STREAMING FEATURE SELECTION

Dilbar Isakova - dilbar.isakova@estudiantat.upc.edu
Linhan Wang - linhan.wang@estudiantat.upc.edu
Padova, Italy - eBISS 2024

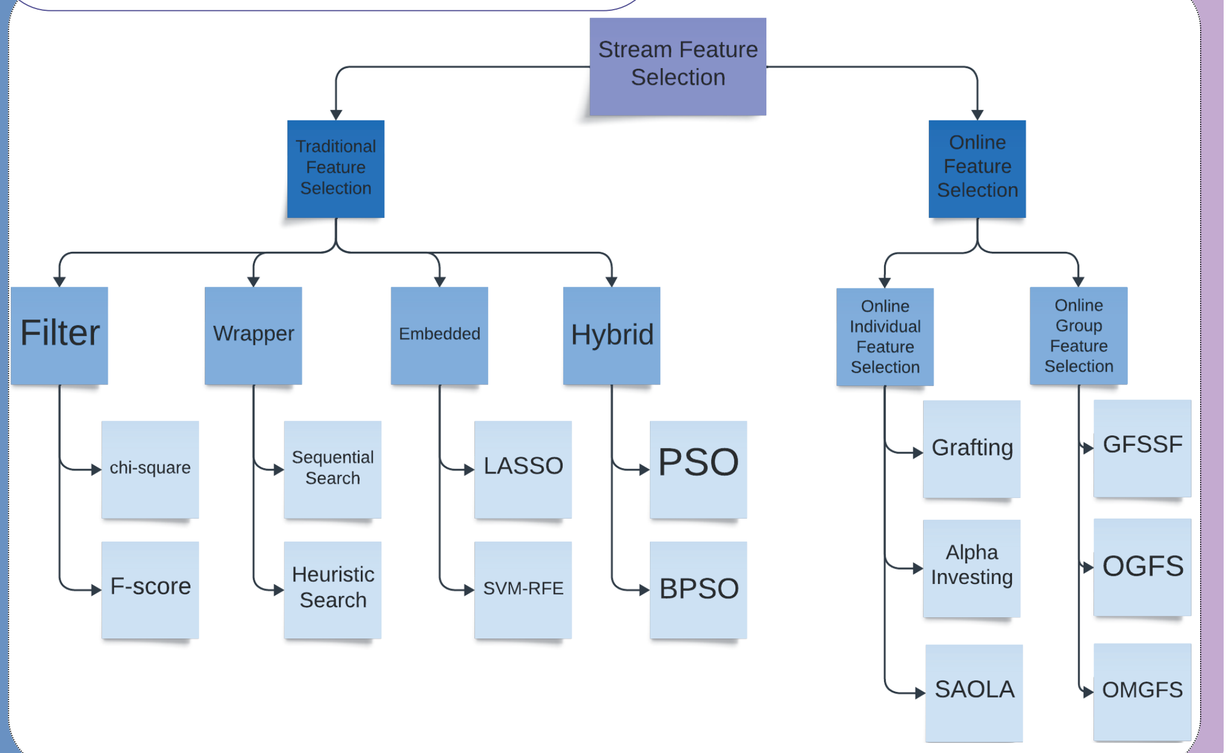
INTRODUCTION

Data science faces the challenge of efficiently managing high-dimensional streaming data, and this study focuses on streaming feature selection (SFS), a technique crucial for real-time data insight extraction. SFS reduces data dimensionality by selecting relevant features, enhancing the efficacy of machine learning algorithms. The study compares various SFS methods with traditional techniques, illustrating their suitability for streaming data scenarios. Grounded in a review of foundational studies, it discusses the evolution and current state-of-the-art SFS approaches, addressing issues like feature drift and scalability, and suggests future research directions.

