Enhancing Feature Selection Strategies for Imbalanced and High-dimensional Data

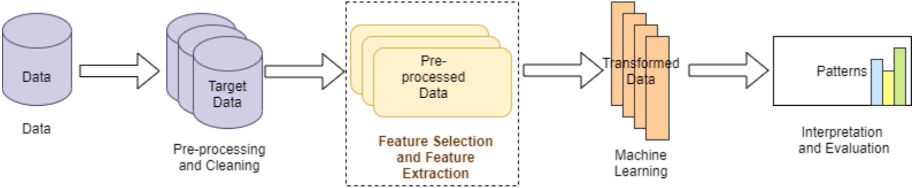
Surani Matharaarachchi, Ph.D. [1]
Joint work with:
Dr. Saman Muthukumarana [2] Dr. Mike Domaratzki [3]

[1]New York Institute of Technology [2]University of Manitoba [3]Western University

WORLD STATISTICS CONGRESS 2025 THE HAGUE

Feature Selection

Definition. Given $(X_1, \ldots, X_p)$ and response Y, *feature selection* seeks an index set $S \subset \{1, \ldots, p\}$ with $|S| \ll p$ such that a predictor $f_S : \mathbb{R}^{|S|} \to \mathcal{Y}$ achieves low generalization risk.



Motivation for Feature Selection

- **High dimensionality:** In modern datasets ($p \gg n$), model complexity grows rapidly, causing the *curse of dimensionality*.
- **Redundancy and noise:** Irrelevant features obscure true signals and weaken predictive accuracy.
- **Overfitting:** Using all features fits noise rather than structure; selection serves as an implicit regularization step.
- **Efficiency:** Fewer parameters reduce variance and improve model stability.
- **Interpretability:** A compact subset enhances understanding and scientific insight.

Main types of feature selection methods

Table: Differences between Filter and Wrapper

Filter Method	Wrapper Method
Measure the relevance of features.	Measure the usefulness of a subset of features.
Use statistical methods for evaluation of a subset of features.	Evaluates on a specific machine-learning algorithm to find optimal features.
Much faster.	Computationally expensive .
Less prone to over-fitting.	High chance of over-fitting.
Sometimes may fail to select best features.	Better performance.
Eg: Pearson's Correlation, LDA, ANOVA, Chi-Square	Eg: Forward selection, Backward elimination, RFE

Contribution

A Unified Approach for Feature Selection [5]

Machine Learning with Applications 6 (2021) 100170

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Assessing feature selection method performance with class imbalance data

Surani Matharaarachchi [a,*], Mike Domaratzki [b,1], Saman Muthukumarana [a]

[a] Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
[b] Department of Computer Science, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada

Identifying a Method that Extracts the Most Informative Features

- Feature selection plays a crucial role in high-dimensional settings, improving interpretability, reducing variance, and avoiding overfitting.

- Our objective is twofold:

  1. To identify the most effective **feature ordering mechanism**, capable of ranking features by informativeness.

  2. To develop a **unified feature subset selection procedure** that optimally balances dimensionality reduction and predictive accuracy.

What is the Best Feature Ordering Technique? I

We compared four feature ordering techniques with the aim of identifying the most stable and informative ranking across different data settings.

1. **Model-Based Feature Importance**
   - Derived from supervised models that incorporate variable regularization or splitting criteria.
     1. **Coefficient-based Models:** Logit or SVM-Linear - magnitude of standardized coefficients $|\beta_j|$ as feature importance.
     2. **Tree-based Models:** Decision Trees, Random Forests, Gradient Boosting - use impurity reduction (Gini/entropy) or information gain [4].
   - These are inherently data-adaptive but may be sensitive to imbalance and feature scaling.

What is the Best Feature Ordering Technique? II

2. **Univariate Feature Selection (ANOVA F-Value Classification)**
   - Each feature is independently evaluated against the response variable using a one-way ANOVA F-test.
   - The F-statistic quantifies the ratio of between-class to within-class variability:

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}}.$$

   - Higher F-values indicate stronger discriminatory power; features are ranked accordingly.

What is the Best Feature Ordering Technique? III

3. **Absolute Correlation with the Response Variable**
   - For continuous or binary responses, the point-biserial correlation coefficient r_{pb} is computed:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_X} \sqrt{\frac{n_1 n_0}{n^2}}.$$

   - Features with high $|r_{pb}|$ values exhibit stronger linear association with the target variable.
   - However, this approach ignores inter-feature dependencies and nonlinear effects.

