Assessing Feature Selection Methods and their Performance in High Dimensional Classification Problems

Presented by: Surani Matharaarachchi

Supervised by: Dr. Saman Muthukumarana & Dr. Mike Domaratzki

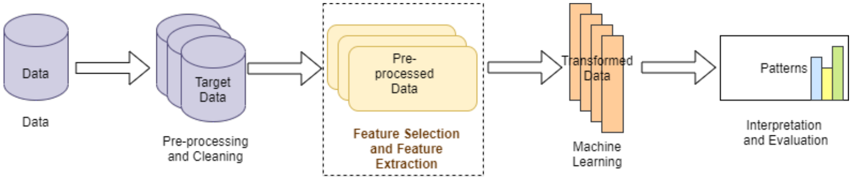
June, 10 2021

Outline

1. Introduction
2. Selecting Minimal Number of Features with Similar Performance
3. Identifying a Method that Extracts the Most Informative Features
4. Combining Proposed Methods
5. Discussion
6. Acknowledgment

Feature Selection

Selecting a subset from the original feature set is called “feature selection".



Motivation

Two main objectives of feature selection:

1. Minimising the number of features
2. Identifying the most informative features

- while achieving higher accuracy [1, 6]

Part I

Selecting Minimal Number of Features with Similar Performance

Motivation

- Wrapper feature selection methods select the subset which gives the maximum score.
- There may be other selections of a lower number of features with a lower-scoring value, yet the difference is negligible.

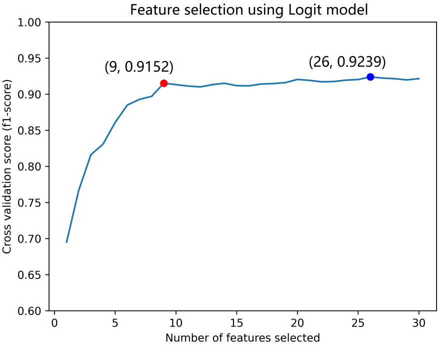


Figure: The blue point indicates the RFE feature selection whereas the red point explains the same for the proposed method.

Suggested Method I

- An extension of the Wrapper feature selection method.
- The exiting Recursive Feature Elimination (RFE) [4] chooses the feature subset giving the best scoring value in cross-validation.
- The suggested method identifies a feature subset under an applicable threshold to obtain the smaller feature subset with minimal loss.

Algorithm

inputs:

- Grid scores: $g = [g_1, g_2, \ldots, g_m]$
- Number of selected features by RFE: n_{rfe}
- Total number of features: n
- Feature importance scores (obtained from the classifier): $i = [i_1, i_2, \ldots, i_{n_{rfe}}]$
- Maximum tolerable F1-score reduction: T (User-defined)

procedure:

Step 1: Consider all the local maximum grid scores (g_j) corresponding to the number of subsets of features selected by RFE which is less than the optimal number of features selected (n_{rfe}) where,

$$g_j > max(g_{j-1}, g_{j+1}), \qquad j < n_{rfe}$$

Step 2: Connect each point with the maximum point and compute each line's gradient values.

Motivation

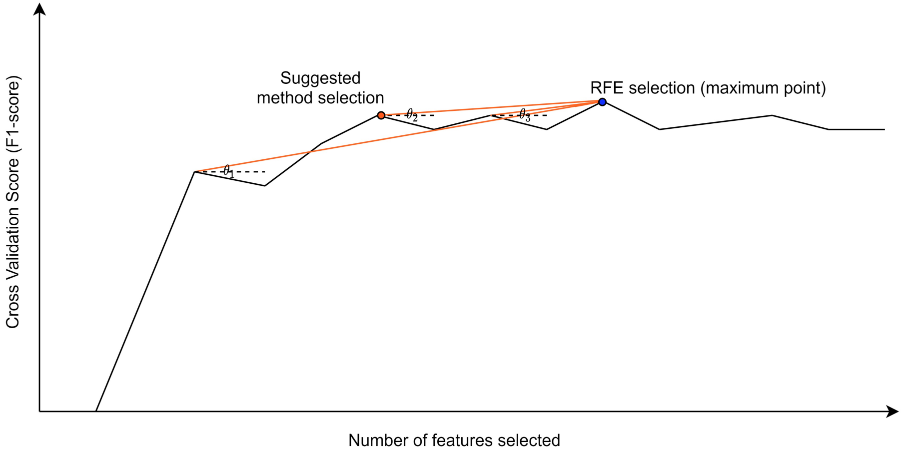


Figure: Graphical view of the suggested algorithm

Algorithm Cont.

