An Evaluation of Big Data Analytics in Feature Selection for Long-lead Extreme Floods Forecasting

Yong Zhuang[1], Kui Yu[2], Dawei Wang[1], and Wei Ding[1]

[1]Department of Computer Science
University of Massachusetts Boston

[2]School of Information Technology and Mathematical Science

University of South Australia

Why we need long lead Extreme Floods Forecasting

Extreme floods are the one of the most destructive hazards on Earth. Despite local efforts and national encouragement to mitigate flood hazards and regulate development in flood-prone areas, flood damages have increased in the United States in the past decades. Long-lead prediction of extreme floods is great important to society for providing support of emergency response.

How to handle this long lead forecasting problem?

Because a type of extreme floods are associated with a sequence of prior heavy precipitation events occurring frequently from over several days to several weeks, long-lead forecasting of extreme floods can be formulated as a classification problem by identifying the precursors to heavy precipitation event clusters.

The challenge of long lead extreme floods forecasting

While a short-term prediction of certain location depends only on variables in near spatial and temporal neighborhood, predictions with long lead time must consider variables in a long time window and large spatial neighborhoods, this means an enormous amount of potentially influencing variables and only a subset of them strongly relate to prediction. Processing a deluge of variables and discovering strongly relevant features pose a significant challenge for big data analytics.

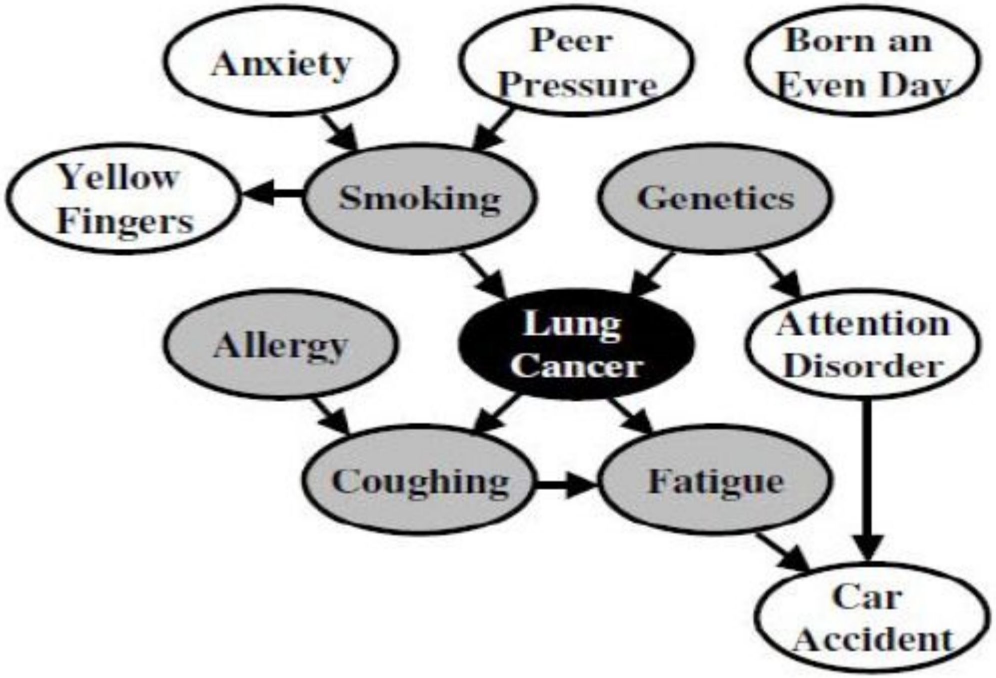
We use streaming feature selection methods to solve this problem.

The challenge of long lead extreme floods forecasting

Extreme floods rarely occurred in a year, so the total number of positive samples (extreme precipitation events) in the experimental data set is much less than the number of negative samples. How to deal with the imbalance problem is another challenge.

To deal with this class imbalance problem, we use the over-sampling method and the under-sampling method.

What is strongly relevant features



A causal Bayesian network for lung cancer

What is streaming feature selection

- The stream of features sequentially added
- The total data observations are fixed
- It aims to select a **subset of strongly relevant** features from the original feature set.
- To achieve simplification of models for easier interpretation, time efficiency, and enhanced generalization by reducing over-fitting.

