From Feature Selection to Instance-wise Feature Acquisition[1]

Tutorial @ SDM 2024

Daphney-Stavroula Zois[1] **Charalampos Chelmis**[2]

[1] *Electrical and Computer Engineering Department*

[2] Computer Science Department

Saturday, April 20th, 2024

[1] This research is based upon work supported by the National Science Foundation grants ECCS-1737443 & CNS-1942330 and a Google AI for Social Good award.

Advanced Topics

Advanced Topics

- Feature Acquisition
  - Interpretability (e.g., [liy23])
  - Dealing with structure (e.g., multidimensional Bayesian network classification [ELZ21, EZ23])
  - Reducing label uncertainty or learning to defer (e.g., dynamic classifier selection [EZC23b, EZC23a])
- Feature Selection
  - Incorporating fairness constraints (e.g., [GSSV22])
  - Feature selection for hierarchical classification (e.g., [ZHZ$^+$19])

Is Instance–wise Feature Acquisition Interpretable? [liy23]

- Using sparse set of features to classify data instances is essential for model interpretability
  - Observe which features contribute to each model output
- Sparsity can be achieved
  - globally by incorporating regularizer to objective function
  - instance–level, e.g., evaluate features along different decision paths in decision trees
- Goal: assess interpretability of IFCO [LZ21]

Interpretability of IFCO

- Model–based interpretability: humans can understand how model behaves and which factors influence its decision–making process
- Post–hoc interpretability: relationships learned by model from given dataset

Dataset & Baselines

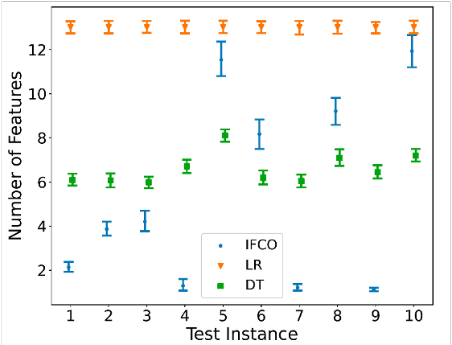
- For demonstration purpose, we use the German credit–risk dataset: classify people as high or low credit risk

Feature	Description	Feature	Description
F_1	Checking account status	F_{11}	Present residence
F_2	Duration in months	F_{12}	Property
F_3	Credit history	F_{13}	Age in years
F_4	Purpose of the credit	F_{14}	Other installment plans
F_5	Credit amount	F_{15}	Housing
F_6	Savings account status	F_{16}	Existing credits
F_7	Present employment (years)	F_{17}	Job
F_8	Installment rate	F_{18}	Number of dependents
F_9	Personal status	F_{19}	Telephone
F_{10}	Other debtors	F_{20}	Foreign worker

- Standard interpretable models:
  - Logistic regression with L1–norm regularizer (LR)
  - Decision tree (DT)

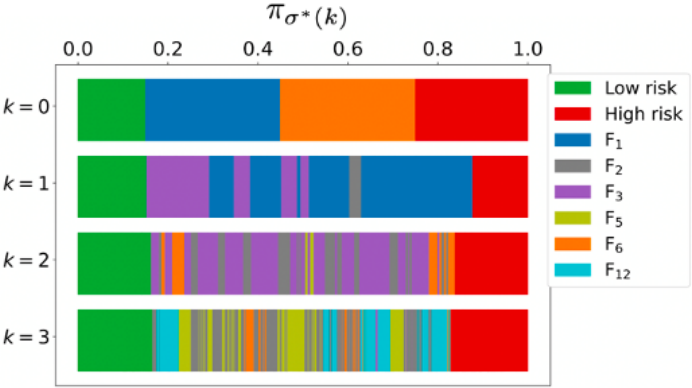
Model–based Interpretability

- Sparsity: use sparse set of features for classification
  - LR: global sparsity by using the L1–norm penalty
  - DT: instance–level sparsity by evaluating features along different branches (greedy learning of tree structure)
  - IFCO: instance–level sparsity by using feature acquisition cost $\sum_{k=1}^{R} e(F_{\sigma(k)})$
- Sparsity stability: interpretations are meaningless if sparsity varies drastically due to small perturbation in training dataset



Model–based Interpretability

- Simulatability: human can reason about decision–making process
  - LR: dot product between feature vector and weight vector
  - DT: hierarchical decision–making
  - IFCO:



Model–based Interpretability

- Modularity: ability to interpret meaningful portions of decision–making process independently
  - LR: affine transformation of input feature space (i.e., $w_i F_i$)
  - DT: each tree node is modular block that contributes to final classification decision
  - IFCO: sequential decision–making process based on sufficient statistic

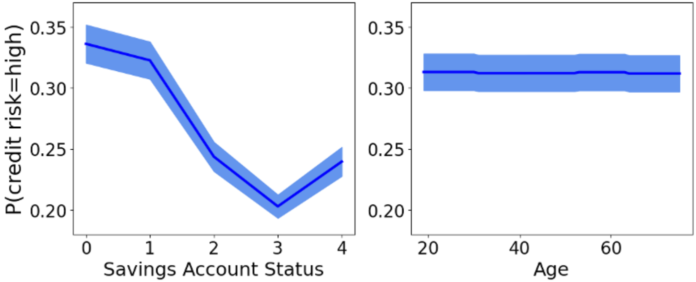
$$\pi_{\sigma^*(k)} = \frac{\Big(\Delta\big(F_{\sigma^*(k)} | F_{\sigma(1)}, \ldots, F_{\sigma^*(k-1)}, \mathcal{C}\big)\Big)\pi_{\sigma^*(k-1)}}{\Delta^T(F_{\sigma^*(k)} | F_{\sigma^*(1)}, \ldots, F_{\sigma^*(k-1)}, \mathcal{C})\pi_{\sigma^*(k-1)}}$$

  - Conditional independence assumption helps to decompose $\pi_{\sigma^*(k)}$ into simple and meaningful portions in terms of $P(F_{\sigma^*(k)}|\mathcal{C})$

Post–hoc Interpretability: Dataset–level Interpretations

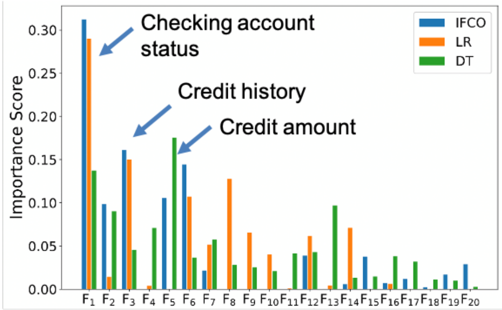
- Partial dependence: marginal effects of individual feature on output of machine learning model

$$PD(F_i) \approx \frac{1}{N}\sum_{n=1}^{N} \hat{f}(F_i, \bar{F}_i^{(n)})$$



Post–hoc Interpretability: Dataset–level Interpretations

- Feature importance: number of times specific feature contributes to specific classification decision



