28 June 2021 Report

Brad Burkman

27 June 2021

Contents

1 Accomplishments This Week

- Worked with George on logistics for getting and processing 5-min weather data, rather than just daily weather data.
- Created new metric, *balanced precision*, and used that to make *balanced f1*.
- Engineered some new features.
  - 'fatal'
  - 'injury'
  - 'pedestrian'
  - 'single_vehicle'
  - 'two_vehicle'
  - 'multi_vehicle'
- Reorganized code to be able to switch between 'fatal' and 'injury' as the independent variable.

2 Questions

- How would I find whether others have used “balanced precision” ? Google and a library search weren’t productive.
-

3 New Metric for Imbalanced Data: “Balanced Precision”

I thought this metric might be the same as scikit-learn’s Precision with the `average=weighted` option, but I checked all of the options, and none of them give this result.

3.1 Reminder of Definitions of Metrics

		Prediction	
		N	P
Actual	N	TN	FP
	P	FN	TP

$$\begin{aligned}
\text{Accuracy} &= \frac{TN + TP}{TN + FP + FN + TP} \\
\text{Recall or TPR} &= \frac{TP}{TP + FN} \\
\text{Specificity, Selectivity, or TNR} &= \frac{TN}{TN + FP} \\
\text{Precision} &= \frac{TP}{TP + FP}
\end{aligned}$$

3.2 Imbalanced Data Set

In an unbalanced data set, the number of actual negatives ($N = TN + FP$) is much different from the number of actual positives ($P = FN + TP$). In our case, if our independent variable is fatal crashes, the negatives are 99.574714% of the data set, and the positives are just 0.425286%.

3.3 The Problem

The standard metrics get thrown off by the imbalance. If we predict that every crash is nonfatal, we have accuracy of 99.57%, which sounds really impressive.

The recall (true positive rate) is not thrown off by an imbalanced data set, because it only works with TP and FN, the actual positives. Similarly for specificity (true negative rate).

The precision is thrown off by an imbalanced data set, because it works with both a subset of the actual positives (TP) and a subset of the actual negatives (FP).

3.4 Balanced Accuracy

There is a metric called *balanced accuracy.* You get it from the definition of *accuracy* by multiplying the actual negative elements (TN and FP) by the ratio of the positives to negatives,

$$\frac{P}{N} = \frac{FN + TP}{TN + FP}$$

so that the total number of actual negatives and total number of actual positives in the sample are equal.

[I suppose you could also get it by multiplying the actual positive elements (FN and TP) by the reciprocal.]

I got this derivation by intuiting about what I would want *balanced accuracy* to mean, and it matches the definition I found in Wikipedia.

`https://en.wikipedia.org/wiki/precision_and_recall#Imbalanced_data`

Wikipedia says [I'm sure I can find a more authoritative source.]

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

$$\begin{aligned}
\text{Recall or TPR} &= \frac{TP}{TP + FN} \\
\text{Specificity or TNR} &= \frac{TN}{TN + FP} \\
\text{Accuracy} &= \frac{TN + TP}{TN + FP + FN + TP} \\
\text{Balanced Accuracy} &= \frac{TN \cdot \frac{P}{N} + TP}{TN \cdot \frac{P}{N} + FP \cdot \frac{P}{N} + FN + TP} \\
&= \frac{TN \cdot P + TP \cdot N}{TN \cdot P + FP \cdot P + FN \cdot N + TP \cdot N} \\
&= \frac{TN \cdot P + TP \cdot N}{(TN + FP) \cdot P + (FN + TP) \cdot N} \\
&= \frac{TN(FN + TP) + TP(TN + FP)}{(TN + FP)(FN + TP) + (FN + TP)(TN + FP)} \\
&= \frac{TN(FN + TP) + TP(TN + FP)}{2(TN + FP)(FN + TP)} \\
&= \frac{TN(FN + TP)}{2(TN + FP)(FN + TP)} + \frac{TP(TN + FP)}{2(TN + FP)(FN + TP)} \\
&= \frac{TN}{2(TN + FP)} + \frac{TP}{2(FN + TP)} \\
&= \frac{TNR + TPR}{2}
\end{aligned}$$

3.5 Balanced Precision

I haven't found *balanced precision* in a brief Google search, although Google knows the kind of stuff I look up and sent me to articles on balanced accuracy. Finding it will take some work, because "balanced precision" has different meanings in other tech fields.

We can make balanced precision the same way we made balanced accuracy, by taking the actual negative results (TN and FP) and scaling them so that the total number of actual negatives equals the total number of actual positives, by multiplying by $\frac{P}{N} = \frac{FN+TP}{TN+FP}$.

Is this related to the G-mean in last week's report? [No]

$$\text{G-mean} = \sqrt{\text{Precision} \times \text{Specificity}}$$

$$\begin{aligned}
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{Balanced Precision} &= \frac{TP}{TP+FP\cdot\frac{P}{N}} \\
&= \frac{TP\cdot N}{TP\cdot N+FP\cdot P} \\
&= \frac{TP(TN+FP)}{TP(TN+FP)+FP(FN+TP)} \\
&= \frac{TP(TN+FP)}{TP(TN+FP)+FP(FN+TP)} \\
&= \ldots
\end{aligned}$$

Giving up here on finding some nice, concise connection between Balanced Precision and other metrics.

4 Top Twenty Features that Correlate with Fatality

Last column is the *balanced f1* score.

DR_COND_CD2	I	DRUG USE - IMPAIRED	0.33
SEC_CONTRIB_FAC_CD	L	CONDITION OF PEDESTRIAN	0.32
PRI_CONTRIB_FAC_CD	L	CONDITION OF PEDESTRIAN	0.25
PRI_CONTRIB_FAC_CD	M	PEDESTRIAN ACTIONS	0.20
VEH_TYPE_CD1	G	OFF-ROAD VEHICLE	0.18
M_HARM_EV_CD1	B	FIRE/EXPLOSION	0.17
DR_COND_CD2	F	APPARENTLY ASLEEP/BLACKOUT	0.17
CRASH_TYPE	C	[Unknown]	0.17
SEC_CONTRIB_FAC_CD	M	PEDESTRIAN ACTIONS	0.16
M_HARM_EV_CD1	O	PEDESTRIAN	0.15
VEH_COND_CD	E	ALL LIGHTS OUT	0.15
F_HARM_EV_CD1	O	PEDESTRIAN	0.15
M_HARM_EV_CD1	F	FELL/JUMPED FROM MOTOR VEHICLE	0.15
F_HARM_EV_CD1	F	FELL/JUMPED FROM MOTOR VEHICLE	0.14
PEDESTRIAN			0.13
VEH_TYPE_CD1	E	MOTORCYCLE	0.13
DR_COND_CD2	G	DRINKING ALCOHOL - IMPAIRED	0.13
CRASH_TYPE	A	[Unknown]	0.13
MOVEMENT_REASON_2	G	VEHICLE OUT OF CONTROL, PASSING	0.12