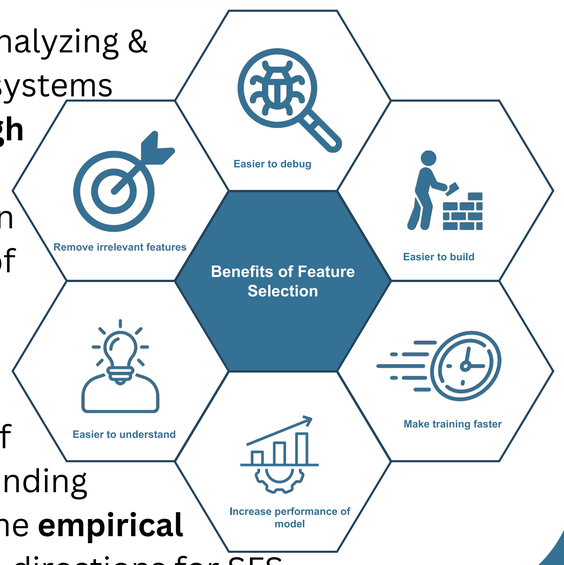
Streaming Feature Selection

Muhammad Rizwan Khalid - Prashant Kumar Gupta

Barcelona, Spain - eBISS 2023

Introduction

Feature Selection is an important task in analyzing & predicting the outcome. But for real-time systems where features are **dynamic**, **velocity is high** and **domain knowledge is limited**. It is not efficient to use traditional feature selection techniques because of streaming fashion of real-time datasets. In this article, we are presenting the methodology (**redundancy** and **relevancy** analysis) of streaming feature selection (SFS) and the evolution of algorithms employing innovative ways of finding **optimal** no. of features. We also discuss the **empirical analysis** of SFS algorithms and future directions for SFS.



Traditional vs Streaming

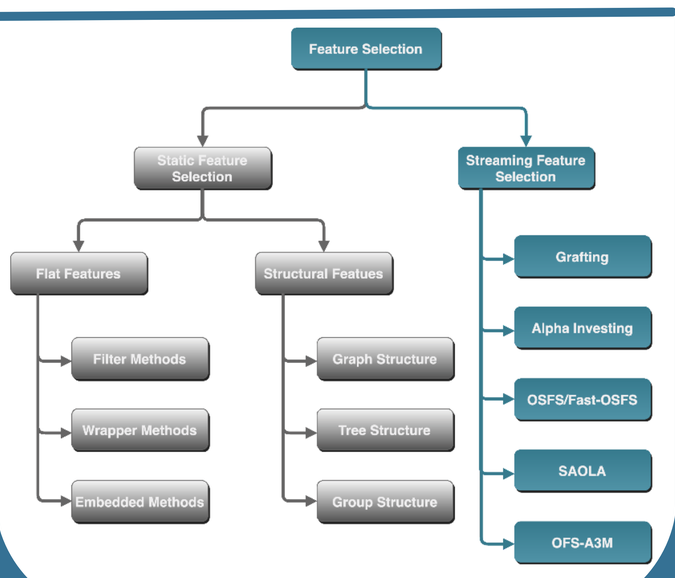
Traditional Feature Selection

1. Assumes the entire dataset is available at once, and aims to select a subset of features most relevant for the learning task.
2. **Offline Processing**: Operates on static data, whole dataset is available before hand.
3. **Batch Computation**: Computes feature relevance based on the entire datase
4. Fixed Feature Set

Streaming Feature Selection

1. Data arrives sequentially or in a streaming fashion. This approach is designed **to adapt** to changing data distributions and evolving feature relevance.
2. **Online Processing**: Operates on data as it arrives, often in real-time.
3. **Incremental Computation**: Updates feature relevance on-the-fly.
4. Dynamic Feature Set

