ScienceDirect®

Knowledge-Based Systems
Volume 161, 1 December 2018, Pages 205-239

Real-time feature selection technique with concept drift detection using adaptive micro-clusters for data stream mining

Mahmood Shakir Hammoodi [a] [b], Frederic Stahl [a], Atta Badii [a]

[a] Department of Computer Science, University of Reading, PO Box 225, Whiteknights, Reading, RG6 6AY, United Kingdom
[b] Department of Computer Center, University of Babylon, Iraq

Received 5 March 2018, Revised 6 August 2018, Accepted 8 August 2018, Available online 9 August 2018, Version of Record 31 October 2018.

Check for updates

Show less

Share Cite

https://doi.org/10.1016/j.knosys.2018.08.007
Get rights and content

Abstract

Data streams are unbounded, sequential data instances that are generated with high *Velocity*. Classifying sequential data instances is a very challenging problem in machine learning with applications in network intrusion detection, financial markets and applications requiring real-time sensor-networks-based situation assessment. Data stream classification is concerned with the automatic labelling of unseen instances from the stream in real-time. For this the classifier needs to adapt to concept drifts and can only have a single pass through the data if the stream is fast moving. This research paper presents work on a real-time pre-processing technique, in particular feature tracking. The feature tracking technique is designed to improve Data Stream Mining (DSM) classification algorithms by enabling and optimising real-time feature selection. The technique is based on tracking adaptive statistical summaries of the data and class label distributions, known as Micro-Clusters. Currently the technique is able to detect concept drifts and identify which features have been influential in the drift.

Introduction

Velocity in Big Data Analytics [1] refers to data that is generated at ultra high speed and is live-streamed whereupon the processing and storing of it in real-time (Data Stream Mining, DSM) constitute significant challenges to current computational capabilities [2]. Thus data stream mining algorithms that are capable of learning over a single-pass through the training data are necessary. The general area of Data Stream Mining covered by this paper is Data Stream Classification, which is the prediction of class labels of new instances in the data stream in real-time. Potential applications that need real-time data stream classification techniques are for data streams in the chemical process industry [3], intrusion detection in telecommunications [4], etc. In order to keep as high a predictive accuracy as possible, data stream classification techniques need not only be able to learn incrementally but also be able to adapt to concept drifts.

A concept drift occurs if the pattern encoded in the data stream changes. DSM has developed various real-time versions of established predictive data mining algorithms that adapt to concept drift and keep the model accurate over time, such as CVFDT [5] and G-eRules [6]. The benefit of classifier independent concept drift detection methods is that they give information about the dynamics of data generation [7]. Common drift detection methods are for example ADaptive sliding WINdow (ADWIN) [8], Drift Detection Method (DDM) [9] and the Early Drift Detection Method (EDDM)[10]. However, no drift detection method devised to-date, can provide potentially highly valuable insights as to which features are involved in the concept drift. For example, if a feature is contributing to a concept drift it can be assumed that the feature may have become either more or less relevant to the current concept. This causal responsibility theoretic perspective of the evaluation of concept drift has inspired the development of a real-time feature tracking method based on feature contribution information for the purpose of feature selection to identify features that have become (more) relevant or irrelevant due to concept drift. Thus, an approach for detecting causality of drifts, providing the feature contribution information for the purpose of tracking features and identifying the relevant features for classification for the purpose of feature selection in real-time has been developed in this research.

Common feature selection techniques are for example Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA), Multi-View CCA, Principal Component Analysis (PCA) [11], [12], and Support Vector Machine (SVM) based techniques. These techniques can be applied on a sample of the data stream before commencing the training and adaptation of a data stream classifier. However, this would not account for changes in the relevance of features for the classification task at hand due to concept drift which can be dealt with by re-running the above methods to update the feature rankings in order to accommodate any drifts. However, this can potentially be an expensive procedure especially if there are many dimensions in the data, but it also depends on the user settings of how frequently this feature re-ranking is performed. Hence, the rationale for a single-pass method requiring the re-evaluation of only the features where the classification relevance has changed since the last pass.

