International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 5th Jan 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 01

DOI: 10.48047/IJIEMR/V12/ISSUE 01/21

Title **A Contemporary Review of Literature on Concept Drift Detection in Data Stream Mining**

Volume 12, ISSUE 01, Pages: 212-220

Paper Authors
Gollanapalli V Prasad, Dr Kapil Sharma

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A Contemporary Review of Literature on Concept Drift Detection in Data Stream Mining

Gollanapalli V Prasad[1], Dr Kapil Sharma[2]
[1]Research Scholar in Computer Science and Engineering, Amity University, Gwalior, M.P., India and Senior Assistant Professor in Computer Science and Engineering at CVR College of Engineering, Affiliated to JNTU Hyderabad, India
Email: gvenkataprasad81@cvr.ac.in / prasad.venkata8@gmail.com
[2] Professor in Computer Science and Engineering, Amity University, Gwalior, M.P., India, Email: ksharma@gwa.amity.edu

Abstract
The streams come from various sources, at varying speeds and volumes, and flow into a single, continuous, combined stream. Predicting variations in the underlying distribution of streaming data over time is referred to as concept drift. With more and more data being organized as data streams rather than static databases, the concept drift problem is becoming more and more important in machine learning and data mining. The approaches for learning about concept drift have noticeably become more systematic as a result of the rapid development of "concept drift" in recent years. Concept drift detection, concept drift understanding, and concept drift adaptation are the three primary parts of concept drift learning.

Keywords*:* Drift detectors; ensemble classifiers; data stream mining.

Introduction

The pervasive influence of Digital Era communication impacts the generation and capture of value knowledge-based society. Data, Information, and Knowledge have a significant role in every sphere of human activity, and these are vital parameters in any decision-making process. Data mining is a knowledge discovery process that analyzes high volumes of data from several perspectives and summarizes it into useful information.

With the advancements in information and communication technologies, extensive usage of social media, and paradigm shift in digital transfers the concept of data streaming came into existence (Giuseppe Aceto, 2018). The is data generated by an infinite amount of internet sources,

hardware sensors, servers, mobile devices, applications, web, browsers, and an increase in user availability and accessibility (Bhavani, S., Subhash Chandra, N. (2022).

If such data can be analysed promptly, the data streams can be a source of significant qualitative data. Stream data analysis concurrently challenges include the limitless data, varying speed, and variety of data, over the past 10 years, this field of research has received a lot of attention since it deals with the not related data properties of coming occurrences from a data stream. (Scott Wares, 2019). Online data processing techniques need to address more accurately in real-time is another dimension to compete with the hardware and cutting-edge algorithmic solutions (AhmedQussous, 2018). However, incremental predictive models were developed to address these problems partially. The varying capacity of data characteristics in a continuously changing domain or time creates another inherent issue of concept drift (Indre Zliobaite, 2018). The proposed research focuses to address a novel solution for data stream mining challenges and concept drift detection algorithms in unknown characteristics data.

Traditional machine learning algorithms are seldom applicable in eventualities with streaming knowledge. Most algorithms were designed for offline settings, i.e., the whole knowledge set has to be scanned and processed (multiple times), before a choice is created (Soppari, K., Chandra, N.S.,2022)

Adaptive machine learning algorithms will analyze the data streams continuously. Advancements in the state-of-the-art algorithms effectively improve the predictive models through drift (B.Ramakrishna,2018)

Concept Drift in Data Streaming

Identifying the change in data distribution is the major concern for stream data mining techniques. Concept drift is the term used to describe the shift in the distribution of data coming from the data stream. It takes place over time, during which the drifts may alter. The following four types of drifts exist:

1. **Sudden Drift** or Abrupt Drift results from a fulminate modification within the knowledge distribution. It takes place once data is suddenly replaced by another concept.
2. **Incremental Drift** or Stepwise drift consists of a sequence of tiny changes. It is often known solely over an

associated extended amount of your time, as a result of tiny changes over time.