What is the Best Feature Ordering Technique? IV

4. **Summation of the Absolute Values of Principal Component (PC) Loadings (PCL)** [1]
   - In Principal Component Analysis (PCA), each component is a linear combination of the standardized variables:

     $$PC_k = w_{k1}\underline{X}_1 + w_{k2}\underline{X}_2 + \ldots + w_{kp}\underline{X}_p,$$

     where w_{kj} denotes the loading of variable j on component k.
   - The absolute magnitude of w_{kj} represents the contribution (importance) of feature X_j to the variance captured by the k-th component.
   - To assess the overall influence of a variable, we sum the absolute loadings across the first k principal components:

     $$\text{Score}(X_j) = \sum_{i=1}^{k} |w_{ij}|.$$

Simulation Design

- A Monte Carlo simulation was conducted to compare feature ranking techniques under varying data conditions.
- Experimental factors:
  - Sample size: $n \in \{200, 500, 1000\}$
  - Number of informative features: p_{inf}
  - Class imbalance: balanced, moderate, and severe
- Each scenario was replicated 100 times for stability and reproducibility.
- **Performance metric:**

$$\text{Informative Selection Rate} = \frac{\text{Mean number of correctly identified informative features}}{p_{\text{inf}}}.$$

- The expected selection range corresponds to the true number of informative features embedded in the dataset.

Simulation Results

- PCL (blue line) consistently identifies the highest proportion of informative features.
- Logit-based model (red dashed) performs comparably for large n but less stable in imbalanced settings.
- ANOVA F-value and absolute correlation overlap, both sensitive to multicollinearity.
- Empirically, the PCL method provides a robust ordering mechanism by capturing joint variance contributions.

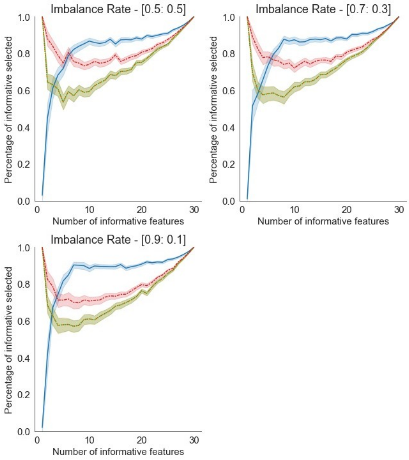


Figure: Feature selection accuracy for $n = 200$.

Deriving the Most Informative Feature Subset

- After establishing the most reliable ordering mechanism, we propose a systematic subset extraction algorithm - the **Principal Component Loading Feature Selection (PCLFS)** method.

Principal Component Loading Feature Selection (PCLFS)

- The theoretical rationale integrates variance-based ranking with performance-based selection.

Step 1: Perform PCA on standardized training data to obtain loading matrix $W_{k\times p}$.

Step 2: Compute feature importance scores $\sum_{i=1}^{k} |w_{ij}|$ and order features accordingly.

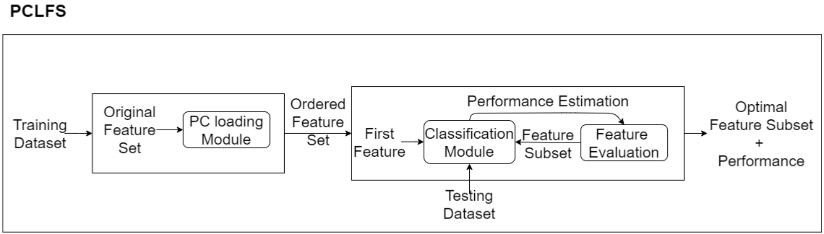
Step 3: Iteratively fit a classification model (e.g., Logistic Regression) starting from the top-ranked feature and cumulatively add one feature at a time.

Step 4: Evaluate performance using F1-score on the validation or test set.

Step 5: Select the subset size p^* that maximizes the F1-score:

$$p^* = \arg\max_{p} F1(p).$$

Principal Component Loading Feature Selection (PCLFS)



The PCLFS method combines the interpretability of PCA with predictive validation, yielding a stable, data-driven approach to feature selection that respects feature correlations and maximizes generalization performance.

