Evolving Systems (2013) 4:1–2
DOI 10.1007/s12530-012-9070-5

EDITORIAL

Introduction to the special issue on handling concept drift in adaptive information systems

Mykola Pechenizkiy · Indre Zliobaite

Received: 4 December 2012 / Accepted: 6 December 2012 / Published online: 29 December 2012

Modern information systems collect data from multiple sources, process it, extract information and use it for decision support or decision making. Predictive modeling is an important component of an information system that makes the system intelligent.

This special issue focuses on adaptive information systems that can adjust their behavior relying on additional mechanisms continuously monitoring the operational setting and/or the performance of predictive models. Adaptive information systems have become ubiquitous in various application areas including online businesses, personal information access, industry, medicine, education, defence, and in which predictive analytics is an important decision making or decision support component.

In the real world data is often non stationary. In predictive analytics, machine learning and data mining the phenomenon of unexpected change in underlying data over time is known as concept drift. Changes in underlying data may occur due to changing personal interests, changes in population, adversary activities or they can be attributed to the complex nature of the environment. When data drifts, predictions may become less accurate as the time passes or opportunities to improve the accuracy may be missed. Thus, the learning models need to be able to adapt automatically to changes over time.

The problem of concept drift is of increasing importance in machine learning and data mining since more and more data is organized in the form of data streams rather than static databases, and it is rather unusual that concepts and data distributions stay stable over long periods of time. It is not surprising that the problem of concept drift has been studied in several research communities including but not limited to machine learning and data mining, data streams, information retrieval, and recommender systems. Different approaches for detecting and handling concept drift have been proposed in the literature, and many of them have already proved their potential in a wide range of application domains, e.g. fraud detection, adaptive system control, user modeling, information retrieval, text mining, biomedicine. Moreover, a fast growing scope of applications that rely on data arriving in real time where very often the problem of data drift is observed and recognized to be important, helped to identify and shape a number of new important research challenges that have not been well-studied in the research communities yet.

This special issue includes selected contributions from the first and the second International Workshops on Handling Concept Drift in Adaptive Information Systems; HaCDAIS[1] at ECMLPKDD 2010 and HaCDAIS[2] at IEEE ICDM 2011. The papers address both methodological issues and practical challenges for handling concept drift, such as (a) label availability, (b) recurring concepts, (c) systematic handling of event detection, and (d) mining changes in customer profiling and medical anesthesia domains.

We hope you will find the following papers interesting for reading.

The first paper "Drift Detection Using Uncertainty Distribution Divergence" by Patrick Lindstrom, Brian Mac Namee and Sarah Jane Delany addresses the problem of

M. Pechenizkiy
Eindhoven University of Technology, Eindhoven,
The Netherlands
e-mail: m.pechenizkiy@tue.nl

I. Zliobaite (✉)
Bournemouth University, Poole, UK
e-mail: zliobaite@gmail.com; izliobaite@bournemouth.ac.uk

[1] http://wwwis.win.tue.nl/hacdais2010/

[2] http://wwwis.win.tue.nl/hacdais2011/

label availability. While most existing approaches for handling concept drift make the key (yet often unrealistic) assumption that the labeled data will be available at no labeling cost shortly after classification, this work proposes the Confidence Distribution Batch Detection (CDBD) technique that explicitly detects changes in the data rather than in the decision boundary and does this without using labeled data. CDBD compares the distribution of classifier confidences in a batch of test instances to a reference distribution to generate an indicator stream correlated to changes in concept. The authors also show how this indicator stream combined with a trigger and a rebuild policy can maintain classifier accuracy while using a limited amount of labeled data. CDBD and its extension was evaluated on several dataset from the text classification domain and compared to other drift handling approaches. The results suggest that the proposed approach is comparable in effectiveness to existing techniques while requiring a smaller amount of labeled instances.

The second paper "Using a Classifier Pool in Accuracy Based Tracking of Recurring Concepts in Data Stream Classification" by Mohammad Javad Hosseini, Zahra Ahmadi and Hamid Beigy proposes an ensemble approach for learning and tracking recurring concepts by exploiting previous knowledge obtained in the learning process under concept drift. The proposed approach called Pool and Accuracy based Stream Classification is rather straightforward; each classifier is used to describe an existing concept, consecutive batches of instances are classified by the current pool of active or weighed classifiers, when the true labels become available, the pool is updated leading either to the construction a new or to an update of an existing classifier. Experimental results conducted by the authors on real and artificial datasets show the effectiveness of the proposed approach leading to a quicker adaptation of the learner whenever a concept reappears.

The third paper "EVE—A Framework for Event Detection" by Iris Adä and Michael Berthold introduces a generic framework for event detection, where events can include outliers, model changes and data drifts. The authors try to unify various existing and foreseen methods for event detection by analyzing their prior assumptions and by providing a common basis for their consideration, consisting of generic types of time slots, measures of similarity between these time slots, and progress mechanisms. To demonstrate the generality of the proposed framework, the authors provide several illustrative examples demonstrating how EVE can fit existing algorithms.

The fourth paper "Detecting and Visualizing the Change in Classification of Customer Profiles based on Transactional Data" by Edward Apeh and Bogdan Gabrys presents the use of the data binning process (based on the number of items purchased) for constructing customer profiles. The authors investigate on the real dataset of purchasing transactions, with multiple transactions per customer, how customer behavior or their profiles may change over time. This is done by learning and monitoring the performance of decision tree ensembles maintained over different window sizes. From the application perspective, visualization and analysis of detected changes allows to better understand customer behavior and how stable it is or how (frequently) it changes over time (class label changes), as well as to identify possible inaccuracies in the original data labeling process.

The fifth paper "Real-time Algorithm for Changes Detection in Depth of Anesthesia Signals" by Raquel Sebastião, Margarida M. Silva, Rui Rabiço, João Gama and Teresa Mendonça presents a real-time algorithm for changes detection in depth of anesthesia signals. The algorithm is based on the Page–Hinkley test with a forgetting mechanism (PHT-FM) giving more weight to more recent samples. Experimental evaluation of PHT-FM was performed first offline on historical data collected during general anesthesia, on which the authors studied how to adjust the forgetting mechanism. Then, PH-TFM was embedded in a real-time software and evaluated in the surgery room. The clinician in the operational real-time settings judged the automatically detected changes with PHT-FM as true positives (68 %), false positives (26 %) or false negatives (5 %). These results encourage the authors to suggest the inclusion of the proposed PHT-FM in a real-time decision support system in the clinical practice, since the PHT-FM can effectively alert the clinician for changes in the anesthetic state of the patient, allowing a more prompt action.

Acknowledgments First of all we would like to thank all the authors who contributed to the special issue and all the reviewers who helped to shape this issue. We would also like to thank the editorial office of the *Evolving Systems* journal for their support. The following two funding sources are greatly acknowledged. (1) This research is supported by the Netherlands Organisation for Scientific Research NWO HaCDAIS project. (2) The research leading to these results has received funding from the European Commission within the Marie Curie Indus-try and Academia Partnerships and Pathways (IAPP) programme under Grant Agreement No. 251617.