Feature Selection

Bruno Iochins Grisci

Generative AI Academy

Part 3: Evaluation and Visualization of Feature Selection

Overview

Selection Accuracy

Selection Accuracy

How well does the method recover truly informative features?

- Especially meaningful for synthetic datasets.
- Requires ground truth knowledge of informative features.

Two key metrics:

- Percentage of Informative Features Selected (PIFS)
- Percentage of Selected Features that are Informative (PSFI)

PIFS: Percentage of Informative Features Selected

- Measures **recall**: how many informative features were recovered.
- Defined as:

$$\text{PIFS} = \frac{|S_{\text{selected}} \cap S_{\text{informative}}|}{|S_{\text{informative}}|}$$

- S_{selected}: Features selected by the algorithm
- $S_{\text{informative}}$: Known informative features

High PIFS means the method retrieves many useful features.

PSFI: Percentage of Selected Features that are Informative

- Measures **precision**: how many selected features are truly relevant.
- Defined as:

$$\text{PSFI} = \frac{|S_{\text{selected}} \cap S_{\text{informative}}|}{|S_{\text{selected}}|}$$

- Focuses on avoiding selection of irrelevant or noisy features.

High PSFI means the method selects mostly relevant features.

Predictive Power

Predictive Power

How well do selected features support prediction?

- Evaluate using model accuracy, F1-score, AUC, etc.
- Compare performance with full vs. selected feature set.
- Useful when ground truth informative features are unknown.

Goal: Maintain or improve prediction quality with fewer features.

Redundancy

Redundancy

Redundant features do not add new information.

- May be correlated with already selected features.
- Can cause overfitting and increase model complexity.
- Ideally, selected features should be minimally redundant.

Common measures:

- Pearson correlation
- Mutual Information among features
- Redundancy term in mRMR

$$R(f, S) = \frac{1}{|S|} \sum_{x_i \in S} I(f, x_i) \tag{1}$$

Where f is a random variable, S is a set of random variables, x_i is the i-th value in the variable x, and $I(x, f)$ is the Mutual Information between x and f.

Stability and Reliability

Stability and Reliability

Do selected features vary significantly with small changes in data?

- Critical for reproducibility and scientific trust.
- Evaluate by comparing results from multiple dataset variations.
- Typical method: bootstrap resampling or subset sampling.

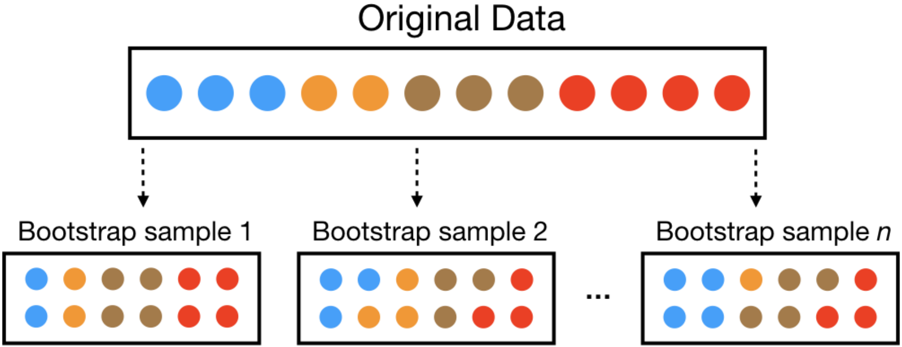
Measuring Stability

Desired properties of stability measures:

- Value in $[0, 1]$ where $1 =$ identical selections.
- Symmetric and permutation invariant.
- Handle different list lengths and partial overlaps.

Apply on results from k resampled subsets.

Bootstrap



https://datasciencedojo.com/blog/bootstrap-sampling/

Properties of Stability Metrics

Important criteria for evaluating stability indices:

- **Fully Defined:** The measure is computable for different sizes of selected subsets.
- **Bounds:** Result lies in a fixed interval, often $[0, 1]$ or $[-1, 1]$.
- **Maximum:** The measure achieves maximum stability when selections are identical.
- **Correction for Chance:** Adjusted for overlap that may occur randomly.
- **Monotonicity:** Value decreases as overlap between sets decreases.
- **Input Type:** Subsets, Rank, or Weights

Kuncheva Index

Stability measure accounting for overlap by chance.

$$KI = \frac{|A \cap B|n - k^2}{k(n-k)}$$

- A, B: feature subsets of size k
- n: total number of features

Range: $[-1, 1]$, with 1 indicating perfect agreement.

