International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

A study on concept drift detection algorithms for real-world data streams

Abdul Razak M S [1], Naseer R [1], Sreenivasa B R [2,*] and Nirmala C R [1]

[1] *Department of CSE, Bapuji Institute of Engineering & Technology, Davangere, India.*
[2] *Department of ISE, Bapuji Institute of Engineering & Technology, Davangere, India.*

International Journal of Science and Research Archive, 2024, 11(02), 1301–1305

Publication history: Received on 27 February 2024; revised on 06 April 2024; accepted on 09 April 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.11.2.0593

Abstract

The arrival of data has changed in today's digital environment, becoming more dynamic. Dynamic data is characterized by its speed, variety, and infinite size. Data streams are one category of dynamic data. To address the issues with data streams, several strategies and AI models were developed. One such problem is concept drift, which results from changes in the data's distribution and eventually lowers the performance of the AI model. This means that regular updates to the model are required. In our work, we will analyse the performance using evaluation metrics and compare the effectiveness of the current error-based methods with window-based methods for real-world datasets.

Keywords: Concept Drift; Data Stream; Drift Algorithms; Classification; Prediction Error.

1. Introduction

Data is information and facts about the things we do every day. The amount of data in today's digital world is growing. Various sources such as banks, stock markets, hospitals, credit card transactions, smartphones, weather, sensor data, social media, and many more generate quintillion bytes of data every period [1]. Improved business can result from the correct storage and analysis of large amounts of data that can assist several enterprises. In the modern digital world, data no longer arrives in a static form but rather as a stream. Data streams, in contrast to static data, cannot be saved and examined repeatedly. The data stream needs to be processed right away in one pass.

The field of knowledge discovery and data mining, or KDD, focuses on methods for drawing valuable insights from data. There is an enormous demand for KDD techniques due to the internet's ever-growing volume of online data and databases' widespread use. The goal of KDD is to retrieve predictive information that is hidden from large databases. Businesses can make proactive and informed decisions by using data mining techniques that forecast future events and patterns [2].

Concept drift, a problem where the target concept changes with time, usually after a minimal stability period [3], is another challenge with data streams. Over the past few years, concept drift has drawn a lot of attention, primarily because it hurts the performance of classifiers that are trained on historical data.

In our work, we used the Pearson correlation technique to pick features from streaming data blocks and compared them to a comprehensive feature list of streaming data blocks for the human activity recognition dataset (HAR). The suggested method produces encouraging results in terms of accuracy (94.85%) throughout the entire feature list [4].

The article provides information on the relevant experimental data, an analysis of the results, and a comparison of concept drift detection strategies, including error-based and window-based methods. More accurately, accuracy and detections are assessed for four different drift detector configurations in a fully labelled context (supervised learning).

* Corresponding author: Sreenivasa B R

The most efficient methods now in use are suggested by the study's findings. A summary and elaboration of the experiments are given.

2. Literature Review

The strategies for concept drift identification that are based on performance are examined here. These strategies can be classified into multiple groups based on the mechanism utilized to detect performance declines.

2.1. Error-based Methods

To assess the effectiveness of the learning process, it monitors the base learners' online error rate progression. Concept drift is presumed if the model's performance is below the significance test threshold.

Example: DDM, EDDM

- Drift Detection Method (DDM): DDM looks for variations in the classifier's error rate [5]. It views error as a binomial-distributed Bernoulli random variable. It tracks Pt, the standard deviation St across time, and the misclassification probability at time t.

$$\text{Warning level} = P_i + S_i \geq P_{min} + 2 * S_{min}$$
$$\text{Drift level} = P_i + S_i \geq P_{min} + 3 * S_{min}$$

- Early Drift Detection Method (EDDM): A method for identifying concept drift that is compatible with gradual modifications [6]. This method employs the distance between them (number of examples between two classification errors) instead of classification errors. Maintaining a fixed distance stops drift. It determines the average separation between two errors using the standard deviation (St) and probability of misclassification (Pt).

$$\text{Warning level} = (P'_i + 2\,S'_i) / (P'_{max} + 2 * S'_{max}) < \alpha$$
$$\text{Drift Level} = (P'_i + 2\,S'_i) / (P'_{max} + 2 * S'_{max}) < \beta$$
$$\alpha, \beta = \text{thresholds} = 0.95, 0.90$$
$$P_i' = \text{The mean separation between two errors}$$
$$S_i' = \text{Standard Deviation}$$

2.2. Window-based Methods

Incoming data instances (or a window) are grouped together using this approach. Two windows are typically present in window-based systems. New instances are added later, with data stream instances remaining in the initial frame. The shift and the disparity in the data distribution were both made clear by comparing these two windows. You can alter or fix the size of the window.