Step 3: Compare the gradient values with a threshold value.

$$\text{gradient}(Tan(\theta_j)) = \frac{(\Delta y)_j}{(\Delta x)_j} < \text{Threshold}$$

The threshold (t) can be interpreted as the tolerable reduction of the F1-score to reduce one feature,

$$\text{Threshold }(t) = \frac{\text{Maximum tolerable F1score reduction}}{\text{Total number of features}} = \frac{T}{n}$$

Algorithm Cont.

Step 4: Obtain the F1- score which gives the smallest number of features ($n_{proposed}$).

Note: If there is no value found for the given condition, return the same RFE results.

Step 5: To get the relevant feature subset, use feature importance scores (i).

Then obtain the best $n_{proposed}$ number of features as the smallest feature subset with similar performance (s).

output:

- The smallest number of features with minimum scoring loss: $n_{proposed}$
- Relevant feature subset: s

Part II

Identifying a method that extracts the most informative features

Identifying a method that extracts the most informative features

1. Identifying the best feature ordering technique.

2. Identifying a method that extract the best informative feature subset.

What is the best feature ordering technique? I

We used four different feature ordering methods to compare the feature ordering behavior.

1. **Summation of the absolute values of PC loadings (PCL)**
   - The PC loadings [3] are the coefficients of the linear combination of the original variables.
   - In PCA, with n sample and p variables, the first k principal components are given by,

$$\begin{aligned} PC_1 &= w_{11}\underline{X}_1 + w_{12}\underline{X}_2 + \ldots + w_{1p}\underline{X}_p \\ PC_2 &= w_{21}\underline{X}_1 + w_{22}\underline{X}_2 + \ldots + w_{2p}\underline{X}_p \\ &\vdots \\ PC_k &= w_{k1}\underline{X}_1 + w_{k2}\underline{X}_2 + \ldots + w_{kp}\underline{X}_p. \end{aligned}$$

   - Compute the sum of the absolute values of the two PC loadings for each feature and order features accordingly.
   - That is for $\underline{X}_i$, it is $\sum_{j=1}^{k} |w_{ji}|$, where $i = 1, \ldots, p$.

What is the best feature ordering technique? II

2. **Univariate feature selection (ANOVA F value classification)**
   - Conduct a F test and order feature according to the set of F values (p values).
3. **Absolute correlation of features with the response variable $|r|$**
   - We consider the point biserial correlation to measure the relationship. between a binary variable, Y, and a continuous variable, X
   - This coefficient also varies between -1 and +1 where 0 implies no correlation.
4. **Classification model based feature importance**
   1. Feature importance from model coefficients (Logit, SVM-Linear) [9].
   2. Feature importance from decision trees (Decision trees, Random Forest, Gradient boosting algorithms) [8].

Simulation Study

- We repeatedly generated 100 data sets for each scenario to meet different practical situations by changing,
  - Sample size
  - Number of informative features
  - Class imbalanced rate
- Calculated the percentage of selecting informative features using,

$$\text{percentage of informative selected} = \frac{\text{average number of informative selected within the expected range}}{\text{number of informative in the sample}}$$

- The expected range is the total number of informative given in the data set.
- PCL method picks most informative features within the range of given informative features.

Simulation Results

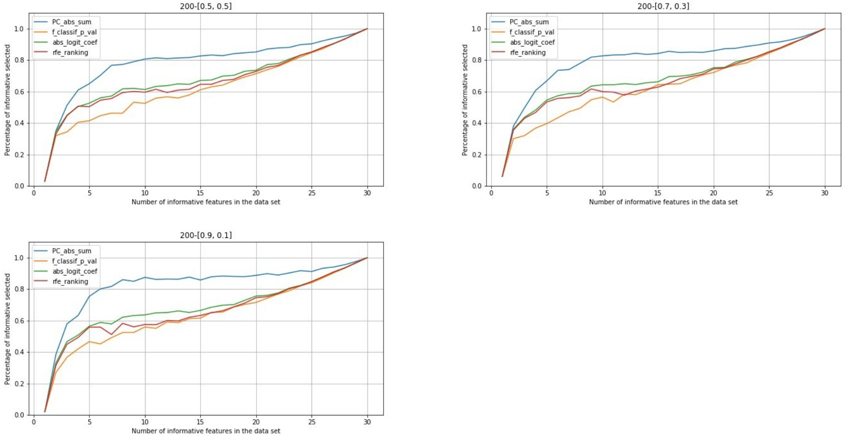


Figure: Mean percentages of informative features selected by each ordering technique in different class imbalanced levels with 200 sample size

Simulation Results Cont.

