Class prediction for high-dimensional class-imbalanced data

Lara Lusa and Rok Blagus

Institute for Biostatistics and Medical Informatics – Faculty of Medicine
University of Ljubljana, Slovenia

Lara.Lusa at mf.uni-lj.si
Class prediction for high-dimensional class-imbalanced data

Develop a rule that can be used to predict the class membership of new samples.

Variables $p \gg n$

Class 1 samples $n_1$

Class 2 samples $n_2$

Class
How to perform high-dimensional classification

• Design a developmental study (training set) to answer a specific clinical question
  – Select $n$ subjects (patients, customers, ...)
  – Measure $p$ variables (genes, SNPs, mRNAs, characteristics of a customer)
  – High-dimensional data if $p >> n$ :
    • Select a subset of variables (variable selection)

Obtain a RULE based on the values of the variables for the classification of new samples

The RULE must be completely specified

• variables included
• how to combine them
• normalization of data
• thresholds for classification
• ...
Class prediction rules

• Many different methods are available to specify the classification rule
  – Discriminant (linear diagonal) analysis \((\text{DLDA, DQDA})\)
    • Prediction analysis of Microarrays \((\text{PAM})\)
  – Support vector machines \((\text{SVM})\)
  – Penalized regression methods \((\text{PLR-L1, PLR-L2})\)
  – Classification and regression trees \((\text{CART})\) and random forests \((\text{RF})\)
  – k-nearest neighbor methods \((\text{k-NN})\)
  – ...

• Risk of overfitting and spurious findings
• Simple methods perform well with high-dimensional data \((\text{Dudoit et al, 2002})\)
• Variable selection - reducing the number of variables usually leads to more accurate classification

• No single method is optimal in every situation
  – No Free Lunch Theorem: in absence of assumptions we should not prefer any classification algorithm over another
  – Ugly Ducking Theorem: in absence of assumptions there is no “best” set of features
Overfitting with high-dimensional data

The model will not work well on new data. Can be easily obtained with high-dimensional data.

The problem is generally reduced by variable selection and use of simple methods.
Class-imbalance: a “problem” for classification

- **Class-imbalanced data** = The number of samples in each class is not equal
- Most classifiers trained on imbalanced data
  - do not accurately predict the samples from the minority class
  - tend to classify all the samples in the majority class

*Learning from Imbalanced Data – He and Garcia, 2009, IEE Transactions on Knowledge and Data Engineering*
Example on the effect of class prevalence on the accuracy of a classifier (null case)

| Training set | Test set |
|--------------|---------|
| Bad=50       | Bad=80  |
| Good=50      | Good=20 |
| Bad=80       | Bad=80  |
| Good=20      | Good=20 |

The classifier is uninformative if \( PA_1 = 1 - PA_0 \)

\[ P(\hat{Y}=1|Y=1) = P(\hat{Y}=1|Y=0) \]
Why does it happen?

Why does it happen?

Overall Predictive Accuracy if \( p_1 = p_2 \) in test set

\[ P(Pr_1 \mid H_0) = ? \]

\( p_i \) = proportion of samples in class \( i \)

Theoretical values and expected behavior

Easy to calculate for 3-NN (using hypergeometric distribution)

Variable selection increases the bias towards the majority class

Consequences of class imbalance for other classifiers can be more difficult to understand …
Does it happen also when there are some differences between the classes and using real data?

Breast cancer classification and prognosis based on gene expression profiles from a population-based study

Christos Sotiropoulos, Soek-Ying Neo, Lisa M. McShane, Edward L. Korn, Philip M. Long, Amir Jazaeri, Philippe Martiat, Steven G. Fox, Adrian L. Harris, and Edison T. Liu

PNAS | September 2, 2003 | vol. 100 | no. 18 | 10393–10398

7650 clones
99 samples

3-NN classifier, 40 genes with largest t-statistic, missing values replaced by 0

Test set: 20 ER+ / 20 ER-

Training set: 5 ER+ / 5 ER-, 10 ER+ / 5 ER-, 20 ER+ / 5 ER-
Our initial questions

• Does the high-dimensionality of the data further exacerbate the class-imbalance problem?
• Are there any classification methods that are more robust than others?
• Are the methods commonly used to deal with the class-imbalance problem effective if the data are high-dimensional?
• Can we get some theoretical insights on the nature of the class-imbalance problem?
• ...
Simulations

Training n=80

Test n=20

Class 1

Class 2

Feature Selection

1000 genes

gene expression matrix

Rule

Genes: 40 genes with largest t-statistic

1-NN, 3-NN, 5-NN, DLDA, DQDA, RF, SVM, PAM, PLR (L2)

Normalization *samples or genes

Differentially expressed genes

20 DE genes, magnitude of difference: \(g_{1i} \sim i.d. N(1, 1), g_{2i} \sim i.d. N(0, 1)\)
Null Hypothesis

$p=40$, $n_{\text{train}}=80$, same class imbalance in training and test set

What happens if the number of variables increases and we use the 40 most different variables for classification?

