OPTIMUM FEATURE ORDERING FOR DYNAMIC INSTANCE–WISE JOINT FEATURE SELECTION AND CLASSIFICATION

Yasitha Warahena Liyanage, Daphney–Stavroula Zois
Electrical and Computer Engineering Department
University at Albany, SUNY, Albany
{yliyanage, dzois}@albany.edu

Paper ID: 1536

IMAgINE Lab
http://goo.gl/KzUNR8

Introduction

- In many real–world applications (e.g., medical diagnosis)
  - **time–sensitive** and **interpretable** decisions are needed
  - **features** are **not freely available** to acquire
- Example: doctor wants to diagnose patient
  - must diagnose (**classification decision**) quickly by conducting minimum number of tests (**features**)
  - different set of tests may be appropriate for each individual patient (**data instance**)
  - order by which tests are conducted (**feature ordering**) is important

ONLINE ☑
OFFLINE ☐

Related Work

- **Feature selection** methods
  - features used are **same for all instances**
- **Instance–wise feature selection** methods
  - **reveal** all feature assignments and **do not scale** for large feature spaces
- Our **prior work**
  - **order** by which features are reviewed is **fixed**
- In contrast, **proposed method**
  - **optimizes** both **order** by which feature are reviewed and **number of features** used per data instance
  - **dynamically** selects features and **scales** for large feature spaces

Problem Description

- $F \triangleq \{F_1, F_2, \ldots, F_K\}$ set of features
- $\mathcal{C} \in \{c_1, \ldots, c_L\}$ class variable
- $e(F_k)$ cost of evaluating features
- Q_{ij} misclassification cost of selecting class when class c_j is true c_i
- σ order by which features are reviewed (**feature ordering**)
  - e.g., if $K = 3, \sigma = (F_3, F_1, F_2)$ is a valid feature ordering
- $\sigma(R)$ feature at which the sequential process stops (**stopping feature**)
  - e.g., $\sigma(R = 2) = F_2$ framework stops after reviewing the second feature F_2
- $D_{\sigma(R)}$ **classification strategy**
  - e.g., $\{D_{\sigma(R=2)} = 1\}$ deciding in favor of class c_1 based on $\{f_3, f_1\}$

Solution

Optimization Problem

$$\underset{\sigma, \sigma(R), D_{\sigma(R)}}{\text{minimize}} \; J\big(\sigma, \sigma(R), D_{\sigma(R)}\big)$$

$$J\big(\sigma, \sigma(R), D_{\sigma(R)}\big) = \mathbb{E}\left\{\sum_{k=1}^{R} e\big(F_{\sigma(k)}\big)\right\} + \sum_{j=1}^{L}\sum_{i=1}^{L} Q_{ij} P\big(D_{\sigma(R)} = j, \mathcal{C} = c_i\big)$$

Cost of evaluating features | Misclassification cost

Optimum Classification

- Optimum classification strategy:

$$D^*_{\sigma(R)} = \arg \min_{1 \leqslant j \leqslant L} \left[Q_j^T \pi_{\sigma(R)}\right]$$

Optimum Stopping

- Optimum solution via dynamic programming (DP):

$$\underbrace{\widehat{J}_k(\pi_{\gamma_k})}_{\text{Optimal cost–to–go}} = \min\Big[\underbrace{g(\pi_{\gamma_k})}_{\text{Cost of stopping}}, \underbrace{\widehat{\mathcal{A}}_k(\pi_{\gamma_k})}_{\text{Cost of continuing}}\Big]$$

$$g(\pi_{\gamma_k}) = \min_{1 \leqslant j \leqslant L} \left[Q_j^T \pi_{\gamma_k}\right]$$

$$\widehat{\mathcal{A}}_k(\pi_{\gamma_k}) \triangleq \min_{F_{k+1} \in Z_k} \left[e(F_{k+1}) + \sum_{F_{k+1}} \Delta^T(F_{k+1}|F_{\gamma_1}, \ldots, F_{\gamma_k}, \mathcal{C}) \pi_{\gamma_k} \widehat{J}_{k+1}(\pi_{\gamma_{k+1}})\right]$$

Theoretical Results

Lemma

- Function $g(\varpi)$ is **continuous, concave, and piecewise linear** and represented by set $\{Q_j^T\}_{j=1}^L$ of L vectors.