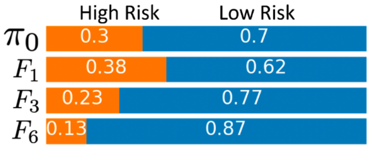
Post–hoc Interpretability: Dataset–level Interpretations

- Accuracy stability: test accuracy should be stable for any perturbations in training data

Method	Accuracy	Feat.
IFCO	**0.754±0.040**	**5.85**
DT	0.702±0.044	6.78
LR	0.740±0.034	14.0
XGB	**0.755**±0.037	19.9

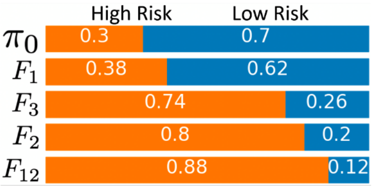
- Gradient boosted trees (XGB) (black box) requires 3.4 times more features for a just 0.1% improvement

Post–hoc Interpretability: Prediction–level interpretations



- bad checking account status
- good credit history
- good savings account status

Correctly predicted → low credit–risk

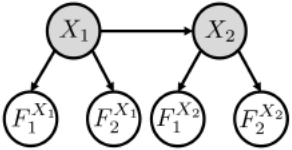


- bad checking account status
- bad credit history
- credit history of 36 months
- no known property

Correctly predicted → high credit–risk

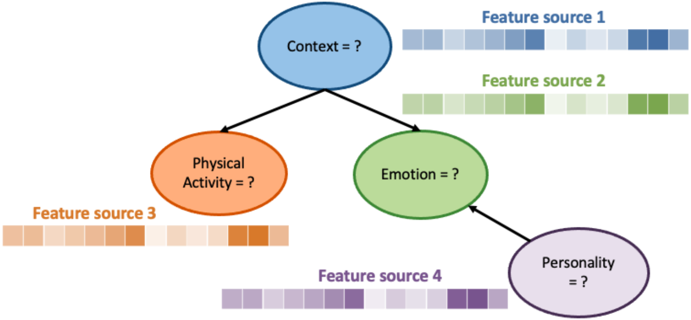
Instance–wise Multidimensional Classification [ELZ21, EZ23]

- Many real–world applications (e.g., medical diagnosis, behavioral analysis)
  - Bayesian networks used to describe relationships between variables
  - Variables not directly observable but can be inferred via features
- Multi–dimensional Bayesian network classification [GBBL21] learns underlying unknown Bayesian network structure between variables in X and features in F, and then performs inference to compute the values of variables in X



Problem Statement

- What happens if features are acquired at a cost?
- Goal: accurately classify each data instance during testing, while keeping total feature acquisition cost minimum when data instance label corresponds to known Bayesian network of multiple class variables



Optimization Setup

- $\mathcal{G} = (X, E)$: known Bayesian network structure
- $X \triangleq \{X_1, X_2, \ldots, X_n\}$: set of nodes corresponding to categorical variables
- E: set of directed edges to represent relationships between categorical variables
- $F \triangleq \{F_1^{X_1}, \ldots, F_{K_1}^{X_1}, F_1^{X_2}, \ldots, F_{K_2}^{X_2}, \ldots, F_1^{X_n}, \ldots, F_{K_n}^{X_n}\}$: set of features, where $F_k^{X_i}$ is kth feature associated with variable X_i
- e_k^i: cost of acquiring kth feature associated with variable X_i
- $C_l^{X_i}$: class value for variable X_i

Optimization Setup

- Introduce random variables
  - $R_i \in \{0, \ldots, K_i\}$: last feature acquired before classification decision for variable X_i
  - $D_{R_i} \in \{1, \ldots, N_i\}$: classification decision based on R_i features for variable X_i

$$\min_{\mathbf{R},\mathbf{D_R}} J(\mathbf{R}, \mathbf{D_R})$$

$$J(\mathbf{R}, \mathbf{D_R}) = \mathbb{E}\left\{ \sum_{i=1}^{n} \sum_{k=1}^{R_i} e_k^i + \sum_{\mathbf{j}} \sum_{\mathbf{m}} M_{\mathbf{mj}} P(\mathbf{D_R} = \mathbf{j}, \mathbf{C} = \mathbf{c_m}) \right\}$$

- The computational complexity of directly solving the above problem is high

Alternative Approach

- Determine features to be acquired and classification decision for each categorical variable X_i in $\mathcal{G}$

$$J(R_i, D_{R_i}) = \mathbb{E}\left[\sum_{k=1}^{R_i} e_k^i + \sum_{l=1}^{N_i}\sum_{m=1}^{N_i} M_{lm}^i P(D_{R_i} = l, \mathcal{C}_i = C_m^{X_i})\right],$$

- How to account for relationships between categorical variables? propagate decisions across $\mathcal{G}$
  - Initially, acquire features and make classification decisions for in–degree 0 nodes
  - Use such decisions to drive feature acquisition and classification decisions for each in–degree greater than 0 node

ISEC Algorithm

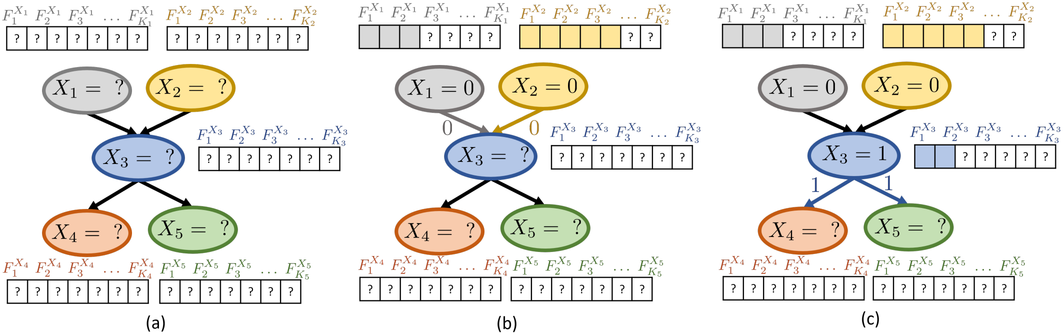


Figure 6: (a) Original Bayesian network; (b) Feature acquisition and classification for variables of in-degree 0; (c) Feature acquisition and classification for variables of in-degree > 0

Some Results

TABLE II: Comparison of global accuracy (GA), mean accuracy (MA), and the average number of features (AF). The highest and the second highest accuracy values are bolded and gray–shaded, and gray–shaded, respectively. The smallest and the second smallest AF values are bolded and gray–shaded, and gray–shaded, respectively.

