Accident prediction in time and space

Timothée Guédon, Antoine Hébert

Concordia University
Computer Science and Software Engineering Department

April 15, 2019

Table of Contents

Table of Contents

Road accident prediction

Context:

- Road accidents result in millions of people deaths and injuries every year.
- Human and economic impact on society.
- Master's thesis of Antoine Hébert in a different context.

Our goals:

- Dataset assembling.
- Predict the risk of road accident at a given time on a given road.
- Finding key insights in accident prediction.
- Identifying issues related to class imbalance and geo-spatial data.

Datasets

We used open datasets from Open Canada.

National Collision Database (NCDB):

- All vehicle collisions reported by the police from 2012 to 2017 in Montreal.
- Date and time, severity, number of death and injury, number of vehicles, localization, speed limit etc.

National Road Network:

- Shape of the road, how roads are connected together, the mail addresses on the street, etc.

Climate data - Hourly Data Report:

- The dataset provides the temperature, the humidity, the wind speed and direction, the visibility, the atmospheric pressure etc. at different weather stations in Montreal.

Tree-based algorithms

We chose to focus on tree-based algorithms because it has proven its efficiency on accident prediction.

Related work on accident prediction:

- Extensive use of decision trees in several form. (Theofilatos [2017], Abellán et al. [2013], Lin et al. [2015], Chang and Chen [2005])
- Recently, deep learning algorithms (LSTM and CNN architectures for example) (Yuan et al. [2018], Chen et al. [2016])

Chosen algorithms:

- Random Forest
- Balanced Random Forest
- Gradient Boosted Trees (GBT) - XGBoost implementation

Table of Contents

Generating the examples

Positive samples:

- Identification of road for each accident: Computation of the distances between roads and accidents using their GPS coordinates (taking into account earth curvature)
- Extrapolation of missing points on the roads for more precise road identification

Negative samples:

- Every time an accident did not happen is considered a negative sample.
- We extracted a subsample from the cartesian product between all dates (between 01/01/2012 and 31/12/2017) and all roads.
- Generation of 5 million negative samples for about 200 000 positive samples.

Some feature engineering

- Transformation of the dates and times into a cyclic format
- Interpolation of weather statistics at a given point by averaging data collected from several stations
- Road features: street type and length

Area under PR curve

- ROC Curve : false positive rate VS true positive rate

$$FPR = \frac{FP}{FP+TN}, TPR = \frac{TP}{TP+FN}$$

- PR Curve : true positive rate VS precision

$$TPR = \frac{TP}{TP+FN}, Precision = \frac{TP}{TP+FP}$$

- Data imbalance imply decrease of TP and increase of FN (none of them appears in the FPR)
- Therefore, area under PR exhibits larger differences than area under ROC (Davis and Goadrich [2006])

Table of Contents

Remainder: Random Forests

Remainder

- Random Forests (Breiman [2001]) are combinations of fully-grown decision trees.
- Each tree is trained on a random sample of the data and a random sample of the features.
- The final prediction is given by averaging the votes of the trees.

Weighted Random Forests (WRF):

- A solution to imbalance (Chen and Breiman [2004]).
- Add weights to the classes when growing the tree (in the Gini coefficient).
- Add weights to the classes at the voting time.

Solution to imbalance

Balanced Random Forests (BRF):

- Another solution to imbalance (Chen and Breiman [2004]).
- Combine undersampling and ensemble method to migitate the loss of information.
- For each tree : Randomly draw samples of same size from majority (with replacement) and minority class.
- The rest of the algorithm is identical to Random Forests

BRF versus WRF:

- No clear winner in terms of prediction power ! But...
- WRF - More vulnerable to noise on minority class (Chen and Breiman [2004])
- BRF - More computationally efficient because use less data (subsampling) (Chen and Breiman [2004])
- BRF - Easier to implement in Spark (see next slide)

Implementation of BRF in Spark

Existing implementation of RF in Spark

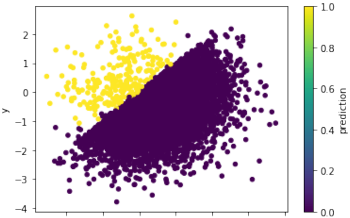
- To manage the subsampling of the dataset, Spark's scala implementation use the Poisson distribution.
- The Poisson distribution is used to compute the probability of an event to occur at a given time.
- To know how many times a given sample will be in the subsample of a given tree (can be 0 times), we use "distribution sampling" of 1 element. (Distribution sampling: Get a random sample following a given distribution)

Our modifications:

- Instead of creating one Poisson distribution (with the "p" parameter and a seed) we create two Poisson distributions:
  - One for the distribution of the "positive" samples,
  - Another for the distribution of the "negative" samples.

Experimenting with our BRF implementation

- Area Under PR : 98,1% for RF VS 98,7% for BRF
- F1 score : 99.7% for RF VS 98.8% for BRF (which proves that F1 score can be misleading in the context of class imbalance)
- With BRF, we detect more accidents but it increases the false positive rate.