Lemma

- Functions $\widehat{\mathcal{A}}_k(\varpi), k = 0, \ldots, K-1$ are **continuous, concave, and piecewise linear**

$$\widehat{\mathcal{A}}_k(\varpi) = \min_{F_{k+1} \in Z_k} \left[\beta_k^{F_{k+1}} \varpi\right]$$

$$F_{\gamma_{k+1}} = \arg \min_{F_{k+1} \in Z_k} \left[\beta_k^{F_{k+1}} \varpi\right]$$

Theorem

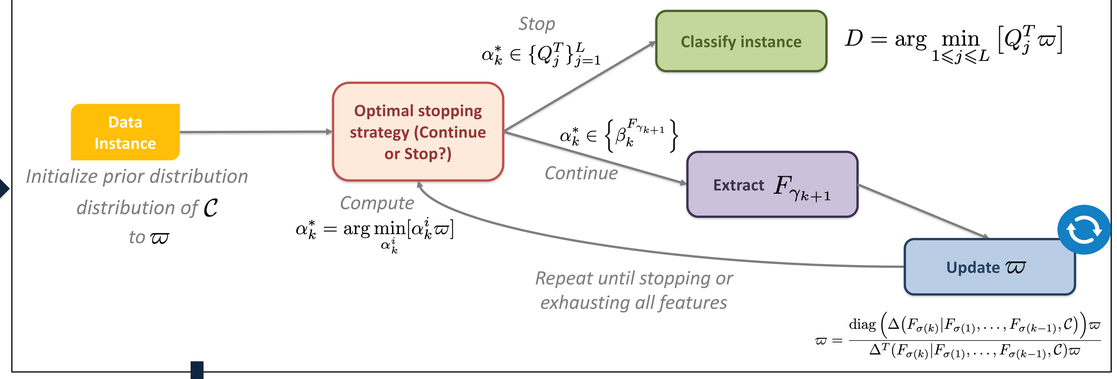
- At every stage $k \in \{0, \ldots, K\}$, there exists a **finite set** $\{\alpha_k^i\}$ of vectors such that

$$\widehat{J}_k(\varpi) = \min_i [\alpha_k^i \varpi]$$

$$\{\alpha_k^i\} = \left\{\left\{\beta_k^{F_{\gamma_{k+1}}}\right\} \cup \{Q_j^T\}_{j=1}^L\right\}, k \in \{0, \ldots, K-1\}$$

$$\{\alpha_K^i\} = \{Q_j^T\}_{j=1}^L$$

IFCO Algorithm



Results

Method	MLL		Spambase		Lung2		Car	
	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat
IFCO	**1.00**	**3.20**	**0.813**	**3.01**	**0.887**	**3.94**	**0.857**	**8.64**
MB [Liyanage2020]	**1.00**	**4.88**	0.741	3.08	0.842	3.96	0.539	5.63
ASSESS [Liyanage2019]	**1.00**	**5.07**	0.847	7.47	0.882	15.6	0.810	12.91
OFS-Density [Zhou2019]	0.960	11.0	0.787	7.60	**0.912**	**16.2**	0.597	6.80
SAOLA [Yu2014]	0.867	28.0	0.824	24.6	0.882	28.2	0.798	41.4
OSFS [Wu2014]	0.800	3.00	0.801	33.8	0.847	5.80	0.556	5.20
FAST–OSFS [Wu2014]	0.800	5.00	0.801	33.8	0.842	9.40	0.608	8.40
Lasso	**1.00**	**4.00**	0.902	29.6	0.685	9.40	0.551	28.8
Tree [Geurts2006]	0.933	100	0.947	18.2	0.897	207	0.752	429
PCA	0.667	36.0	0.693	1.00	0.897	88.4	0.391	91.0
SVM–G	**1.00**	**All**	0.834	All	0.788	All	0.563	All
R–Forest	**1.00**	**All**	0.940	All	0.911	All	0.758	All
XG–Boosting	0.733	All	**0.955**	**All**	0.906	All	**0.844**	**All**

- **IFCO** requires **less features** to achieve competitive accuracy compared to baselines
- IFCO **classification decisions** are **interpretable**, since we have access to the features used per each data instance

Conclusions

- Contributions
  - framework to select both **order** and **number of features** for **each data instance** individually
  - **properties** of **optimum solution**
  - **IFCO** algorithm and validation of its performance on real–world datasets
- Future directions
  - extend framework to **regression settings**

This research is based upon work supported by the National Science Foundation (NSF) under Grants ECCS–1737443 & CNS–1942330.