Dataset	Metric	ISEC	IC–NB	IC–ETANA	PC–NB	BCC	MD–KNN	IC–SVM	PC–SVM
Edm	GA	**0.5905**	0.3890	0.4668	0.5443	0.3905	0.3864	0.3578	0.4483
	MA	**0.7401**	0.6491	0.6500	0.7101	0.6952	0.6209	0.6755	0.7013
	AF	**5.8654**	16.0000	8.6333	16.0000	16.0000	16.0000	16.0000	16.0000
Voice	GA	**0.8753**	0.6897	0.8224	0.6824	0.2735	0.8359	0.7663	0.7220
	MA	**0.9364**	0.8243	0.8748	0.8343	0.5210	0.9142	0.8780	0.8514
	AF	2.5127	19.0000	**2.2719**	19.0000	19.0000	19.0000	19.0000	19.0000
Jura	GA	**0.4402**	0.3036	0.3481	0.4010	0.1588	0.2591	0.2562	0.2393
	MA	**0.6352**	0.5405	0.5845	0.6016	0.4764	0.4889	0.5307	0.4830
	AF	**7.0517**	9.0000	8.2394	9.0000	9.0000	9.0000	9.0000	9.0000
Song	GA	0.3299	0.2114	0.2509	0.2611	0.3082	**0.4229**	0.3471	0.3548
	MA	0.7134	0.6012	0.6709	0.6360	0.6802	**0.7565**	0.6728	0.6724
	AF	**16.3172**	98.0000	16.6072	98.0000	98.0000	98.0000	98.0000	98.0000
Flare	GA	0.8173	0.0277	0.7800	0.0463	**0.8204**	0.7802	0.8202	0.8202
	MA	0.9205	0.2194	0.8906	0.5736	**0.9226**	0.9035	0.9225	0.9225
	AF	**1.3040**	10.0000	7.0573	10.0000	10.0000	10.0000	10.0000	10.0000
Student	GA	**0.6099**	0.5742	0.5914	0.0815	0.5469	0.5208	0.5334	0.5021
	MA	**0.7409**	0.7227	0.5529	0.5418	0.6522	0.6546	0.6560	0.6084
	AF	**8.4940**	30.0000	14.9458	30.0000	30.0000	30.0000	30.0000	30.0000
Emotion	GA	0.3121	0.1820	0.2378	0.2731	0.0000	0.1164	0.2631	**0.3203**
	MA	0.7783	0.7391	0.7641	0.7700	0.6885	0.7026	**0.7934**	0.7718
	AF	**8.5983**	72.0000	15.3432	72.0000	72.0000	72.0000	72.0000	72.0000
Child	GA	**0.5620**	0.5509	0.5350	0.4800	0.3910	0.5098	0.3909	0.3909
	MA	**0.8197**	0.8156	0.8069	0.7783	0.7106	0.7799	0.7106	0.7106
	AF	**4.4293**	17.0000	5.8147	17.0000	17.0000	17.0000	17.0000	17.0000
Hepar2	GA	0.4200	0.0900	0.4170	0.0350	0.4180	0.4150	**0.4230**	0.4150
	MA	0.7807	0.4260	0.7757	0.4193	0.7813	0.7792	**0.7813**	0.7747
	AF	**12.6213**	67.0000	31.9470	67.0000	67.0000	67.0000	67.0000	67.0000
Sachs	GA	**0.7920**	0.7770	0.6000	0.3000	**0.7920**	0.7880	**0.7920**	**0.7920**
	MA	**0.8420**	0.8345	0.7250	0.5765	**0.8420**	0.8399	**0.8420**	**0.8420**
	AF	**1.8575**	9.0000	8.4295	9.0000	9.0000	9.0000	9.0000	9.0000
Insurance	GA	**0.8270**	0.6920	0.8100	0.6150	0.4320	0.6062	0.7240	0.7310
	MA	**0.9050**	0.8350	0.9030	0.7840	0.5870	0.7841	0.8540	0.8520
	AF	**2.9115**	25.0000	5.4565	25.0000	25.0000	25.0000	25.0000	25.0000

Joint Feature Acquisition & Classifier Selection [EZC23b, EZC23a]

- ML models cannot accurately predict all test instances
- Problematic, especially in risk–sensitive applications (e.g., autonomous vehicles, medical diagnosis)
- To the best of our knowledge, instance–wise feature acquisition assumes single loss function
- How to jointly acquire the subset of features based on which each example is to be classified and the appropriate classifier to be used for this task?
  - Assess difficulty of classifying data instances to guide decision making process
  - Easy–to–classify data instances: few features and simple classifier
  - Hard–to–classify data instances: more features and powerful classifier

Problem Description

- $X \triangleq [X_1, \ldots, X_F]^\top$: feature vector containing F features
- c_f: cost of acquiring fth feature
- $Y \in \{1, \ldots, N\}$: label
- $C \triangleq \{C_1, \ldots, C_Z\}$: set of Z classifiers

Objective: jointly determine subset of features to be acquired, classifier to be used and the label of each example

Optimization Setup

- Introduce random variables
  - $S \in \{0, \ldots, F\}$: last feature acquired before label assignment
  - $U_S \in \{0, \ldots, Z\}$: classifier selected after S features have been acquired
  - $D_S \in \{1, \ldots, N\}$: classification decision for data instance under consideration based on S features

$$\min_{S,U_S,D_S} L(S, U_S, D_S)$$

$$L(S, U_S, D_S) = \mathbb{E}\Bigg\{ \sum_{f=1}^{S} c_f + \sum_{z=1}^{Z} \lambda_z \mathbb{I}_{\{U_S=z\}} h_S^z + \gamma \mathbb{I}_{\{U_S=0\}} \times \sum_{j=1}^{N} \sum_{i=1}^{N} \Omega_{ij} P(D_S = j, Y = i) \Bigg\},$$

Optimum Solution

- $\phi_f \triangleq [\phi_f^1, \ldots, \phi_f^N]^T$: posterior probability vector with $\phi_f^i \triangleq P(Y = i | x_1, \ldots, x_f)$
- Optimum label assignment strategy

$$D_S^* =_{1 \leq j \leq N} [\mathbf{\Omega}_j^T \phi_S].$$

- Optimum classifier selection strategy

$$U_S^* =_{0 \leqslant t \leqslant Z} [\lambda_t H_S^t(\phi_S)].$$

- Optimum feature acquisition strategy via dynamic programming

$$\begin{aligned} \bar{L}_f(\phi_f) &= \min\left[l(\phi_f), \bar{I}_f(\phi_f)\right] \\ l(\phi_f) &= \min_{0 \leqslant t \leqslant Z} [\lambda_t H_f^t(\phi_f)] \\ \bar{I}_f(\phi_f) &= c_{f+1} + \sum_{x_{f+1}} \bar{L}_{f+1}(\phi_{f+1}) \Pi_{f+1}^T(x_{f+1}) \phi_f \end{aligned}$$