Subsets must be of the same size.

Other Set Similarity Metrics

- **Jaccard Index:** $\frac{|A \cap B|}{|A \cup B|}$
- **Hamming Index:** Measures symmetric difference
- **Ochiai Index:** $\frac{|A \cap B|}{\sqrt{|A||B|}}$
- **Dice Index:** $\frac{2|A \cap B|}{|A|+|B|}$
- **Percentage of Overlapping Features:** $\frac{|A \cap B|}{|A|}$

Used to compare binary or ranked feature selection results.

Distance and Correlation Metrics

- **Canberra Distance:** Weighted absolute differences

$$D(x, y) = \sum \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

- **Spearman's Rank Coefficient:** Rank correlation
- **Pearson Correlation Coefficient:** Linear correlation of scores

Properties of stability measures

Metric	Fully Defined	Bounds	Maximum	Correction for Chance	Monotonicity	Result Type
Jaccard	✓	✓	✓		✓	Subsets
Hamming	✓	✓	✓		✓	Subsets
Dice	✓	✓	✓		✓	Subsets
Ochiai	✓	✓	✓		✓	Subsets
POG	✓	✓	✓		✓	Subsets
Kuncheva		✓	✓	✓	✓	Subsets
Canberra	✓	✓		✓		Rank
Spearman	✓	✓		✓	✓	Rank
Pearson	✓	✓		✓	✓	Weights

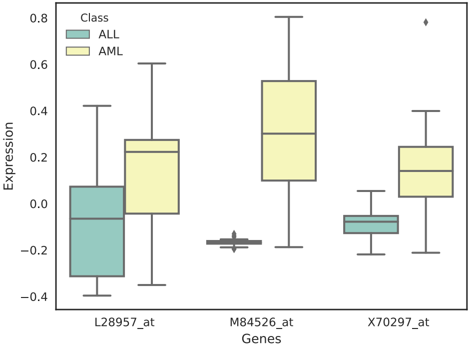
Visualization

Boxplot

Visualize distribution of feature values across classes.

- Highlights outliers, medians, and variability.
- Useful for single-feature analysis.

Boxplot

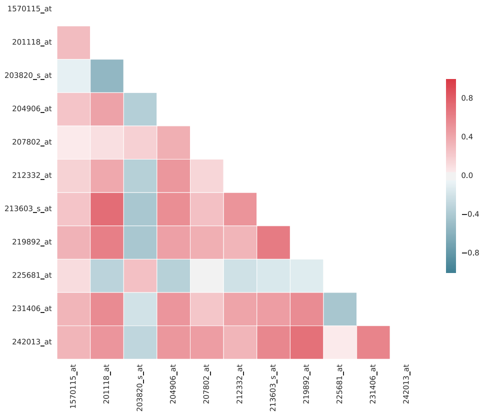

Class
ALL
AML
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Expression
L28957_at
M84526_at
X70297_at
Genes

Heatmap

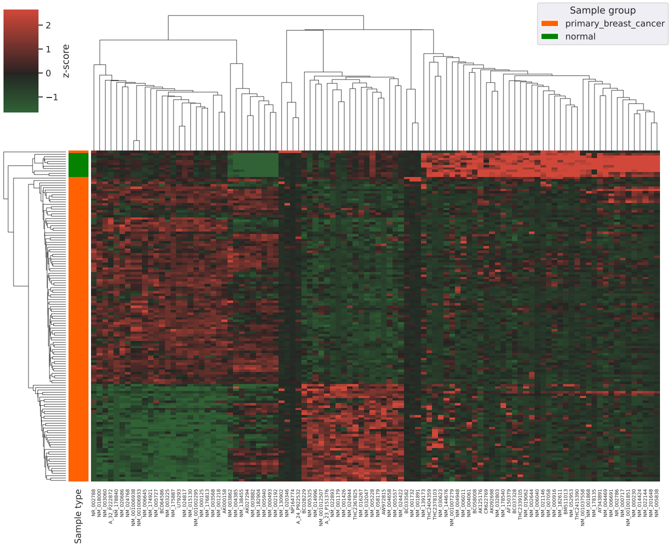
Visualize pairwise correlations or relevance scores.