Class 1 predictive accuracy in the test set

Proportion of samples from Class 1 in the training set
Sampling variability and class imbalance

Null case: distribution of the sample means

\[ \bar{X}_{maj} \sim N(0, 1/\sqrt{72}) \]

\[ \bar{X}_{min} \sim N(0, 1/\sqrt{8}) \]

\[ X_{maj} \sim N(0, 1); \ n=72 \]

\[ X_{min} \sim N(0, 1); \ n=8 \]
Null Case Results

(p=1000, G=40, n_{train} =80, same class imbalance in training and test set)

All the classifiers are non-informative (PA_1=1-PA_0)
Alternative Case Results – moderate class differences

(p=1000, G=40, delta=0.7, n_{train}=80, same class imbalance in training and test set)
| **High-dimensionality** | **Additionally biases classification towards majority class for most classifiers. Mostly due to large sampling variability of the minority class** |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| **Variable selection**  | **🙂 Improves the PA**  
|                        | **-routingly biases classification towards majority class for some classifiers (k-NN)**  
|                        | **🙂 Most variable selection methods share the same problems seen here (using t-test)** |
| **Classification methods** | **🙂 DLDA, DQDA, PLR, RF**  
|                        | **-routingly k-NN, PAM, SVM** |
| **Matching the prevalence in training and test set** | **🙂 Does not remove the problem** |
| **Variable normalization** | **routingly Further increases the bias if $p_{\text{train}} = p_{\text{test}}$**  
|                        | **🙂 $p_{\text{train}} = p_{\text{test}}$** |
Proportion of samples from **Class 1** in the training set

**Alternative hypothesis:** the classes are different, and prediction using the balanced training set is easy.
Does **multiple downsizing** work also on real data?
| Variable selection | ☺ Improves the PA  
|⏰ Additionally biases classification towards majority class for some classifiers  
|☺ Most variable selection methods share the same problems seen here (t-test) |
|---|---|
| Classification methods | ☺ DLDA, DQDA, PLR, RF  
|⏰ k-NN, PAM, SVM |
| Matching the prevalence in training and test set | ☺ Does not remove the problem |
| Variable normalization | ☕ Further increases the bias if $p_{\text{train}} \neq p_{\text{test}}$  
|☺ $p_{\text{train}} \neq p_{\text{test}}$ |
| Solutions | ☎/☺ Built-in solutions (SVM, RF, PAM)  
|☺ Downsizing  
|⏰ Oversampling  
|☺ Multiple downsizing |
SMOTE
Synthetic Minority Over-sampling Technique

$$x^{SMOTE} = x + u \times (x^{NN} - x)$$
Does SMOTE work? (null case, $p=1000$)
Does SMOTE work if we perform variable selection? (null case, p=1000, 40 selected)
Does SMOTE work if there is a difference between classes?
(alternative case, p=1000, 40 selected)
Could we expect it?

Theoretical results

- SMOTE does not change the expected value of the minority class
  \[ E(X_{SMOTE}) = E(X) \]

- The variance of the minority class is reduced if SMOTE is used to balance the class-distribution
  \[ \text{var}(X_{SMOTE}) = \frac{2}{3} \text{var}(X) \]

- When \( p \) is large new samples are expected to be closer (in terms of Euclidean distance) to SMOTE samples than to original samples
  \[ E(d(X^{test}, X^{original})) > E(d(X^{test}, X^{SMOTE})) \]
  \[ 2p \cdot \text{var}(X) > 2p \cdot \frac{5}{6} \cdot \text{var}(X) \]

- It does not affect much the classifiers that rely on mean values (DLDA)

- It negatively affects the classifiers that use class-specific variances (DQDA). Care with variable selection methods!

- If data are high-dimensional: classifiers that base their classification rule on the Euclidean distance tend to classify most samples in the MINORITY class (k-NN without variable selection)
Other theoretical results

- SMOTE introduces correlation between some samples

\[ \rho \left( X_{j}^{SMOTE_1}, X_{j}^{SMOTE_2} \right) = \begin{cases} 
\frac{3}{4} & \text{If they “share” two original samples} \\
\frac{3}{8} & \text{If they “share” one of the original samples} \\
0 & \text{Otherwise} 
\end{cases} \]

\[ \rho \left( X_{j}^{SMOTE_1}, X_{j}^{o} \right) = \begin{cases} 
\frac{\sqrt{3}}{2\sqrt{2}} & \text{If the original sample was used to generate the SMOTE sample} \\
0 & \text{Otherwise} 
\end{cases} \]

Can we still reliably use classification methods and variable selection methods that assume independence between samples (discriminant analysis methods, PLR, two sample t-test, ...)?
Effect of SMOTE on two-sample t-test P-values

Null case, p=1000, n=10+90

Original data

After “SMOTE-augmenting” the data
Do we see the same things on real data?