Intuition

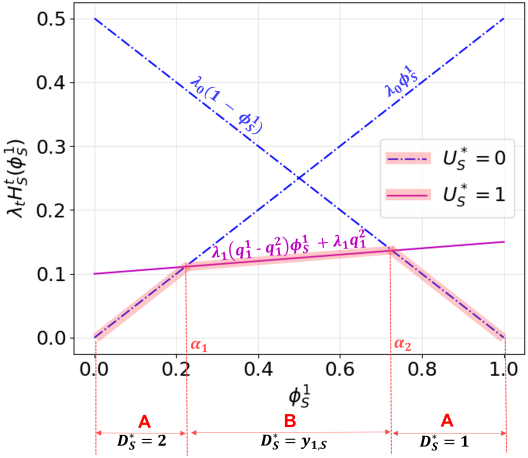
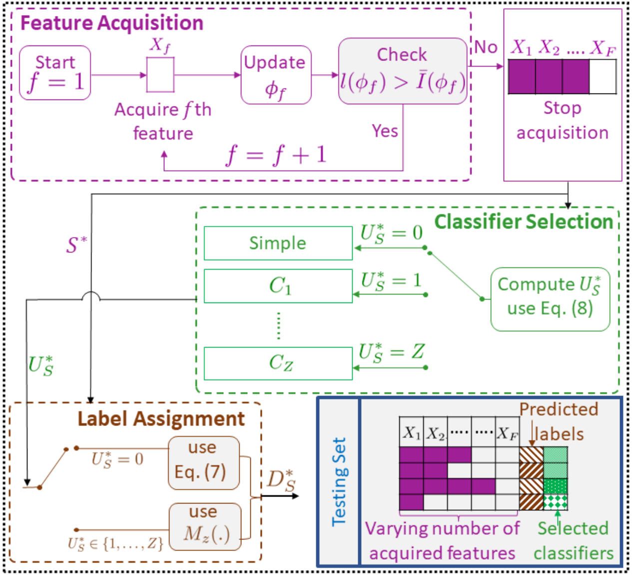


Figure 7: Illustration of classifier selection and label assignment processes in the case of two label values (i.e., $N = 2$), a simple classifier (region A), and a single powerful classifier (region B).

SFCS Algorithm

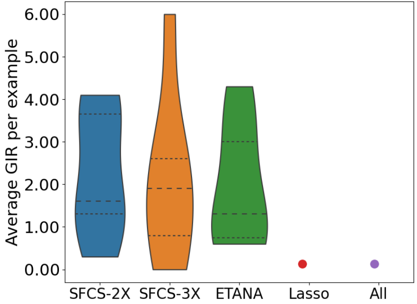


Some Results

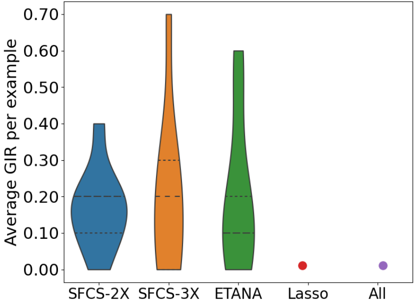
Method	Monks Problem		Diabetes		EEG Eye State		MagicTelescope		Student		German Credit	
	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat	Acc	Feat
SFCS–SVM	0.536	5.722	**0.753**	6.056	0.536	**3.315**	0.794	6.316	0.864	4.656	0.732	12.081
SFCS–DT	0.795	5.722	0.753	6.056	0.485	**3.315**	0.807	6.316	**0.869**	4.656	0.732	12.081
ETANA	0.529	**5.188**	0.749	**5.935**	0.500	12.261	0.775	**6.302**	0.864	**4.617**	0.714	**11.846**
NB	0.591	6.000	0.751	8.000	0.437	14.000	0.727	11.000	0.827	32.000	0.700	20.000
SVM	0.657	6.000	0.674	8.000	**0.551**	14.000	0.806	11.000	0.787	32.000	0.700	20.000
DT	**0.922**	6.000	0.706	8.000	0.475	14.000	**0.819**	11.000	0.838	32.000	0.664	20.000
Lasso	0.654	4.800	0.766	8.000	**0.551**	13.400	0.789	9.000	0.851	14.600	**0.734**	17.800

- Good balance between accuracy and average number of acquired features
- Classifier selection in instance–wise feature acquisition enhances accuracy, but in most cases, increases average number of acquired features
- Why does SFCS–DT performs worse than DT?

Some Results



(a) Diabetes dataset.



(b) Magic dataset.

Figure 8: Distribution of average Gini impurity reduction (GIR) per example. "All" denotes baselines that use all features (e.g., SVM, DT)

- Feature with higher GIR is more significant than a feature with lower GIR, since latter cannot be used to effectively separate labels

Some Results

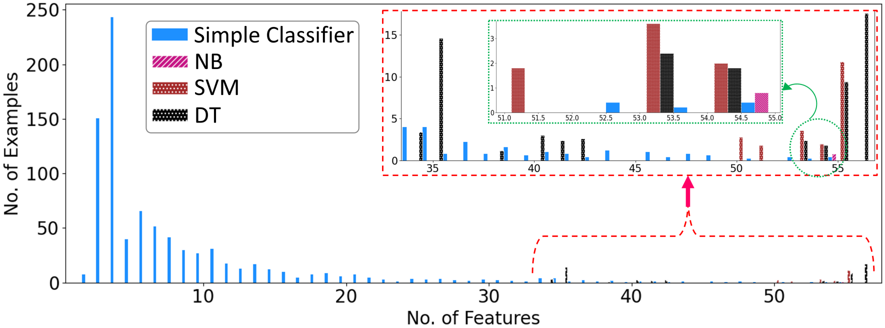


Figure 9: Distribution of number of acquired features during testing for the Spambase dataset using SFCS–3X (NB, SVM, DT).

- Classify most instances using simple classifier with few features
- When number of acquired features increases, SFCS switches to other classifiers (difficult-to-classify instances)

Causal Feature Selection for Algorithmic Fairness [GSSV22]

- Algorithmic fairness is critical when supervised classification models are used to support decisions in high–stake domains
- Not discrimination–aware feature selection methods prefer features that improve accuracy
- Goal: identify subset of new features to include in a dataset without worsening its biases against protected groups
  - Meant to be used during training dataset creation time
  - Key challenge: one or more non–protected features can facilitate reconstruction of protected information (e.g., infer race from zip code)
  - Main idea: perform conditional independence tests between different subsets of features

Causal Feature Selection for Algorithmic Fairness [GSSV22]

- Input dataset comprises:
  - Target variable Y (e.g., credit score)
  - Set of protected/sensitive features $\mathbf{S}$ (e.g., gender and race)
  - Set of admissible features $\mathbf{A}$ (e.g., expected monthly usage)
    - Protected variables can affect the outcome through admissible features
  - Features that are neither admissible nor sensitive (e.g., age and education)

