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

MSc Thesis Defense

Presented by: Surani Matharaarachchi

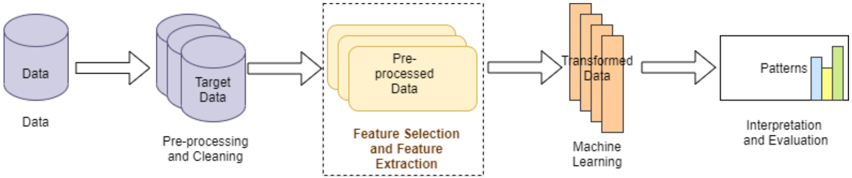
July, 05 2021

Faculty of Science
Department of Statistics

Outline

Feature Selection

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



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

Motivation

Two main objectives of feature selection:

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

- while achieving higher accuracy

[Cervante et al., 2013, Kuhn, Kuhn]

The Class Imbalance Issue

- A machine learning classification issue.
- Number of instances in the minority class is far less than the total number of instances in the majority class.
- Standard classifiers tend to be overwhelmed by the majority classes.
- The minority class is more relevant and more important.
- Affected applications: anomaly detection, fraud detection, medical diagnosis/monitoring, churn prediction.
- Solution applied: Synthetic Minority Over-Sampling TEchnique (SMOTE) [Chawla et al., 2002].

Synthetic Minority Over-Sampling TEchnique (SMOTE)

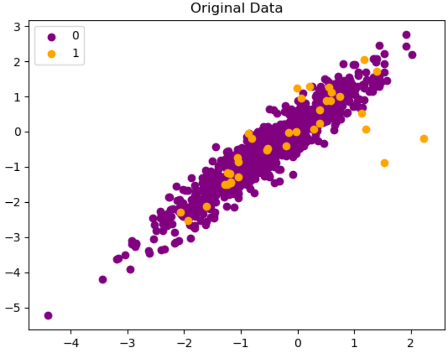


Figure: Scatter plot for originally imbalanced data.

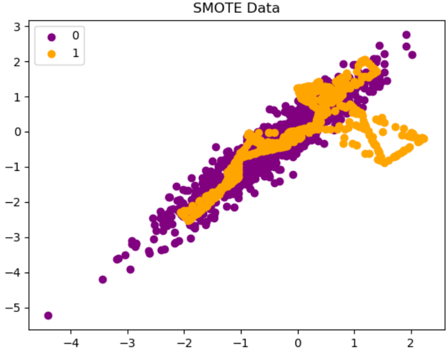


Figure: Scatter plot for SMOTE balanced data.

Synthetic Data Generation

- Synthetic simulations and computations were done using python.
- 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.
- Total No. of features = No. of informative features + No. of non-informative features.

Performance Evaluation Matrices

Table: Model confusion matrix

		Predictions	
		Class 1	Class 0
Actual	Class 1	TP_{model}	FN_{model}
	Class 0	FP_{model}	TN_{model}

Table: Feature selection confusion matrix

	selected	not selected
informative	TP_{fs}	FN_{fs}
non-informative	FP_{fs}	TN_{fs}

$$\text{Precision} = \frac{TP_{model}}{TP_{model} + FP_{model}}$$

$$\text{Recall} = \frac{TP_{model}}{TP_{model} + FN_{model}}$$

$$\text{F1-score} = 2\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

$$TPR_{fs} = \frac{TP_{fs}}{N_{informative}} = Recall_{fs}$$

$$FPR_{fs} = \frac{FP_{fs}}{N_{informative}}$$

$$TNR_{fs} = \frac{TN_{fs}}{N_{non\_informative}}$$

$$FNR_{fs} = \frac{FN_{fs}}{N_{non\_informative}}$$

$$Correct\%_{fs} = \frac{TPR_{fs} + TNR_{fs}}{2}$$

Part I

Selecting Features with Similar Performance

Motivation

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

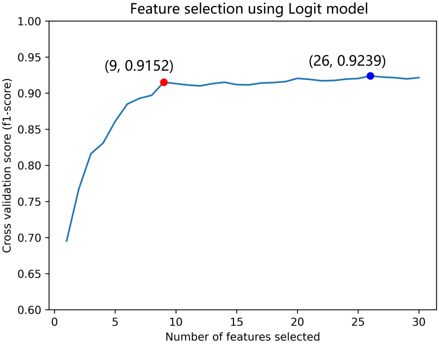


Figure: The blue point indicates the RFE feature selection whereas the red point explains the same for the proposed method.

