Clearing the Confusion: Unbalanced Class Data

Norm Matloff
University of California, Davis
(talk + discussion)

Presentation to Data Science Initiative
University of California, Davis
January 23, 2020

URL for these slides (repeated on final slide):
`http://heather.cs.ucdavis.edu/DSI.pdf`

The Setting

The Setting

In a classification setting, unequal numbers of data points in each class.

The Setting

In a classification setting, unequal numbers of data points in each class.

Example: Credit card fraud data

- 284807 card transactions
- only 492 cases of fraud (class 1)
- 284315 cases of nonfraud (class 0)

What Are They Worried About?

What Are They Worried About?

- Say fit logit model, neural nets, whatever.
- Fit will always predict class 0.
- So, never catch the fraudsters.

Norm Matloff
University of California, Davis
(talk + discussion)

“Remedy”

Norm Matloff University of California, Davis (talk + discussion)

“Remedy”

- Alarmingly common — even standard — remedy:

Norm Matloff
University of California, Davis
(talk + discussion)

"Remedy"

- Alarmingly common — even standard — remedy:

  Artificially *equalize the class sizes.*

Norm Matloff
University of California, Davis
(talk + discussion)

“Remedy”

- Alarmingly common — even standard — remedy:

  Artificially *equalize the class sizes.*

- Downsample: Throw out (precious) data.

“Remedy”

- Alarmingly common — even standard — remedy:

  Artificially *equalize the class sizes.*

- Downsample: Throw out (precious) data.
- Upsample: Create artificial new data to augment the smaller class.

“Remedy”

- Alarmingly common — even standard — remedy:

  Artificially *equalize the class sizes.*

- Downsample: Throw out (precious) data.
- Upsample: Create artificial new data to augment the smaller class.
- Resample: Do a resampling of the data, like bootstrap, but with a weighted scheme so that the new class sizes come out equal.

Who Is Worried?

Who Is Worried?

Examples of methodology/advocacy:

- Torgo, *Data Mining with R*, CRC, 2011; see also his many citations to AI literature
- Kuhn and Johnson, *Feature Engineering and Selection*; see also various short courses at useR!

Packages

Packages

- **caret**
- **DMwR2**
- **imbalance**
- **mlr3** (Machine Learning in R: Next Generation)
- **ROSE** (Random Oversampling Examples)
- etc.

Norm Matloff
University of California, Davis
(talk + discussion)

Model Fit Issues

Model Fit Issues

- Advocates of rebalance also cite poor model fit.

Model Fit Issues

- Advocates of rebalance also cite poor model fit.
- We might be “fitting to the dominant class.”

Model Fit Issues

- Advocates of rebalance also cite poor model fit.
- We might be "fitting to the dominant class."
- Actually, it is probably the opposite; rare cases will have high leverage.

Model Fit Issues

- Advocates of rebalance also cite poor model fit.
- We might be "fitting to the dominant class."
- Actually, it is probably the opposite; rare cases will have high leverage.
- But there is no inherent reason that rebalancing will fix a bad model.

Norm Matloff
University of California, Davis
(talk + discussion)

Model Fit Issues

- Advocates of rebalance also cite poor model fit.
- We might be "fitting to the dominant class."
- Actually, it is probably the opposite; rare cases will have high leverage.
- But there is no inherent reason that rebalancing will fix a bad model.
- Studies use questionable criteria for "success," e.g. AUC. Relevant to one's actual application?

How Were the Data Generated?

How Were the Data Generated?

3 cases:

A Sample from overall pop., class sizes approx. reflect pop. values.

B Sample evenly from each class, known class priors. (Not subjective Bayesian!)

C Sample even from each class, unknown priors.

What Your ML Algorithm Is Thinking

What Your ML Algorithm Is Thinking

- If you rebalance, the algorithm thinks the true pop. priors are about even.

What Your ML Algorithm Is Thinking

- If you rebalance, the algorithm thinks the true pop. priors are about even.
- Question: Do you want the alg. to think this?

What Your ML Algorithm Is Thinking

- If you rebalance, the algorithm thinks the true pop. priors are about even.
- Question: Do you want the alg. to think this? Do you have any rationale for that?

Some Indeed Have Objected

Some Indeed Have Objected

Frank Harrell, prominent biostatistician:

> *For this reason the odd practice of subsampling the controls is used in an attempt to balance the frequencies and get some variation that will lead to sensible looking classifiers (users of regression models would never exclude good data to get an answer). Then they have to, in some ill-defined way, construct the classifier to make up for biasing the sample. It is simply the case that a classifier trained to a 12 [$q = 1/2$] prevalence situation will not be applicable to a population with a 11000 [$p = 1/1000$] prevalence.*

Norm Matloff
University of California, Davis
(talk + discussion)

A Simpler, More Direct Approach

Norm Matloff
University of California, Davis
(talk + discussion)

A Simpler, More Direct Approach

- For Sampling setting A above.
- We don't need to do formal classification.

A Simpler, More Direct Approach

- For Sampling setting A above.
- We don't need to do formal classification.
- Just flag the cases of interest, i.e. those for which $P(Y = 1|X) > \text{threshold of interest}$

A Simpler, More Direct Approach

- For Sampling setting A above.
- We don't need to do formal classification.
- Just flag the cases of interest, i.e. those for which $P(Y = 1|X) > \text{threshold of interest}$
- E.g. credit card fraud. Instead of flagging those for which prob. > 0.50, may set threshold at 0.20.