Processing one feature at a time

$$U_i = \underset{U'}{\operatorname{argmin}}\{|U'| : U' = \underset{K \subseteq \{U_{i-1} \cup f_i\}}{\operatorname{argmax}} P(C|K)\}$$

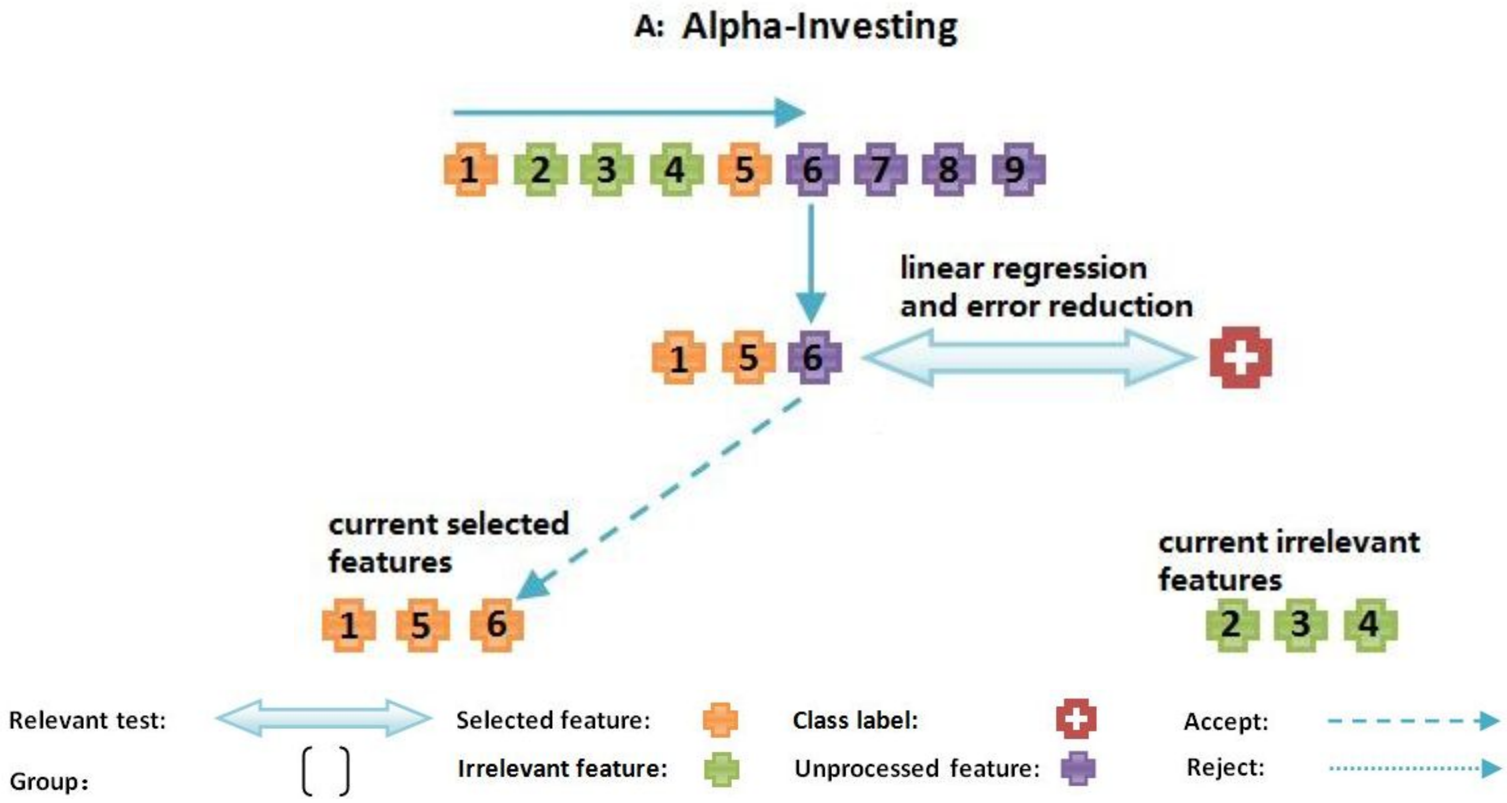
- This corresponds to find the optimal set of relevant features Ui for class C. Especially, when operating a new coming feature (fi), the currently selected feature set Ui will be updated dynamically.

