A comparison of statistical methods for improving rare event classification in medicine

Thesis Defense

Alyssa Forber
April 18th 2018

Colorado School of Public Health
University of Colorado, Anschutz Medical Campus
Department of Biostatistics & Informatics

Outline

Imbalanced Learning Problem

- When one outcome class greatly outnumbers the other, the outcome is said to be imbalanced
- Developing a classification model for imbalanced data results in over learning the majority class and under learning the minority class, and subsequently in low sensitivity and high specificity
- Need to improve predictive performance
- Many examples in medicine: rare diseases, rare adverse reactions to medications, etc.

Aims

- Improve predictive performance for imbalanced dataset
- Utilizing measures of sensitivity, specificity, accuracy, and ROC analysis to evaluate performance
- Compare two methods to handle imbalance
    - Informed probability cutpoints for predicted probabilities
    - Sampling techniques to balance datasets
- Evaluate these methods in two clinical data sets and in a simulation study

Methods

- Split data into training and test sets using a 2/3:1/3 temporal split
- Create balanced datasets through sampling on training set
- Fit model to sampled data
- Get predicted probabilities on the hold-out test data
- Optimize probability cutoff for outcome
    - Note: we are not doing cost-sensitive learning

Model

Cross-validated lasso regression - Least Absolute Shrinkage and Selection Operator

- Lasso:
  - Shrinks estimates by penalizing size of coefficients
  - Performs variable selection by shrinking some to 0

$$\hat{\beta}_{lasso} = argmin \sum_{j=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$ where t is the tuning parameter.

Cross Validation

- Find the best "tuning parameter" for determining amount of shrinkage
- Split data into k parts and then train on all except one, which is held out for testing
- Then pick the tuning parameter that minimizes error

Figure 1: Source: Zhang Y, Wang S. (2015) Detection of Alzheimer's disease by displacement field and machine learning.

Advantages and Disadvantages of Lasso

- Advantages:
  - Lower variance of the predicted values
  - More accurate predictions
  - Reduces the number of predictors
- Disadvantages:
  - Biased coefficients, inference not same as logistic regression
  - No standard errors or p-values out of the model
  - When two or more independent variables are highly correlated, lasso arbitrarily chooses only one to keep

ROC & Cutoff Probabilities

ROC (with pROC package):

- Receiver Operating Characteristics
- ROC curve plots sensitivity vs 1- specificity
- Each point on curve corresponds to a decision cutoff for classification
- Youden's J statistic is used a measure to choose an optimal cutoff
  - J = sensitivity + specificity - 1

ROC Curve

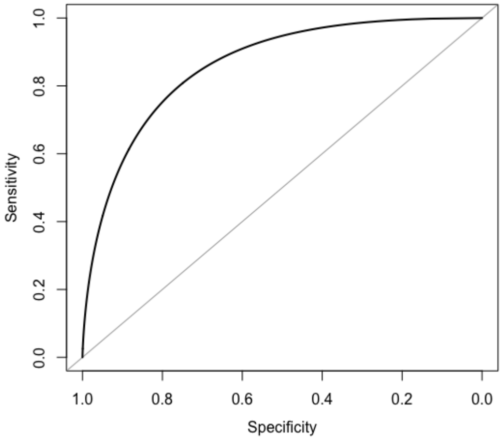


Figure 2: ROC Curve

Confusion Matrix

Correctly identify those w/ outcome:

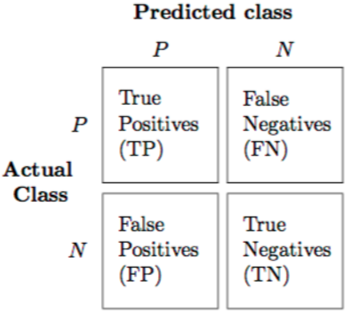
$$Sensitivity = \frac{TP}{TP + FN}$$

Correctly identify those w/o outcome:

$$Specificity = \frac{TN}{TN + FP}$$

Correctly identify either group:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$



Picture Source: "Confusion Matrix." rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/.