3. **Gradual Drift** results from a slow transition from one knowledge distribution to consecutive. That is, the 2 patterns might be at the same time. It's characterized by a transitioning window wherever instances from the new construct become predominant and instances from the previous construct a less frequent.
4. **Recurring Drift** refers to the case once an antecedent construct reappeared when it slowed or it happens whenever ideas keep continual each thus usually or willy-nilly. The return of drifts can be cyclic.

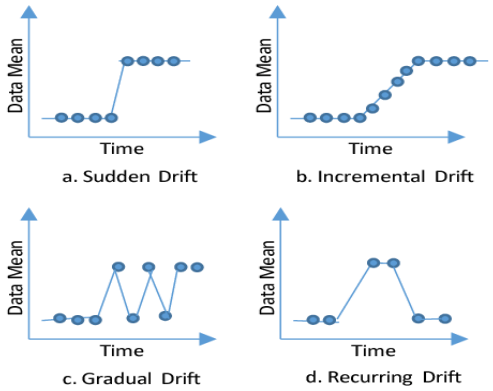


Fig 1. Types of Concept Drifts

The underlying assumption behind how traditional machine learning algorithms work is that the information distribution is static. Due to the inherent temporal structure of knowledge streams, the distribution of internal instances may change over time. Due to this, outdated batch-learning algorithms are inappropriate for use in applications that learn from knowledge streams.

2. Review of Literature

The process of analyzing the hidden patterns data into important information is most important for security applications and business expansion. The data collected and kept in data warehouses are used for data analysis through advanced algorithms. Data-stream mining systems should conjointly manage to miss and corrupt data—noisy communication lines, human error, experimental style, and failing sensors will all alter and interrupt knowledge streams. In online learning systems, each observation and response may be missing or corrupted at any time. wheezy and missing observations are the topic of intensive analysis. Observation noise is expressly sculptural by learning procedures, and numerous imputation techniques are planned for handling missing values.

Classification algorithms for Stream data :

We all know data stream is reported very fast and also has huge size. There are

different varieties of algorithms to store and streaming methods used for training the systems.

A variety of algorithmic rules for the classification of the fixed dataset, however, these techniques are not fit for streaming-data.

Table 1 Classification Algorithms Used On Streaming Data

SNo	AuthorName & Year	Algorithm	Limitation
1	C.Berkaman, and Jeffery A,1997. "Decision Tree Induction based on Efficient Tree Restructuring".	Incremental Induction (ITI)	Requires huge storage but is not proper for a huge knowledge Stream–Tree-based.
2	Domingo's, Pedro, and Geoff Hulten,2000 "Mining High-Speed Data Streams"	Very Fast Decision Tree(VFDT)	Deep tree growth (skewed growth)
3	Janardan, Dr.Shikha Mehta, 2017 "Concept drift in Streaming Data Classification: Algorithms, Platforms, and Issues"	Concept Adapting very fast decision tree(CVFDT)	Encroachment of VFDT with conversion for idea drift
4	Street,W. Nick, and YongSeog Kim-2001 "A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification"	Ensemble Streaming Algorithms (ESA)	It will hold the idea drift however not sensible with high-speed knowledge streams.
5	Cohen, Lior, Gil Avrahami, and Mark Last(2004) "Incremental Info-Fuzzy Algorithm for Real-Time Data Mining of Non-Stationary Data Streams".	On-Line Information Network (OLIN)	Needs less significant space and uses the info. fuzzy network for concept drift adaptation-tree-based.
6	Wang, Haixun, et al 2003 **"Mining Concept-Drifting Data Streams"**	Weighted Classifier Ensemble (WCE)	Deals with idea drift by victimization grouping of the weighted classifier.
7	Aggarwal, CharuC., et al,2004 **"On Demand Classification of Data Streams"**	One order Classifier	Dynamic or changeable window size for higher order classification is missing.
8	Gama, Joao, Pedro Medas, and Ricardo Rocha(2004) "Forest trees for on-line data"	Ultra-Fast Forest Tree System (UFFT)	Supports binary tree classification model
9	Law, Yan-Nei, and Carlo Zaniolo,2005 "An Adaptive Nearest Neighbor Classification Algorithm for Data Streams"	The progressive	Progressive rule for adaptive learing with low value through the Nearest neighbor method.
10	Cohen, Lior, et al,2008 "Info-fuzzy algorithms for mining dynamic data streams"	Incremental Online-Information Network (IOLIN)	The tree-based technique has skewed generation.
11	Bifet, Albert, et al, 2009 "A Survey on Concept Drift Adaptation"	ADWIN Bagging, Adaptive-Size Hoeffding Tree(ASHT)	Employs ADWIN rule to sight changes, moreover estimating the load through the reinforcing method.