UMASS BOSTON

Alpha-investing

(Zhou et al. in "The Journal of Machine Learning Research" 2006)

- The idea is to dynamically update the relevant feature set by adding a new feature as addition into the current selected feature set if the new feature is correlated with the class feature

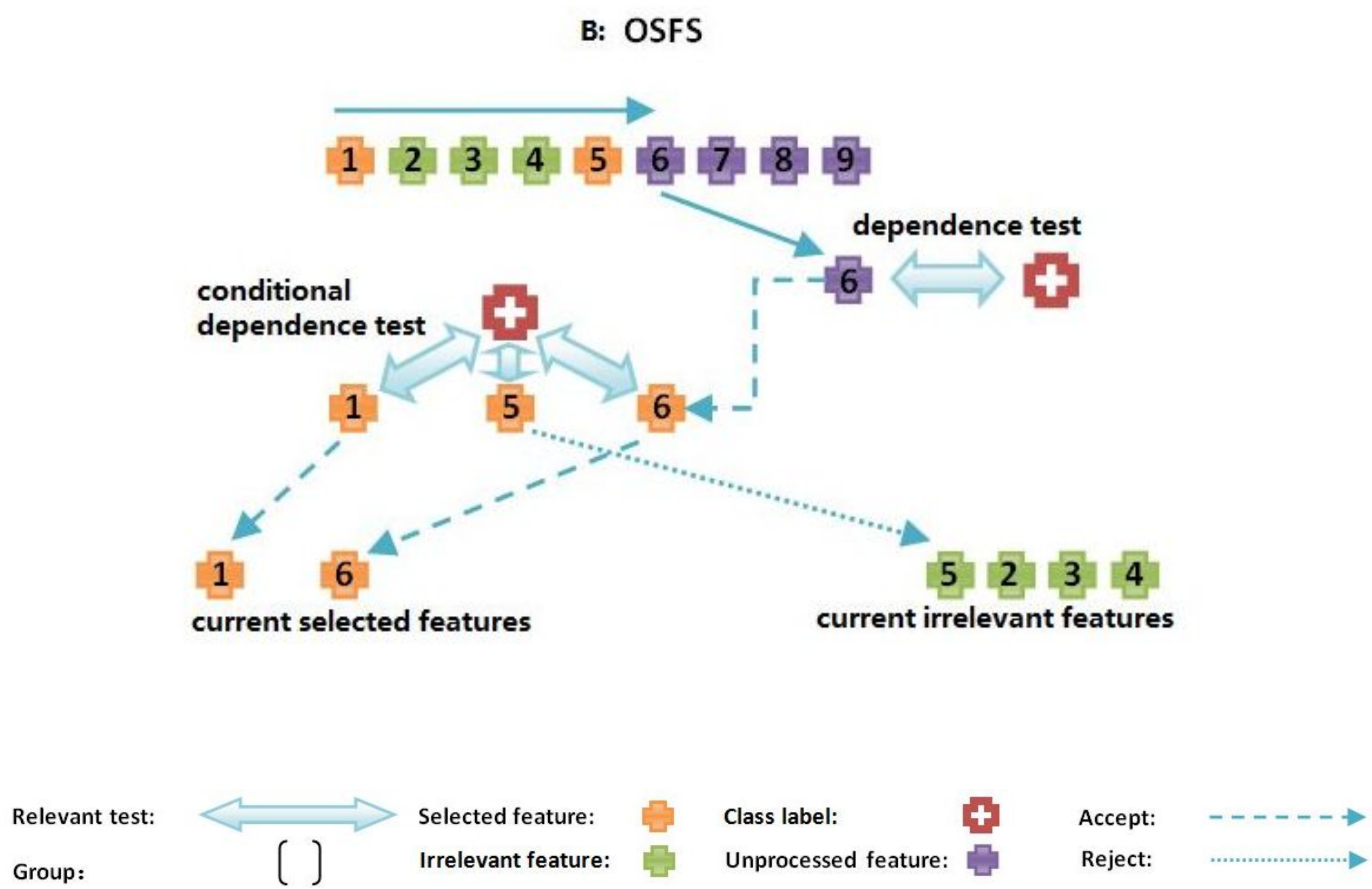


Online Streaming Feature Selection (OSFS)

