International Journal of Innovative Computing, Information and Control
Volume **8**, Number **7(B)**, July **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 4995–5010

ADAPTIVE DATA REUSE FOR CLASSIFYING IMBALANCED AND CONCEPT-DRIFTING DATA STREAMS

Hien M. Nguyen[1], Eric W. Cooper[2] and Katsuari Kamei[2]

[1]Graduate School of Science and Engineering
[2]College of Information Science and Engineering
Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan
hiennm@spice.ci.ritsumei.ac.jp; cooper@is.ritsumei.ac.jp; kamei@ci.ritsumei.ac.jp

Received May 2011; revised September 2011

Abstract. *Mining data streams has recently been the subject of extensive research efforts. However, most of the works conducted in this field assume a balanced class distribution underlying data streams. In this paper, therefore, we propose a new method for learning from imbalanced data streams. To deal with the problem of class imbalance, we select and reuse past data to improve the representation of the minority class. Different from previous methods, our method has the ability to automatically adapt data selection for concept drift. A data stream may experience a complicated concept drift, making data selection more difficult. Therefore, we consider several different candidate solutions of data selection, each of which is possibly more appropriate for certain data streaming conditions. In other words, no one of them is the best at all times. We make comparisons and identify the best candidate solution by cross-validation on the most recent training data. By experimental evaluations on simulated and real-world data streams, we show that our method achieves better performance than previous methods, especially when concept drift occurs.*

Keywords: Adaptive data reuse, Data selection, Class imbalance, Concept drift, Data stream, Ensemble learning

1. **Introduction.** Mining data streams has recently been the subject of extensive research efforts [1]. Some examples include detection of fraudulent credit card transactions [2], web filtering [3], and click-stream clustering [4]. Basically, there are two approaches to learning from data streams. The first is incremental learning in which a single model is built from an initial training set, and then is continuously updated with new training instances upon their arrival [5]. The second is ensemble learning in which a series of base models is trained on consecutive chunks of a training data stream, and forms an ensemble to classify previously unseen instances [6].

One of the biggest challenges facing data stream learning is to deal with "concept drift", i.e., the concept to learn is changing over time. For example, in a task of online news filtering, some user might expand his/her reading interests with a new topic such as "nuclear safety" after the Fukushima nuclear accident occurred in Japan. In this example, the concept to learn (or target concept) is the reading interests of a specific user. Concept drift could cause a classification model to perform poorly if this model is not updated appropriately to reflect changing data distributions. To deal with concept drift, one solution is to regularly retrain the classification model on a sliding window covering only the most recent training instances [7]. Another solution is to use ensemble learning in which the base models are weighted using their classification accuracy on the newest