Norm Matloff University of California, Davis (talk + discussion)

A Simpler, More Direct Approach

- For Sampling setting A above.
- We don't need to do formal classification.
- Just flag the cases of interest, i.e. those for which $P(Y = 1|X) > \text{threshold of interest}$
- E.g. credit card fraud. Instead of flagging those for which prob. > 0.50, may set threshold at 0.20.
- Could set up formal loss function, etc. — but no point to it.

A Simpler, More Direct Approach

- For Sampling setting A above.
- We don't need to do formal classification.
- Just flag the cases of interest, i.e. those for which $P(Y = 1|X) > \text{threshold of interest}$
- E.g. credit card fraud. Instead of flagging those for which prob. > 0.50, may set threshold at 0.20.
- Could set up formal loss function, etc. — but no point to it.
- Actually, **mlr3** docs do suggest this as an alternative to rebalancing.

Norm Matloff
University of California, Davis
(talk + discussion)

Example: glm()

Example: glm()

```
> glmout <- glm(Class ~ ., data=ccf, family=binomial)
> condprobs <- predict(glmout, ccf, type='response')
> tocheck <- which(condprobs > 0.25)
> names(tocheck) <- NULL
> head(tocheck)
[1]  542 6109 6332 6335 6337 6339
```

Example: Random Forests

Example: Random Forests

```
> ccf$Class ← as.factor(ccf$Class)
> rfout ← randomForest(Class ∼ .,data=ccf)
> predout ← predict(rfout,ccf,type='response')
> treeguesses ←
    predout$individual  # class guesses, each tree
> tgs ← as.matrix(treeguesses)
# tgs[i,] has guesses for case i,
# '1's and '0's, from each tree
> probs ← apply(tgs,1,
    function(rw) mean(as.numeric(rw)))
> tocheck ← which(probs > 0.25)
> head(tocheck)
[1]   70  542  624 1747 4921 6109
```

Other Packages

Other Packages

- Most packages will output those estimated condit. probs. as an option.

Other Packages

- Most packages will output those estimated condit. probs. as an option.
- E.g. **gbm** is similar to **glm()** case.

Other Packages

- Most packages will output those estimated condit. probs. as an option.
- E.g. **gbm** is similar to **glm()** case.
- E.g. for **neuralnet** package, call **compute()** then take the **net.result** component.

Norm Matloff
University of California, Davis
(talk + discussion)

Sampling Setting B

Norm Matloff University of California, Davis (talk + discussion)

Sampling Setting B

- Classes were set the same size by sample *design*.
- Example: UCI Letter Recognition Data.
- 26 letters, approx. equal frequency.
- Yet actual frequency is E 12.02%, T 9.10%, A 8.12% etc.

Norm Matloff University of California, Davis (talk + discussion)

An Adjustment Formula, Setting B

An Adjustment Formula, Setting B

Def. $f_i(t)$ = density of X within class i.

$$P(Y = 1|X = t) = pf_1(t)/[pf_1 t) + (1-p)f_0(t)]$$

$$P(Y = 1|X = t) = 1/[1 + (1-p)/pf_0(t)/f_1(t)]$$

- In sample setting B, $p = 0.5$ (artificially).

An Adjustment Formula, Setting B

Def. $f_i(t)$ = density of X within class i.

$$P(Y = 1|X = t) = pf_1(t)/[pf_1 t) + (1 - p)f_0(t)]$$

$$P(Y = 1|X = t) = 1/[1 + (1 - p)/pf_0(t)/f_1(t)]$$

- In sample setting B, $p = 0.5$ (artificially).
- We have the LHS from output.

An Adjustment Formula, Setting B

Def. $f_i(t)$ = density of X within class i.

$$P(Y = 1|X = t) = pf_1(t)/[pf_1 t) + (1 - p)f_0(t)]$$

$$P(Y = 1|X = t) = 1/[1 + (1 - p)/pf_0(t)/f_1(t)]$$

- In sample setting B, $p = 0.5$ (artificially).
- We have the LHS from output.
- Solve for $f_0(t)/f_1(t)$.

An Adjustment Formula, Setting B

Def. $f_i(t)$ = density of X within class i.

$$P(Y = 1|X = t) = pf_1(t)/[pf_1 t) + (1-p)f_0(t)]$$

$$P(Y = 1|X = t) = 1/[1 + (1-p)/pf_0(t)/f_1(t)]$$

- In sample setting B, $p = 0.5$ (artificially).
- We have the LHS from output.
- Solve for $f_0(t)/f_1(t)$.
- Now recalculate RHS with the real value of p, to get the real LHS.

Sample Setting C

Sample Setting C

- Not much we can do.
- We are finding

$$\arg\max_i \text{cond. density of } X|Y = i$$

Sample Setting C

- Not much we can do.
- We are finding

$$\arg\max_i \text{cond. density of } X|Y = i$$

- I.e. which i makes our X most likely?

Sample Setting C

- Not much we can do.
- We are finding

$$\arg\max_i \text{cond. density of } X|Y = i$$

- I.e. which i makes our X most likely?
- It's the MLE!

Norm Matloff University of California, Davis (talk + discussion)

Sample Setting C

- Not much we can do.
- We are finding

$$\arg\max_i \text{cond. density of } X|Y = i$$

- I.e. which i makes our X most likely?
- It's the MLE!
- But of question value. We want $P(Y|X)$, not $P(X|Y)$.

Conclusion

Conclusion

As usual:

No perfect solutions, but better understand the problem, and have some reasonable remedies.