First Approach

No Sampling, Optimize Cut-off:

- Use original unsampled data and get predictions from the lasso model
  - Predictions return probability between 0 and 1 for each observation
- Use 0.5 standard probability cutoff to compare
- Find "best" probability cutoff
  - Youden's Index

Second Approach

Sampling:

- Create sampled data sets that are balanced
  - Down sample
  - Up sample
  - SMOTE
- Predict and use Youden's Index as cutoff

Down Sampling

- Under-sample majority to equal minority

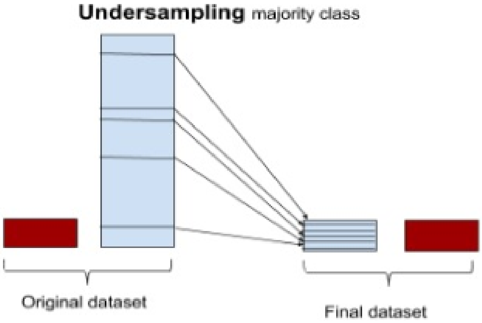


Figure 3: Down sample

Picture Source: "Learning from Imbalanced Classes." Silicon Valley Data Science, 25 Sept. 2017, svds.com/learning-imbalanced-classes/.

Up Sampling

- Over-sample minority to equal majority

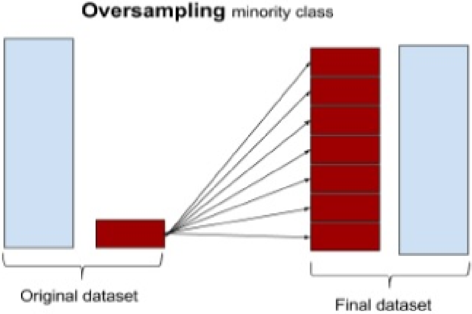


Figure 4: Up sample

Picture Source: "Learning from Imbalanced Classes." Silicon Valley Data Science, 25 Sept. 2017, svds.com/learning-imbalanced-classes/.

SMOTE

- Synthetic Minority Over-sampling Technique

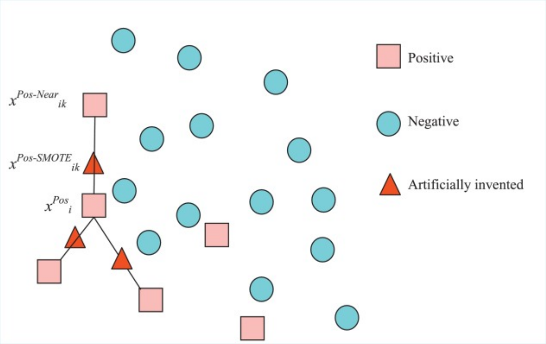


Figure 5: SMOTE

Picture Source: Sun, Jie, et al. "Imbalanced Enterprise Credit Evaluation with DTE-SBD: Decision Tree Ensemble Based on SMOTE and Bagging with Differentiated Sampling Rates."

Models

- Model 1: No Sampling
  - 0.5 cutoff
  - Youden's Index
- Model 2: Down Sampling
  - Youden's Index
- Model 3: Up Sampling
  - Youden's Index
- Model 4: SMOTE
  - Youden's Index

Case Studies

We used these methods on two case studies to see how results may differ between two real world examples with rare outcomes.

- Case Study 1: Predicting chronic opioid therapy in hospitalized patients (5%)
- Case Study 2: Predicting surgical site infections in hospitalized patients (3.4%)

Case Study 1

- Data from Denver Health electronic health records (EHR) 2008 to 2014
- Definition of Chronic Opioid Therapy (COT) one year following the index hospital discharge:

  Receipt of $\geq$ 90-day supply of opioids with $<$ 30-day gap in supply over a 180-day period or receipt of $\geq$ 10 opioid prescriptions over one year.

- 27,705 patients where 5% developed COT within a year
- 35 explanatory variables
  - Ex: age, race, history of chronic pain, discharge diagnosis

Case Study 1:

Describe some of the 35 variables, especially the important ones Do I put a mini table 1 or just some descriptions? And am I doing this for case study 2 as well or is this mostly for Susan?