12	Abdulsalam, Hanady, David B. Skillicorn, and Patrick Martin,(2011) ”classification Using Streaming Random Forests”	Random Forest based frequently classifier	Handles changing knowledge streams with irregular tagged knowledge case advents in unit step. Routines Entropy to sight idea Drift-Tree based.
13	Prasad, **B**akshiRohit, and SonaliAgarwal, (2016) “Critical parameter analysis of Vertical Hoeffding Tree for optimized performance using SAMOA”	Vertical Hoeffding tree(VHT)	A dissimilarity of VFDT that performs strewn similar intended by columns divided knowledge sets-Tree-based.
14	Wang, Lei, Hong-Bing Ji, and Yu Jin,2013 “Fuzzy Passive-Aggressive classification: A robust and efficient algorithm for online classification problems”	Uncertain Passive-aggressive cataloging	A complete unique stra-line association generates, an appropriate for repetition with an unavailable outlier in online classification issues.
15	Mena-Torres, Dayrelis, and Jesús S. Aguilar-Ruiz,2014 **“A similarity-based approach for data stream classification”**	Similarity-based mostly on know Similarity-basedsifier (SimC)	Uses new addition /deletion approaches for rapidly taking and on behalf of modifications in knowledge to enhance attainment-Rule based.
16	Brzezinski, Dariusz, and Jerzy Stefanowski, 2014 **“Reacting to different types of concept drift: the accuracy Updated Ensemble algorithm”**	Prequential AUC based mostly classifier	It works higher with extremely unbalanced knowledge streams-Rule based.
17	Loo, HuiRu, and Muhammad N. Marsono,2015 **“Online Data Stream Learning and Classification with Limited Labels”**	Incremental partial supervised learning is implemented in the Online stream classifier.	Utilizes the selective self-training-based semi-supervised learning approach.
18	Ángel, Abad Miguel, Gomes Joao Bartolo, and Menasalvas Ernestina,2016 **“Predicting recurring concepts on data-streams by means of a meta-model and a fusingrity function”**	Classify the recursive hypothesis using the fuzzy likeness method.	Applied the rule-based learning method.
19	Jędrzejowicz, Joanna, and Piotr Jędrzejowicz, 2015 **“Concept drift in Streaming Data Classification: Algorithms, Platforms, and Issues”.**	Distance-based collection of the live Classifier using kernel clustering	A collection of classifiers is built on the idea of a portfolio of distance life. Ensemble Technique used for classification.

20	Zeng Li, Yan Xiong, Wenchao Huang,2020 "Drift-detection Based Incremental Ensemble for Reacting to Different Kinds of Concept Drift".	Drift-Detection-Based Incremental Ensemble (Die) Algorithm.	This It can't perform with imbalanced data streams.
21	Osama A. Mahdi (2020) Fast Reaction to Sudden Concept Drift in the Absence of Class Labels	Diversity Measure as a Drift Detection Method (DMDDM)	The current method may fail to measure the differences between classifiers that incorrectly predict the same instance using different labels for multiclass cataloging problems.
22	Mashail Althabiti and Manal Abdullah (2019) Streaming Data Classification with Concept Drift.	Data Stream Mining(DSM)	DSM components including the I/O, estimation methods, and classification algorithms with concept drift have been presented.
23	B. Ramakrishna (2018) Attribute Pattern Weights(APW): A scale to detect concept drift in Data Stream Mining Models	Attribute Pattern Weights (APW)	These results depict that in the future there's a ton of scope for evaluation, to attain excessive potency of likelihood, and machine learning-based totally category over streaming data continues to exist as open analysis.