This research therefore describes a concept drift detection method for data stream classification algorithms with the feature tracking capability. This enables linking features to concept drifts over a statistical sliding window for feature selection purposes. The method only needs to examine features that have potentially changed their relevance and only when there is an indication that the relevance of a feature may have changed. The proposed method can be used with any learning algorithm either as a real-time wrapper or a batch classifier or realised inside a real-time adaptive classifier [6], [13]. Previous work of the authors has developed a feature tracking technique [14], however, the techniques was not used for feature selection purposes as it suffered from over-fitting on noise and outliers. Thus the contributions of this paper are:

1. A new improved concept drift detection technique with feature tracking capabilities.

2. A feature selection technique based upon the causality of drifts obtained through the developed feature tracking method.

This paper is organised as follows: Section2 describes related works, Section3 summarises the MC-NN classifier whose data representation has been used and modified for the real-time feature selection method presented here. Section4 introduces the drift detection method developed in this research and Section5 explains the developed feature tracking method. Section6 takes these developments forward to devise a real-time feature selection approach. Section7 provides an empirical evaluation of the developed methods and concluding remarks are given in Section8.

Section snippets

Concept drift detection techniques

A concept drift occurs if the pattern encoded in the data stream changes over time. The gathered data changes or shifts, after a stability period. Identifying a drift point as distinct from noise or outlier, is the first and most challenging task for drift detection algorithms [7], [15]. Thus analytics algorithms need to adapt. This issue of concept drift needs to be considered in order to mine relevant data with appropriate accuracy. At least four types of drift can be identified; gradual,…

Micro-Cluster Nearest Neighbour (MC-NN)

This section summarises the previously mentioned MC-NN approach [39]. MC-NN was originally developed for predictive data stream analytics, however, the underlying Micro-Cluster structure of MC-NN has been adapted and extended in this research in order to develop a feature tracker for online feature selection purposes. Thus MC-NN is discussed in greater detail. Essentially there are three operations to adapt MC-NN to concept drifts: (1) absorption of data instances into nearest Micro-Clusters,…

Real-time concept drift detection technique using adaptive Micro-Clusters

The MC-NN algorithm aims to keep a recent and accurate summary of the data stream using Micro-Clusters. Significant changes to these summaries are used in this research to detect concept drift.…

Real-time feature tracking technique using adaptive Micro-Clusters

The *Velocity* or *Variance* of a feature can be derived from MC-NN Micro-Clusters and are calculated once a drift is detected as described in Section4. The *Velocity* and *Variance* can then be analysed to identify features that have been involved in the drift. This information can be used for feature selection purposes which will be explained in Section6. The tracking of features is potentially influenced by feature-bias, outlier and noise. Thus our method incorporates approaches to counter these…

Real-time feature selection technique using adaptive Micro-Clusters

Three main tasks will be explained in this section which are feature analysis, feature selection, and monitoring the relevance of selected features. After detection of a concept drift, the statistical information of the features (i.e., *Velocity* and *IQR*) is analysed to identify which features were involved in the drift. Loosely speaking only features that had a significant change of their statistical information are re-examined for feature selection using Information Gain in each statistical…

Experimental evaluation

This section first provides information about the experimental setup and then presents an extensive empirical evaluation of the proposed techniques.…

Discussion and conclusions

This paper has investigated the problem of real-time feature selection. At present the focus of data stream mining lies in the development of data mining algorithms rather than on pre-processing methods. Thus at present there are no developments for truly real-time feature selection given a data streaming input space. This is important as features may potentially change their relevance for data mining tasks based on certain measures of relevance such as Information Gain. Thus the three…

References (55)

D. Marrón *et al.*

Data stream classification using random feature functions and novel method combinations

J. Syst. Softw. (2017)

C. Leys *et al.*

Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median

J. Exp. Soc. Psychol. (2013)

C.C. Aggarwal *et al.*

A framework for projected clustering of high dimensional data streams

Proceedings of the Thirtieth international conference on Very large data bases-Volume 30(2004)

C.C. Aggarwal *et al.*

A framework for clustering evolving data streams

Proceedings of the 29th international conference on Very large data bases-Volume 29(2003)

G.J. Ross *et al.*

Exponentially weighted moving average charts for detecting concept drift

Pattern Recognit. Lett. (2012)

P. Kadlec *et al.*

Data-driven soft sensors in the process industry

Comp. Chem. Eng. (2009)

