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 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

4 Approach

5 Anticipated Results

6 Significance of the Proposed Research

McKinsey Global Institute [1] 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 [2]. 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] "Big data: The next frontier for innovation, competition and productivity," 2011, accessed: 2014-09-26.

[2] 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.