UNIVERSITY OF CAMBRIDGE

Harry Graham
Christ's College
hg402

Computer Science Tripos - Part II Project Progress Report

Deep Learning Techniques for Credit Card Fraud Detection

Progress Report

February 2, 2018

Project Supervisor: Dr M. Jamnik & B. Dimanov

Project Originator: H. Graham & B. Dimanov

Director of Studies: Dr R. Mortier

Project Overseers: Dr S. B. Holden & Dr N. R. Krishnaswami

Summary

So far the project is on schedule, as set out by the project proposal document. Some of the michaelmas term work was done ahead of schedule which compensated for other parts of the work that took longer than expected. I am also in the final stages of the second CNN model which was planned mainly for the start of Lent term.

Work Done

- Experiments with sampling techniques
  – I experimented with under-sampling, random oversampling and SMOTE methods. I used a linear regression classifier.
- Implemented custom cross validation
  – I created a custom cross-val method, as built in library functions do not provide enough granularity for the project's interests. This was to ensure no overfitting and to correctly ensure that we resample the data during cross validation in a systematic way. By implementing a procedure based on SKLearn's KFOLD cross validation, I could retrieve a lot more metrics (F1, precision, recall etc) and also implement the oversampling/SMOTE inside the KFold loop, to ensure that the validation test set is always a preserved portion of the original data and not part of the oversampled data.
- Experiments relating to systematic training and testing of classifiers
  – I researched and experimented thoroughly the effects of test results when over-sampling before and during cross validation.
- Baseline models implemented and summarised
  – I have collated results for the baseline models, for the metrics described in the proposal document and F1 score.
- First CNN model implemented and trained
  – I have created the first CNN model whereby the input shape of the network is the length of the vectors by 1 and the kernel size is also the length of the vectors. So in this case (29,1) and 29 respectively. I have trained this model and collated results.
- Scaffolding for second CNN model created
  – The outline of the second CNN model has been implemented. This is the model whereby principally we pass in batches of 100 vectors into the network, input shape (29,100) and the kernel size is (29,5) with a stride of 1, striding over the data in a vertical dimension of 2-D space to exploit temporal ordering.
- Rough plans for GAN work have been established
  – I have thought about how to approach this portion of the work.

Work to be Done

Work to be done is set out in the project proposal document. This includes:

- Data preprocessing and training second CNN model
  – I need to pad my data to ensure that the network is happy with the input shape. That is to say, that there are multiples of 100 for each batch. I then need to train the model to retrieve results. This, however, will need care when cross validating because this is now a time series style of data and so we want to do CV appropriately, by looking systematically at future data points and not random old data points, like other CV functions do.
- Summarising CNN work
  – I will summarise results of both CNN models and draw comparisons from each, as detailed in the proposal document.
- Implement GAN network model and training
  – As described in proposal
- Extension work if time permits
  – As described in proposal

Difficulties

There have not been significant difficulties in completing the work thus far. The time to setup project work took longer than expected and experimentation with libraries (for example gpu use in tensorflow and setting up virtual environments) also ate more time than estimated.

An unforeseen workload was library functions for cross validation not being enough for the work being done. I had to implement my own cross validation (Based on standard Kfold in SKLearn) in order to achieve the project's needs. For example retrieving more complex scoring data from the CV function and also ensuring that we are oversampling correctly inside the cross validation loop and not overfitting. The work was a result of research and experimentation and although was not something foreseen at the time of project proposal, was a good learning experience.