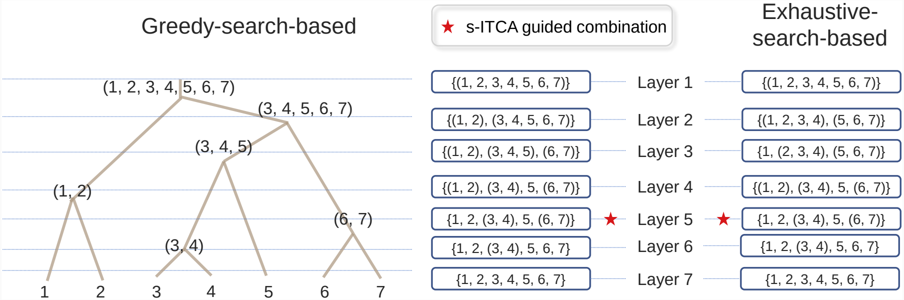


ITCA consistently leads to **more balanced** levels and a **more significant improvement** from the best guess (assigning every patient to the level that has the most patients)

ITCA induces multi-layer prediction frameworks

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

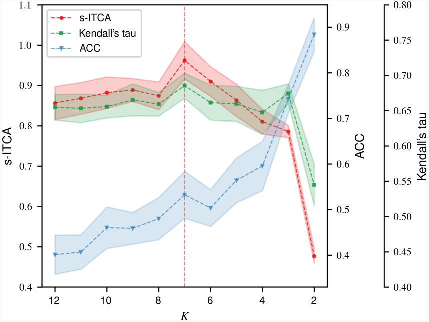
- **Nested-search-based**: classes in each layer are combined from the classes in the layer below
- **Exhaustive-search-based**: no nested constraint



ITCA boosts the 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



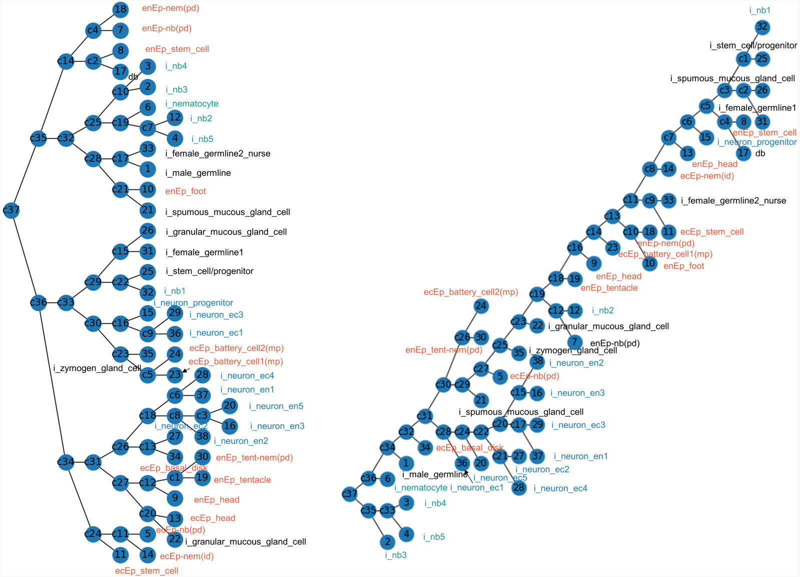
ITCA ($\mathcal{C} = $ NN) vs. ACC vs. Kendall's tau

ITCA-guided classification model achieves the best performance

- 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)	$\mathbf{0.9623 \pm 0.0464}$	$\mathbf{0.6855 \pm 0.0178}$	**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

scRNA-seq hydra Cell-type hierarchies built by the greedy-search-based ITCA



Conclusion and discussion

- A principled criterion ITCA guides the combination of ambiguous outcome labels
- Extensive simulation studies verify the effectiveness of ITCA
- Multiple real-world applications demonstrate the application potential of ITCA
- Future: use ITCA to help determine the number of clusters

Acknowledgements

- Prof. Jingyi Jessica Li, UCLA
- Prof. Shihua Zhang, AMSS
- Dr. Yiling Elaine Chen, UCLA

- **Publication**
  Journal of Machine Learning Research, 2022
  `https://www.jmlr.org/papers/v23/21-1150.html`
  Journal of Computational Biology, 2023
  `https://doi.org/10.1089/cmb.2023.0191`
- **Software** – `https://github.com/JSB-UCLA/ITCA`
  `>>> pip install itca`

Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster, *Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy*, Ophthalmology **125** (2018), no. 8, 1264–1272.

Appendix

Censored cross entropy (CCE)

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

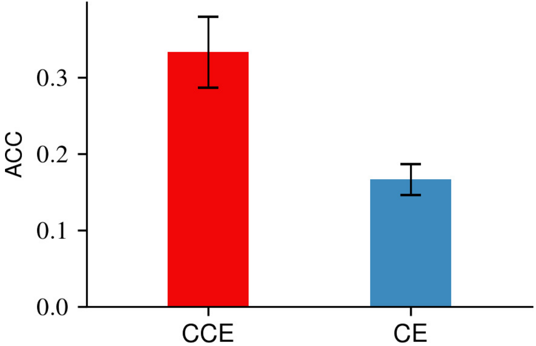
$$\text{CE} = -\sum_{i=1}^{K} I(Y_i = k) \log[\phi(X_i)]_k,$$

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

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

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.

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 $\phi^{\mathcal{C},\mathcal{D}_t}_{\pi_K}$ and $\phi^{\mathcal{C},\mathcal{D}_t}_{\pi_{K-1}}$ satisfies

$$\sum_{k\in[K]\setminus\{i,j\}} [-\mathbb{P}(\pi_K(Y)=k)\log\mathbb{P}(\pi_K(Y)=k)]\cdot\mathbb{P}(\phi^{\mathcal{C},\mathcal{D}_t}_{\pi_{K_0}}(\boldsymbol{X})=\pi_K(Y)|\pi_K(Y)=k)\geq$$

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

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:*

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

- RHS ≥ 1, p-ITCA cannot be improved by combing 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