Algorithm

inputs:

- Total number of features: n
- Number of selected features by RFE: n_{rfe}
- Grid scores: $\mathbf{g} = [g_1, g_2, \ldots, g_m]$
  - g_i corresponds to the average CV score of the i^{th} feature subset with i remaining features.
  - m is the total number of feature subsets.
- Feature importance scores (obtained from the classifier): $\mathbf{i} = [i_1, i_2, \ldots, i_{n_{rfe}}]$
- Maximum tolerable F1-score reduction: T (User-defined)

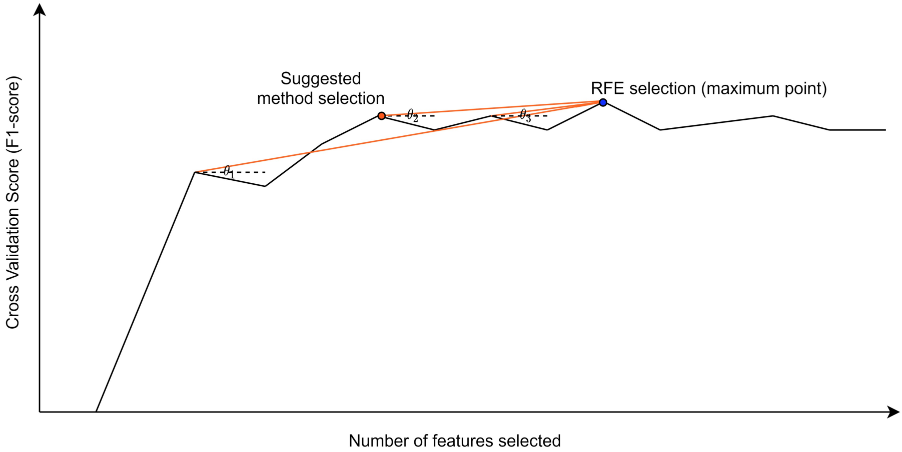


Figure: Graphical view of the suggested algorithm. θ_i is the angle between the horizontal dotted line (a line parallel to the number of features selected axis) and the red line, which combines the i^{th} point with the maximum point.

Algorithm Conti.

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.

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

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

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**).

outputs:

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

Role of threshold (t)

- Simulation trials to determine the factors affect the behavior of the threshold.
- A numerical cut-off value as the threshold reduces the same amount of F1-score regardless of the number of features removed.
- Having many features reduces the F1-score significantly unless the cut-off is extremely small.
- Considered a tolerable F1-score decrease per feature as the threshold.

Part II

A Unified Approach for Feature Selection

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

Four different feature ordering methods to compare the feature ordering behavior.

1. **Summation of the absolute values of PC loadings (PCL)**
   - The PC loadings [Dunteman, 1989] 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**
   - Consider the point biserial correlation.
   - 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) [Tsuruoka et al., 2009].
   2. Feature importance from decision trees (Decision trees, Random Forest, Gradient boosting algorithms) [Ryzin, 1986].

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.

Simulation Results

- Blue line - the sum of the absolute values of PCL.
- Red dashed line - the Logit model-based feature importance.
- Overlapped green and orange dashed lines - ANOVA F value classification and the absolute correlation.
- PCL method picks most informative features.

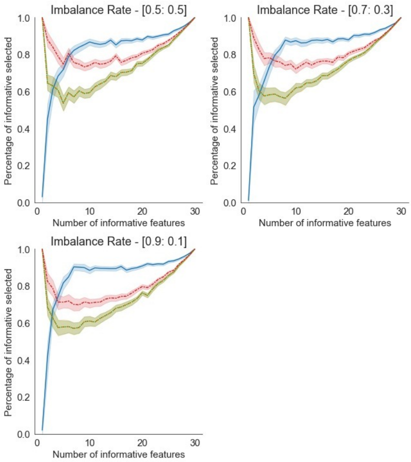


Figure: With 200 sample size

Which method extracts the best informative feature subset?

- Next challenge is to obtain the most informative feature subset.
- **Suggested method**,

  Step 1: Run PCA for the training dataset.

  Step 2: Identify the loadings and order original features according to the summation of absolute loadings of first k PCs.

  Step 3: Start from the first feature in the ordered list fit data on classification method.

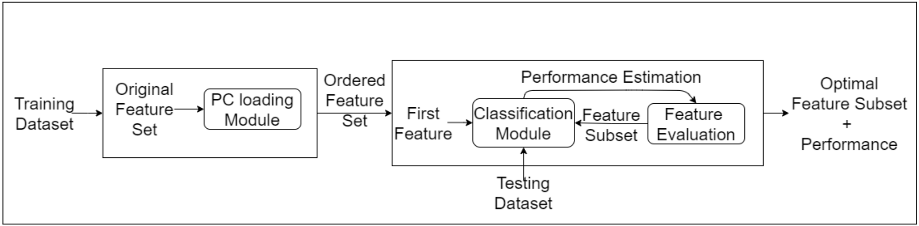
  Step 4: Get the score value (F1-score) by comparing values with the test set.