- Often used for redundancy analysis.
- Also applies to expression levels or score matrices.

Heatmap: Correlation



Heatmap: Feature Values



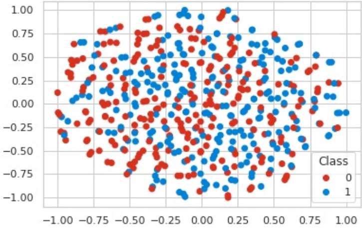
Heatmap: Importance

	SCORE	0	1	2			...			...			...
REL002	0,38	0,36	0,75	0,20	1,20	-0,55	...	0,11	-0,94	...	2,33	0,20	...
RED005	0,27	0,14	0,24	0,57	-1,67	-1,83	...	-0,76	-1,21	...	4,01	1,46	...
REL001	0,26	0,18	0,17	0,55	-2,11	-2,25	...	-1,61	-1,72	...	0,19	0,99	...
REL003	0,22	0,52	0,20	0,11	-1,06	0,74	...	0,66	1,34	...	2,11	0,50	...
RED004	0,07	0,04	0,09	0,09	-0,22	-0,15	...	-0,25	-0,14	...	-0,93	-0,08	...
IRR083	0,03	0,02	0,02	0,05	-0,14	-0,27	...	-1,86	1,12	...	-0,84	1,23	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
IRR801	0,02	0,02	0,02	0,03	0,50	-0,89	...	-0,50	0,86	...	0,67	-2,03	...
IRR082	0,01	0,01	0,01	0,00	-0,42	-0,27	...	0,50	0,65	...	0,57	-0,43	...

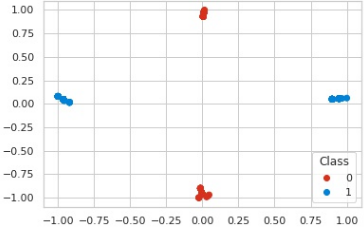
t-SNE (t-distributed Stochastic Neighbor Embedding)

Non-linear dimensionality reduction technique.

- Preserves local structure in 2D or 3D plots.
- Helps visualize class separability.



(a) All features



(b) Informative features

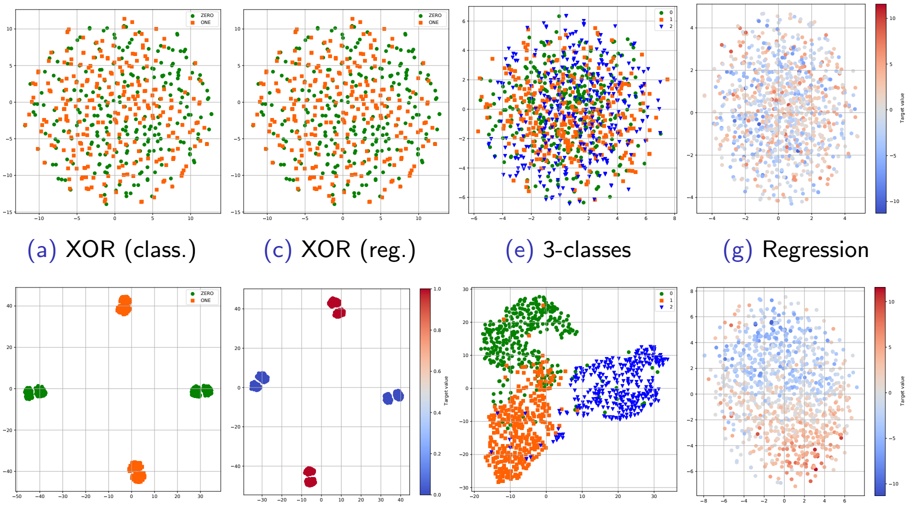
Weighted t-SNE

Modified t-SNE using feature relevance as weights.