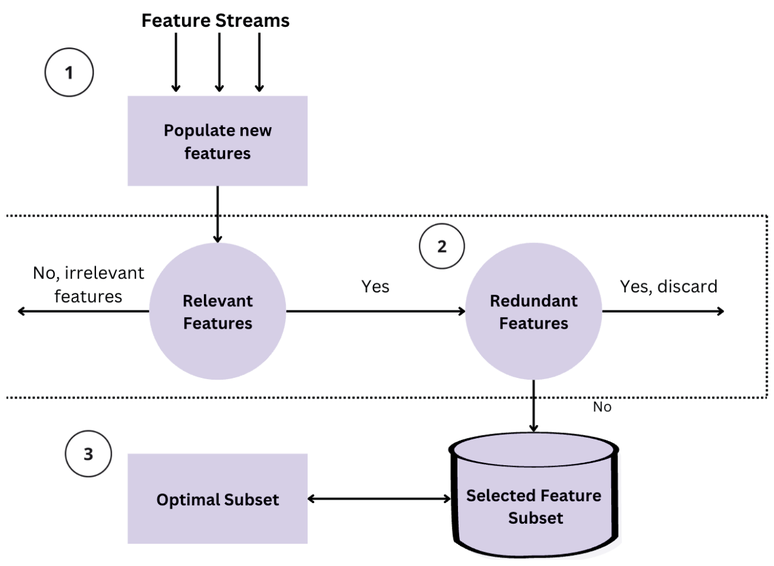
TRADITIONAL VS STREAMING

Feature	Traditional Feature Selection	Stream Feature Selection
Data Processing	Static, batch processing	Dynamic, continuous processing
Data Availability	All data available at once	Data arrives in a sequential stream
Computational Cost	High for large datasets	Typically lower, real-time processing
Memory Usage	Requires loading entire dataset into memory	Operates with limited memory
Feature Evaluation	One-time evaluation	Continuous, incremental evaluation
Relevance Detection	Single, fixed relevance computation	Adapts to changes in data over time
Redundancy Handling	Static redundancy removal	Continuous redundancy analysis
Example Techniques	PCA, LASSO, Genetic Algorithms	Alpha-Investing, Online Group Feature Selection (OGFS)
Adaptability	Limited, requires retraining for new data	High, adapts to new data in real-time
Suitability	Suitable for static datasets	Suitable for applications with continuous data streams (e.g., real-time analytics, IoT)

TAXONOMY



FLOW



- Goal: Find the optimal final feature subset
- Relavant analysis: the relevance of each feature to the target class is evaluated by statistical methods
- Redundant analysis: the similarity between features is evaluated to determine how much adding a new feature can improve the accuracy of a machine learning model