  Step 5: Repeat step 3 & 4 by adding one feature at a time from the ordered list.

  Step 6: 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

Simulation Study

1. Simulation was done for original data and for SMOTE data applying PCLFS, PCLFS-extended and RFE methods.
2. For the PCLFS-extended method, grid scores are the F1-scores such that g_i corresponds to the F1-score of the i^{th} feature subset with the first i features of the PCLFS ordered feature list.
3. The rest of the notations are the same as for the RFE.
4. The maximum tolerable F1-score reduction was taken as 0.05 for all samples.
5. Illustrated the results of the logistic regression model.

Scenarios

Variable	No. of levels	Levels
Methods	3	RFE, PCLFS, PCLFS-Ext
Classification Models	5	Logit, SVM-Linear, Decision Tress RFC, 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 to 30 increasing by 1
Sample Sizes	2	200, 1000
Performance Evaluation Matrices		F1-score$_{model}$, Correct_Percentage$_{fs}$, TPR$_{fs}$
Repeat Samples	100	

Simulation Results

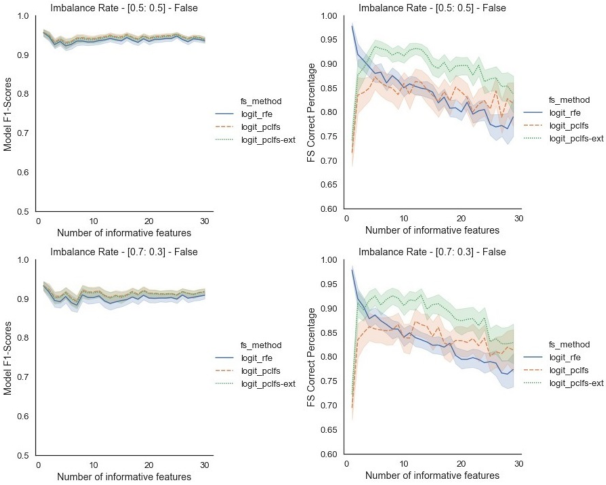


Figure: Without SMOTE

Simulation Results

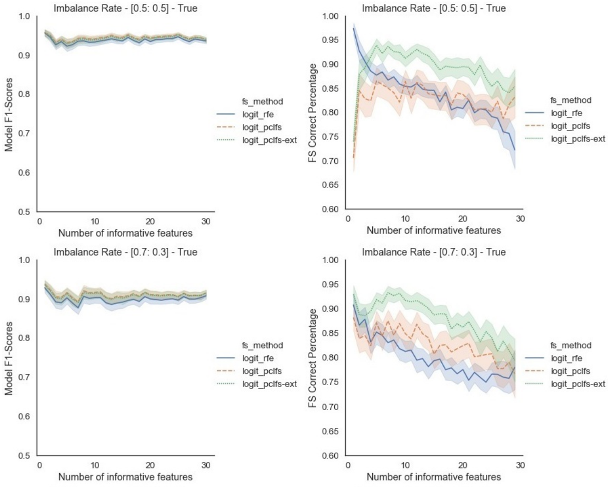


Figure: With SMOTE

SPECTF heart data

1. Publicly available Single-photon emission computed tomography (SPECT) heart data set. [Krzysztof et al., 1997, Kurgan et al., 2001]
2. Diagnosing cardiac abnormalities using SPECT.
3. Each of the patients into two categories: normal and abnormal.
4. 267 SPECT image sets with 44 continuous feature patterns for each patient.
5. Data set is divided into 75% training samples and 25% test samples.
6. Class-imbalanced rate is 80%:20%, 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

- First proposed method receives the most important smallest number of features and the feature subset.
- 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.
- 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".

Discussion Cont.

- Proposed methods makes a reasonable improvement over RFE results.
- Proposed methods are important contributions.
- Python and WestGrid facility was used.
- Manuscript is submitted based on, “Assessing Feature Selection Method Performance with Class Imbalance Data."

Future Work

- Use of loss function concept as the threshold.
- Optimize the number of principal components k.
- Apply proposed method on non-linear classification models.
- Examine the impact of adding redundant and repeated features in simulation.

References

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.

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

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.

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

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

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.

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–.

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 my supervisors Dr. Saman Muthukumarana & Dr. Mike Domaratzki for their excellent guidance.
- To Dr. Max Turgeon and Dr. Carson Leung for being in the advisory committee and for their time and advice.
- To the department of Statistics, the faculty of graduate studies for funding and resources.
- To my family and friends for the continuous support.

Thank You!