Simulation Scenarios

Variable	No. of Levels	Levels / Descriptions
Methods	3	RFE, PCLFS, PCLFS-Extended
Classification Models	5	Logit, SVM-Linear, Decision Trees, Random Forest (RFC), LightGBM (LGBM_C)
Training Sets	2	Original, SMOTE
Imbalance Rates	3	50%:50%, 70%:30%, 90%:10%
No. of Features	1	30
No. of Informative Features	30	1–30 (increment of 1)
Sample Sizes	2	200, 1000
Performance Evaluation Metrics	3	$\text{F1-score}_{\text{model}}$, $\text{Correct\_Percentage}_{\text{fs}}$, TPR_{fs}
Repeat Samples	100	Each scenario replicated 100 times

Simulation Results

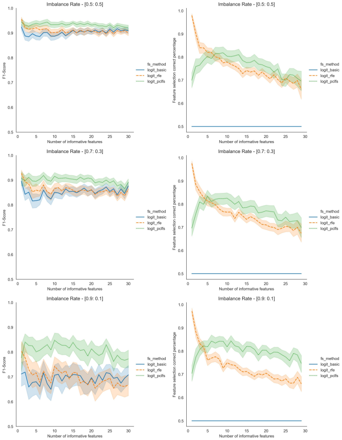


Figure: Final model F1-scores and feature selection correct percentages for the Logit model, without SMOTE when sample size is 1000 and threshold is 0.0017.

SPECTF Heart Data

- The **SPECTF Heart Dataset** [2, 3] is a publicly available benchmark for diagnosing cardiac abnormalities using *Single Photon Emission Computed Tomography (SPECT)* imaging.
- Each record corresponds to one patient and is labeled as either:
  - **Normal**, or
  - **Abnormal** (presence of cardiac abnormality).
- The dataset contains:
  - 267 patient samples (image-derived feature sets)
  - 44 continuous diagnostic features per patient
- Data were randomly divided into:

  $$\text{Training: } 75\% \quad \text{and} \quad \text{Testing: } 25\%.$$

- The dataset is **class-imbalanced** with a ratio of 80%:20%, where the minority class represents patients with abnormal cardiac function.

Application Results Comparison

Table: Final F1-score comparison between RFE and proposed methods (PCLFS).

SMOTE	Method	Basic		**RFE**		**PCLFS**		Feature reduction%/ (increment%)	F1-score (reduction)/ increment
		#Features	F1-scores	#Features	F1-scores	#Features	F1-scores		
TRUE	Logit	44	0.6809	36	0.6957	24	0.6957	56.8%	(0.0018)
	LGBM	44	0.6667	27	0.6286	13	0.7027	31.8%	0.0741
	Decision Tree	44	0.5556	44	0.5556	9	0.6667	93.2%	0.1110
	RFC	44	0.6486	38	0.6111	42	0.7059	59.0%	0.0731
	SVM-Linear	44	0.6511	30	0.6977	12	0.7727	40.9%	0.0750
FALSE	Logit	44	0.5455	30	0.5000	44	0.5455	(31.8%)	0.0455
	LGBM	44	0.6250	15	0.5455	15	0.6250	0.0%	0.0795
	Decision Tree	44	0.5294	27	0.5161	9	0.5946	40.9%	0.0785
	RFC	44	0.2609	9	0.3704	11	0.4444	(4.5%)	0.0740
	SVM-Linear	44	0.5946	21	0.5882	37	0.6316	(36.4%)	0.0434

Discussion

- Using the summation of the absolute values of principle component loadings, features can be ordered from most informative to the least.
- Feature ordering is entirely independent of the classification model.
- Combined results returns "The most informative feature subset with minimal number of features with similar performance".
- Proposed methods makes a reasonable improvement over RFE results.
- Python and Digital Research Alliance of Canada facility was used.
- An extended version of the PCLFS is published in the Journal of PeerJ Computer Science [6].

References

[1] Dunteman, G. (1989). Using principal components to select a subset of variables. In *Principal Components Analysis*, Quantitative Applications in the Social Sciences. Newbury Park: SAGE Publications, Inc.

[2] Krzysztof, J. C., K. W. Daniel, and L. Ning (1997). Clip3: Cover learning using integer programming. *26*(5).

[3] Kurgan, L. A., K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. S. Goodenday (2001). Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial Intelligence in Medicine 23*(2), 149–169.

[4] Ryzin, J. V. (1986). Breiman, leo, friedman, jerome h., olshen, richard a., and stone, charles j., "classification and regression trees" (book review). *Journal of the American Statistical Association 81*(393), 253–.

[5] **Matharaarachchi, S**, M. Domaratzki, and S. Muthukumarana (2021). Assessing feature selection method performance with class imbalance data. *Machine learning with applications 6*, 100170–.

[6] **Matharaarachchi, S**, M. Domaratzki, and S. Muthukumarana (2022). Minimizing features while maintaining performance in data classification problems. *PeerJ Computer Science 8*, e1081.

Thank You!

Contact: smathara@nyit.edu
Personal Website: https://suranimatharaarachchi.com