Taxonomy



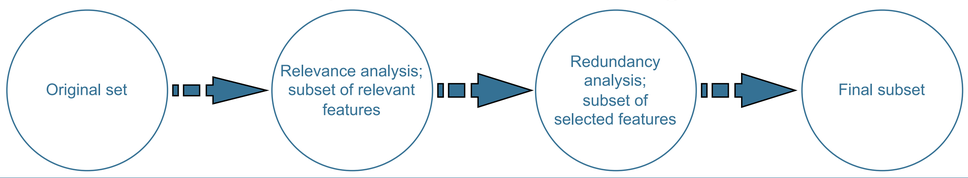
Methodology

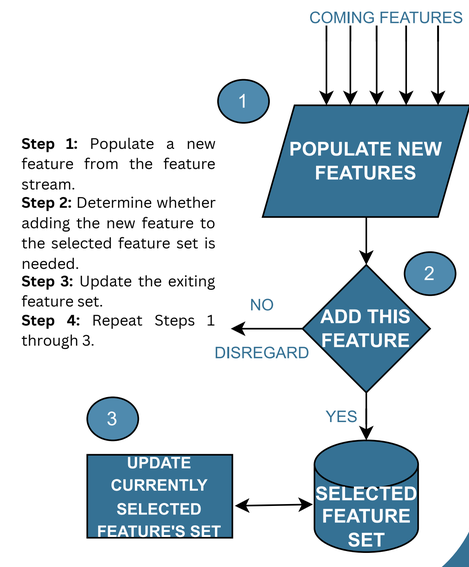
OnLine Relveance Analysis

During this stage, features are categorised into three groups: strong relevance, weak relevance, and irrelevance. If the feature is **strongly or weakly relevant**, it is included in the feature subset. However, if the feature is deemed **irrelevant**, it is **disregarded**.

OnLine RedunDancY Analysis

Identifies and **removes redundant features** by checking subsets of the **Best Candidate Features (BCF)** and ensures that newly added features are optimal by examining if they create conditional independence with existing features and the class attribute.





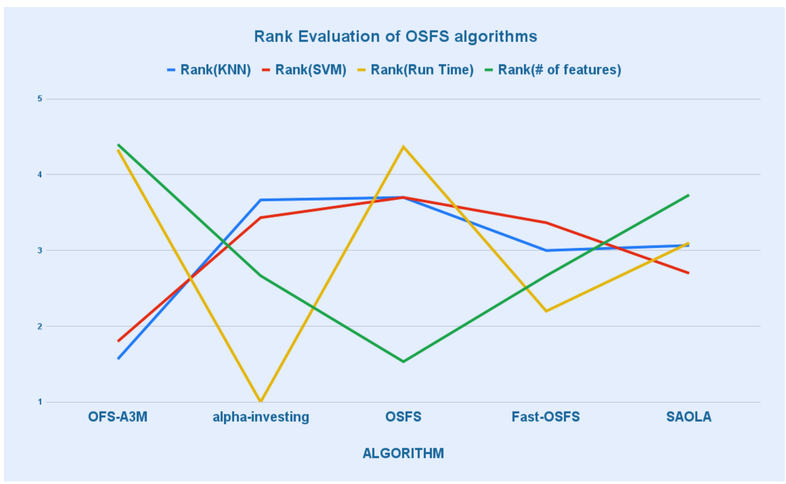
Step 1: Populate a new feature from the feature stream.
Step 2: Determine whether adding the new feature to the selected feature set is needed.
Step 3: Update the exiting feature set.
Step 4: Repeat Steps 1 through 3.

Evolution of Algorithms

Result Analysis

The graph shows the rank of performance for OFS algorithm on **15 classification** datasets. It shows:

- OFS-A3M has the **best** predictive **accuracy** for KNN
- Alpha-Investing is the **fastest algorithm** in term of running time
- OSFS selects the **least** number of **feature** and potentially misses critical information
- SAOLA has comparable performance but is **sensitive** to **noisy** datasets



Conclusion

Existing approaches to streaming feature selection typically involve testing **individual features** to select an optimal subset, which may not be effective for extremely high-dimensional big data, necessitating the development of more innovative approaches. In real life, feature may not arrive individually but in **group fashion** as well. Latest SFS algorithms usually suffer from high time complexity. The **scalability** challenge presents a future research direction for online feature-selection algorithms, considering the difficulty of loading the entirety of big data into memory during a **single scan** and obtaining reliable relevance scores for features without sufficient **density** around each sample.

References

1. Isabelle Guyon and Andre Elisseeff. 2003. "An introduction to variable and feature selection". J. Mach. Learn. Res. 3, null (3/1/2003), 1157-1182
2. Zhou, Jing, et al. "Streamwise feature selection." Journal of Machine Learning Research 7.9 (2006).
3. Hu, Xuegang, et al. "A survey on online feature selection with streaming features." Frontiers of Computer Science 12.3 (2018): 479-493
4. AlNuaimi, Noura, et al. "Streaming feature selection algorithms for big data: A survey." Applied Computing and Informatics 18.1/2 (2020): 113-135
5. Zhou, & Xuegang, Hu & Li, Peipei & Wu. (2018). "Online Streaming Feature Selection Using Adapted Neighborhood Rough Set". Information Sciences. 481. 10.1016/j.ins.2018.12.074.