Example: ADWIN, STEPD

- ADWIN: To detect concept drift, adaptive windowing [7] looks at the distribution between two windows. The mean error rate of each partition is compared to a threshold based on Hoeffding. A sub-partition loses its window value if it goes over this limit. Remove the final component if

$$|\mu_0 - \mu_1| > \theta_{hoeffding}$$

μ_0 = W0's error rate

μ_1 = W_1's error rate

- STEPD: Evaluating both general and present accuracy is a fundamental principle [8]. We make two assumptions: first, the accuracy of a classifier for the last W examples is equal to its total accuracy from the start of learning; second, a significantly lower recent accuracy indicates that the concept is changing. Compiling the following statistic is how the test is run.

$$T(r_0, r_r, n_0, n_r) = \frac{\left|\frac{r_0}{n_0} - \frac{r_r}{n_r}\right| - 0.5(\frac{1}{n_0} + \frac{1}{n_r})}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{(1/n_0 + 1/n_r)}}$$

Other existing algorithms for concept drift detection include (PHT) [9], HDDM [10,11], and other statistical based techniques like WSTD [12], FTDD [13], CSDD [14], MDDM [15], KS-test [16].

3. Datasets and Description

The popular dataset that was examined and characterized by Gama and M. Harries is electricity. The Australian New South Wales Electricity Market provided the data for this analysis. In this market, supply and demand have an impact on prices, which are not fixed. They have a five-minute interval set. 45, 312 occurrences can be found in the ELEC dataset with 8 features and 2 class labels. The price change about the last 24-hour moving average is indicated by the class designation.

The Airlines Dataset drew from Elena Ikonomovska's regression dataset. It contains 7 features and 2 class labels with 53000 samples. Using the information about the scheduled departure, the aim is to determine if a certain flight will be delayed.

4. Experimental Set Up

All pertinent information about the experiments described in this article is provided in this section. As the most often used classifiers in the field and because their implementations are available within the MOA framework [17], all of the concept drift detection approaches have been tested using both Naive Bayes (NB) and Hoeffding Tree (HT) as base learners.

The following default settings are used in MOA during execution:

- DDM (min instances = 30, warning level = 2 and out control level = 3)
- ADWIN (delta adwin = 0.002)
- STEPD (Window Size = 30, alpha drift = 0.003, alpha warning = 0.005)

Lastly, the Prequential methodology—the default in MOA—was used for the accuracy evaluation, and a sliding window of size 1,000 served as the forgetting mechanism. According to this methodology, each incoming instance is first utilized for training and then testing. The accuracy of the system is determined by adding up all of the sequential errors over time, or the loss function that separates the observed values from the predictions. The preceding 1,000 occurrences are the portion of the data that is taken into account in each calculation in its sliding windows variation.

5. Results and Discussions

The following Table 1 describes the results of the airline dataset using error-based algorithms (DDM, EDDM) and window-based algorithms (ADWIN, STEPD) for the above experimental settings. We can derive that the window-based approach algorithm ADWIN is hyperactive in notifying about changes in the data. ADWIN has detected maximum changes for the airline dataset i.e. 254 and the model trained on ADWIN has a prediction error rate of 0.44. From this, we can conclude the window-based algorithm is well suited for Airline datasets over other error rate-based algorithms.