Wu et al. in ICML 2010

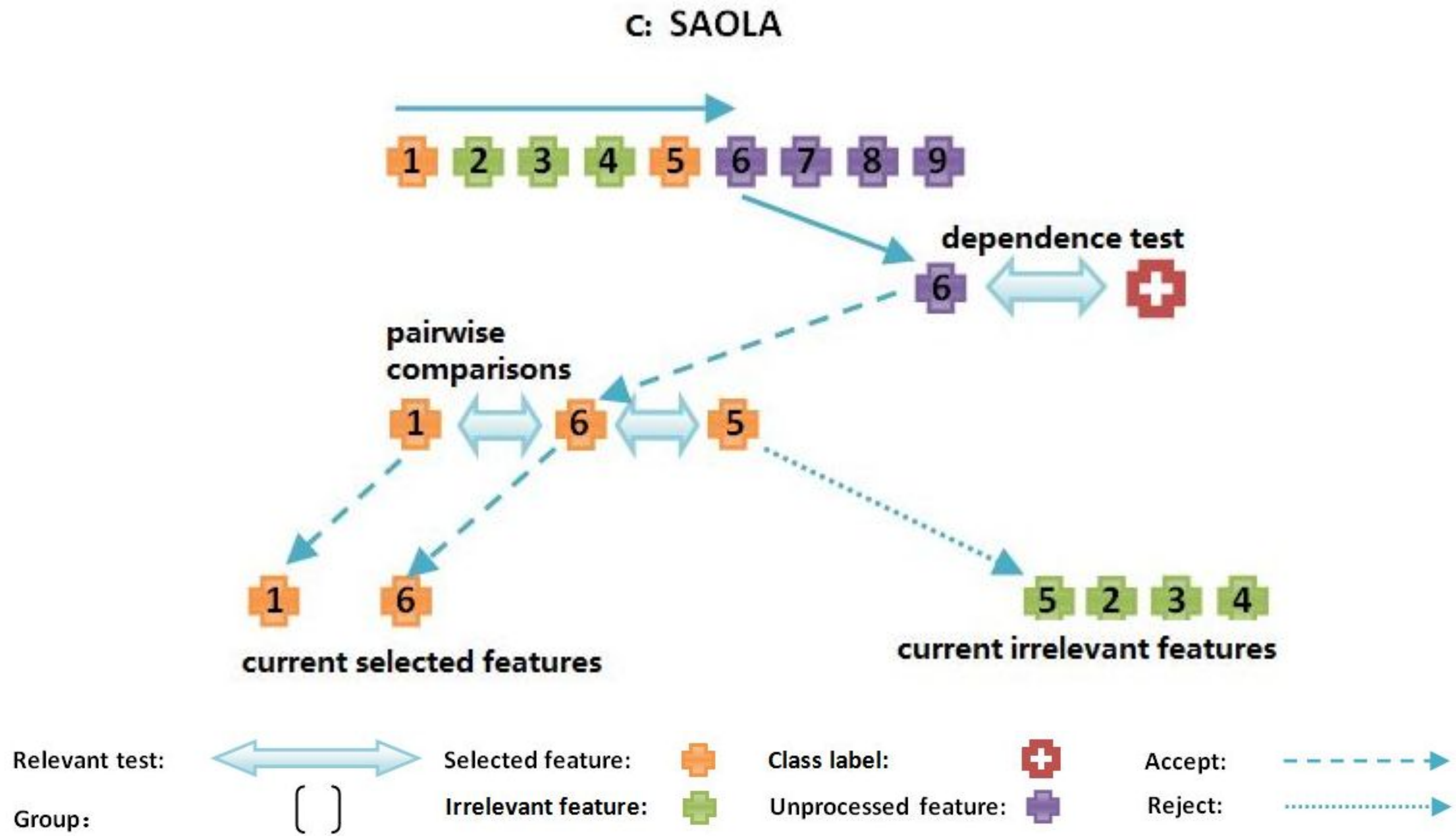
- The idea is to find the best so far relevant feature set from the original feature set by two steps:
- Step 1. Calculate whether the new coming feature is relevant to the class feature.
- Step 2. Analyze whether there exists redundancy among the selected feature set currently once the new coming feature is added.

