Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models

Debopam Sanyal, Nigel Bosch, and Luc Paquette

What are feature selection metrics?

- When deciding what features to include in a model, we must decide what a "good" feature is using some metric
- What is a reasonable choice of metric?

Wrapper forward feature selection

1) Build every possible model with only 1 feature

2) Keep the best feature, as measured by model accuracy according to some **selection metric**

3) Build every possible model with the best feature plus one other feature

4) Keep the best additional feature

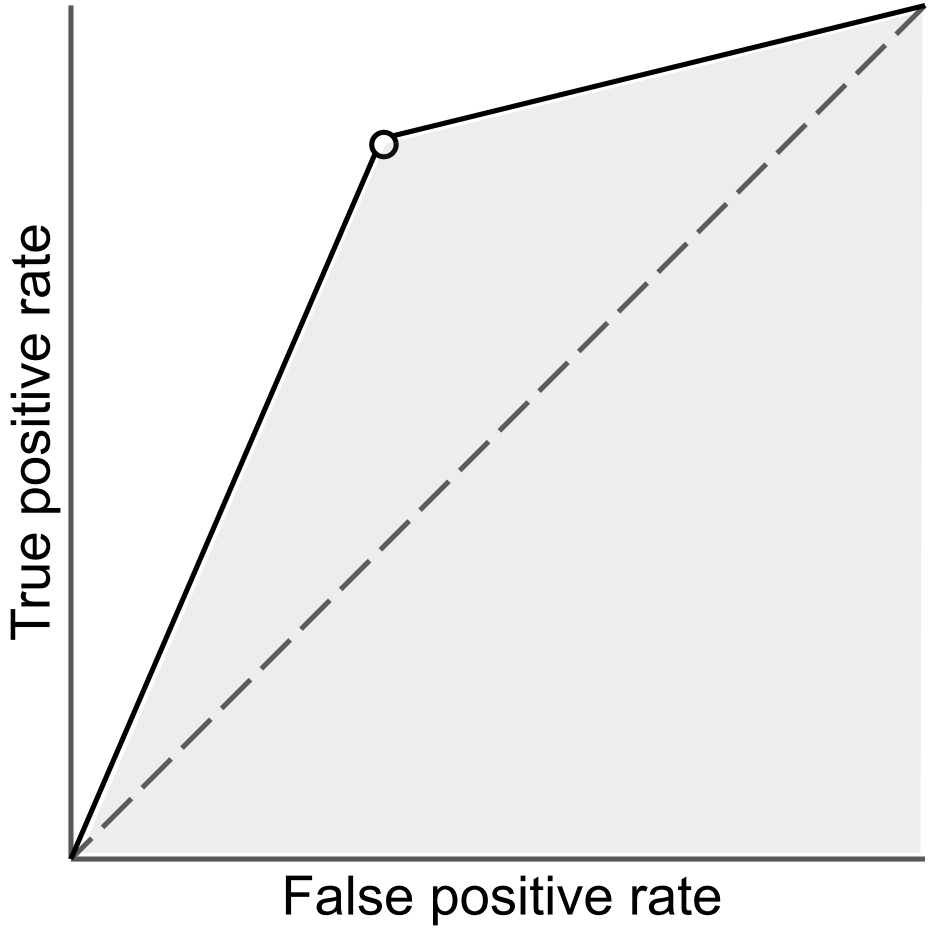
5) Repeat until model accuracy stops improving

Metrics we considered

Common in EDM	Less common
▪ AUC	▪ MPAUC
▪ Recall	▪ MCC
▪ Precision	
▪ F_1	
▪ RMSE	
▪ Cohen's kappa	

Minimum proper AUC (MPAUC)

AUC approximation calculated from a confusion matrix



Matthews correlation coefficient (MCC)

- Equivalent to Pearson's *r* with two binary variables
- Robust to class imbalances

Research question 1

- When selecting features based on a specific metric, how do the results vary in terms of the other metrics?

Classifiers

- Random forest
- SVM
- Naïve Bayes
- Logistic regression

Results were relatively similar, so we averaged over classifiers

Datasets (11)

- Gaming the system (2)
- Video-based affect detection (6)
- Final grade prediction (3)

Number of features ranged from 7 up to 2,304

Model training process

- 4-fold student-level cross-validation
- Nested 4-fold cross-validation for hyperparameter selection (including feature selection)
- 11 months of computation
  - There were many hyperparameters
  - Wrapper feature selection is slow!

