Online Group Feature Selection from Feature Streams

Haiguang Li[1] **and Xindong Wu**[1] **and Zhao Li**[2] **and Wei Ding**[3]
[1]Department of Computer Science, University of Vermont
[2]TCL Research America
[3]Department of Computer Science, University of Massachusetts Boston
{hli, xwu}@cems.uvm.edu, zhaoli01@tcl.com, ding@cs.umb.edu

Abstract

Standard feature selection algorithms deal with given candidate feature sets at the individual feature level. When features exhibit certain group structures, it is beneficial to conduct feature selection in a grouped manner. For high-dimensional features, it could be far more preferable to online generate and process features one at a time rather than wait for generating all features before learning begins. In this paper, we discuss a new and interesting problem of online group feature selection from feature streams at both the group and individual feature levels simultaneously from a feature stream. Extensive experiments on both real-world and synthetic datasets demonstrate the superiority of the proposed algorithm.

1 Introduction

We target at the problem domain where features possess certain group structures. The most common example is the multi-factor ANOVA problem, in which each factor may have several levels and can be expressed through a group of dummy variables (Yuan and Lin 2006). A dummy variable uses the value 0 or 1 to indicate the absence or presence of some categorical effect. In this case, feature/factor selection corresponds to the selection of groups rather than individual dummy features/variables, as each group corresponds to one measured feature/factor and is directly related to the measurement cost. As generating features in different groups may require different domain knowledge, measurements, procedures, etc., the candidate features are very likely to appear in the form of a feature steam, in which features are generated dynamically and arrive one by one and group by group. This situation appears in many practical applications. For instance, only on one run, the next-generation sequencing techniques can generate data with several giga features with values in {A, C, G, T} (Liu et al. 2010); thus, each feature having four levels and being represented by 4 dummy features. The storage cost is expensive to keep those features, and it is impractical to wait until all features have been generated before learning begins, thus it could be far more preferable to generate candidate features one at a time for all observations (Wu et al. 2010).

Many feature selection algorithms can effectively perform feature selection from a given candidate feature set or a feature stream, however, without considering the group structures, most feature selection algorithms always try to select features with sparsity (a small percentage of the original candidate feature set) only in the individual feature level. Obviously, when group structures exist, sparsity in both the group level and the individual feature level is more preferable.

In this paper, we propose a new online algorithm to perform group feature selection from feature streams named GFSFS, which performs feature selection at both the group level and the individual feature level simultaneously from the features generated and arrived so far.

2 Online Group Feature Selection

Given a candidate feature set $\mathbb{X}$ and a target feature Y, from the perspective of information theory, the task of feature selection is to find a subset of features $\mathbb{F} \subseteq \mathbb{X}$, such that (1) $\mathbb{F}$ and Y share as much information as possible, where the information shared by $\mathbb{F}$ and Y is represented by their mutual information $I(\mathbb{F}; Y)$; and (2) the residual uncertainty of Y is as little as possible given $\mathbb{F}$, where the uncertainty of Y is represented by the entropy $H(Y)$. According to the definitions of entropy, mutual information, and conditional mutual information in information theory, we can easily discover the relationships among individual features and feature groups. Obviously, the task of group feature selection is removing feature groups that are irrelevant to the target feature and irrelevant features in relevant groups, and selecting the feature groups that are relevant to the target feature without redundancy at both the individual feature level and group level.

The proposed GFSFS algorithm Algorithm 1 assures that only a single pass over the feature stream is able to complete feature selection at the individual feature and group levels simultaneously.

3 Experiments

MEMset (Yeo and Burge 2004) (28 features in 7 groups, available at http://genes.mit.edu/burgelab/maxent/ssdata/) and SYNTHETIC (664 features in 166 groups) are employed for testing. SYNTHETIC was built from the real UCI dataset Musk (v2) by introducing 4 dummy features for each original feature: those continuous features were discretized into nominal ones, and then each feature was replaced by 4

Algorithm 1 GFSFS:group feature selection from feature streams

```
repeat
    repeat
        Get the new arrived feature from the feature stream, check its *irrelevance* and *redundancy* (mutual information, conditional mutual information);
        If the new arrived feature is selected, *deselect* the previously selected ones that become *redundant* (conditional mutual information) to the target feature;
    until {This is the last arrived feature of a feature group}
    *Check redundancy* (conditional mutual information) to decide select the new selected feature group or not;
    If the new selected feature group is selected, then *deselect* the previously selected ones that become *redundancy* (conditional mutual information);
until {The selected features cover a certain percentage of the target feature's entropy (entropy, mutual information)}
return The selected feature groups
```

dummy features. Balanced training datasets were built by random sampling without replacement, and the rest were used for testing. Existing group feature selection algorithms use the group lasso regression model, and features are all available before learning begins for Group Lasso (Yuan and Lin 2006). GFSFS does not require all the features to be present. To simulate feature streams, features arrive one at a time in a random order. Classifiers NaiveBayes, k-NN, C4.5, and Randomforest, were chosen to evaluate the selected subsets, and the best accuracy was reported.

