STM JOURNALS

Current Trends in
Signal Processing

https://journals.stmjournals.com/ctsp

ISSN: 2277-6176 (Online)
ISSN: 2321-4252 (Print)
Volume 15, Issue 3, 2025
September–December
DOI (Journal): 10.37591/CTSP

Review CTSP

A Study on Feature Subset Selection in Feature Streams of Dynamic Data

Priyadarshini[1]*, Shirish S. Sane[2]

Abstract

As the use of real-time data with high dimensions continues to expand across various domains, selecting important features from the dataset is a key step to improve the predictive accuracy and time taken to build a machine learning model. In datasets where not all features are available at the same time and we are unaware of the total number of features, and features arrive at different time stamps, for example, in real-time patient monitoring in a hospital's intensive care unit (ICU), feature selection becomes even more challenging. In an ICU, patients are continuously monitored using various medical instruments with sensors for monitoring heart rate, blood pressure, oxygen saturation, and electrocardiograms (ECG), etc. These devices stream data at different intervals, often providing updates asynchronously. The goal is to detect early signs of deterioration in a patient's condition and alert medical staff in real-time. In such cases, an efficient feature stream selection algorithm is needed to select features from the available set and incorporate new ones as they arrive. This paper provides a study of the various methods available for online feature stream selection, along with the methodologies used, and then identifies various issues and challenges that need to be addressed for feature stream selection. To provide practical insights, several existing approaches are implemented and analyzed. Additionally, key challenges and future directions in feature stream selection are identified and discussed.

Keywords: Feature stream, feature selection, machine learning, real-time

INTRODUCTION

Streaming data often arrives continuously at high speed. They are classified into data and feature streams, each presenting unique processing considerations [1]. A data stream refers to the continuous arrival of a large amount of data records, and the features of each record remain the same. Each record may represent a complete set of features (attributes) at a given time. Example: A stream of sensor readings, where each reading contains all relevant measurements (e.g., temperature, humidity, and pressure) at a specific time. A feature stream refers to the continuous flow of individual features. Different features may arrive at different times. In a smart factory, various machines are equipped with multiple sensors that monitor different aspects of machine health, such as temperature, vibration, sound, and operational speed. These sensors stream data at different intervals, and not all features are available at the same time. The prediction of potential machine failures in real-time needs to be performed to prevent downtime. Feature Selection is a data preprocessing step in which a decision is made regarding the inclusion or exclusion of the newly arrived feature. Feature selection in feature streams, where not all features are available at the same time, and we are unaware of the total number of features, is more challenging than in data streams

***Author for Correspondence**
Priyadarshini
E-mail: ilppriyadarshini@gmail.com

[1]Research Scholar, Department of Computer Engineering MET's Institute of Engineering Bhujbal Knowledge City (BKC), Nashik, Maharashtra, India
[2]Principal and Professor, Department of Computer Engineering, Gokhale's Education Society R H Sapat College of Engineering, Nashik, Maharashtra, India

Received Date: June 12, 2025
Accepted Date: July 10, 2025
Published Date: September 26, 2025

Citation: Priyadarshini, Shirish S. Sane. A Study on Feature Subset Selection in Feature Streams of Dynamic Data. Current Trends in Signal Processing. 2025; 15(3): 26–32p.

with fixed features. Our study focused on dynamic feature streams. Selecting the relevant number of features speeds up the data mining algorithm and improves the predictive accuracy of the classification model [2, 3, 4].

The remainder of this paper is organized as follows. First, a literature survey is presented along with the issues and challenges of feature stream selection. Next, a general framework describing the methodology of feature stream selection is outlined, followed by the implementation results and observations from a few existing feature stream algorithms. Finally, the paper concludes with a summary of the study.

LITERATURE REVIEW

The techniques carried out for feature stream selection were analyzed based on the availability of class labels for learning, categorizing them into supervised and unsupervised approaches.

Supervised Algorithms

The various algorithms under supervised learning are discussed below.

In the grafting-based approach, the feature selection technique involves adding features to an existing model by adjusting one or more weights in the weight vector [5]. Each added weight (wj) incurs a regularization penalty, which is computed as $\lambda|wj|$. After incorporating a feature, the mean loss decrease of the model was evaluated. If the reduction in error is greater than the penalty imposed by regularization, the feature is retained in the model. If the reduction in error is smaller than the penalty, the feature is excluded because its contribution is not sufficiently significant to justify its inclusion [5]. *Issue:* This approach is supervised, and parameter tuning is very difficult.