Online Streaming Feature Selection (OSFS)



Scalable and Accurate OnLine Approach (SAOLA)
Yu et al. in ICDM 2014

- It employs online pairwise comparisons between features in the currently selected feature set once a new coming feature is included.



Processing grouped features sequentially

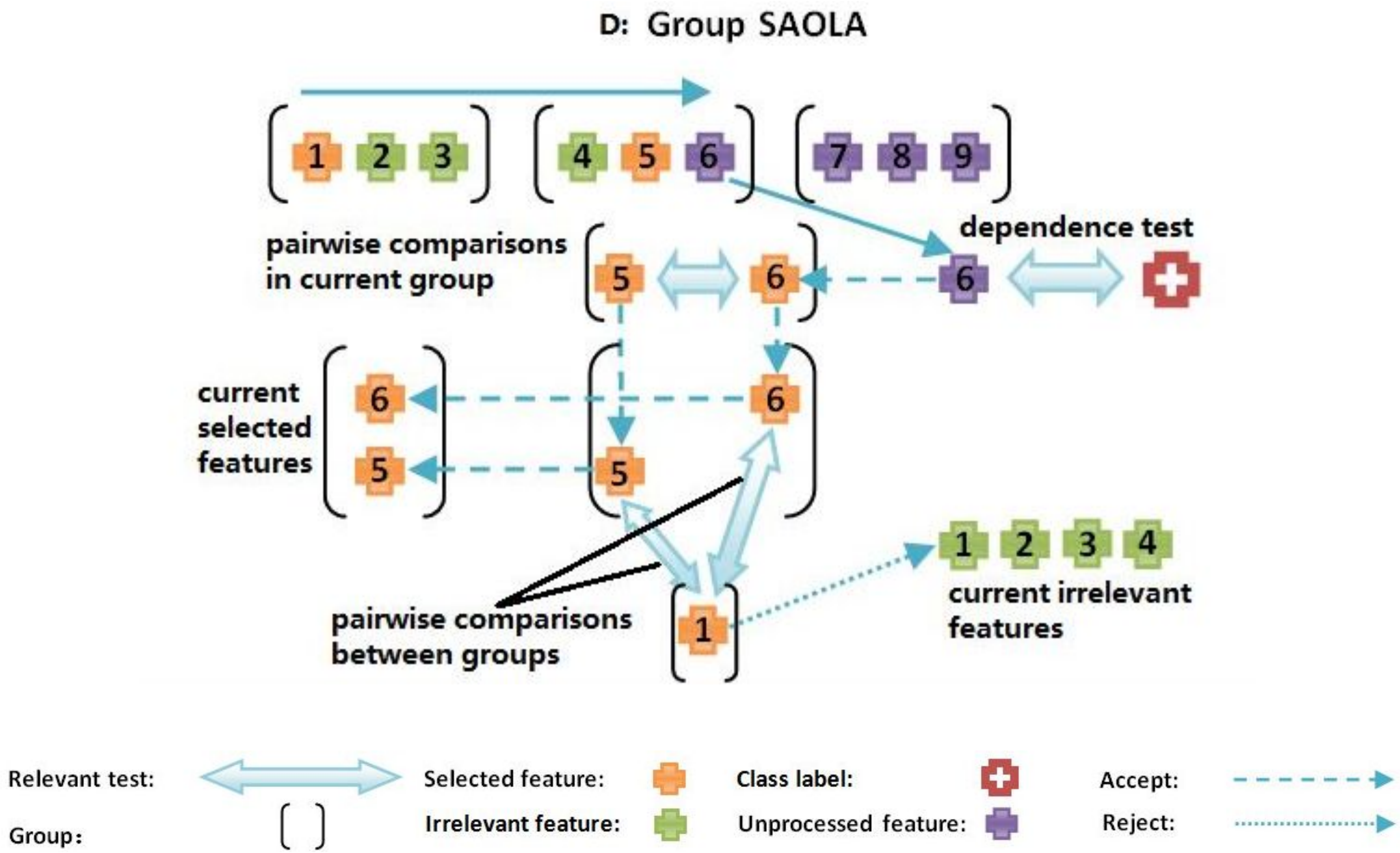
$$U_{G_i} = \underset{G' \subseteq \{U_{G_{i-1}} \cup G_i\}}{\operatorname{argmax}} P(C|G')$$

