Proposal for Compsci 760 Group Project:
Impact of drift detectors and sampling methods on classification of streaming imbalanced data with underlying concept drift

By Robert Anderson, Kylie Chen and Eric Jin

26/09/2014

1 Introduction

This proposal aims to explain our group project for Compsci 760. First, we detail our motivation for our research - why is this area important to study? We then clearly set our research question, and explore the hypotheses we will test. The next section explains our approach to our research, with each step detailed and justified. After this, we discuss our anticipated results, with the support of related works in the field. Finally, we demonstrate the contribution this makes to the community as a whole.

For the purposes of this paper, 'streaming data' refers to large amounts of data, arriving over time. 'Underlying concept drift' is defined as the distribution that generates the data changing over the period of the stream. 'Classification' is a method that uses attributes of a data item to label it with a class systematically. For the purposes of this project, we limit our scope to binary-class problems. 'Drift-detectors' are mechanisms that detect changes in the underlying distribution and adapt the classification approach if changes occur. 'Imbalanced' data is characterized by having one prominent class and one rare class, in the binary-class context. 'Sampling methods' are techniques that re-use or remove underlying data records in some manner so as to improve the results of classification.

Many of the techniques we discuss have been referenced in our submitted literature review, "Classification of Imbalanced Data Streams with Concept Drift: a Literature Review". Where we have referenced other sources of information, we have cited them within this paper.

2 Motivation

The cost of infrastructure required to collect data has dropped dramatically, while systems are capable of collecting much more information. Data stream analysis is more important than ever, providing understanding of these huge swathes of data without having to store and process such a large amount of data over a long period. Effective and robust data stream analysis can allow near real-time understanding of a domain, permitting a more effective response to what the data shows. In short, effective data stream analysis provides value that delayed, static dataset analysis rarely can.

However, there is a significant downside when analysing streaming data. Without a prior knowledge of the characteristics of our data, we cannot hand-pick an approach that will provide strong analysis of a particular dataset. Classification of imbalanced data is a difficult problem even in a static environment, as it is difficult to develop a meaningful understanding of a class with few examples. As streams run over time, concept drift also becomes an issue: we may be classifying data in a manner that is no longer current and is therefore inaccurate. In our research, we seek to test current solutions to these problems that have been optimized to work in a data stream. As imbalanced streams are just as likely to have concept drift as normal streams, it is important to consider both issues in conjunction. What is more, concept drift in a rare dataset can be much more difficult to detect, so current methods of concept drift detection need to be tested in an environment that may interfere with their effectiveness.

Finally, we will not always know whether data is class-imbalanced or suffers from concept drift. It is important to measure how drift detection methods and sampling techniques, which seek to solve the problems above, impact on the performance of analyses of data streams that suffer from neither problem.

3 Research Question

We propose to evaluate the effects of combining sampling methods (none/SMOTE/potentially NCL) with drift detectors (none/ADWIN/PHT) for a single base learner (Hoeffding trees) on classification of imbalanced data streams with concept drift. We would like to investigate how these different hybrid approaches perform on datasets with different characteristics such as data with gradual drifts, abrupt drifts, and varying levels of class imbalance. The performance measures of interest to us include memory usage, runtime, overall accuracy, recall, precision and F-score. Our hypotheses are:

- that applying sampling techniques to drift detectors will greatly improve the performance of classifiers for datasets with high and moderate degrees of class imbalance compared to drift detectors without sampling.
- that applying more layers of complexity (i.e. sampling, drift detection) to the base learner will increase the time and memory requirements, but will not always guarantee better performance.
- that there is no statistically significant difference in accuracy or recall between using different drift detectors.

4 Approach

MOA and Weka libraries include a lot of ready-to-use algorithms for stream data mining. We will use MOAs API or command line to run our experiment programmatically in Java. We will generate 9 synthetic data streams with no/some/high class imbalance (through manual manipulation) and with no/gradual/abrupt concept drift (through the stream generators listed below). On each synthetic dataset, we will run our classifier with each combination of drift detector and sampling method 30 times to ensure representative measures of the underlying performance. Through the testing suite we will develop, we will record the measurements listed above per run, and use ANOVA in R on the results to compare each metric for different combinations of drift detector and sampling method.

4.1 System architecture

Our system architecture consists of our stream generators, sampling modules, drift detection modules, learning module and evaluation module. The drift detection module and learning module may be combined where MOA offers a combined implementation e.g. the Adaptive Hoeffding Tree which combines the Hoeffding Tree and ADWIN. Our testing module will output metrics from all of our experiments to one CSV file which we will then use for evaluation in R. Our final result will compare results for the possible algorithm combinations for our nine separate datasets.

4.1.1 Sampling module

SMOTE(Synthetic Minority Over-Sampling TEchnique) is a popular approach which increases new, non-replicated minority class instances to the training dataset. We will apply SMOTE sampling to our ARFF datasets to re-balance them, so that after pre-processing, the class distribution of the training dataset is balanced. We also want to see the performance without sampling applied: for these experiments, we won't apply any sampling techniques. Given enough time, we will try NCL as an alternative sampling technique.

4.1.2 Stream generator

Streams will be generated using the following MOA classes:

- NoChangeGenerator
- SEAGenerator/AbruptChangeGenerator for abrupt concept drift
- GradualChangeGenerator

We will combine the ConceptDriftStream objects above with an imbalanced ArffFileStream to generate an imbalanced dataset with concept drift.