Some top variables:

Gender, race, insurance status, three year history of chronic pain/acute pain/depression/anxiety, discharge diagnosis of chronic pain, Past Year Receipt of Non-Opioid Analgesics (NSAIDs, neuropathic agents, topical capsaicin & lidocaine), Past Year Number of Opioid Prescriptions Filled, Receipt of Opioid at Discharge, Milligrams of Morphine Per Hospital Day

Case Study 1: Results

Table 1: Results for Chronic Opioid Therapy

Model	Threshold	Sensitivity	Specificity	NPV	PPV	Accuracy	AUC	Coefficients
Unsampled 0.5	0.5	8	99	96	35	96	86	31
Unsampled	0.043	85	73	99	12	73	86	31
Down sampled	0.401	85	73	99	12	74	86	34
Up sampled	0.399	85	74	99	12	74	87	34
SMOTE	0.472	74	84	99	17	84	86	33

ROC Plot

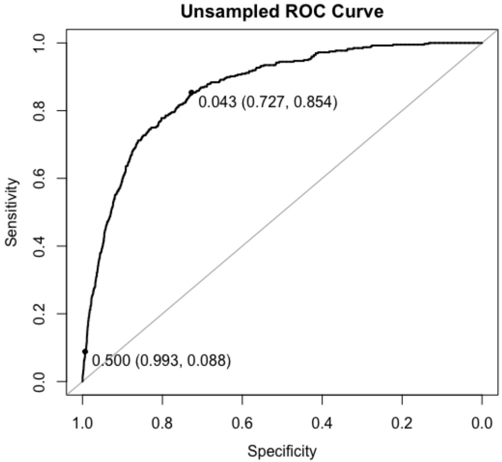


Figure 6: ROC for Original Data: Youden's Index and 0.5 cutoffs

Case Study 2

- Need to identify post-operative complications to supplement manual chart reviews by nurses
- Surgical site infections (SSI) are the most common postoperative complication
- $>$50% of SSIs occur after patient discharge
- Data from 6,840 patients at the University of Colorado Hospital from 2013-2016
- 136 independent variables, mostly binary indicators
    - Ex: Antibiotic prescriptions, procedure codes, ICD-9 codes

Case Study 2: Results

Table 2: Results for Surgical Site Infections

Model	Threshold	Sensitivity	Specificity	NPV	PPV	Accuracy	AUC	Coefficients
Unsampled 0.5	0.50	23	99.9	97	88	97	89	35
Unsampled	0.04	80	90	99	24	90	89	35
Down sampled	0.48	82	87	99	20	87	89	20
Up sampled	0.45	79	91	99	24	90	89	123
SMOTE	0.15	89	79	99	14	80	88	88

Simulation Study

- We conducted a simulation study to look at how the methods performed at a greater range of prevalences
- We chose 3%, 5%, 10%, 20%, 40%, and 50% prevalences
- Goals:
  - Evaluate performance of methods described previously
  - Determine at what point it's no longer necessary to worry about imbalance

Simulation Methods

- Using the data from case study 1, we selected 10 of the strongest predictors to fit a logistic regression model
- Generate simulated COT outcome using coefficients from that logistic regression
- The new outcome was simulated with a logistic distribution

$$F(x) = \frac{e^x}{1 + e^x}$$

- Controlled prevalence by adjusting the intercept
- Included an additional 20 predictors (30 total) to implement sampling and lasso regression

Simulation Results Part 1

Table 3: Results for 3, 5 and 10%

	Threshold	Sensitivity	Specificity	Accuracy	AUC	Coefficients
3%						
Unsampled 0.5	0.50	2	100	97	79	6
Unsampled	0.03	70	75	74	79	6
Down Sampled	0.49	70	74	74	78	9
Up Sampled	0.48	70	74	74	78	27
SMOTE	0.41	70	74	74	78	15
5%						
Unsampled 0.5	0.50	4	100	95	78	7
Unsampled	0.05	69	74	74	78	7
Down Sampled	0.49	69	74	74	78	9
Up Sampled	0.49	69	74	74	78	23
SMOTE	0.41	69	74	74	78	16
10%						
Unsampled 0.5	0.50	8	100	91	77	8
Unsampled	0.10	68	74	73	77	8
Down Sampled	0.49	68	74	73	77	9
Up Sampled	0.49	68	74	73	77	18
SMOTE	0.42	68	73	73	77	19