$$\begin{aligned}
&s.t. \\
&(a) \forall f_i \in U_j, U_j \subset U_{G_i}, \\
&P(C|\{U_j - \{f_i\}, f_i\}) \neq P(C|\{U_j - \{f_i\}\}) \\
&(b) \forall U_j \subset U_{G_i}, \\
&P(C|\{U_{G_i} - U_j, U_j\}) \neq P(C|\{U_{G_i} - U_j\}).
\end{aligned}$$

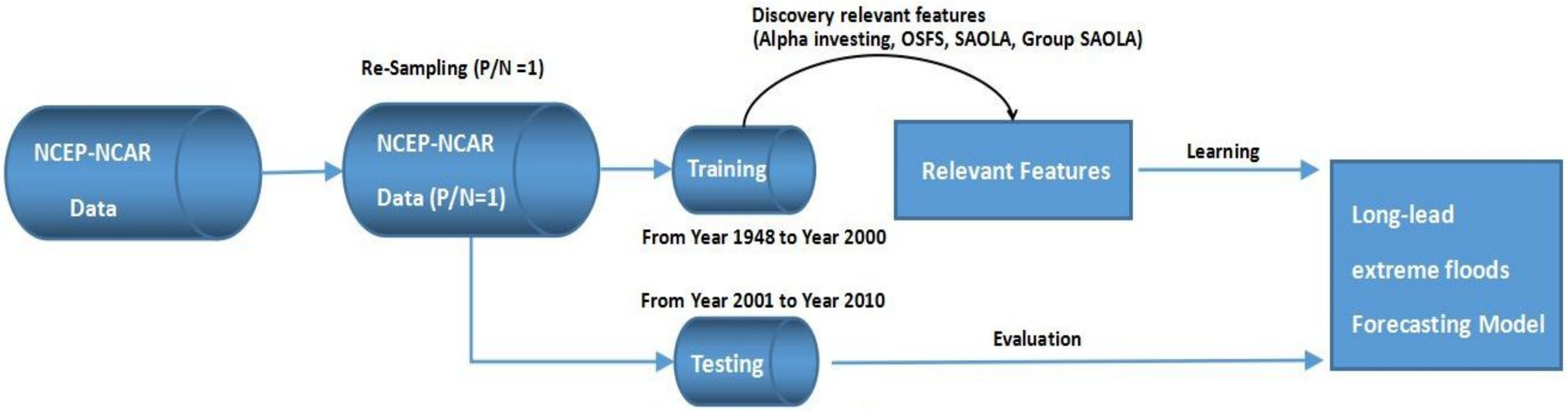
- This objective corresponds to find the optimal set of feature groups U_Gi for class C.
- (a) aims to find the minimal number relevant features in each group.
- (b) aims to remove redundant features in currently selected set.

Group SAOLA Yu et al. in 2015

- It utilizes the prior group to maximize each group's predictive performance for classification.



Experiment - Flow Chart



Experiment - Data

- We choose several variables from the historical meteorological data collected in the State of Iowa, the United States from January 1st, 1948 to December 31st, 2010
- We pick the samples collected during the rainy season (April to October) every year, which might have correlation with precipitation events.
- The samples in (1948 - 2000) are used as training set to learn the forecasting model.
- The samples in (2001 - 2010) data are used as test set to evaluate the forecasting model.