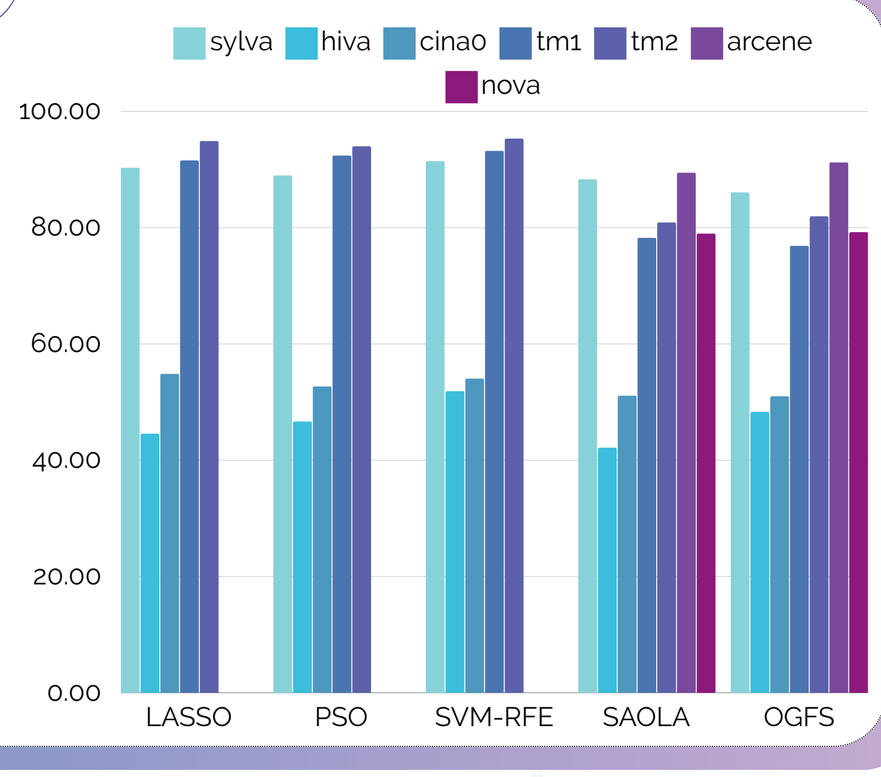
OVERVIEW OF ALGORITHM CAPABILITIES

Technique	Ref. Year	Classifier	DataSet	Platform	Structure	Strategy	Supervision	Classification Type
Sequential Forward Selection	1971	Any classifier	Varies: commonly used in feature selection tasks	Multiple	Single	Feature Selection	Supervised	Binary, Multiclass
LASSO	1996	Linear models with L1 regularization	Varies: commonly used in statistics and machine learning	Multiple	Single	Feature Selection	Supervised	Binary, Multiclass
Grafting	2003	Support Vector Machine (SVM)	Three Datasets (DS), A and B are synthetic, C is real-world	Matlab	Single, Streaming	Data Streaming	Supervised	Binary
SMOTE	2002	Any classifier	Datasets with class imbalance	Multiple	Single	Data Sampling	Supervised	Binary, Multiclass
PSO	2001	Any classifier	Varies: optimization in various domains	Multiple	Single	Optimization	Unsupervised	Binary, Multiclass
Alpha investing	2006	Applied stream-wise regression	Datasets from UCI: Ionosphere, Wine, etc.	R	Streaming	Data Streaming	Supervised	Binary
OSFS	2010	K-NN, J48, Random Forest (RF)	10 public challenge Datasets: Breast, Prostate, etc.	C Language	Single	Feature Selection	Supervised	Binary
SAOLA	2014	J48 and KNN	High-dimensional biomedical and NIPS 2003 Datasets	Matlab	Single	Feature Selection	Supervised	Multiclass
OGFS	2015	K-NN, J48, RF	Eight biomedical Datasets from UCI: Ionosphere, etc.	Not mentioned	Single, Group	Feature Selection	Supervised	Multiclass
Group-SAOLA	2016	J48, SVM, KNN	Ten high-dimensional Datasets: Leukemia, Lung, etc.	Matlab	Single, Group	Feature Selection	Supervised	Multiclass
OSFS-KW	2020	KNN, SVM, and RF	Used a high dimensional medical dataset: Hill, Hill (Noise), Wdbc, Lym phoma, Dlbcl, Colon, Tumor, Car, Glioma Srbct, Leu, Mll, Prostate, Lung Std, Arcene, Madelon, Lung, Breast Cancer, Ovarian Cancer and Sido0	Matlab	Single	Data Streaming	Supervised	Binary

RESULTS & ANALYSIS

We measure the feature selection methods by calculating the accuracy, precision, recall and F1-score of machine learning algorithms. This figure shows the F1-score of SVM under different feature selection methods. The results are:

- only streaming algorithms (SAOLA and OGFS) are capable of processing feature space for very large datasets(acrene and nova). Other algorithms are unable to complete their work due to time or memory limits
- As three batch algorithms (LASSO, PSO and SVM-RFE) search the whole feature space to evaluate each candidate feature, they are unable to process large datasets. However, for small and medium-scaled datasets, such a search is an advantage and results in highly accurate subsets.



CONCLUSIOIN

- This study presented a comprehensive survey of recent FS algorithms for both static and dynamic environments across various domains, along with a taxonomy categorizing these methods based on their search strategy, evaluation process, and feature structure. Initially, the study reviewed existing traditional and online FS methods, providing a qualitative analysis of their strengths and weaknesses. The proposed taxonomy also includes a quantitative analysis of these techniques based on their category and publication timeline.
- This survey aims to enhance the efficiency of learning state-of-the-art FS methods and assist researchers in understanding and applying key characteristics of FS in Big Data. It also helps identify limitations and research gaps in current FS methods.
- While our paper provided a detailed discussion on online methods, the experimental analysis was limited to SVM and 5 stream selection methods. In future work, we plan to include experimental analysis of more classifiers and more feature selection methods.

REFERENCES

FULL PAPER