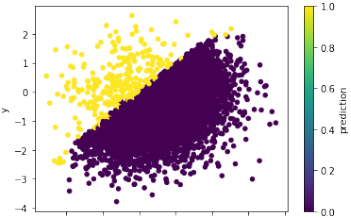
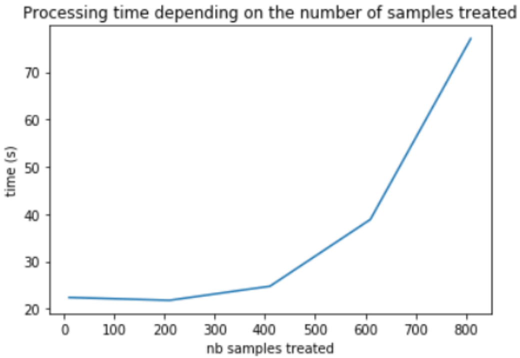


Table of Contents

Dataset generation

- Using cluster to increase the dataset generation's speed.
- Tricks to increase the generation speed and reduce the memory consumption.



Spark VS Dask

Advantages of using Dask over Spark

- Mimic the pandas dataframe API
- Lighter than Spark
- Designed to be more flexible in terms of applications and algorithms

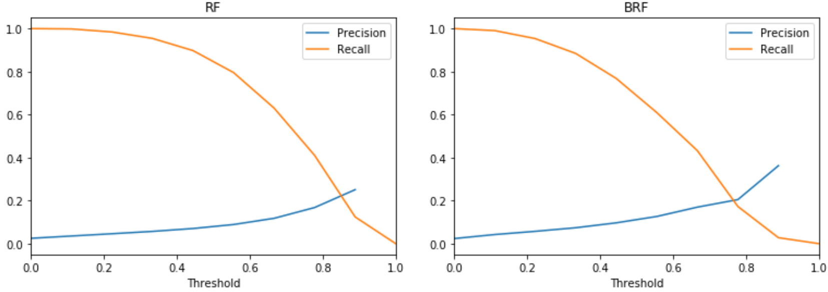
Why we switched for Spark

- We found ourselves writing a lot more code using Dask than using Spark
- We found Spark easier to use for processing the dataset in batch
- Spark's map-reduce and shuffling operations are very handy

Table of Contents

Results

- Random forest: 15.3% area under PR curve
- Balanced random forest: 17.2% area under PR curve



Key features identified

- The most important features seem to be the time and location of the vehicle, followed by some road's features like the type of road (highway for example) and the street length.
- Finally, come some weather features like the visibility, the wind and the humidity.

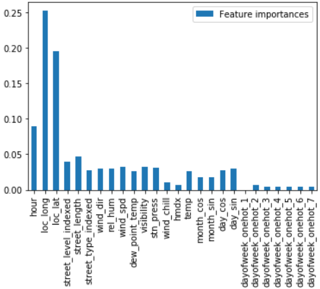


Table of Contents

Conclusion

- We found some key features that can explain the occurence of accidents.
- We identified some solution to deal with class imbalance and big data-related issues.
- We implemented BRF in Spark.
- Future work would include tuning the XGBoost classifier, adding some datasets like a road works dataset.
- The code can be found on Github at https://github.com/GTimothee/accident-prediction-montreal

References I

Joaquín Abellán, Griselda López, and Juan de Oña. Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Applications, 40(15):6047 – 6054, 2013. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2013.05.027. URL http://www.sciencedirect.com/science/article/pii/S0957417413003138.

Leo Breiman. Random forests. Machine Learning, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Li-Yen Chang and Wen-Chieh Chen. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36(4):365 – 375, 2005. ISSN 0022-4375. doi: https://doi.org/10.1016/j.jsr.2005.06.013. URL http://www.sciencedirect.com/science/article/pii/S0022437505000708.

Chao Chen and Leo Breiman. Using random forest to learn imbalanced data. University of California, Berkeley, 01 2004.

Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. Learning deep representation from big and heterogeneous data for traffic accident inference. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pages 338–344. AAAI Press, 2016. URL http://dl.acm.org/citation.cfm?id=3015812.3015863.

References II

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. pages 233–240, 2006. doi: 10.1145/1143844.1143874. URL http://doi.acm.org/10.1145/1143844.1143874.

Lei Lin, Qian Wang, and Adel W. Sadek. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies, 55:444 – 459, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2015.03.015. URL http://www.sciencedirect.com/science/article/pii/S0968090X15000947. Engineering and Applied Sciences Optimization (OPT-i) - Professor Matthew G. Karlaftis Memorial Issue.

Athanasios Theofilatos. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. Journal of Safety Research, 61:9 – 21, 2017. ISSN 0022-4375. doi: https://doi.org/10.1016/j.jsr.2017.02.003. URL http://www.sciencedirect.com/science/article/pii/S0022437517301378.

Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. pages 984–992, 2018. doi: 10.1145/3219819.3219922. URL http://doi.acm.org/10.1145/3219819.3219922.