The state-of-art feature selection algorithms MIFS, JMI, mRMR and RELIEF cannot deal with datasets with group structures. As Figure 1 shows, using the whole training set, their best accuracies are only close to 0.7 on both MEMset and SYNTHETIC.

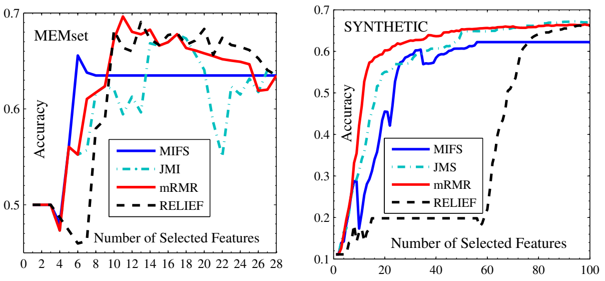

Figure 1: Four Other Feature Selection Algorithms

The comparisons between Group Lasso and GFSFS in accuracy, the # of selected features, and the # of selected groups (the Y axis of two sub-figures in the bottom) are given in Figure 2. On the one hand, the accuracies of both algorithms increase with the number of training instances. On the other hand, we can easily observe that GFSFS clearly outperforms Group Lasso but higher accuracies with less features and groups on both MEMset and SYNTHETIC.

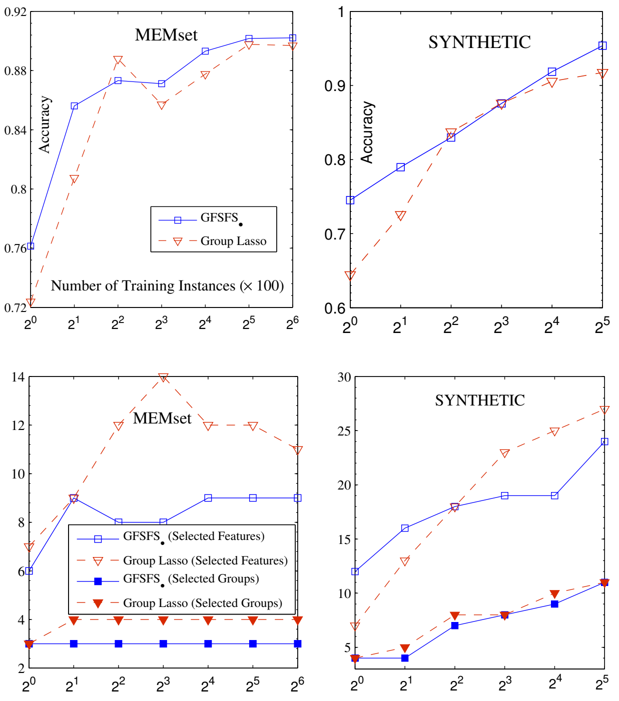

Figure 2: GFSFS vs Group Lasso

4 Conclusion

We have proposed a new online algorithm for group feature selection from feature streams, which performs feature selection at both the group and individual feature levels. Comprehensive experiments on both real-world and synthetic datasets demonstrated that the proposed method can select less features and groups for higher classification accuracies.

Acknowledgments

This work is supported by the US National Science Foundation (NSF) under grant CCF-0905337 and Vermont Agencies of Transportation under grant No. 000025425.

References

Liu, H.; Motoda, H.; Setiono, R.; and Zhao, Z. 2010. Feature selection: An ever evolving frontier in data mining. *J. of Machine Learning Research* 10:4–13.

Wu, X.; Yu, K.; Wang, H.; and Ding, W. 2010. Online streaming feature selection. In *Proc. of the 27th Intl. Conf. on Machine Learning*, 1159–1166.

Yeo, G., and Burge, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *J. of Computational Biology* 11:377–394.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *J. of the Royal Statistical Society, Series B* 68:49–67.