Simulations Results Part 2

Table 4: Results for 20, 40, and 50%

	Threshold	Sensitivity	Specificity	Accuracy	AUC	Coefficients
20%						
Unsampled 0.5	0.50	21	97	82	76	8
Unsampled	0.20	66	73	72	76	8
Down Sampled	0.49	66	73	72	76	9
Up Sampled	0.49	66	73	72	76	13
SMOTE	0.42	66	72	71	75	23
40%						
Unsampled 0.5	0.50	49	84	70	74	9
Unsampled	0.39	66	71	69	74	9
Down Sampled	0.49	66	71	69	74	9
Up Sampled	0.49	65	71	69	74	10
SMOTE	0.41	64	71	68	73	28
50%						
Unsampled 0.5	0.50	63	72	68	73	9
Unsampled	0.49	63	72	68	73	9

Sensitivity and Specificty by Prevalence

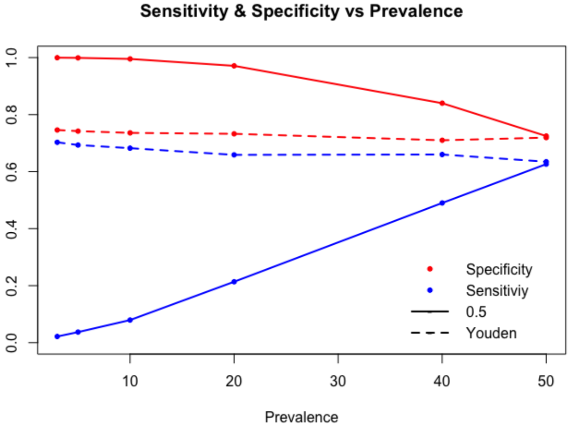


Figure 7: Comparing Youden's Index vs 0.5 cutoff results across prevalence

Youden's Index by Prevalence

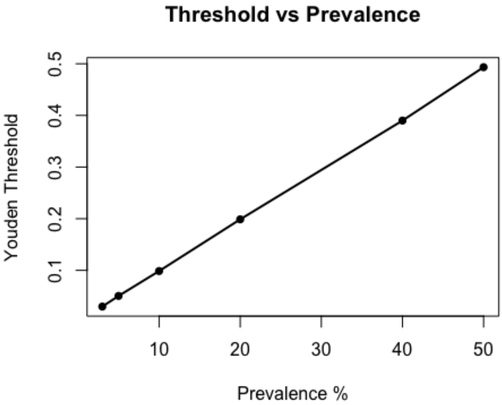


Figure 8: Youden's Index matches prevalence

Discussion

- Not much difference between Youden's + sampling vs. Youden's alone
- Over sampling and SMOTE had highest number of coefficients
- Threshold equals prevalence
- Sensitivity low even in less extreme imbalanced data for 0.5 cutoff
- Costs are unknown

Conclusion

- Always address imbalances in data, these are useful measures to evaluate performance
- When costs are known, these approaches may be used with more intentional decision making
- Each dataset may behave differently depending on the predictors
- Going Forward:
  - Choosing cutoff with training set rather than test set

Acknowledgments

Thesis Adviser Katie Colborn, thank you for all your time and mentoring

Committee members Elizabeth Juarez-Colunga and Susan Calcaterra, thank you for your time and expertise

Thank you for attending.

Case Study 1

- Design: Denver Health retrospective analysis electronic health record (EHR) data from 2008 to 2014.
- Patients: Hospitalized patients at an urban, safety-net hospital.
- Definition of Chronic Opioid Therapy (COT) one year following the index hospital discharge:

Receipt of $\geq$ 90-day supply of opioids with $<$ 30-day gap in supply over a 180-day period or receipt of $\geq$ 10 opioid prescriptions over one year.

Case Study 1: Patient Population

- 27,705 patients
- Majority had incomes <185% of the Federal Poverty Level
- 70% were ethnic minorities
- 5% with COT
- Excluded Patients:
  - <15 or >85 years old
  - Those in prison, jail, or police custody
  - Those who died within one year following their index hospitalization
  - Patients with <2 healthcare visits to Denver Health three years preceding their index hospitalization
  - Undocumented persons receiving emergent hemodialysis
  - Obstetric patients

Case Study 2

Add more details on this case study here