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 they are capable of collecting much more information. Analysis of data streams is more important than ever, as it allows 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 (SMOTE and potentially NCL) with drift detectors (ADWIN and PHT) and a single base learner (Hoeffding trees) on 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, F-score, and accuracy of minor classes (kappa statistics??). 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.
- applying more layers of complexity (i.e. sampling, drift detection) to the base learner will increase the time and memory requirements, but does not always guarantee better performance.
- 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 may use MOAs API or command line to run our experiment programmatically in Java. The code should be performing following steps in order: first we will generate some synthetic data streams with imbalanced class distribution using data feeding module below, each type of data stream will be generated 30 times with different seed; save stream to arff data set; we use sampling module to re-balance the class distribution for those data sets; we use classification module to test and train models using the data streams; evaluation module will be used to extract useful measurements.

4.1 System architecture

Our system architecture consists of sampling module, data feeding module, drift detection module, learning module and evaluation module. The drift detection module and learning module could be combined if the learning algorithm is capable of drift detection. In one particular module, when there are two or more algorithms that want to compare their performance with each other, we will run them one by one and record their performance metrics. And our final result will contain all the possible algorithm combinations between different modules.

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. When we have external arff dataset, we will apply SMOTE sampling on arff dataset to re-balance as part of the experiment procedure, so that after pre-processing the class distribution of training dataset is balanced. We also want to see the performance without sampling approach presented in the system, in which case we wont apply apply any sampling techniques.

4.1.2 Data feeding module

Streams will be generated using the following MOA classes:

- NoChangeGenerator
- SEAGenerator/AbruptChangeGenerator for abrupt concept drift
- GradualChangeGenerator
- ConceptDriftStream+ArffFileStream

Each generator will be used to generate 30 streams with different random seed. Non-Drifted data stream could be joining concept drift stream to form a concept-drifted stream.

4.1.3 Classification module

- NoChange Or HoeffdingTree(VFDT)
  This classifier is chosen because we want to see how the classifier performs when it cannot handle concept drift, and it will be our baseline classifier for reference point. A Hoeffding tree is an incremental, anytime decision tree induction algorithm that is capable of learning from data streams, assuming that the distribution generating examples does not change over time. Hoeffding trees exploit the fact that a small sample can often be enough to choose an optimal splitting attribute. This idea is supported mathematically by the Hoeffding bound, which quanties the number of instances needed to estimate some statistics within a prescribed precision. A theoretically appealing feature of Hoeffding Trees not shared by other incremental decision tree learners is that it guarantees the performance.

- SingleClassifierDrift(optional)
  This is a wrapper of any base learner also capable of handling concept drift datasets any drift detection method could be used along with it, e.g. DDM, EDDM, ADWIN.

- Hoeffding Adaptive Tree
  Hoeffding Window Tree that uses ADWIN as a change detector. Hoeffding Adaptive Tree as a new method that evolving from Hoeffding Window Tree, 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.

4.1.4 Evaluation module

We will use an evaluation method implemented in MOA: 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: accuray, time, memory usage, AUC, Kappa-statistic, Precision, Recall, F-Measure. Recall is the True Positive rate. Precision is the Positive Predicted Value. F-measure is a combination of recall and precision. This represents a harmonic mean between recall and precision. In practice, high F-measure value ensures that both recall and precision are reasonably high. The area under a ROC (Receiver Operating Characteristic)curve (AUC) provides a single measure of a classifiers performance for evaluating which model is better on average.

5 Anticipated Results

We expect ADWIN to have better performance in terms of runtime 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 also expect that the accuracy of ADWIN would be higher than that of PHT and methods without drift detector, but the difference between ADWIN and PHT would be negligable as Goncalves et al showed that there is no difference between the performance of classifiers 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.

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.

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 a critical technique in handling the volume and velocity of data in today's academic, business and governmental context. 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 very relevant 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.