Ranking metrics in terms of other metrics

- Train a model by selecting features based on metric X
- Measure the model performance with all 9 metrics

- Train a model by selecting features based on metric Y
- For each metric, how does this new model compare to the first one?
  - Repeat for each metric, and rank each **selection metric** in terms of the **result metrics**

Dataset	Accuracy	AUC	F1	Kappa	MCC	MPAUC	Precision	RMSE	Recall
CTA-C	6.319	2.653	2.153	2.681	3.167	4.181	4.778	3.528	6.542
CTA-PF	5.403	2.222	4.903	4.625	3.986	2.208	6.069	3.736	2.847
VIDEO-ANU-C	2.958	4.069	3.583	3.403	4.194	3.806	6.556	5.250	2.181
VIDEO-HR-C	3.847	5.111	4.514	3.764	3.556	4.181	5.153	2.333	3.542
VIDEO-LBP-C	3.806	3.389	3.986	3.528	3.319	3.986	5.875	4.097	4.014
VIDEO-ANU-R	4.931	3.306	3.431	5.139	2.931	3.264	4.764	3.542	4.694
VIDEO-HR-R	2.000	4.389	4.111	4.333	3.583	4.722	5.319	3.306	4.236
VIDEO-LBP-R	3.833	2.361	6.458	2.528	4.056	3.069	4.597	3.306	5.792
EPM	3.222	5.319	3.208	2.458	3.125	2.694	6.056	3.556	6.361
MATH	5.556	3.569	1.583	3.333	2.222	2.653	6.472	4.056	6.556
PORTUGUESE	4.819	1.764	3.028	3.792	3.708	3.472	6.750	2.125	6.542
Mean	4.245	3.468	3.723	3.598	3.441	3.476	5.672	3.530	4.846
Std. dev.	1.282	1.172	1.327	0.862	0.573	0.776	0.781	0.843	1.605

RQ1: Ranking results

Mean ranking interpretation:

On average, there were X out of 9 result metrics for which some other selection metric was better

- **Best**
  - MCC (*M* = 3.441; best in 2/11 datasets)
  - AUC (*M* = 3.468; best in 2/11 datasets)
- **Worst**
  - Precision (*M* = 5.672; worst in 6/11 datasets)
  - Recall (*M* = 4.846; worst in 3/11 datasets)

Research question 2

- How do different feature selection metrics impact model calibration?

CAL score

- Calibration: how well a model's predictions correspond to the probability that it is right or wrong *(confidence – correctness)*
- Out of all the predictions where a model says "70% probability of the positive class", it **should** be wrong 30% of the time
- CAL score: bin predictions, measure difference between actual proportion correct vs. expected in each bin

RQ2: Calibration results

- **Best (lower is better; 0 is best)**
  - RMSE (*M* = .166; best on 8/11 datasets)
  - Kappa (*M* = .189; best on 1/11 datasets)
- **Worst**
  - Recall (*M* = .243; worst on 5/11 datasets)
  - Precision (*M* = .228; worst on 6/11 datasets)

Research question 3

- How do different feature selection metrics impact the predicted rates of models?

RQ3: Predicted rates

- **Best (closest to true base rate)**
  - Accuracy (mean diff = .079; best on 5/11 datasets)
  - RMSE (mean diff = .080; best on 5/11 datasets)
- **Worst**
  - Recall (mean diff = .233; worst on 5/11 datasets)
  - Precision (mean diff = .210; worst on 5/11 datasets)

Research question 4

- Do some feature selection metrics tend to result in more parsimonious models (fewer features) than others?

RQ4: Number of features

- **Most features**
  - RMSE (*M* = 10.523; highest on 8/11 datasets)
  - AUC (*M* = 10.006; highest on 3/11 datasets)

- **Fewest features**
  - Precision (*M* = 4.173; lowest on 4/11 datasets)
  - Recall (*M* = 5.173; lowest on 5/11 datasets)

Note: RMSE and AUC are the only two metrics calculated not from a confusion matrix, but from the probability predictions

Takeaways

- MCC produced the best average results in terms of all metrics (by a small margin)
- RMSE yielded the best-calibrated models
- Precision and recall were not good choices
- F_1 worked well even though it can be easily inflated by over-predicting the positive class (like accuracy/recall)

Thanks!