4.1.3 Classification module

- HoeffdingTree(VFDT)

  We have chosen to use the HoeffdingTree as a base learner. A Hoeffding tree is an incremental, anytime decision tree induction algorithm that is capable of learning from data streams. We have detailed this tree in our submitted literature review. We have chosen to use Hoeffding trees as they are popular within data streaming implementations. Importantly, it has no intrinsic characteristics to help it adjust to concept drift.

- SingleClassifierDrift(optional)

  This is a wrapper for any base learner capable of handling concept drift datasets. It allows any drift detection method to be used along with it, such as PHT and ADWIN.

- Hoeffding Adaptive Tree

  This is derived from the Hoeffding Window Tree and uses ADWIN as a change detector. These adaptively learn from data streams that change over time without needing a fixed size of sliding window. The optimal size of the sliding window is a very difficult parameter to guess for users, since it depends on the rate of change of the distribution of the dataset. The Hoeffding Adaptive Tree sidesteps this issue.

4.1.4 Evaluation module

We will use an evaluation method implemented in MOA: either Interleaved Test-Then-Train or Prequential. Each individual example is used to test the model before it is used for training, and from this the accuracy is incrementally updated. This scheme has the advantage that no holdout set is needed for testing, making maximum use of the available data.

The following measurements will be used: memory usage, runtime, overall accuracy, recall, precision and F-score. Precision is the Positive Predicted Value. F-measure is a combination of recall and precision, representing a harmonic mean. In practice, high F-measure values ensure that both recall and precision are reasonably high.

We may use ROC (Receiver Operating Characteristic)curves, which provides a single measure of a classifiers performance for comparing models' mean performance.

5 Anticipated Results

We expect that the accuracy of ADWIN and PHT will perform similarly, and better than without a drift detector. This is supported by Goncalves et al, who showed that there is no difference between the performance of classifiers using either method at a p-level of 0.05, based on the Nemenyi test [1]. However as their results did not focus on imbalanced datasets, the conclusions can not be generalised to datasets with skewed distributions. We expect drift detectors to provide little improvement when there is class imbalance without sampling methods, as drift detectors have trouble detecting fluctuations in rare classes.

We expect ADWIN to have a better runtime than PHT, as ADWIN has been previously shown to have the lowest evaluation time for most datasets (with a variety of properties) at a 95% confidence interval by Goncalves et al [1]. We expect to have the lowest runtime when not using drift detectors.

The SMOTE sampling technique has been shown to be an effective way to deal with data with imbalanced classes [2] and creates better decision boundaries by shifting the boundary away from positive instances when combined with support vector machines - SVMs (Akbani et al, 2004 as cited by [3]). Other studies such as Hulse et al's paper [4] have evaluated the effects of combining different sampling methods with different learners for imbalanced datasets, but did not examine datasets with underlying concept drift. Hulse et al. showed that there is an interaction between the sampling technique and learner, and the effectiveness of a particular sampling technique depends on the type of learner that is used [4]. Thus it is probable that the performance of sampling techniques also depends on the type of drift detector that is used in streams with concept drift. We hope and expect that using sampling techniques on our data will allow concept drift detectors to function near-optimally.

6 Significance of the Proposed Research

McKinsey Global Institute [5] have deemed 'Big Data' to be 'the next frontier for innovation, competition and productivity'. They state that it is relevant to '/textbfevery business and industry function', and that it will become 'a key basis of competition and growth for individual firms'.

This research sits at the forefront of data stream mining, which is critical for handling the volume and velocity of data for organisations today. However, data stream mining is prone to issues that are not addressed in traditional data analysis. A robust, reliable technique will need to account for concept drift and imbalanced datasets, classifying accurately while still performing acceptably in time and memory usage. These issues are, in fact, common in data streams [6]. If streaming analysis is not robust in the presence of these issues, then it will not be fit to analyse real-world data that comes from an unknown distribution, and cannot help in solving the problems that 'Big Data' analysis creates.

More specific to the field, the algorithms we test in this paper are recent and highly-regarded. The research community will benefit from having a clear analysis of how well they function under the particular conditions we have set. This work examines algorithms that we have found to be influential and important within our literature review. By testing them under our specific conditions, we will test their robustness. Our results will show what they do effectively and may highlight areas of improvement for fellow researchers.

7 Conclusion

Through this report, we have outlined our intended research into data streaming with concept drift and imbalanced classes. We describe why we are motivated to study this area. We explain our research question, and divide it into three clear hypotheses that we will test. We have described the process by which we plan to implement the tools we need to answer our research question, and describe how we will test and evaluate our results. We have discussed what we expect from our results, and justify our opinions through citing similar research. Finally, we have clearly spelt out the contribution we will be making to the field of data stream analysis through our research.

References

[1] P. M. G. Jr., S. G. de Carvalho Santos, R. S. Barros, and D. C. Vieira, "A comparative study on concept drift detectors," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8144 – 8156, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414004175

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *arXiv preprint arXiv:1106.1813*, 2011.

[3] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined {SMOTE} and {PSO} based {RBF} classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456 – 3466, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231211003559

[4] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 935–942. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273614

[5] The McKinsey Global Institute, "Big data: The next frontier for innovation, competition and productivity," http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, 2011, accessed: 2014-09-26.

[6] S. Wang, L. Minku, D. Ghezzi, D. Caltabiano, P. Tino, and X. Yao, "Concept drift detection for online class imbalance learning," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, Aug 2013, pp. 1–10.