M. Ebbers *et al.*

Addressing data volume, velocity, and variety with IBM infosphere streams V3. 0

(2013)

B. Babcock *et al.*

Models and issues in data stream systems

Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (2002)

A. Jadhav *et al.*

A novel approach for the design of network intrusion detection system (NIDS)

Sensor Network Security Technology and Privacy Communication System (SNS & PCS), 2013 International Conference on (2013)

G. Hulten *et al.*

Mining time-changing data streams

Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (2001)

View more references

Cited by (21)

A hybrid deep learning classifier and Optimized Key Windowing approach for drift detection and adaption

2023, Decision Analytics Journal

Show abstract

Non-IID data and Continual Learning processes in Federated Learning: A long road ahead

2022, Information Fusion

> *Citation Excerpt :*
>
> …The algorithms that implement a response to these drifts will be reviewed in Section 5. Many concept drift detection strategies have been proposed to attach the situations of virtual and real drifts [8,125–133]. These approaches are often classified as Data Distribution-based or Error Rate-based methods correspondingly.…

Show abstract

Feature selection for online streaming high-dimensional data: A state-of-the-art review

2022, Applied Soft Computing

> *Citation Excerpt :*
>
> …Concept drift behaviours require investigation in terms of both features and instances as it is considered another branch of OFS that has not been explored in depth [58,59]. Generally, this is addressed as a feature drift issue, in which the data stream is prone to change in terms of feature characteristics or in which disparities exist in feature importance, such that a feature that has been important may later become meaningless, or vice versa [60]. One common example of data with concept drift and feature drift behaviour is text streams [61].…

Show abstract

Prosumer in smart grids based on intelligent edge computing: A review on Artificial Intelligence Scheduling Techniques

2022, Ain Shams Engineering Journal

> *Citation Excerpt :*
>
> …MDC contributes to the EC's ultimate goal and can be seen as "how" the EC tends to function [115]. The authors in [116] reported that MDCs near Access Points (AP) could reduce cloud and congestion costs. In [117], the authors considered

> MDCs to be cloudlet and sometimes designed to meet the needs of low-bandwidth, battery life limits or latency.…

Show abstract

Discovering three-dimensional patterns in real-time from data streams: An online triclustering approach

2021, Information Sciences

> *Citation Excerpt :*
>
> …Afterwards, the model has to evolve to include knowledge from the new stream sample. In traditional clustering techniques applied to streaming, as for example K-Means, a common strategy is to add the new instance to its nearest "centroid" if it is within the boundary of the cluster [42]. The STriGen algorithm computes different updating options and selects the one with the highest GRQ value (Eq. 1), that corresponds to the best graphical quality of each tricluster.…

Show abstract

An adaptive algorithm for dealing with data stream evolution and singularity

2021, Information Sciences

> *Citation Excerpt :*
>
> …Mining the hidden information in data streams is a topic of theoretical interest and also of great relevance to applications [1,2]. However, the adaptation of most data mining models is gradually declining for data streams, which are characterized by unknown, continuous and high speed characteristics [3]. Meanwhile, the characteristics of data streams often change as time passes, i.e., there is a so-called concept drift [4] or a singularity also exists in a dynamic environment [5,6].…

Show abstract

View all citing articles on Scopus

Recommended articles (6)

Research article

Short-term traffic volume prediction by ensemble learning in concept drifting environments

Knowledge-Based Systems, Volume 164, 2019, pp. 213-225

Show abstract

Research article

Concept drift detection based on Fisher's Exact test

Information Sciences, Volumes 442–443, 2018, pp. 220-234

Show abstract

Research article

Scalable real-time classification of data streams with concept drift

Future Generation Computer Systems, Volume 75, 2017, pp. 187-199

Show abstract

Research article

A buffer-based online clustering for evolving data stream

Information Sciences, Volume 489, 2019, pp. 113-135

Show abstract

Research article

Handling adversarial concept drift in streaming data

Expert Systems with Applications, Volume 97, 2018, pp. 18-40

Show abstract

Research article

Merit-guided dynamic feature selection filter for data streams

Expert Systems with Applications, Volume 116, 2019, pp. 227-242

Show abstract

View full text