Classification Algorithms

A classification algorithm depicts the streaming of data, using rule or tree-based algorithms. By exploiting the nearest neighborhood and applying mathematical approaches algorithms few were developed. These results describe that in the future there's a ton of possibility for analysis, to achieve high potency of likelihood, based on machine learning classification over streaming data continues to exist as an open analysis.

- The below algorithms are suitable for finding concept drift.

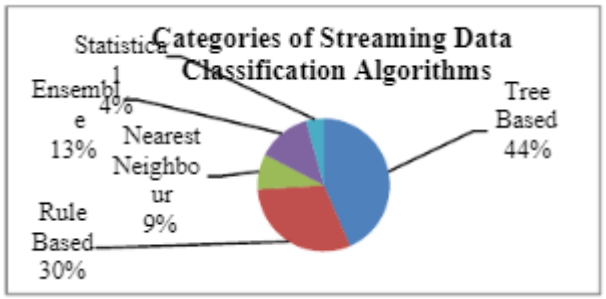


Fig.2 Classification methods used on streaming data

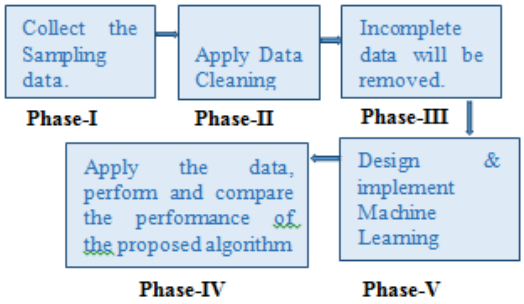
The proposed research needs to address the performance issues in data stream mining. The evaluation parameters play a vital role in the assessment of the proposed algorithm.

In data stream mining, accurate computation is done incrementally by holding checks set for every example-(Gama, J, 2010). Accuracy inherits the weakness of ancient exactness, i.e., the

discrepancy with relevance category distribution and promoting majority category predictions.

framework for proposal

To attain the objectives of future research, the following phases are to be completed.



Conclusion

- A novel machine learning algorithm for drift detection with high efficiency.
- Incorporating temporal dependency and other data anomalies into drift detection algorithms.
- To reduce dependency over time and improve accurate feedback.

References

1. Janardan, Shikha Mehta (2017)Concept drift in Streaming Data Classification: Algorithms, Platforms and Issues. Volume 122, Pages 804-811
2. Prasad, BakshiRohit, and SonaliAgarwal (2016). Stream Data Mining: Platforms, Algorithms, Performance Evaluators and Research Trends, International Journal of Database Theory and Application 9.9: 201-218.
3. Utgoff, Paul E., Neil C. Berkman, and Jeffery A (1997). Clouse. Decision tree induction based on efficient tree restructuring, Machine Learning 29.1: 5-44.
4. Domingos , Pedro, and Geoff Hulten (2000). Mining high-speed data streams, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
5. Street, W. Nick, and YongSeog Kim (2001). A streaming ensemble algorithm (SEA) for large-scale classification, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
6. Cohen, Lior, Gil Avrahami, and Mark Last(2004). Incremental info-fuzzy algorithm for real-time data mining of non-stationary data streams, TDM Workshop, Brighton UK. Vol. 43.
7. Wang, Haixun, et al (2003). Mining concept-drifting data streams using ensemble classifiers, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
8. Aggarwal, CharuC., et al (2004). On demand classification of data streams, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
9. Gama, Joao, Pedro Medas, and Ricardo Rocha (2004). Forest trees for on-line data, Proceedings of the 2004 ACM symposium on Applied computing. ACM.
10. Law, Yan-Nei, and Carlo Zaniolo (2005). An adaptive nearest neighbor classification algorithm for data streams, European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg.
11. Giuseppe Aceto(2018). The role of Information and Communication Technologies in

Healthcare: Taxonomies, Perspectives, and Challenges, 2018, Pages 125-154 Scott Wares(2019), Data stream mining: methods and challenges for handling concept drift, Article number :1412 (2019)