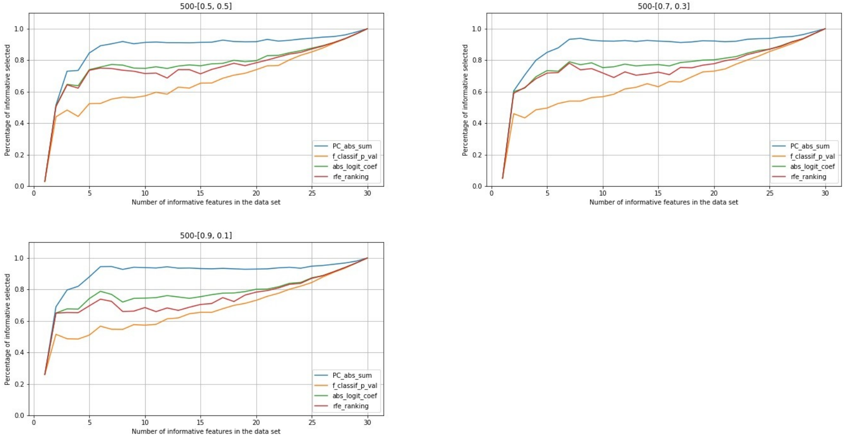


Figure: Mean percentages of informative features selected by each ordering technique in different class imbalanced levels with 500 sample size

Simulation Results Cont.

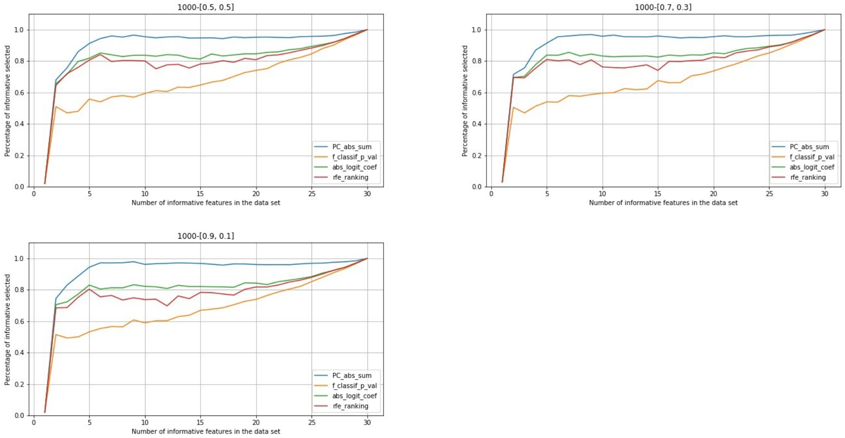


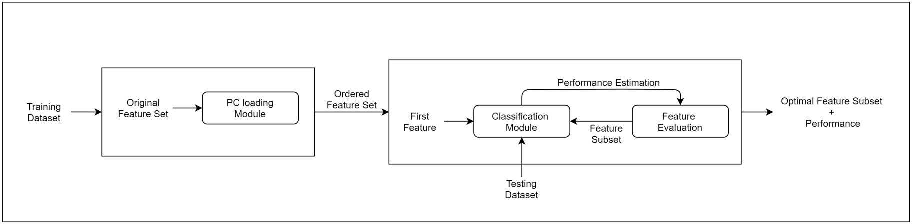
Figure: Mean percentages of informative features selected by each ordering technique in different class imbalanced levels with 1000 sample size

Which method extracts the best informative feature subset?

- Next challenge is to obtain the most informative feature subset
- **Suggested method**,
  1. Run PCA for the training test.
  2. Identify the loadings and order to the summation of absolute loadings.
  3. Start from the first feature in the ordered list and get the score value (F1-score) by comparing values with the test set.
  4. Repeat step 3 by adding one feature at a time from the ordered list.
  5. Obtain the subset which gives the maximum F1-score.

Principal Component Loading Feature Selection (PCLFS)

PCLFS



Combination

The most informative feature subset with minimal number of features and similar performance

Simulated Data

- Synthetic simulations, computations and related experiments were done using python.
- WestGrid facility was used due to the computer intensity.
- In simulation, each class is formed of several Gaussian clusters, each located around the vertices of a hypercube in a subspace of dimension number of informative.
- Informative features are drawn independently from Normal(0, 1) distribution for each cluster and then randomly linearly combined within each cluster to add covariance.
- Remaining non informative features are filled with random noise.
- Simulation was done for original data and for SMOTE [2] data applying PCLFS, PCLFS-extended and RFE methods.

Simulation Study

1. One hundred samples are simulated from each scenario.
2. Number of informative features is increased from 1 to the total number of features (30).
3. The results were obtained for different synthetic data sets with a sample size of 1000.
4. The relationship of $n\_features = n\_informative + n\_non\_informative$ is maintained.
5. We generated data for 50%:50% balanced and two other imbalance rates, 70%:30% and 90%:10%.
6. Illustrated the results of the logistic regression model.
7. The maximum tolerable F1-score reduction was taken as 0.05 for all samples.

