22 June 2021 Report

Brad Burkman

20 June 2021

Contents

1 Relevant Article

I found an article that does a really good job of analyzing crash data with solid tools and good explanation. [1] They have different data (driving simulator), but their treatment of the data is really good.

2 Imbalanced Data Sets

2.1 Library

There is a Python library, `imbalanced-learn`, that augments `scikit-learn` for ML on imbalanced data sets.

2.2 Example

`https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets`

This Jupyter Notebook works through how to deal with an imbalanced dataset much like ours.

3 Five-Minute Weather Data

At the bottom of this page

`https://www.ncdc.noaa.gov/data-access/land-based-station-data`

is an FTP link

`ftp://ftp.ncdc.noaa.gov/pub/data/asos-fivemin/`

to weather data every five minutes for every land-based weather station in the US.

3.1 Not User Friendly

The data is not in a user-friendly format. The .txt and .dat files are in some weird encoding such that my text editor GUI and Excel can't open them, but I can open them in a terminal window.

Here's what two typical lines look like.

```
13970KBTR BTR201901312250113 01/31/19 22:50:31  5-MIN KBTR 010450Z 10008KT 10SM CLR
     12/07 A3022 -200 71 -600 090/08 RMK AO2 SLP231 T01170067 $
13970KBTR BTR201901312255109 01/31/19 22:55:31  5-MIN KBTR 010455Z 09008KT 10SM FEW090
     12/07 A3022 -190 71 -600 090/08 RMK AO2 T01170067 $
```

- 13970 is the WBAN (Weather Bureau, Army, Navy) number of the weather station.
- KBTR is the ICAO call sign of a weather station in Baton Rouge.
- BTR is the station call sign.
- 2019 01 31 22 55 is the date and local time.
- 0113 is the length of this record, starting at the end of 0113.
- 01/31/19 is the date.
- 22:50:31 is the local time.
- 5-MIN is the data type.
- KBTR is the station call sign.
- 01 04 55 Z is the day and time in GMT, in the format (ddhhmmZ). Note that it's "01" because when it's late evening in Baton Rouge, it's the next day in London.
- AUTO Present in some records. Means the station is in AUTO mode.
- 09008KT, 07007KT, VRB05KT "VRB" may be 'variable.' 09 is wind direction in [0,36] in tens of degrees from true north. 00 would indicate "calm," while 36 would be North. 008 is wind speed in knots in [0,125].
- 10SM, 4SM, 5SM ????
- FEW090, CLR, FEW120, OVC100, BKN100, SCT120 Sky condition with cloud height in hundreds of feet above the ground.
  - CLR - Clear
  - SCT - Scattered clouds
  - OVC - Overcast
  - BKN - Broken?
  - CLRBLO120 - Clear below 12,000 ft.

- 12/07, 09/04, 09/05, 14/07, 02/M01, 23/05, M01/M04 Wind direction and wind speed? Wind direction would be in tens of degrees from true north. Perhaps "M" is "Manual" observation. Wind speed would be in knots.
- A3022, A2998, A3029 ?
- -190, -270, 70, 50, ?
- 71, 82, 85, 96, ?
- -600 ?
- 090/08 ?
- RMK ?
- AO2 ?
- T01170067 ?
- $ Maintenance check indicator.

3.2 Tools Online

I found one GitHub site that had a script to convert this data to Matlab-readable format. I'm sure there are other tools out there, but I haven't looked hard yet. There must be a Python script that will turn a list of .dat files into a .csv file with similar data collected in columns.

4 G-mean, a metric for imbalanced data

Let us review the evaluation metrics.

Precision is the positive predictive value.
Recall is the sensitivity, hit rate, or true positive rate.
Specificity is the selectivity or true negative rate.
F1 is the harmonic mean of Precision and Recall.
Gmean is the geometric mean of Precision and Specificity.

$$
\begin{aligned}
Precision &= \frac{TP}{TP+FP} \\
Recall &= \frac{TP}{TP+FN} \\
Specificity &= \frac{TN}{TN+FP} \\
Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\
F1 &= \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2TP}{2TP+FP+FN} \\
Gmean &= \sqrt{Precision \times Specificity} = \sqrt{\frac{TP}{TP+FP} \times \frac{TN}{TN+FP}}
\end{aligned}
$$

"Gmean ... is considered as a metric of stability between correct classification of positive class and negative class viewed independently. It is usually adopted in order to resist the imbalances in the dataset (Kubat et al., 1997). As this is a class imbalance problem, both F1 score and G-mean have been calculated to evaluate base classifiers." [1]

The Kubat article is available by ILL.

5 References

[EMA20] Zouhair Elamrani Abou Elassad, Hajar Mousannif, and Hassan Al Moatassime. "A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems". In: *Transportation Research Part C: Emerging Technologies* 118 (2020), p. 102708. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102708. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306239.