- Highlights relevant features in distance computation.
- Enhances visual interpretability post-FS.

$$d(p,q) = \sum_{i=1}^{n} \sqrt{(w_i(q_i - p_i))^2} \quad (2)$$

Weighted t-SNE



(a) XOR (class.) (c) XOR (reg.) (e) 3-classes (g) Regression

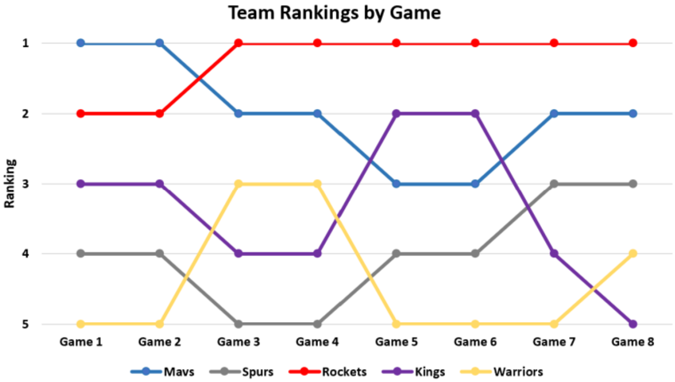
(b) XOR (class) wgt (d) XOR (reg) wgt (f) 3-classes wgt. (h) Regression wgt.

Bump Chart

Displays feature importance or selection order.

- Provides visualization of the stability of FS methods.
- Horizontal bar chart or line ranking plots.

Bump Chart


Team Rankings by Game
Ranking
1
2
3
4
5
Game 1
Game 2
Game 3
Game 4
Game 5
Game 6
Game 7
Game 8
Mavs
Spurs
Rockets
Kings
Warriors

Visualization of Feature Selection

Why use visualization?

- Understand feature relevance patterns.
- Inspect class separability and structure.
- Reveal bias, redundancy, or dataset artifacts.

Examples:

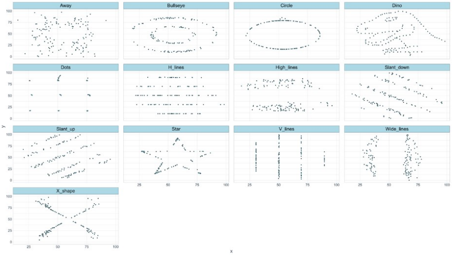
- Datasaurus dozen: identical stats, different plots.
- Random forest importance plots.

Same stats, different figures

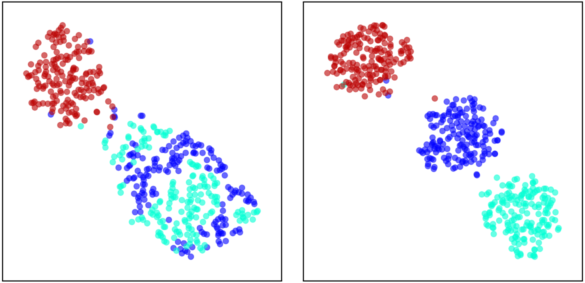
```
dataset      x     y
<chr>    <dbl> <dbl>
dino      55.4  97.2
dino      51.5  96.0
dino      46.2  94.5
dino      42.8  91.4
dino      40.8  88.3
dino      38.7  84.9
dino      35.6  79.9
dino      33.1  77.6
dino      29.0  74.5
dino      26.2  71.4
ℹ 1,836 more rows
```

Common statistical values for each group in the dataset

	Summary statistics					*Regression results*	
Dataset	Mean x	Mean y	Std Dev x	Std Dev y	Corr x y	Intercept	Coefficients
Away	54.27	47.83	16.77	26.94	-0.06	53.43	-0.10
Bullseye	54.27	47.83	16.77	26.94	-0.07	53.81	-0.11
Circle	54.27	47.84	16.76	26.93	-0.07	53.80	-0.11
Dino	54.26	47.83	16.77	26.94	-0.06	53.45	-0.10
Dots	54.26	47.84	16.77	26.93	-0.06	53.10	-0.10
H_lines	54.26	47.83	16.77	26.94	-0.06	53.21	-0.10
High_lines	54.27	47.84	16.77	26.94	-0.07	53.81	-0.11
Slant_down	54.27	47.84	16.77	26.94	-0.07	53.85	-0.11
Slant_up	54.27	47.83	16.77	26.94	-0.07	53.81	-0.11
Star	54.27	47.84	16.77	26.93	-0.06	53.33	-0.10
V_lines	54.27	47.84	16.77	26.94	-0.07	53.89	-0.11
Wide_lines	54.27	47.83	16.77	26.94	-0.07	53.63	-0.11
X_shape	54.26	47.84	16.77	26.93	-0.07	53.55	-0.11