Table 1 Results of Error-based and Window based algorithms on Airline Dataset

Dataset	Algorithm	Detected Changes	Detected Warnings	True Changes	Prediction error
Airline	DDM	100	147	100	0.45
Airline	EDDM	247	90	100	1.39
Airline	ADWIN	254	0	100	0.44
Airline	STEPD	210	113	100	0.49

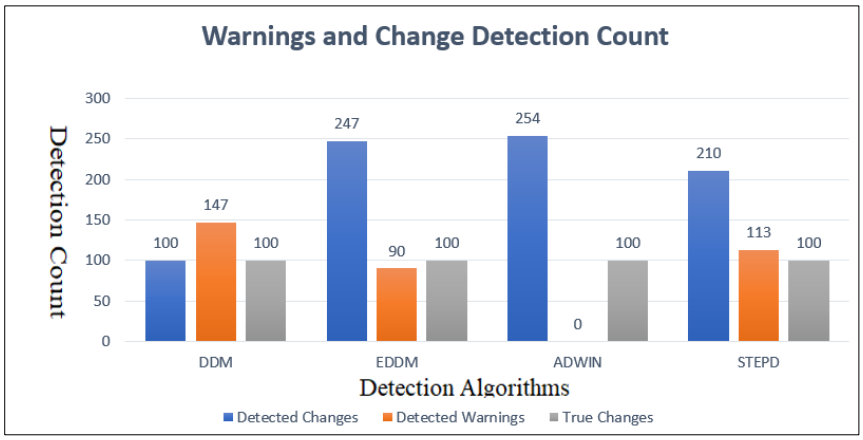

Figure 1 Error-based and Window based algorithms on Airline Dataset

The results of the electrical dataset utilizing window-based algorithms (ADWIN, STEPD) and error-based methods (DDM, EDDM) for the aforementioned experimental settings are shown in Table 2. It may be inferred that the window-based approach algorithm ADWIN is extremely proactive in informing users of any modifications to the data. Concerning the electricity dataset, ADWIN has identified the most changes 231, and the model trained on ADWIN has a prediction error rate of 0.44. This leads us to the conclusion that, in comparison to other error rate-based methods, the window-based technique works well even with electricity datasets.

Table 2 Results of Error-based and Window based algorithms on Electricity Dataset

Dataset	Algorithm	Detected Changes	Detected Warnings	True Changes	Prediction error
Electricity	DDM	90	133	90	0.45
Electricity	EDDM	223	81	90	1.41
Electricity	ADWIN	231	0	90	0.44
Electricity	STEPD	189	102	90	0.49

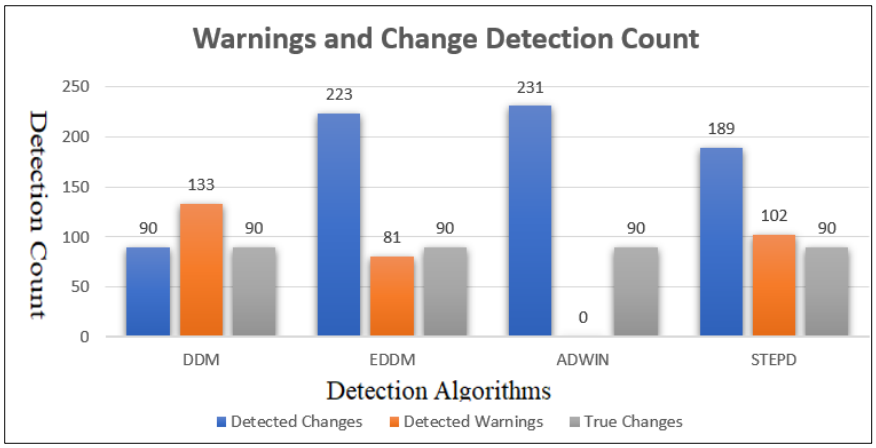

Figure 2 Error-based and Window based algorithms on Electricity Dataset

6. Conclusion

The article dives into a head-to-head comparison of two main categories of concept drift detection methods: error-based and window-based. This evaluation is conducted using real-world data, not simulated datasets. To make the comparison more robust, the researchers employ two different machine learning algorithms for classification - Naive Bayes and Hoeffding Tree. The ultimate goal is to identify which concept drift detection method performs best, considering both

how well it predicts outcomes (prediction error) and how effectively it identifies changes in the data (change detections).