Simulation Results

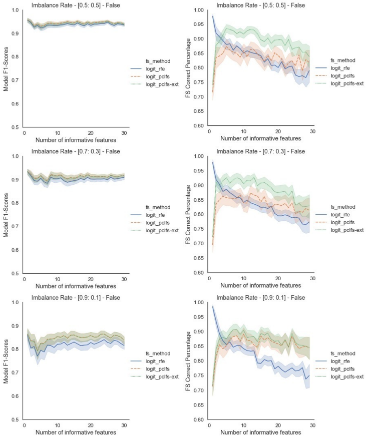


Figure: Final model F1-scores and feature selection correct percentages for the Logit model, without SMOTE when sample size is 1000 and threshold is 0.0017.

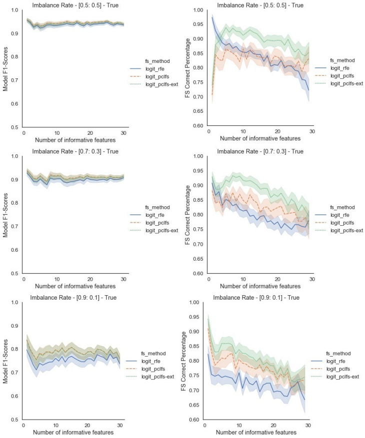


Figure: Final model F1-scores and feature selection correct percentages for the Logit model, with SMOTE when sample size is 1000 and threshold is 0.0017.

SPECTF heart data

1. Consider the publicly available Single-photon emission computed tomography (SPECT) heart data set. [5, 7]
2. It describes diagnosing cardiac abnormalities using SPECT.
3. The data set has classified each of the patients into two categories: normal and abnormal, by considering the diagnosis of images.
4. The data set has 267 SPECT image sets (patients) with 44 continuous feature patterns for each patient.
5. Data set was divided into 75% training samples and 25% test samples.
6. The class-imbalanced rate for the data set is 80%:20%, where the minority class represents the abnormal patients.

Application Results Comparison

Table: Final F1-score comparison between RFE and proposed methods (PCLFS/PCLFS-Extended (t=0.00455)).

SMOTE	Method	Basic		**RFE**		PCLFS		**PCLFS-Extended**		Feature reduction%/ (increment%)	F1-score (reduction)/ increment
		#Features	F1-scores	#Features	F1-scores	#Features	F1-scores	#Features	F1-scores		
TRUE	Logit	44	0.6809	36	0.6957	24	0.6957	11	0.6939	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	3	0.6666	93.2%	0.1110
	RFC	44	0.6486	38	0.6111	42	0.7059	12	0.6842	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

- Existing methods identify the feature subset which gives the best scoring values.
- Some other feature subsets practically reduce the number of features with a minimal loss of scoring value.
- First proposed method receives the most beneficial smallest number of features and the feature subset with a tolerable scoring value deduction.
- The threshold plays a vital role in the introduced algorithm.
- Using the summation of the absolute values of principle component loadings, features can be ordered from most informative to the least.
- We should consider the underlying assumptions of the Principal Component Analysis when using the method.

Discussion Cont.

- Feature ordering features are entirely independent of the classification model.
- Combined both methods to achieve objectives of feature selection.
- Final results returns “The most informative feature subset with minimal number of features with similar performance".
- Simulated and application results showed that the proposed method makes a reasonable improvement over RFE results.
- Proposed method is an important contribution, especially if we have to collect data from costly sources.
- Two manuscripts are submitted based on,
  - “Selecting Minimal Number of Features with Similar Performance".
  - “Assessing Feature Selection Method Performance with Class Imbalance Data"

References

[1] Cervante, L., B. Xue, L. Shang, and M. Zhang (2013). A multi-objective feature selection approach based on binary pso and rough set theory. In M. Middendorf and C. Blum (Eds.), *Evolutionary Computation in Combinatorial Optimization*, Berlin, Heidelberg, pp. 25–36. Springer Berlin Heidelberg.

[2] Chawla, N. V., K. Bowyer, L. Hall, and W. P. Kegelmeyer (2002). Smote: Synthetic minority over-sampling technique. *ArXiv abs/1106.1813*.

[3] 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.

[4] Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning 46*(1), 389–422.

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

[6] Kuhn, M. *Applied predictive modeling* (1st ed. 2013. ed.). New York, New York: Springer.

[7] 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.

[8] 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–.

[9] Tsuruoka, Y., J. Tsujii, and S. Ananiadou (2009). Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, USA, pp. 477–485. Association for Computational Linguistics.

Acknowledgment

I would like to express my special thanks of gratitude to,

- To my supervisors Dr. Saman Muthukumarana & Dr. Mike Domaratzki for their excellent guidance
- To the department of Statistics and the staff for funding and resources
- To my family and friends for the continuous support

Thank You!