Experiments

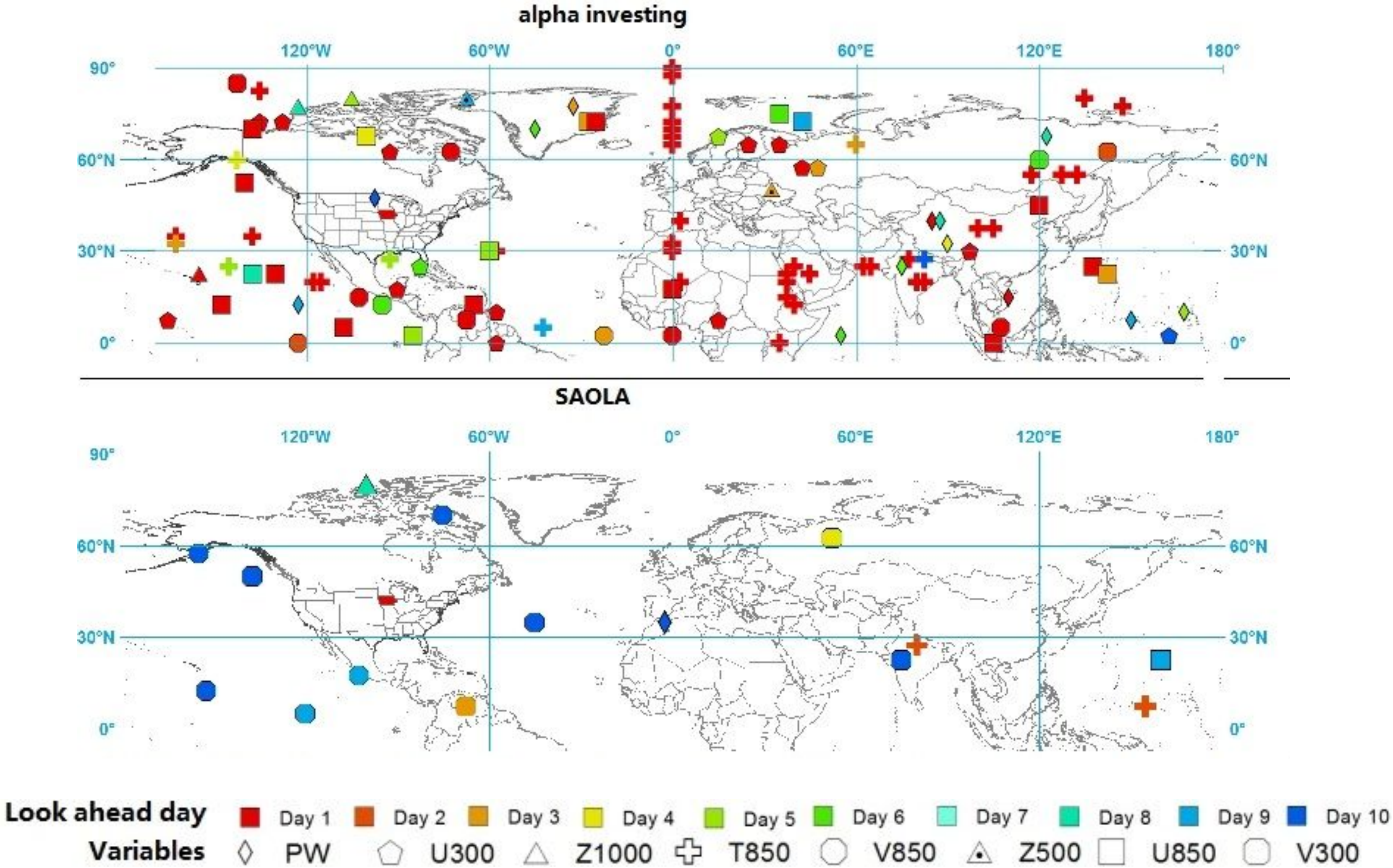
- Experiment 1: Four online streaming feature selection methods + original data + KNN. The aim of this experiment is to check the effect of four online streaming feature selection methods on imbalanced data.
- Experiment 2: Four online streaming feature selection methods + over-sampled data + KNN. It aims to check the effect of four online streaming feature selection methods on the data balanced by over-sampling method.
- Experiment 3: Four online streaming feature selection methods + under-sampled data + KNN. This experiment aims to check the effect of four online streaming feature selection methods on the data balanced by the under-sampling method. We do this experiment 10 times with randomly under sampled balanced data sets. Then we calculate the mean values of the static measures.