In the feature stream method using alpha investing [6], "a p-value is used to determine if a new feature should be added to the model. Linear regression was performed to assess the updated model, with the null hypothesis suggesting that the new feature did not improve the model's predictive performance. A p-value is generated through hypothesis testing, and if it is lower than a predetermined significance level (α_i), the feature is incorporated into the model" [6]. *Issue:* Alpha investing does not account for feature redundancy, as it evaluates each feature only once.

OSFS (online streaming feature algorithm) classifies features into four categories: "irrelevant, redundant, weakly relevant but non-redundant, and strongly relevant features" [7]. This classification is based on conditional independence principles from probability theory [7]. For example, probability $P(A \mid B, C) = P(A \mid C)$, which indicates that knowing B does not alter the probability of A once C is known, signifying that A and B are conditionally independent, given C. This principle helps eliminate redundant features and assesses whether new features contribute valuable information to the target variable T. If a feature X and class attribute C are conditionally independent for a given subset of features S, then feature X is considered redundant and discarded. *Issue:* Redundancy analysis in OSFS is computationally expensive, and as the feature set grows, it becomes a bottleneck, leading to reduced performance.

Fast-OSFS (fast-online streaming feature algorithm) method improves upon OSFS by assessing both redundancy and interactions between newly added features and previously selected features, thereby speeding up the feature selection process [8]. However, it still incurs high computational costs, particularly with large high-dimensional datasets. *Issue:* Even with optimization, computational costs can be prohibitive when dealing with datasets containing millions of features.

SAOLA (scalable and accurate online approach): The theory of mutual information (MI) is used to identify the relevant and redundant features [9]. It first decides a relevance threshold and then it calculates the mutual information between the newly arrived feature X_t and the class variable Y. The

new feature is considered to be a relevant feature only if the calculated mutual information $I(X_t, Y)$ is greater than the decided threshold. In the redundancy removal step, if $I(Xt, X_j) >= I(X_t, Y)$ or $I(X_j, Y)$, where X_j and X_j are features of the selected feature set. In such a case, the feature with the lowest mutual information with respect to the class variable is removed. *Issue:* Deciding the threshold for mutual information is tedious.

In the approach mentioned in the online scalable streaming feature selection via dynamic decision [10], normalized mutual information (NMI) is used to select relevant features. NMI is defined as:

$$NMI (X, Y) = 2MI (X, Y)/H(X) + H(Y),$$

Where, mutual information (MI) and entropy (H) are used to assess the relationship between features [10]. It dynamically adjusts thresholds based on global statistics to categorize new features as relevant, redundant, or requiring further evaluation. This provides more accurate predictions but remains computationally expensive.

Unsupervised Feature Selection Algorithm

Feature selection for dynamic data in which class labels are unknown poses a great challenge [3]. Various unsupervised learning algorithms are discussed below.

Unsupervised feature selection for dynamic features method extends k-means clustering to decide whether newly arrived features should be selected or discarded [11]. It uses metrics, such as the Pearson correlation coefficient (PCC), least squares regression error (LSRE), and maximal information compression index (MICI), to assess feature relevance [11].

Online unsupervised streaming feature selection through dynamic feature clustering uses density-based clustering to select features [12]. Although the time complexity of this algorithm is relatively high compared to supervised methods, it is capable of handling both continuous and discontinuous feature streams.

This method dynamically adjusts thresholds based on global statistical information to categorize new features as relevant, redundant, or requiring further evaluation. This approach provides more accurate predictions but can still be computationally expensive.

Most existing studies used statistical approaches in feature stream selection to select relevant features, as shown in Table 1. Most of these methods require the selection of an appropriate threshold in their test to filter the attribute [13–15]. The selection of an appropriate threshold value requires extensive experimentation. Redundancy analysis of the feature subsets is a time-consuming task. Most studies have been performed on supervised datasets where class labels are available. Most existing streaming feature selection methods are suitable for features of a single data type, or they have provided different versions of algorithms for categorical and numerical features [16]. Therefore, feature stream selection algorithms that work for mixed-data-type feature streams are required.