Compliance with ethical standards

Acknowledgment

The authors would like to thank the Bapuji Institute of Engineering and Technology for providing the infrastructure required to carry out the research work.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] R. Devakunchari Analysis on big data over the years International Journal of Scientific and Research Publications, Volume 4, Issue 1, January 2014.

[2] Gama, J, 2010, Knowledge Discovery from Data Streams. CRC Press, Boca Raton, USA.

[3] Alexey Tsymbal 'The Problem of Concept Drift: Definitions and Related Work', 2004

[4] M. S. AR, Nirmala CR, Aljohani M, Sreenivasa BR, ' A novel technique for detecting sudden concept drift in healthcare data using multi-linear artificial intelligence techniques'. Frontiers in Artificial Intelligence 5, 2022. https://doi.org/10.3389/frai.2022.950659

[5] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, 'Learning with drift detection', in Proc. 17th Brazilian Symp. Artificial Intelligence, ser. Lecture Notes in Computer Science. Springer, pp. 286–295,2004

[6] M. Baena-Garc´ıa, J. del Campo-Avila, R. Fidalgo, A. Bifet, ´ R. Gavalda, and R. Morales-Bueno, 'Early drift detection 'method',in Proc. 4th Int. Workshop Knowledge Discovery from Data Streams, 2006.

[7] Bifet and R. Gavalda, 'Learning from time-changing data with adaptive windowing', Proceedings of the Seventh SIAM International Conference on Data Mining, vol. 7, 2007.

[8] Nishida, K., Yamauchi, K. 'Detecting Concept Drift Using Statistical Testing', Proceedings Discovery Science, 10th International Conference, DS, vol 4755, October 1-4,2007.

[9] A. Qahtan, B. Alharbi, S. Wang, and X. Zhang, 'A PCA-based change detection framework for multidimensional data streams', in Proc. 21th Int. Conf. on Knowledge Discovery and Data Mining. ACM, Conference Proceedings, pp. 935–944,2015.

[10] R. M. S, N. C. R, C. B. B, M. Rafi and S. B. R, Online feature Selection using Pearson Correlation Technique, *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, MANGALORE, India, 2022, pp. 172-177, doi: 10.1109/ICRAIE56454.2022.10054267.

[11] Fr´ıas-Blanco, J. d. Campo-Avila, G. Ramos-Jim´enez, R. Morales-Bueno, A. ´ Ortiz-D´ıaz and Y. Caballero-Mota, 'Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds',in IEEE Transactions on Knowledge and Data Engineering , vol. 27, no. 3, pp. 810-823, 1 March 2015.

[12] Roberto Souto Maior de Barros, Juan Isidro Gonz´alez Hidalgo, Danilo Rafael de Lima Cabral, 'Wilcoxon Rank Sum Test Drift Detector',Neurocomputing , Volume 275, Pages 1954-1963, ISSN 0925-2312, 2018.

[13] Danilo Rafael de Lima Cabral, Roberto Souto Maior de Barros, 'Concept drift detection based on Fisher's Exact test', Information Sciences, Volumes 442–443, Pages 220-234, ISSN 0020-0255, 2018.

[14] Hidalgo, J.I.G., Mari˜no, L.M.P., de Barros, R.S.M. 'Cosine Similarity Drift Detector', Artificial Neural Networks and Machine Learning – ICANN: Text and Time Series. Lecture Notes in Computer Science, vol 11730, Springer, Cham. 2019.

[15] Ali Pesaranghader, Herna Viktor, Eric Paquet 'McDiarmid Drift Detection Methods for Evolving Data Streams', International Joint Conference on Neural Networks (IJCNN)), 2018.

[16] Wang, Z., Wang, W. 'Concept Drift Detection Based on Kolmogorov–Smirnov Test', Artificial Intelligence in China. Lecture Notes in Electrical Engineering, vol 572. Springer, Singapore 2020.

[17] Albert Bifet, Geoff Holmes, Richard Kirkby, Bernhard Pfahringer (2010); MOA: Massive Online Analysis; Journal of Machine Learning Research 11: 1601-1604