R. Anderson, K. Chen and S. Jin

***Abstract*—**

I. Introduction

In recent years, researchers have drawn many attentions to data stream mining in the field of data mining, because the computing power has increased dramatically. An imbalanced dataset has a skewed class distribution, with usually one majority class distribution and one or more minority class distribution. The imbalance problem has been well researched in static data mining area. Imbalanced data stream exists in many applications such as network traffic data and credit card transactions. Concept drift refers to the underlying class distribution of data stream potentially changing over time. Concept drift has also been well studied in the research community of data stream. An example of concept drift is online shopping customer preferences changing over time. A drift can happen abruptly or gradually. To solve the problem of concept drift, predictive models needs to be updated, which is also called online adaptive learning. Concept drift is difficult to detect because the detector may be confused with noise and outliers. There are very few researches combining the two problems of concept drift and imbalanced data stream. One representative paper is A General Framework for Mining Concept-Drifting Data Streams with Skewed Distributions[1], in which Gao[1] proposed a framework applying both under-sampling and ensemble techniques demonstrating better prediction performance.

As part of our assignment for Compsci 760, in our literature review paper, we have reviewed: various sampling methods for solving the problem of imbalance; many types of drift detection methods to solve the problems of concept drift; ensemble of base learning algorithms to solve the problems of concept drift; evaluation metrics to measure the systems overall performance.

Our experiments objective is to answer some hypotheses that we raised in our project proposal: 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, especially when the data is highly imbalanced. 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; there is no statistically significant difference in accuracy or recall between using different drift detectors.

MOA and Weka libraries include a lot of ready-to-use algorithms for stream data mining. In our experiment, we use MOA and WEKAs API to interact with MOAs stream mining capability and SMOTEs sampling capability. We have modified some synthetic data stream generators to provide us 3 different levels of desired class imbalance proportion: 0.01, 0.1, 0.5. For each data stream generator we use 30 different seed, which will ensure that our experiment result is not due to random chance and our result could be re-produced by setting the same seed; we have prepared 3 different concept drift detection methods to be used along with the HoeffdingTree base learner; we have set up a test suite to record some evaluation metrics for every possible combination setting among different dataset, class imbalance ratio, concept drift detection methods while running our experiment with and without SMOTE sampling presented. Every time a batch sample passing to SMOTE, our code analyses the current class ratio and an increased percentage is calculated and used to instruct SMOTE to give exact desired class balance ratio, e.g. 1. This feature gives advantage over fixed the increased percentage, and could be applied to multi-class problems easily. We also use Student t test to make sure that the result is reliable to answer those hypothesis.

This paper is organized as follows: next section briefly presents the previous work we have done for our project; section 3 discusses related research in the field of imbalanced data stream mining; section 4 presents the approach to carry out our experiments; section 5 presents our experiment result; section 6 discusses future work and presents conclusion.

II. Related Work

A few papers discuss their approaches to solve the imbalance problem for concept-drifted data stream, classifier ensemble along with sampling approach is the most common one which has been presented in [**?**] and [1]. Gao[1] has shown that reconstruct a model on new data reduces expected error theoretically and experimentally and the proposed algorithm employing both sampling and ensemble techniques generate reliable probability estimate and reduce misclassification error rate. However, these two methods have used different sampling techniques. SMOTE[**?**], SERA[**?**], MUSERA[**?**] , REA[**?**] and Learn++SMOTE[**?**] are over-sampling based approaches.

Generally, classifier ensemble is believed to perform better on the imbalanced data stream. Ensembles consist of many components (typically different models or learners) which are combined by voting or averaging the outcomes from each component. Ensemble methods can often improve the accuracy of a single classifying algorithm. A drawback is the increased memory and time requirement.

In Mining data streams with skewed distribution by static classifier ensemble, Zhang et al. proposed a clustering-sampling based ensemble algorithm with weighted majority voting for learning skewed data streams, he uses k-means clustering algorithm for selecting minority instances to represent minority class, the centroid of each cluster is used as an instance for the class.

In [**?**] , Andrea et al. avoided sampling by using Hellinger Distance Decision tree (HDDT[**?**]) as a base learner for data stream, which showed superior performances in terms of prediction accuracy, recall rate, computational time and resources. HDDT theory derives a new decision tree splitting criteria based on Hellinger distance that is skew insensitive.