Issues and Challenges in Feature Stream Selection

1. Feature streams can have high-dimensional feature spaces, which leads to computational and storage challenges.
2. Feature streams may contain noisy or redundant features that can degrade the model's performance. Identifying and filtering irrelevant or redundant features is crucial.
3. The feature space itself changes with the introduction of new features, and others may become irrelevant. Designing algorithms that can handle dynamic feature spaces is challenging.
4. Feature selection in feature streams often needs to be performed in real-time to ensure timely model updates and decision-making. This requires efficient and rapid algorithms.
5. The optimal number of features needs to be selected that will lead to improvement in the predictive accuracy and efficiency of the machine learning model.

Table 1. Approaches of feature stream selection.

Citation	Scoring mechanism	Statistical approach used
[17, 18]	Statistical Tests	p-values, t-tests, chi-square tests
[19, 20]	Information Theory	Entropy, mutual information, and conditional independence
[20, 21, 22]	Model Weights	Feature importance from linear regression, decision trees (e.g., Gini index, SHAP values)

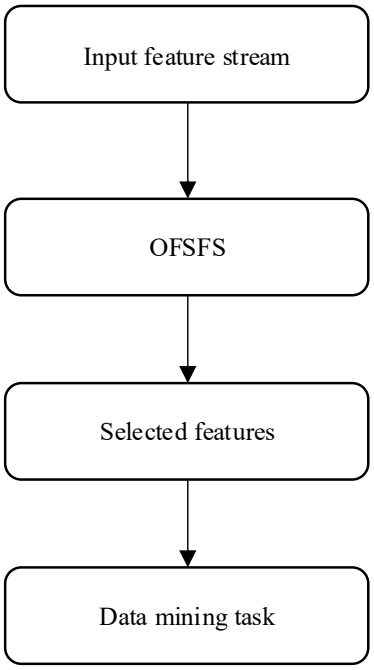

Figure 1. Framework of feature stream selection.

METHODOLOGY

The general methodology used in feature stream selection is shown in Figure 1.

Input Feature Stream

The input to the algorithm is the new features arriving at different timestamps.

Online Feature Selection in Feature Stream

Feature selection was done in two stages. The first stage is relevance analysis, and the second stage is redundancy removal. In relevance analysis, many measures are used to determine the relevance of a newly arrived feature to the class. Various approaches, such as chi-squared, gain ratio, information gain, and probability theory, can be used to determine the attribute's significance in predicting the class. The relevant features are selected, and if the attribute is evaluated as irrelevant, it is discarded. Once the relevant attributes are selected, the redundancy in attributes created owing to the recent selection of features is evaluated using measures such as correlation. If a high correlation is found between any two features, then any one of these can be retained, and the other is considered redundant and discarded. After discarding the features that are considered irrelevant and removing the redundant features, the optimal number of features is selected and used for the data mining task.

Data Mining Task

The performance of the feature selection algorithm is evaluated using the selected features in building the machine learning model, such as k-nearest neighbors (KNN), support vector machine (SVM), and classification and regression tree (CART), and observing the average prediction accuracy of the model built with selected features, the time taken by the feature selection algorithm, and the average number of features selected from each dataset.

Experimental Setup

The experiments conducted by most researchers were performed on benchmark datasets, as shown in Table 2, and the simulation of the feature stream was performed by taking a single feature at a time or using a sliding window to simulate multiple features simultaneously [23, 24]. Some of the work was carried out on synthetic data generated specifically for feature stream simulations. R, MATLAB, and Python are the preferred programming languages used by the majority of researchers. Additionally, Weka and the massive online analysis (MOA) tools are commonly used open-source tools for mining purposes. The performance of the system was measured using the following parameters:

1. Average number of selected features.
2. Time required to select the features.
3. Prediction accuracy of the model built with selected features.