- Two–phase method using conditional independence tests
  - Identify features that do not capture information about sensitive attributes
  - Ensure fairness even if features capture some information about sensitive attributes

Causal Feature Selection for Algorithmic Fairness [GSSV22]

Algorithm 1 SeqSel

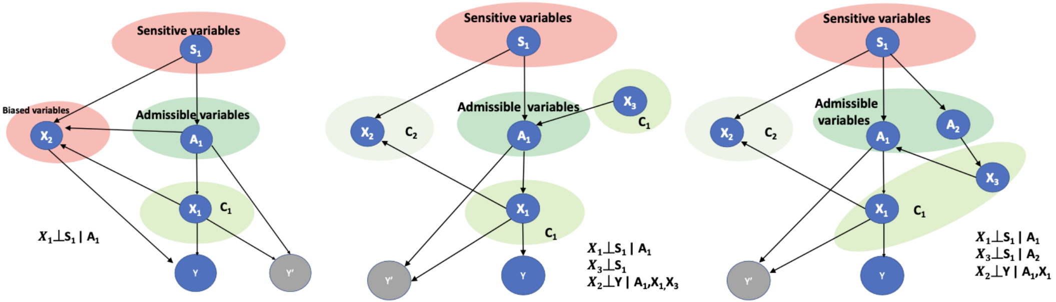
Input: Variables $\mathbf{A}, \mathbf{S}, \mathbf{X}, Y$
$\mathbf{C}_1 \leftarrow \phi$
for $X \in \mathbf{X}$ **do**
  if $\exists A \subseteq \mathbf{A}$ such that $(X \perp \mathbf{S} | A)$ **then**
    $\mathbf{C}_1 \leftarrow \mathbf{C}_1 \cup \{X\}$
$\mathbf{C}_2 \leftarrow \phi$
$\mathbf{X} \leftarrow \mathbf{X} \setminus \mathbf{C}_1$
for $X \in \mathbf{X}$ **do**
  if $(X \perp Y | \mathbf{A} \cup \mathbf{C}_1)$ **then**
    $\mathbf{C}_2 \leftarrow \mathbf{C}_2 \cup \{X\}$
return $\mathbf{C}_1 \cup \mathbf{C}_2$

- Find variables X_i independent of $\mathbf{S}$ by performing conditional independence test
- Variables whose paths from $\mathbf{S}$ are blocked by $\mathbf{A}$ do not provide any new information about $\mathbf{S}$
  - Check if X_i is conditionally independent of $\mathbf{S}$ given $\mathbf{A}$
- Variables X_i not independent of $\mathbf{S}$ even given $\mathbf{A}$ can leak sensitive information
  - If independent of Y given $\mathbf{A}$, no effect on the classifier
- Any variable that is not independent of $\mathbf{S}$ and Y even after intervening on $\mathbf{A}$ should not be added

Causal Feature Selection for Algorithmic Fairness [GSSV22]

- Causal DAG G captures functional dependencies between variables
  - Variable X_1 causes X_2 iff $X_1 \to X_2$ in G
  - Joint probability distribution can be decomposed similar to Bayesian networks
- Variables X and Y are $d-$separated given Z, if all paths between X and Y are blocked by Z
  - Ideally, the prediction and protected attributes should be $d-$separated in G
- do-operator: assign value x to variable X $(do(X) = x)$ in G' induced by G, with the difference that all incoming edges of X have been removed
- A classifier is considered fair if for any collection of values α of $\mathbf{A}$ and output y' $P(Y' = y|do(\mathbf{S}) = \mathbf{s}, do(\mathbf{A}) = \alpha) = P(Y' = y|do(\mathbf{S}) = \mathbf{s}', do(\mathbf{A}) = \alpha), \forall \mathbf{A}, \mathbf{S}, Y'$
- Testing for causal fairness requires fully specified causal graphs (not available in practise)
  - Use conditional mutual information instead

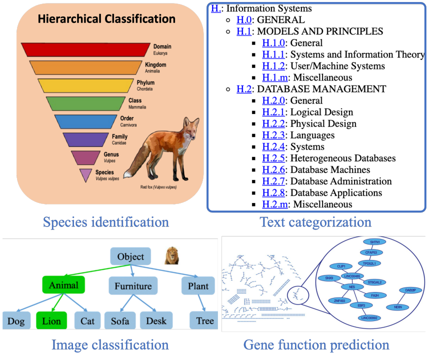
Causal Feature Selection for Algorithmic Fairness [GSSV22]



(a) X_2 is a biased variable (b) X_1, X_2, X_3 ensure fairness (c) X_1, X_2, X_3 ensure fairness

- Given $\mathbf{A}$, $D = \mathbf{A} \cup \mathbf{T}$ is causally fair if the Bayes optimal predictor Y', trained on D satisfies causal fairness with respect to sensitive attributes $\mathbf{S}$
- Goal: identify largest subset $\mathbf{T}$ such that Y', trained using these variables is fair
- New node Y' is added to G
- All features that impact the classifier output are made parents of Y'

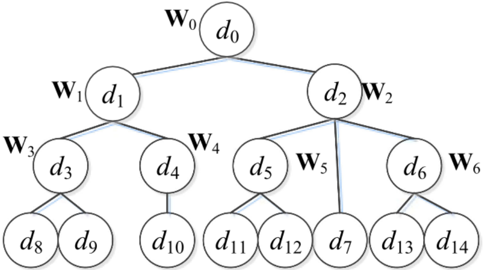
Feature Selection for Hierarchical Classification [ZHZ+19]



- Large–scale classification tasks comprise hundreds, thousands, or even tens of thousands of class labels
- Class labels are structured (often in a tree)
  - Class hierarchy divides the classification task into small and easy subtasks
- Goal: Feature selection for hierarchical classification tasks
  - Relevant features may differ among classes
  - Need to select different features for different subtasks

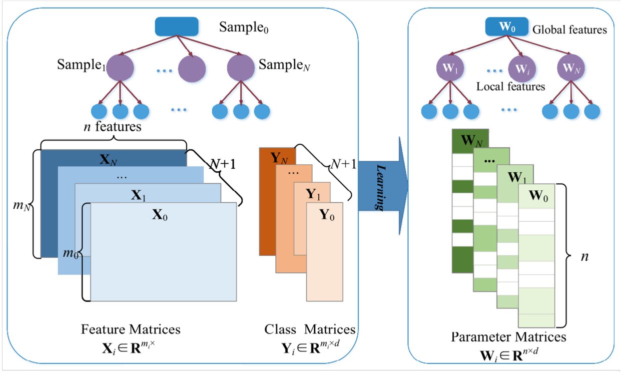
Feature Selection for Hierarchical Classification [ZHZ+19]

