Classification/evaluation metrics for highly imbalanced data

Asked 3 years, 1 month ago Active 11 months ago Viewed 22k times

21

11

I deal with a fraud detection (credit-scoring-like) problem. As such there is a highly imbalanced relation between fraudulent and non-fraudulent observations.

http://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html provides a great overview of different classification metrics. `Precision and Recall` or `kappa` both seem to be a good choice:

> One way to justify the results of such classifiers is by comparing them to those of baseline classifiers and showing that they are indeed better than random chance predictions.

As far as I understand, `kappa` could be the slightly better choice here, as *random chance* is taken into account. From Cohen's kappa in plain English I understand that `kappa` deals with the concept of information gain:

> [...] an Observed Accuracy of 80% is a lot less impressive with an Expected Accuracy of 75% versus an Expected Accuracy of 50% [...]

Therefore, my questions would be:

- Is it correct to assume `kappa` to be a better-suited classification metric for this problem?
- Does simply using `kappa` prevent the negative effects of imbalance on the classification algorithm? Is re-(down/up)-sampling or cost-based learning (see http://www.icmc.usp.br/~mcmonard/public/laptec2002.pdf) still required?

classification unbalanced-classes precision-recall cohens-kappa model-evaluation

edited Apr 13 '17 at 12:44
Community ♦
1

asked Jul 7 '16 at 8:42
Georg Heiler
235 1 2 11

up/down sampling your data is something you should do when training data is imbalanced your data and can *sometimes* help prevent classifiers from ignoring the minority class(s). Its inappropriate (and a little fraudulent) to use resampled data when evaluating your classifier -- you'll be reporting a performance that your classifier doesn't have when its applied on a sample identically distributed to your original test data. – user48956 Mar 13 '17 at 23:08

Related: stats.stackexchange.com/questions/284515/... – Anton Tarasenko Jan 18 '18 at 17:07

3 Answers

Using a metric like Kappa to measure your performance will not necessarily increase how your model fits to the data. You could measure the performance of any model using a number of metrics, but how the model fits data is determined using other parameters (e.g. hyperparameters). So you might use e.g. Kappa for selecting a best suited model type and hyperparametrization amongst multiple choices for your very imbalanced problem - but just computing Kappa itself will not change how your model fits your imbalanced data.

For different metrics: besides Kappa and precision/recall, also take a look at true positive and true negative rates TPR/TNR, and ROC curves and the area under curve AUC. Which of those are useful for your problem will mostly depend on the details of your goal. For example, the different information reflected in TPR/TNR and precision/recall: is your goal to have a high share of frauds actually being detected as such, and a high share of legitimate transactions being detected as such, and/or minimizing the share of false alarms (which you will naturally get "en mass" with such problems) in all alarms?

For up-/downsampling: I think there is no canonical answer to "if those are required". They are more one way of adapting your problem. Technically: yes, you could use them, but use them with care, especially upsampling (you might end up creating unrealistic samples without noticing it) - and be aware that changing the frequency of samples of both classes to something not realistic "in the wild" might have negative effects on prediction performance as well. At least the final, held-out test set should reflect the real-life frequency of samples again. Bottom line: I've seen both cases where doing and not doing up-/or downsampling resulted in the better final outcomes, so this is something you might need to try out (but don't manipulate your test set(s)!).

edited Jul 8 '16 at 9:52

answered Jul 8 '16 at 9:44
geekoverdose
3,161 2 9 22

But is a cost-based approach like DOI 10.1109/ICMLA.2014.48 more suitable because the overall business impact is considered? – Georg Heiler Jul 20 '16 at 6:27

14

Besides the AUC and Kohonen's kappa already discussed in the other answers, I'd also like to add a few metrics I've found useful for imbalanced data. They are both related to *precision* and *recall*. Because by averaging these you get a metric weighing TPs and both types of errors (FP and FN):

- F1 score, which is the **harmonic mean** of *precision* and *recall*.
- G-measure, which is the **geometric mean** of *precision* and *recall*. Compared to F1, I've found it a bit better for imbalanced data.
- Jaccard index, which you can think of as the $TP/(TP+FP+FN)$. This is actually the metric that has worked for me the best.

Note: For imbalanced datasets, it is best to have your metrics be **macro-averaged**.

answered Sep 20 '18 at 21:27
Johnson
196 1 5

1 What do you mean by 'better' when referring to G-measure and the Jaccard Index? – Narfanar Jan 2 at 12:12

8

For imbalanced datasets, the Average Precision metric is sometimes a better alternative to the AUROC. The AP score is the area under the precision-recall curve.

Here's a discussion with some code (Python)

Here's a paper.

Also see Peter Flach's Precision-Recall-Gain curves, along with a discussion about the shortcoming of AP curves.

edited Sep 20 '18 at 20:43

answered Mar 13 '17 at 23:06

user48956

318 2 10