Random Forest



(a) Random Forest A   (b) Random Forest B

Libraries

- Matplotlib
- seaborn
- Plotly
- openTSNE
- ggplot2 (R)

Time

Time and Efficiency

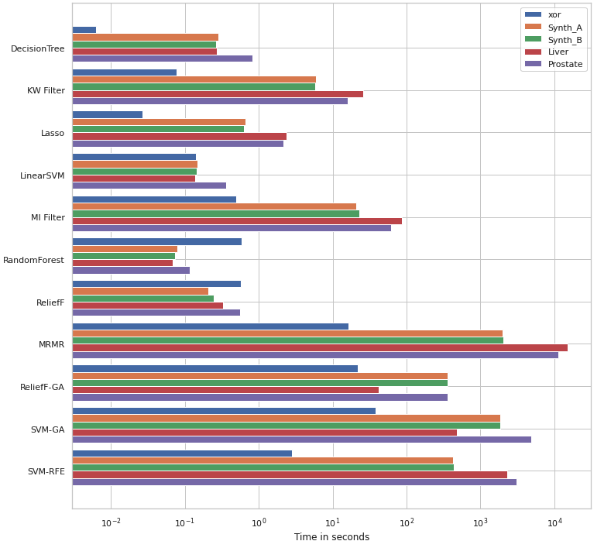
Why evaluate execution time?

- FS methods vary widely in computational cost.
- Essential when scaling to large datasets.
- Balance trade-offs between speed and performance.

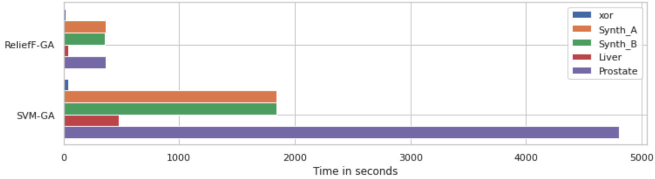
Compare time across:

- Filter, wrapper, embedded methods
- Simple vs metaheuristic-based approaches

Time Comparison I



Time Comparison II



Comparison of FS Algorithms on Classification Tasks

ReliefF

Strengths:

- Fast filter method.
- Good selection accuracy, prediction, stability, and reliability.
- Effective even with strong feature correlation.
- Well-suited for small sample sizes.

Drawbacks:

- Does not consider redundancy between features.

Random Forest

Strengths:

- Fast and robust with nonlinear data.
- Comparable to ReliefF in nonlinear tasks.
- Faster on small datasets.

Drawbacks:

- Performs well primarily on specific tasks (e.g., XOR dataset).
- Less consistent on general datasets.

SVM-RFE

Strengths:

- Best performance for high-dimensional datasets.
- Handles redundancy reasonably well.

Drawbacks:

- Poor on XOR-type nonlinear problems.
- Prone to overfitting with small datasets.

SVM Embedded and LASSO

Strengths:

- Good balance between accuracy and computational cost.
- Suitable for large datasets.

Drawbacks:

- Less accurate than SVM-RFE.
- Does not explicitly manage redundancy.

SVM-GA (Genetic Algorithm)

Strengths:

- Competitive for small-sample datasets.
- Less overfitting than SVM-RFE or embedded methods.
- Flexible execution time based on parameters.

Drawbacks:

- Computational cost varies with configuration.
- More expensive than filters.

mRMR (Minimum Redundancy Maximum Relevance)

Strengths:

- Best for handling redundancy.

Drawbacks:

- Weak performance in prediction.
- Computationally expensive — sometimes slower than wrappers.

Decision Tree (Embedded)

Strengths:

- Good stability on prostate and liver datasets.

Drawbacks:

- Stability possibly biased by fixed subset sizes.
- Less effective in high-dimensional data.

Mutual Information and Kruskal-Wallis Filters

Strengths:

- Simple and intuitive statistical filters.

Drawbacks:

- Computationally slower than expected.
- Performance limited by MI estimation or library inefficiencies.

General Computational Limitations

Current limitations of the implementations:

- Cannot handle incomplete data.
- No support for regression tasks.
- No multi-threading or parallelism.

Parallelization opportunities:

- ReliefF, Kruskal-Wallis, MI Filter, mRMR (instance-wise or pair-wise).
- Random Forest and ensembles (tree creation and prediction).
- Genetic Algorithms (crossover, mutation, fitness evaluation).
- GPU acceleration for vectorized computations.