RESULTS AND DISCUSSION

To understand the working of the existing approach, algorithms such as alpha investing, online feature selection (OFS), and SOALA are implemented with fine-tuned parameters. The experiments were conducted using Python 3.10 on Google Colab, which leverages graphics processing unit (GPU) runtime for computational efficiency. For alpha investing, α=0.05, $\alpha\Delta$ =0.5, was chosen [6]. For OFS Level of Significance was set at 0.05. For scalable and online approach for learning algorithms (SOALA), alpha(α)=0.05 was taken. The results are presented below: Regarding accuracy, alpha investing and OFS perform better than SOALA, as shown in Table 3. SAOLA requires less computation time for feature selection when compared to the other two approaches, as listed in Table 4, and OFS selects fewer features, as indicated in Table 5. Additionally, Figure 2 shows the accuracy results of the stated above.

Table 2. Benchmark datasets.

Data set name	Instance	Features	Classes
SRBCT_std [7, 10]	63	2308	4
Lukemia_std [9, 10]	72	7129	2
Prostate [10]	102	6033	2
Arcene [10]	900	10000	2
Lymphoma [10]	62	4026	3
DLBCL [10]	77	7129	2
Breast [10, 17]	97	24481	2
WDBC [7, 6]	569	30	2

Table 3. Predictive accuracy of KNN.

Data set	Original number of features	Alpha investing	OFS	SOALA
SRBCT	2308	82	92.3	69.23
WDBC	30	92.1	89.4	55.26
Prostrate	6033	90.4	95.23	90.4
Arcene	10000	79	72.5	72.5

Table 4. Time taken in seconds for feature selection.

Data set	Alpha investing	OFS	SOALA
SRBCT	13.3	261.22	4.59
WDBC	0.38	3.85	0.02
Prostrate	48.4	409	11.3
Arcene	31.6	414	29.7

Table 5. Number of features selected.

Data set	Alpha investing	OFS	SOALA
SRBCT	12	4	14
WDBC	20	3	2
Prostrate	8	4	6
Arcene	7	4	22

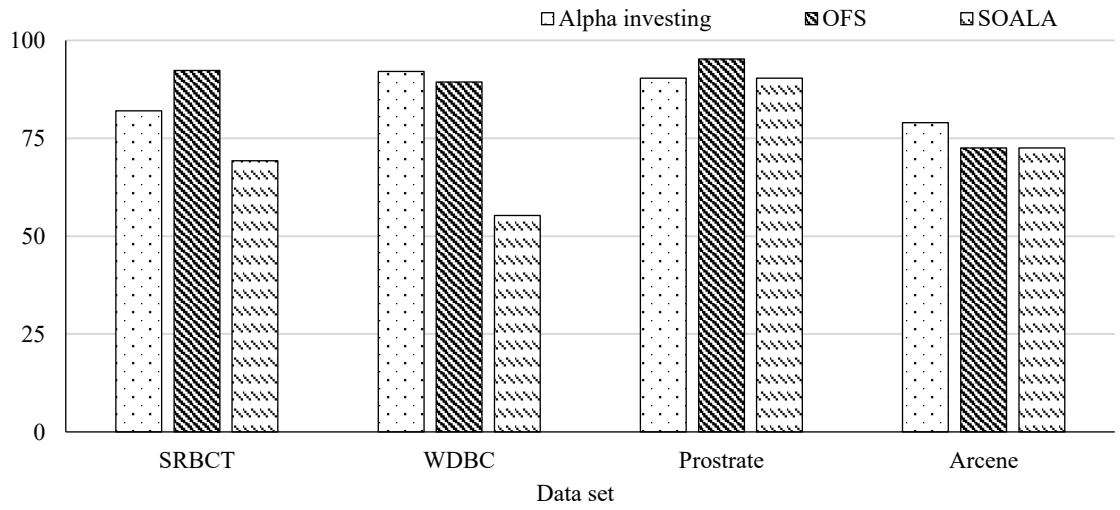

Figure 2. Accuracy using KNN.

CONCLUSION

This study provides an overview of the different techniques and approaches used for feature stream selection, along with issues and challenges. It was observed that most of the work was concentrated on a single feature arrival at a time, and all the features were of a single data type. In real-time, the features can be of mixed data types, and many features arrive at a single instance of time. Feature stream selection for real-time data with high dimensions is still very challenging.

Acknowledgement

The authors are very grateful to MET's Institute of Engineering for providing the opportunity to conduct this research study.

