International Journal of Agricultural and Environmental Information Systems
Volume 11 • Issue 1 • January-March 2020

Concept Drift Detection in Data Stream Clustering and its Application on Weather Data

Namitha K., Artificial Intelligence and Computer Vision Lab, Department of Computer Science, Cochin University of Science and Technology, Kochi, Kerala, India

Santhosh Kumar G., Artificial Intelligence and Computer Vision Lab, Department of Computer Science, Cochin University of Science and Technology, Kochi, Kerala, India

ABSTRACT

This article presents a stream mining framework to cluster the data stream and monitor its evolution. Even though concept drift is expected to be present in data streams, explicit drift detection is rarely done in stream clustering algorithms. The proposed framework is capable of explicit concept drift detection and cluster evolution analysis. Concept drift is caused by the changes in data distribution over time. Relationship between concept drift and the occurrence of physical events has been studied by applying the framework on the weather data stream. Experiments led to the conclusion that the concept drift accompanied by a change in the number of clusters indicates a significant weather event. This kind of online monitoring and its results can be utilized in weather forecasting systems in various ways. Weather data streams produced by automatic weather stations (AWS) are used to conduct this study.

KEYWORDS

Clustering, Concept Drift, Data Streams, Short-Range Weather Forecasting

INTRODUCTION

With the advancement in hardware and software technology, the number of applications producing large volume data streams is ever increasing. These data streams become useful to the industry and society only when the valuable information contained in them is extracted. Various data stream mining algorithms are available for this purpose. Capability to process the data in a single scan, requirement of online and incremental updation of the models, adaptation to the concept changes etc. are some challenges while designing learning algorithms for data streams. Clustering is an important machine learning task applied on data streams to gain useful insights into the natural groupings in data. It also helps to track the development of various phenomena in application fields like meteorology, healthcare and astrophysics (Bhatnagar, 2009).

The main advantage of data stream mining is that it assumes the data source or the process generating the stream is not stationary. According to the changes in the environment that produces

DOI: 10.4018/IJAEIS.2020010104
This article, originally published under IGI Global's copyright on January 1, 2020 will proceed with publication as an Open Access article starting on February 2, 2021 in the gold Open Access journal, International Journal of Agricultural and Environmental Information Systems (converted to gold Open Access January 1, 2021), and will be distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

the stream, the underlying data distributions change over time. Consequently, these changes might affect the inter relationship between input and output variables (Gama, 2014) leading to 'Concept Drift' (Widmer and Kubat, 1996; Gama, 2014). In these situations, the learned model becomes obsolete and the prediction accuracy reduces considerably. Weather data is a classic example where the concept can change over time (Widmer and Kubat, 1996). Depending on the seasons, weather prediction rules might exhibit a change, i.e., the same relation might not hold well in all the seasons.

This paper proposes a framework for the online clustering of data streams. It performs concept drift detection and cluster evolution monitoring to generate a warning on the dynamic changes taking place in the environment of the stream. Studies revealed that concept drift accompanied by clustering structure changes often imply important physical events. Recording such inter-relationships between the occurrence of concept drift and the corresponding physical events over a period of time can help the prediction of these physical phenomena by cluster analysis.

The proposed framework has components to cluster the stream online, detect concept changes and track the evolution of clusters. Before performing the clustering, the best value for the 'number of clusters', k is computed dynamically. This is done with the intention to identify the changes in the clustering structure for the recently arrived data. As the source of the data is highly dynamic, the clustering structure also might exhibit a corresponding change and fixing the value of k limits the ability to capture such changes in the clustering structure.

The utility of this framework is studied by applying it on weather data. Concept drifts, changes in the best value of k and the cluster evolutions are being monitored to understand their relationship with the physical weather phenomena. Nowadays most of the weather monitoring equipment produces high-speed data with a large number of variables. Weather stream produced by the Automatic Weather Station (AWS) at the Advanced Centre for Atmospheric Radar Research (ACARR) of Cochin University of Science and Technology, Kerala, India is used for this study. Data is collected at one-minute intervals and it contains the weather parameters like temperature, relative humidity, wind speed, wind direction, radiation, pressure, rainfall, etc. Monsoons are the most important weather phenomenon as far as the Indian region is concerned. Hence the framework is used to study the interrelationship between the changes in the clustering structure and the evolution of the south-west monsoon.

The paper is structured as follows. Section 2 refers to the background of this work. Section 3 details the proposed framework with subsections on each component of the framework. Section 4 explains the application of the proposed framework on weather data. Section 5 details about the experiments and results followed by 'discussion and future work' in section 6. The paper is concluded in section 7.

BACKGROUND

Detecting Concept Drift in Data Stream Clustering

Concept drift detection and adaptation are studied more in the context of supervised learning. As (Gama, 2014) states in his survey, the problem of concept drift handling has a much wider scope and it is applicable to clustering problems as well. Research in this direction is still in the starting phase. Surveys conducted on data stream clustering (Ghesmoune et al., 2016; Silva et al., 2013) also point to the fact that explicit concept drift detection and adaptation are rarely done in data stream clustering algorithms.

In supervised learning problems, concept drift can be defined as a change in the joint probability distribution P(X, Y), where X denotes a random variable over vectors of attribute values and Y denotes a random variable over class labels (Webb, 2016). But in the case of unsupervised learning, since the instances are not labelled, the definition of concept drift is modified as a change in the probability distribution P(X). Hence statistical methods of change detection are usually used in data

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/concept-drift-detection-in-data-stream-clustering-and-its-application-on-weather-data/244148

Related Content

Detection and Identification of Microbial Volatile Organic Compounds of the Green Mold Disease: MVOC Profile on Different Media
Dalma Radványi, András Geösel, Zsuzsa Jókai, Péter Fodorand Attila Gere (2020). *International Journal of Agricultural and Environmental Information Systems (pp. 14-28).*
www.irma-international.org/article/detection-and-identification-of-microbial-volatile-organic-compounds-of-the-green-mold-disease/249689

Interaction Data: Definitions, Concepts and Sources
John Stillwell, Adam Dennettand Oliver Duke-Williams (2010). *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications (pp. 1-30).*
www.irma-international.org/chapter/interaction-data-definitions-concepts-sources/42718

Decreasing the Digital Divide by Increasing E-Innovation and E-Readiness Abilities in Agriculture and Rural Areas
Miklós Herdon, Szilvia Botosand László Várallyai (2015). *International Journal of Agricultural and Environmental Information Systems (pp. 1-18).*
www.irma-international.org/article/decreasing-the-digital-divide-by-increasing-e-innovation-and-e-readiness-abilities-in-agriculture-and-rural-areas/120469

Collaborative, Stakeholder-Driven Resource Modeling and Management
Howard Passell, Marissa Reno, Jesse Roach, Vince Tidwelland Wael Khairy (2011). *Handbook of Research on Hydroinformatics: Technologies, Theories and Applications (pp. 36-53).*
www.irma-international.org/chapter/collaborative-stakeholder-driven-resource-modeling/45439

Advanced Oxidation Processes (AOPs) in Landfill Leachate Treatment
Mohamad Anuar Kamaruddin (2019). *Advanced Oxidation Processes (AOPs) in Water and Wastewater Treatment (pp. 355-383).*
www.irma-international.org/chapter/advanced-oxidation-processes-aops-in-landfill-leachate-treatment/209310