Experiment - Result

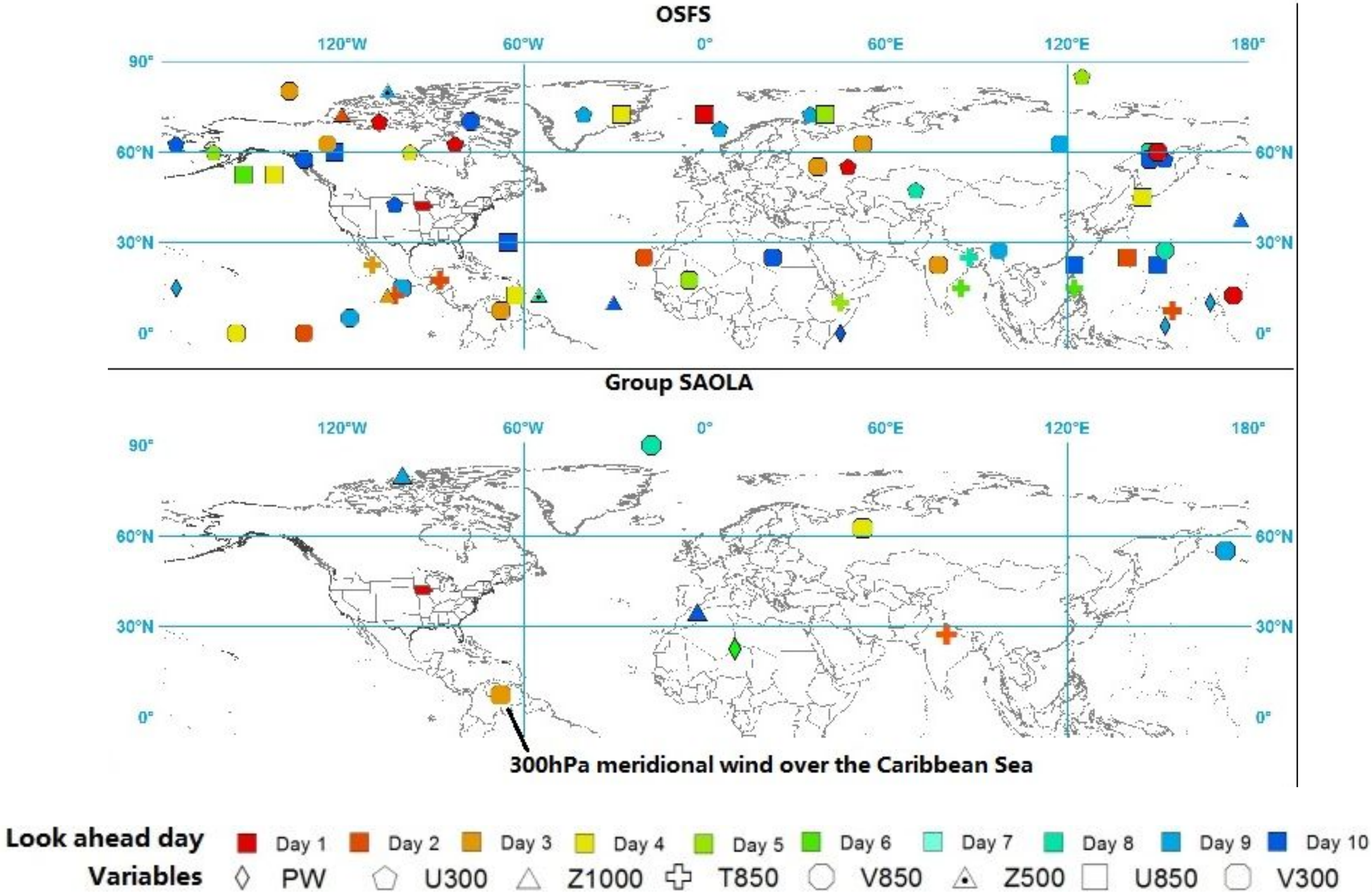
Experiments	Metrics	Alpha investing	OSFS	SAOLA	Group SAOLA
	The size of relevant feature set	112	68	15	8
Experiment 1	Accuracy F-measure	0.8235 0.1284	0.827 0.1128	0.8305 0.1285	0.8435 0.1425
Experiment 2	Accuracy F-measure	0.4766 0.239	0.4789 0.2537	0.4797 0.2594	0.4976 0.2635
Experiment 3	Accuracy F-measure	0.7028 0.7466	0.7696 0.8039	0.712 0.7485	0.7189 0.7589

UMASS BOSTON

Experiment - Map



Experiment - Map



Thank You ！