REFERENCES

1. Villa-Blanco C, Bielza C, Larrañaga P. Feature subset selection for data and feature streams: A review. Artif Intell Rev. 2023;56(Suppl 1):1011-62. doi:10.1007/s10462-023-10546-9.
2. Wang J, Zhao P, Hoi SCH, Jin R. Online feature selection and its applications. IEEE Trans Knowl Data Eng. 2014;26(3):698–710.
3. Patil DV, Bichkar RS. Issues in optimization of decision tree learning: A survey. Int J Appl Inf Syst. 2012;3:13-29.
4. William P, Paithankar DN, Yawalkar PM, Korde SK, Rajendra A, et al. Divination of air quality assessment using ensemble machine learning approach. 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India. 2023. pp. 1-10. doi: 10.1109/ICECONF57129.2023.10083751.
5. Perkins S, Theiler J. Online feature selection using grafting. Proceedings of the Twentieth International Conference on Machine Learning (ICML'03); 2003; Washington, DC, USA. AAAI Press; 2003. p. 592-9.
6. Zhou J, Foster D, Stine R, Ungar L. Streaming feature selection using alpha-investing. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005. p. 384-93. doi:10.1145/1081870.1081914.

7. Wu X, Yu K, Wang H, Ding W. Online streaming feature selection. Proceedings of the 27th International Conference on Machine Learning (ICML'10); 2010; Haifa, Israel. Madison, WI: Omnipress; 2010. p. 1159-66.
8. Hochma Y, Last M. Fast online feature selection in streaming data. Mach Learn. 2025;114:1. doi:10.1007/s10994-024-06712-x.
9. Yu K, Wu X, Ding W, Pei J. Scalable and accurate online feature selection for big data. ACM Trans Knowl Discov Data. 2017;11:1-39. doi:10.1145/2976744.
10. Zhou P, Zhao S, Yan Y, Wu X. Online scalable streaming feature selection via dynamic decision. ACM Trans Knowl Discov Data. 2022;16:1-20. doi:10.1145/3502737.
11. Almusallam N, Tari Z, Chan J, Fahad A, Alabdulatif A, Al-Naeem M. Towards an unsupervised feature selection method for effective dynamic features. IEEE Access. 2021;9:77149-63. doi:10.1109/ACCESS.2021.3082755.
12. Liu H, Setiono R. Chi2: Feature selection and discretization of numeric attributes. Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, Herndon, VA, USA. 1995. p. 388-91. doi:10.1109/TAI.1995.479783.
13. Zubaroğlu A, Atalay V. Data stream clustering: A review. Artif Intell Rev. 2021;54:1201-36. doi:10.1007/s10462-020-09874-x.
14. Zhou P, Hu X, Li P, Wu X. Online streaming feature selection using adapted neighborhood rough set. Inf Sci. 2019;481:258-79. doi:10.1016/j.ins.2018.12.074.
15. Liu J, Lin Y, Du J, Zhang H, Chen Z, Zhang J. ASFS: A novel streaming feature selection for multi-label data based on neighborhood rough set. Appl Intell. 2023;53:1707-24. doi:10.1007/s10489-022-03366-x.
16. Zhou P, Zhang Y, Yan Y, Zhao S. Unknown type streaming feature selection via maximal information coefficient. 2022 IEEE International Conference on Data Mining Workshops (ICDMW), Orlando, FL, USA. 2022. p. 650-7. doi:10.1109/ICDMW58026.2022.00089.
17. Liu H, Setiono R. Feature selection via discretization. IEEE Trans Knowl Data Eng. 1997 Aug 31;9(4):642-5.
18. Miller A. Subset Selection in Regression. Boca Raton, Fla., USA: Chapman & Hall/CRC; 2002. doi:10.1201/9781420035933.
19. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226-38. doi:10.1109/TPAMI.2005.159. PubMed PMID: 16119262.
20. Cover TM. Elements of Information Theory. New Jersey, US: John Wiley & Sons; 1999.
21. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B. 1996;58:267-88. doi:10.1111/j.2517-6161.1996.tb02080.x.
22. Breiman L. Random forests. Mach Learn. 2001;45:5-32. doi:10.1023/A:1010933404324.
23. Sandhiya S, Palani U. A novel hosfs algorithm for online streaming feature selection. 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020. p. 1-6. doi:10.1109/ICSCAN49426.2020.9262401.
24. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17); 2017; Long Beach, California, USA. Curran Associates Inc.; 2017. p. 4768-77.