III. Concept Drift

IV. Class Imbalance

V. Experimental Design

A. Datasets

VI. Results

In this section, we will present the aggregated results of our experiments for our synthetic datasets and discuss each of our hypotheses in turn. We will continue on to showing results for the Electricity dataset [reference] available from the UCI repository, to give an idea of how our approach may work on real world data.

A. Synthetic Datasets

For our synthetic datasets, we ran experiments 30 times for every combination of factors, on streams with 1 million instances generated. We present the mean of those experiments along with a 95% confidence interval given the variance in our experimental results. In graphs following, error bars show the range of this interval (though note that some experiments were so consistent that these error bars are almost invisible).

The data-level factors we varied across these experiments were:

- stream generators, with levels '**No CD/drift**' (STAGGER), '**Gradual**' drift (RBFGenerator) and '**Abrupt**' drift (STAGGER)).
- class-balance ratio, with levels **Balanced** (1:1), **Imbalanced** (1:9) and **Very Imbalanced** (1:99).

The learning-approach factors we varied across these experiments were:

- sampling method, with levels '**SMOTE**' and '**No SMOTE**'.
- drift detection method, with levels '**ADWIN**', '**PHT**' (Page-Hinkley Test) and '**No DDM**' (no drift detection).

Through our experiments in this section, we have chosen to focus on four evaluation measures: memory, runtime, accuracy and precision. Memory refers to model size at the end of the experiment. This is affected heavily by drift detectors, which regularly may choose to dispose of a model after detecting drift. Runtime refers to total time to run the experiment, excluding stream generation and evaluation of the model. Accuracy refers to the proportion of instances correctly classified. F-Score is a weighted average of recall (the true positive rate) and precision (positive predictive value). This measure can usefully distinguish between maximimising accuracy by defining all as the majority class in an imbalanced dataset, and intelligently predicting the minority and majority class. We do not expect SMOTE to outperform no SMOTE in accuracy, but if it is effective, we can expect to see a superior F-Score in experiments.

In our tabulated results, for each set of factors we are comparing, we have put the best result in bold. This can involve some or all comparisons, where the 95% confidence intervals we have used overlap.

1) Impact of drift detectors upon classification: To test our hypothesis that there is no significant difference in accuracy between drift detectors, we have chosen to examine our streams with gradual and abrupt drift with balanced and very imbalanced data(Fig 1 and 2), and include results for the no drift and imbalanced situations in our tabulated results (Fig 3).

Fig. 1. Evaluation measures for drift detectors with gradual drift

Our gradual stream had ten variables, making it more difficult to classify correctly than the STAGGER-based streams that only had three. We can see for balanced data that ADWIN and PHT worked similarly well, with having no difference in F-Score, and ADWIN being marginally better for accuracy. In the very imbalanced situation, ADWIN had a slightly better F-score while PHT had a slightly improved overall accuracy. Both worked significantly better than having no drift detection.

Fig. 2. Evaluation measures for drift detectors with abrupt drift

With the abrupt datastream, no significant difference in accuracy nor F-score could be seen between any type of drift detector until the data became very imbalanced. At this point, PHT outperformed ADWIN very significantly for F-Score, and marginally in accuracy. Having no drift detector did. It may be due to ADWIN changing its model too regularly to be able to clearly classify the rare class, which it sees few instances of. A small number of misclassifications of a rare class can lead to a large impact on an F-score as well.

Analysing our table, we can see that across all testing circumstances, PHT and ADWIN are often indistinguishable, and trade off the top-spot for different measures under different circumstances. There is only one combination where either work markedly worse than having no drift detection (F-score for ADWIN with abrupt drift.) These results bear up our hypothesis reasonably well.

2) Impact of SMOTE upon classification:

3) Impact of SMOTE and drift detectors upon classification:

4) Impact of SMOTE and drift detectors upon cost:

B. Electricity Dataset

VII. Future Work

VIII. Conclusion

Fig. 3. F-Score and Accuracy for combinations of drift, drift detector and balance with no SMOTE

REFERENCES

[1] J. Gao, W. Fan, J. Han, and S. Y. Philip, "A general framework for mining concept-drifting data streams with skewed distributions." in *SDM*. SIAM, 2007, pp. 3–14.