12. Indre Zliobaite(2018), Concept drift over geological times: predictive modeling baselines for analyzing the mammalian fossil record,773–803(2019)
13. Ahmed Qusssous(2018), Big Data technologies: A survey, Volume 30, Issue 4, October 2018, Pages 431-448
14. Cohen, Lior, et al (2008). Real-time data mining of non-stationary data streams from sensor networks, Information Fusion 9.3: 344-353.
15. Bifet, Albert, et al (2009). New ensemble methods for evolving data streams, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
16. Abdulsalam, Hanady, David B. Skillicorn, and Patrick Martin (2011). Classification using streaming random forests, IEEE Transactions on Knowledge and Data Engineering 23.1: 22-36.
17. Wang, Lei, Hong-Bing Ji, and Yu Jin (2013). Fuzzy Passive–Aggressive classification: A robust and efficient algorithm for online classification problems, Information Sciences 220: 46-63.
18. Mena-Torres, Dayrelis, and Jesús S. Aguilar-Ruiz (2014). A similarity-based approach for data stream classification, Expert Systems with Applications 41.9: 4224-4234.
19. Brzezinski, Dariusz, and Jerzy Stefanowski (2014). Prequential AUC for classifier evaluation and drift detection in evolving data streams, International Workshop on New Frontiers in Mining Complex Patterns. Springer International Publishing.
20. Loo, HuiRu, and Muhammad N. Marsono (2015). Online data stream classification with incremental semi-supervised learning, Proceedings of the Second ACM IKDD Conference on Data Sciences. ACM.
21. Jędrzejowicz, Joanna, and Piotr Jędrzejowicz (2015). Distance-based ensemble online classifier with kernel clustering, Intelligent Decision Technologies. Springer International Publishing 279-289.
22. Krawczyk, Bartosz, and MichałWoźniak (2015). One-class classifiers with incremental learning and forgetting for data streams with concept drift, Soft Computing 19.12: 3387-3400.
23. Ángel, Abad Miguel, Gomes Joao Bartolo, and Menasalvas Ernestina (2016). Predicting recurring concepts on data-streams by means of a meta-model and a fuzzy similarity function, Expert Systems with Applications 46: 87-105.
24. Street, W. Nick, and Yong Seog Kim (2001). A streaming ensemble algorithm (SEA) for large-scale classification, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
25. Gama, J (2010). Knowledge Discovery from Data Streams, Chapman and Hall.
26. Gama, João, et al (2010). A survey on concept drift adaptation, ACM Computing Surveys (CSUR) 46.4 (2014): 44. [49]. Žliobaitė, Indrė. Learning under concept drift: an overview, arXiv preprint arXiv:1010.4784.
27. B. Ramakrishna; S Krishna Mohan Rao(2018). Attribute Pattern Weights (APW): A Scale to Detect Concept Drift in Data Stream Mining Models. 2018 International Conference on Computer Communication and Informatics (ICCCI). 10.1109/ICCCI.2018.8441513.IEEE
28. Mashail Althabiti and Manal Abdullah (2019). Streaming Data Classification with Concept Drift. Biosci. Biotech. Res. Comm. Special Issue Vol 12 No

(1) January 2019
29. Zeng Li; Yan Xiong; Wenchao Huang (2020). Drift-detection Based Incremental Ensemble for Reacting to Different Kinds of Concept Drift. 2019 5th International Conference on Big Data Computing and Communications (BIGCOM).
30. Bhavani, S., Subhash Chandra, N. (2022). Histogram-Based Initial Centroids Selection for K-Means Clustering.
31. In: Goswami, S., Barara, I.S., Goje, A., Mohan, C., Bruckstein, A.M. (eds) Data Management, Analytics and Innovation. ICDMAI 2022.
32. Lecture Notes on Data Engineering and Communications Technologies, vol 137. Springer, Singapore. https://doi.org/10.1007/978-981-19-2600- 6_38
33. Soppari, K., Chandra, N.S. Automated digital image watermarking based on multi - objective hybrid meta - heuristic- based clustering approach. Int J Intell Robot Appl (2022). https://doi.org/10.1007/s41315-022-00241- 3