SMOTE works (mostly) better than no-correction, but is outperformed by simple down-sampling (that uses a much smaller sample size)
Summary of the performance of SMOTE on low and high-dimensional data

| Classifier | low-dimensional data | high-dimensional data, without variable selection | high-dimensional data, with variable selection |
|------------|----------------------|--------------------------------------------------|-----------------------------------------------|
|            | NC       | SMOTE   | NC | CO | SMOTE | NC | CO | SMOTE |
| 1-NN       | NC       | ↑       | NC | ≈  | ≈     | NC | ≈  | ↑     |
| 5-NN       | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| DLDA       | ≈       | ≈       | ≈  | ≈  | ≈     | ≈  | ≈  | ≈     |
| DQDA       | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| RF         | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| SVM        | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| PAM        | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| PLR-L1     | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| PLR-L2     | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |
| CART       | ↑       | ↑       | ↑  | ≈  | ↑     | ↑  | ≈  | ↑     |

- **NC**: no correction
- **CO**: classification cut-off calibration
Conclusions

• High-dimensionality of data exacerbates the class imbalance problem
• Some classifiers are less sensitive than others to class imbalance
  – DLDA seems to work well also in this setting if the class-imbalance is not too extreme
• Solutions
  – Down-sizing works surprisingly well (fewer data better than imbalanced data!) but wastes a lot of data
    • Combination of down-sized classifiers works very well
  – Over-sampling is generally a bad idea, SMOTE is not improving over simple downsampling
  – Change
  – Cut-off calibration (ongoing work)
    • Worked reasonably well for RF, PLR and k-NN

• Normalization of variables exacerbated the problems
Ongoing work

• Further explorations for the penalized logistic regression methods (PLR-L1 and PLR-L2)
• Evaluation and adaptation of boosting methods for high-dimensional (and class-imbalanced) data
• cartHD: R package that includes functions to fit classification trees with binary outcomes and to correct the analysis for the class-imbalance problem
  – Multiple downsizing, boosting, cross-validation, ... with fast implementation for high-dimensional data and repeated estimation
Some references

- Japkowicz N, Stephen S: **The class imbalance problem: A systematic study.** *Intell Data Anal* 2002, 6(5):429-449.
- He H, Garcia EA: **Learning from imbalanced data.** *IEEE Trans Knowledge and Data Eng* 2009, 21(9):1263-1284.

- Our published work on the topic
  - **Class prediction for high-dimensional class-imbalanced data.** *BMC Bioinformatics* 2010, 11:253
  - **Impact of Class-Imbalance on Multi-Class High-Dimensional Class Prediction.** *Metodoloski zvezki – Advances in Methodology and Statistics* 2012, 9: 25-45.
  - **SMOTE for high-dimensional class-imbalanced data.** *BMC Bioinformatics* 2013, 14:106
  - **Improved shrunken centroid classifiers for high-dimensional class-imbalanced data.** *BMC Bioinformatics* 2013
Threshold calibration

Develop the rule

PLR: $P(\text{Class 1} \mid X)$

RF: proportion of trees that classify the sample in Class 1

k-NN: proportion of NN from Class 1

Estimated on the training set (by CV)
Alternative hypothesis: the classes are different, and prediction using the balanced training set is easy.
Other functions that can be used to estimate the threshold

Proportion of samples from Class 1 in the training set
How to evaluate the performance of a classifier

• Classification error
  – A sample is classified in a class to which it does not belong
    • \( g(X) \neq Y \)
    • **Predictive accuracy (PA)** = % of correctly classified samples
      – careful interpretation!
  – In a two-class problem, using the terminology from diagnostic tests ("+"=diseased, "-"=healthy)
    • **PA\(_+\)** = **Sensitivity** = \( P(\text{classified }+ | \text{ true } +) \)
    • **PA\(_-\)** = **Specificity** = \( P(\text{classified } - | \text{ true } -) \)
    • **Positive predictive value** = \( P( \text{ true } + | \text{ classified } + ) \)
    • **Negative predictive value** = \( P( \text{ true } - | \text{ classified } -) \)
  – It is important to report all 4 when classes are imbalanced