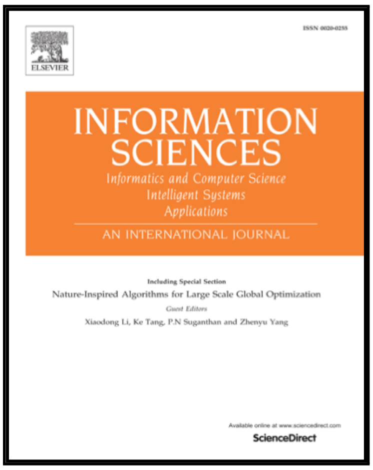
Accepted Manuscript



ISSN 0020-0255

ELSEVIER

INFORMATION SCIENCES

Informatics and Computer Science
Intelligent Systems
Applications

AN INTERNATIONAL JOURNAL

Including Special Section
Nature-Inspired Algorithms for Large Scale Global Optimization
Guest Editors
Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang

Available online at www.sciencedirect.com
ScienceDirect

Knowledge-Maximized Ensemble Algorithm for Different Types of Concept Drift

Siqi Ren, Bo Liao, Wen Zhu, Keqin Li

PII:	S0020-0255(17)31108-8
DOI:	10.1016/j.ins.2017.11.046
Reference:	INS 13274

To appear in:	*Information Sciences*

Received date:	25 June 2016
Revised date:	19 November 2017
Accepted date:	22 November 2017

Please cite this article as: Siqi Ren, Bo Liao, Wen Zhu, Keqin Li, Knowledge-Maximized Ensemble Algorithm for Different Types of Concept Drift, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.11.046

Knowledge-Maximized Ensemble Algorithm for Different Types of Concept Drift

Siqi Ren[a], Bo Liao[a,*], Wen Zhu[a], Keqin Li[a,b]

[a] College of Information Science and Engineering, Hunan University, Changsha 410082, Hunan, China.
[b] Department of Computer Science, State University of New York, New Paltz, New York 12561, USA.

Abstract

Knowledge extraction from data streams has attracted attention in recent years due to its wide range of applications, including sensor networks, web clickstreams, and user interest analysis. Concept drift is one of the most important research topics in data stream mining. Many algorithms that can adapt to concept drift have been proposed. However, most of them specialize in only one type of concept drift and can rarely be used in the environments with a large number of unavailable sample labels. In this study, we propose a new data stream classifier called knowledge-maximized ensemble (KME). First, supervised and unsupervised knowledge are leveraged to detect concept drift, recognize recurrent concepts, and evaluate the weights of ensemble members. Second, the preserved labelled instances in past blocks can be reused to enhance the recognition ability of the candidate member. The final decision for an incoming observation is derived from all the prediction results of the component classifiers. Accordingly, the maximum utilization of the relevant information in a data stream can be achieved, which is critical to models with limited training data. Third, KME can react to multiple types of concept drift by combining the mechanisms of online and chunk-based ensembles. Finally, we compare KME with eight state-of-the-art classifiers on several synthetic and real-world datasets. The comparison demonstrates the effectiveness of KME in various types of concept drift scenarios.

Keywords:
Concept drift, Data stream mining, Ensemble classifier, Unlabelled data.

1. Introduction

1.1. Motivation

In today′s information society, traditional data mining algorithms need to learn from a huge amount of data by means of restricted memory. These algorithms normally require multiple scans of training data, which is unsuitable for mining high-speed data streams [13]. The widespread dissemination of streaming data in many critical real-time tasks has led to a wide range of attention focused on streaming models. Due to the generation speed and the size of data items, it is impossible for streaming models to store the entire observations. Only limited knowledge can be used at each time step, which leads to approximate results. The motivation of this study is to make the results of incremental learning and bath processes as similar as possible by maximizing the usage of relevant knowledge in data streams.

In addition to the overwhelming volumes and high speed, concept drift is an evident characteristic of streaming data. In many real-world applications, the assumption of a fixed data distribution is not truly maintained, thus making most traditional algorithms infeasible. Past observations may become irrelevant or even harmful for the current concept. Therefore, refining or even rebuilding of models is required to remove the obsolete knowledge. Moreover, if the old data are helpful for the current model in the future, then their necessary information should be stored

[*] Corresponding author
Email addresses: siqirenzl@163.com (Siqi Ren), dragonbw@163.com (Bo Liao), syzhuwen@163.com (Wen Zhu), lik@newpaltz.edu (Keqin Li)