- Feature selection as penalized optimization
  - $\min_{\mathbf{W}} L(\mathbf{XW}, \mathbf{Y}) + \lambda R(\mathbf{W})$
  - Empirical loss L (e.g., logistic, hinge, cross–entropy loss)
  - Regularizer R and positive constant λ
  - Structural sparsity with $\ell_{2,1}-$norm



- Goal: minimize $\sum_{i=0}^{N} (\|\mathbf{X}_i\mathbf{W}_i - \mathbf{Y}_i\|_F^2 + \lambda\|\mathbf{W}_i\|_{2,1})$
  - Closed form solution obtained for least squares loss
- Feature weight matrix $\mathbf{W_i}$ is computed for each internal node i
- Data instances of the ith node: $\mathbf{X}_i = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_{m_i}]$

Feature Selection for Hierarchical Classification [ZHZ+19]



Algorithm 1. Hierarchical Feature Selection (Hier-FS)

Input: Input data $\mathbf{X}_i \in \mathbf{R}^{m_i\times n}$ and labels $\mathbf{Y} \in \{0,1\}^{m_i\times d}$, where $i = 0, 1, \dots, N$, and N is the number of internal nodes. To facilitate the calculation, we let d be the maximum number of classes of internal nodes. Regularization parameter is λ, and the maximal iteration number is T.

Output: Matrix $\mathbf{W} \in \mathbf{R}^{n\times d(N+1)}$.

```
Set t = 0 and initialize W_i ∈ R^{n×d} randomly;
W = [W_0, W_1, ..., W_N];
while t < T do
  for i = 0 : N do
    Compute the diagonal matrix D_i^(t) according to
      d^i_jj = 1 / (2||w^i_j||_2);
  end for
  // Update the root node and internal nodes.
  for i = 0 : N do
    Update W_i by W_i^(t+1) = (X_i^T X_i + λD_i^(t))^(-1) (X_i^T Y_i);
  end for
  Update W^(t+1) = [W_0, W_1, ..., W_N];
  t = t + 1;
end while
return W;
```

- Top–down recursive strategy
- Node ith's top–ranked (w.r.t $\|\mathbf{w}_j^i\|_F$) features are selected

Feature Selection for Hierarchical Classification [ZHZ+19]

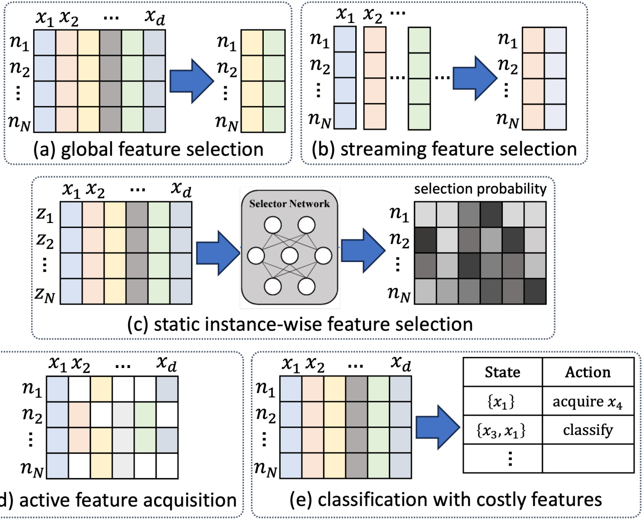
- Hierarchical regularization with parent–child relationship
  - Parent–child classes are similar to each other; should share common features
  - Relationship is incorporated into regularizer: $\sum_{i=1}^{N} \|\mathbf{W}_i - \mathbf{W}_{p_i}\|_F^2$
- Hierarchical regularization with sibling relationship
  - Siblings come from different subtrees
  - Discriminative features must be selected for each sibling
  - Hilbert–Schmidt Independence Criterion to penalize dependence between selected features at sibling nodes
- Hierarchical regularization with family relationship
  - Both parent–child and sibling relationships between categories incorporated into the optimization problem

Summary and Conclusion

Still about Feature Selection vs Feature Acquisition?

- Global Feature Selection
  - Identify, during training, a subset of features (common across instances)
  - Online/streaming methods when full feature set unavailable at training
- Instance–wise Feature Selection
  - Identify, during testing, small subset of features for each data instance (varies between instances)
  - Given a test instance, all of its features must be available
- Active Feature Acquisition
  - During training (related to feature selection with missing values)
  - During testing, learned model is used
- Instance–wise Feature Acquisition
  - Different features acquired, during testing, for each data instance
  - Classification with costly features / Dynamic instance–wise feature acquisition

Feature Selection vs Feature Acquisition Visualized



Key Takeaways

- Traditional feature selection is conducted during training
- Feature acquisition $\neq$ feature selection
  - can be performed either during training or testing
- Instance–wise feature selection $\neq$ instance–wise feature acquisition
- Both feature selection and feature acquisition approaches face significant challenges
- Instance–wise feature acquisition has broader implications to ML

(Non Exhaustive List of) Topics This Tutorial Didn't Cover

- Feature acquisition in both training and testing [DMW10]
- Group feature acquisition during testing [AJD24]
- Multiview/multimodal feature selection [YGSC15, LMF16, KAH20] and acquisition [NZC20]
- Active feature acquisition for time series data [LO21, BBS22, KCV^{+}23]
- Feature selection (prompting) for large language models
- Knowledge–driven feature acquisition
- Causality and feature selection
- Feature selection/acquisition for non–linear models
  - Quantifying feature importance is difficult
  - Interpreting findings becomes challenging

Tutorial Slides

- Our coverage of state–of–the–art and challenges we identify are not exhaustive
- The slides can be found at: `https://www.cs.albany.edu/~cchelmis/tutorials/sdm/2024/`
- Suggested citation:
  Daphney–Stavroula Zois, Charalampos Chelmis, "From Feature Selection to Instance–wse Feature Acquisition", Minitutorial at SIAM International Conference on Data Mining (SDM), Houston, TX, April 2024.

Acknowledgements

- We thank our collaborators and funding agencies for making this work possible
- We thank authors who graciously agreed to provide us with material to include in this tutorial
- We thank the conference tutorial committee for giving us the opportunity to present at SDM
- We thank you, the audience, for your attention
- We welcome your feedback and suggestions

References I

Vedang Asgaonkar, Aditya Jain, and Abir De.
Generator Assisted Mixture of Experts For Feature Acquisition in Batch.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10927–10934, 2024.

Josh Attenberg, Prem Melville, Foster Provost, and Maytal Saar-Tsechansky.
Selective data acquisition for machine learning.
Cost-sensitive machine learning, 101, 2011.

Andrew Arnt and Shlomo Zilberstein.
Attribute measurement policies for time and cost sensitive classification.
In *Fourth IEEE International Conference on Data Mining (ICDM)*, pages 323–326. IEEE, 2004.

Maik Büttner, Christian Beyer, and Myra Spiliopoulou.
Reducing Missingness in a Stream through Cost-Aware Active Feature Acquisition.
In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2022.

Valentina Bayer-Zubek and Thomas G Dietterich.
Integrating learning from examples into the search for diagnostic policies.
Journal of Artificial Intelligence Research, 24:263–303, 2005.

Gabriella Contardo, Ludovic Denoyer, and Thierry Artières.
Recurrent neural networks for adaptive feature acquisition.
In *23rd International Conference on Neural Information Processing (ICONIP)*, pages 591–599. Springer, 2016.

References II

Xiaoyong Chai, Lin Deng, Qiang Yang, and Charles X Ling.
Test-cost sensitive naive Bayes classification.
In *Fourth IEEE International Conference on Data Mining (ICDM)*, pages 51–58, 2004.

Mumin Cebe and Cigdem Gunduz-Demir.
Test-cost sensitive classification based on conditioned loss functions.
In *European Conference on Machine Learning*, pages 551–558. Springer, 2007.

Ziheng Chen, Jin Huang, Hongshik Ahn, and Xin Ning.
Costly features classification using Monte Carlo tree search.
In *International joint conference on neural networks (IJCNN)*, pages 1–8, 2021.

Koby Crammer, Alex Kulesza, and Mark Dredze.
Adaptive regularization of weight vectors.
Machine learning, 91:155–187, 2013.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee.
Learning to maximize mutual information for dynamic feature selection.
In *International Conference on Machine Learning*, pages 6424–6447. PMLR, 2023.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan.
Learning to explain: An information-theoretic perspective on model interpretation.
In *International conference on machine learning*, pages 883–892. PMLR, 2018.

References III

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari.
Datum-wise classification: a sequential approach to sparsity.
In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 375–390, 2011.

Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari.
Sequential approaches for learning datum-wise sparse representations.
Machine learning, 89:87–122, 2012.

Mark Dredze, Koby Crammer, and Fernando Pereira.
Confidence-weighted linear classification.
In *Proceedings of the 25th international conference on Machine learning*, pages 264–271, 2008.

Marie Desjardins, James MacGlashan, and Kiri L Wagstaff.
Confidence-based feature acquisition to minimize training and test costs.
In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 514–524. SIAM, 2010.

Sachini Piyoni Ekanayake, Yasitha Warahena Liyanage, and Daphney-Stavroula Zois.
Dynamic feature selection for classification in structured environments.
In *55th Asilomar Conference on Signals, Systems, and Computers*, pages 140–144, 2021.

Sachini Piyoni Ekanayake and Daphney-Stavroula Zois.
Datum–Wise Inference in Structured Environments.
IEEE Transactions on Artificial Intelligence, 2023.

References IV

Sachini Piyoni Ekanayake, Daphney-Stavroula Zois, and Charalampos Chelmis.
Sequential Datum–wise Feature Acquisition and Classifier Selection.
IEEE Transactions on Artificial Intelligence, 2023.

Sachini Piyoni Ekanayake, Daphney-Stavroula Zois, and Charalampos Chelmis.
Sequential Datum–Wise Joint Feature Selection and Classification in the Presence of External Classifier.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

Santiago Gil-Begue, Concha Bielza, and Pedro Larrañaga.
Multi-dimensional Bayesian network classifiers: A survey.
Artificial Intelligence Review, 54(1):519–559, 2021.

Soham Gadgil, Ian Connick Covert, and Su-In Lee.
Estimating Conditional Mutual Information for Dynamic Feature Selection.
In *The Twelfth International Conference on Learning Representations*, 2023.

Isabelle Guyon and André Elisseeff.
An introduction to variable and feature selection.
Journal of machine learning research, 3(Mar):1157–1182, 2003.

Aritra Ghosh and Andrew Lan.
Difa: Differentiable feature acquisition.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7705–7713, 2023.

References V

Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney.
Causal feature selection for algorithmic fairness.
In *SIGMOD '22: International Conference on Management of Data*, pages 276–285, 2022.

Wenbo Gong, Sebastian Tschiatschek, Sebastian Nowozin, Richard E Turner, José Miguel Hernández-Lobato, and Cheng Zhang.
Icebreaker: Element-wise efficient information acquisition with a bayesian deep latent gaussian model.
Advances in neural information processing systems, 32, 2019.

He He, Hal Daumé III, and Jason Eisner.
Cost-sensitive dynamic feature selection.
In *ICML Inferning Workshop*, 2012.

Xuegang Hu, Peng Zhou, Peipei Li, Jing Wang, and Xindong Wu.
A survey on online feature selection with streaming features.
Frontiers of Computer Science, 12:479–493, 2018.

Shihao Ji and Lawrence Carin.
Cost-sensitive feature acquisition and classification.
Pattern Recognition, 40(5):1474–1485, 2007.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ.
Classification with costly features using deep reinforcement learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3959–3966, 2019.

References VI

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ.
Classification with costly features as a sequential decision-making problem.
Machine Learning, 109(8):1587–1615, 2020.

Majid Komeili, Narges Armanfard, and Dimitrios Hatzinakos.
Multiview feature selection for single-view classification.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10):3573–3586, 2020.

Jannik Kossen, Cătălina Cangea, Eszter Vértes, Andrew Jaegle, Viorica Patraucean, Ira Ktena, Nenad Tomasev, and Danielle Belgrave.
Active Acquisition for Multimodal Temporal Data: A Challenging Decision-Making Task.
Transactions on Machine Learning Research, 2023.

David P Kao, James D Lewsey, Inder S Anand, Barry M Massie, Michael R Zile, Peter E Carson, Robert S McKelvie, Michel Komajda, John JV McMurray, and JoAnn Lindenfeld.
Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response.
European journal of heart failure, 17(9):925–935, 2015.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra.
Planning and acting in partially observable stochastic domains.
Artificial intelligence, 101(1-2):99–134, 1998.

References VII

Athresh Karanam and Sriraam Natarajan.
On test-time active feature selection through tractable acquisition functions.
In *The 6th Workshop on Tractable Probabilistic Modeling*, 2023.

Dugang Liu, Pengxiang Cheng, Hong Zhu, Xing Tang, Yanyu Chen, Xiaoting Wang, Weike Pan, Zhong Ming, and Xiuqiang He.
DIWIFT: Discovering Instance-wise Influential Features for Tabular Data.
In *Proceedings of the ACM Web Conference 2023*, pages 1673–1682, 2023.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra.
Continuous control with deep reinforcement learning.
arXiv preprint arXiv:1509.02971, 2015.

Interpretability in the context of sequential cost-sensitive feature acquisition, author=Liyanage, Yasitha Warahena and Zois, Daphney–Stavroula.
In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

Scott M Lundberg and Su-In Lee.
A unified approach to interpreting model predictions.
Advances in neural information processing systems, 30, 2017.

Hongfu Liu, Haiyi Mao, and Yun Fu.
Robust multi-view feature selection.
In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 281–290, 2016.

References VIII

Yang Li and Junier Oliva.
Active feature acquisition with generative surrogate models.
In *International Conference on Machine Learning*, pages 6450–6459. PMLR, 2021.

Yasitha Warahena Liyanage and Daphney-Stavroula Zois.
Optimum feature ordering for dynamic instance–wise joint feature selection and classification.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3370–3374, 2021.

Yasitha Warahena Liyanage, Daphney-Stavroula Zois, and Charalampos Chelmis.
Dynamic instance-wise classification in correlated feature spaces.
IEEE Transactions on Artificial Intelligence, 2(6):537–548, 2021.

Yasitha Warahena Liyanage, Daphney-Stavroula Zois, and Charalampos Chelmis.
Dynamic instance-wise joint feature selection and classification.
IEEE Transactions on Artificial Intelligence, 2(2):169–184, 2021.

Kachuee Mohammad, Goldstein Orpaz, Kärkkäinen Kimmo, Darabi Sajad, and Sarrafzadeh Majid.
Opportunistic learning: Budgeted cost-sensitive learning from data streams.
In *International Conference on Learning Representations*, 2019.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare.
Safe and efficient off-policy reinforcement learning.
Advances in neural information processing systems, 29, 2016.

References IX

Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney.
An expected utility approach to active feature-value acquisition.
In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang.
EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE.
In *International Conference on Machine Learning*, pages 4234–4243. PMLR, 2019.

Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy.
Instance-wise feature grouping.
Advances in Neural Information Processing Systems, 33:13374–13386, 2020.

Imara Nazar, Daphney-Stavroula Zois, and Charalampos Chelmis.
Knowing When to Stop: Joint Heterogeneous Feature Selection and Classification.
In *54th Asilomar Conference on Signals, Systems, and Computers*, pages 1227–1231, 2020.

Simon Perkins, Kevin Lacker, and James Theiler.
Grafting: Fast, incremental feature selection by gradient descent in function space.
The Journal of Machine Learning Research, 3:1333–1356, 2003.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
"why should i trust you?" explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

References X

Hajin Shim, Sung Ju Hwang, and Eunho Yang.
Joint active feature acquisition and classification with variable-size set encoding.
Advances in neural information processing systems, 31, 2018.

Victor S Sheng and Charles X Ling.
Feature value acquisition in testing: a sequential batch test algorithm.
In *23rd international conference on Machine learning*, pages 809–816, 2006.

John Stamper, Alexandru Niculescu-Mizil, S. Ritter, Geoff Gordon, and Ken Koedinger.
Challenge data set from kdd cup 2010 educational data mining challenge, 2010.
Find it at http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp.

Kirill Trapeznikov and Venkatesh Saligrama.
Supervised sequential classification under budget constraints.
In *Artificial intelligence and statistics*, pages 581–589. PMLR, 2013.

Hado Van Hasselt, Arthur Guez, and David Silver.
Deep reinforcement learning with double q-learning.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Joseph Wang, Tolga Bolukbasi, Kirill Trapeznikov, and Venkatesh Saligrama.
Model selection by linear programming.
In *13th European Conference on Computer Vision (ECCV)*, pages 647–662. Springer, 2014.

References XI

Steve Webb, James Caverlee, and Calton Pu.
Introducing the webb spam corpus: Using email spam to identify web spam automatically.
In *CEAS*, 2006.

Yue Wu, Steven CH Hoi, Tao Mei, and Nenghai Yu.
Large-scale online feature selection for ultra-high dimensional sparse data.
ACM Transactions on Knowledge Discovery from Data (TKDD), 11(4):1–22, 2017.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas.
Dueling network architectures for deep reinforcement learning.
In *International conference on machine learning*, pages 1995–2003, 2016.

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama.
An lp for sequential learning under budgets.
In *Artificial intelligence and statistics*, pages 987–995. PMLR, 2014.

Xindong Wu, Kui Yu, Wei Ding, Hao Wang, and Xingquan Zhu.
Online feature selection with streaming features.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(5):1178–1192, 2012.

Jialei Wang, Peilin Zhao, Steven CH Hoi, and Rong Jin.
Online feature selection and its applications.
IEEE Transactions on knowledge and data engineering, 26(3):698–710, 2013.

References XII

Zhixiang Xu, Matt J Kusner, Kilian Q Weinberger, Minmin Chen, and Olivier Chapelle.
Classifier cascades and trees for minimizing feature evaluation cost.
The Journal of Machine Learning Research, 15(1):2113–2144, 2014.

Qi Xiao, Hebi Li, Jin Tian, and Zhengdao Wang.
Group-wise feature selection for supervised learning.
In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 3149–3153. IEEE, 2022.

Qi Xiao and Zhengdao Wang.
Mixture of deep neural networks for instancewise feature selection.
In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–921. IEEE, 2019.

Wanqi Yang, Yang Gao, Yinghuan Shi, and Longbing Cao.
MRM-lasso: A sparse multiview feature selection method via low-rank analysis.
IEEE transactions on neural networks and learning systems, 26(11):2801–2815, 2015.

Jinsung Yoon, James Jordon, and Mihaela Van der Schaar.
INVASE: Instance-wise variable selection using neural networks.
In *International conference on learning representations*, 2018.

Kui Yu, Xindong Wu, Wei Ding, and Jian Pei.
Scalable and accurate online feature selection for big data.
ACM Transactions on Knowledge Discovery from Data (TKDD), 11(2):1–39, 2016.

References XIII

Jing Zhou, Dean Foster, Robert Stine, and Lyle Ungar.
Streaming feature selection using alpha-investing.
In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 384–393, 2005.

Peng Zhou, Xuegang Hu, Peipei Li, and Xindong Wu.
OFS-Density: A novel online streaming feature selection method.
Pattern Recognition, 86:48–61, 2019.

Peng Zhou, Xuegang Hu, Peipei Li, and Xindong Wu.
Online streaming feature selection using adapted neighborhood rough set.
Information Sciences, 481:258–279, 2019.

Sara Zannone, José Miguel Hernández-Lobato, Cheng Zhang, and Konstantina Palla.
Odin: Optimal discovery of high-value information using model-based deep reinforcement learning.
In *ICML Real-world Sequential Decision Making Workshop*, 2019.

Hong Zhao, Qinghua Hu, Pengfei Zhu, Yu Wang, and Ping Wang.
A recursive regularization based feature selection framework for hierarchical classification.
IEEE Transactions on Knowledge and Data Engineering, 33(7):2833–2846, 2019.

Peng Zhou, Shu Zhao, Yuanting Yan, and Xindong Wu.
Online scalable streaming feature selection via dynamic decision.
ACM Transactions on Knowledge Discovery from Data (TKDD), 16(5):1–20, 2022.