Feature Selection Using Multiple Streams

Paramveer S. Dhillon[1], Dean Foster[2] and Lyle Ungar[1]

[1]CIS, University of Pennsylvania, Philadelphia, PA, U.S.A. [2]Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA, U.S.A.

Grouped Feature Selection

- Data comes in groups (classes).
  - Gene Expression Data ⟶ Gene Families (groups)
  - Word Sense Disambiguation (WSD) ⟶ Adjacent, Previous Words, POS tags etc. (groups)
- Some groups contain highly predictive features; others do not.
- Variety of methods enforce sparsity at level of groups (classes):
  - Group Lasso/ Multiple Kernel Learning (GL/MKL) ⟶ (ℓ_1/ℓ_2) penalty [2].
  - MDL based greedy methods ⟶ (ℓ_0/ℓ_0) penalty (approximate) [1].
- However these methods assume that all the features are known in advance.
- Another line of research deals with streaming feature selection [3].
  - All features need not be known in advance i.e. select features *on-the-fly*.
  - Extremely fast compared to the above "batch" methods

Can we select features *on-the-fly* using the group structure and also dynamically generate new groups?

Our Contribution

1. **A streaming feature selection algorithm which is aware of the group (class) structure.**
   - We propose an extension of Streamwise Feature Selection (SFS) [3] to the case of multiple classes.
     - Divide the wealth (probability of adding spurious features) equally among all the groups (classes) instead of dividing it equally among all the features.
2. **Generate new feature groups (classes) dynamically and select features from them.**
   - Some types of dynamic feature classes that may contain highly predictive features are:
     - Interaction terms of the already selected features with other selected features.
     - Interaction terms of the already selected features with the original set of features.
     - Squares (algebraic) of the selected features.
   - Other types of static classes like PCA, NNMF etc. of original features may also contain "useful" features.

Our Contribution (contd.)

- None of the "batch" methods can consider these set of features for selection as they are not known in advance.
  - The only (and inefficient) way that these methods can consider these features is by considering all the $\binom{p}{2}$ interactions of the original (p) features.

Algorithm

- Our algorithm MSFS selects the next feature from the class which currently has the highest estimated probability of producing a "good feature".

Algorithm 1 MSFS using Alpha-investing

for j = 1 to k **do**
  $w_j = w_0/k$; // initial wealth for j-th class (group)
  $i_j = 1$; // index of features for j-th class
end for
$model = \{\}$; // initially no features in model
while features remain **do**
  // select next class
  $j = argmax_j(w_j/i_j)$; // over all classes with remaining features
  $x = get\_new\_feature(j, i_j)$; // generate new feature on class j
  $\alpha = w_j/2i_j$;
  // is p-value of new feature below threshold?
  if $(get\_p\_value(x, model) \leq \alpha)$ **then**
    // accept
    $add\_feature(x, model)$; // add x to the model
    $w_j := w_j + \alpha_\Delta - \alpha$; // increase wealth
  else
    // otherwise, reject
    $w_j := w_j - \alpha$; // decrease wealth
  end if
  $i_j := i_j + 1$;
end while

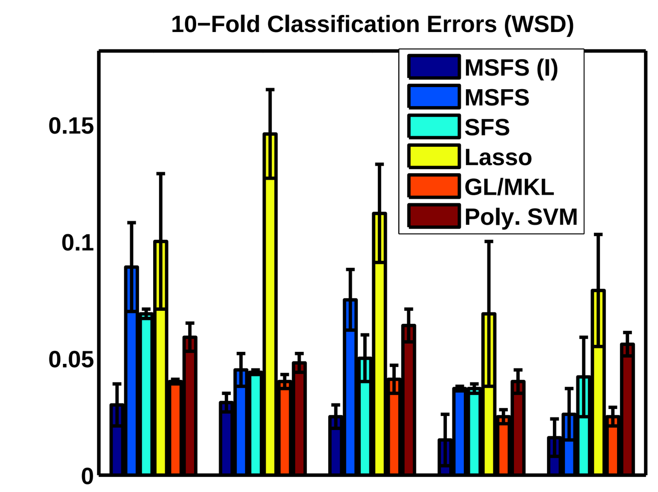
- MSFS has good theoretical properties e.g. in expectation it adds more beneficial features than spurious features.

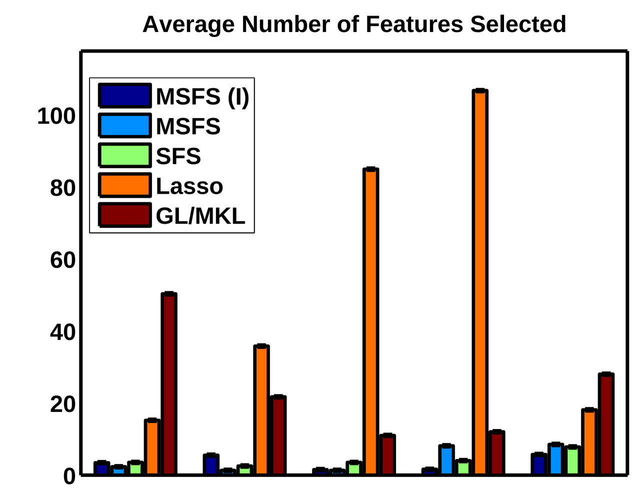
Experimental Results

- Classification Accuracies on WSD (Word Sense Disambiguation) datasets and NIPS 2003 datasets.
  - We consider two versions of MSFS, one without the dynamic interaction terms (MSFS) and one with interaction terms (MSFS (I)).

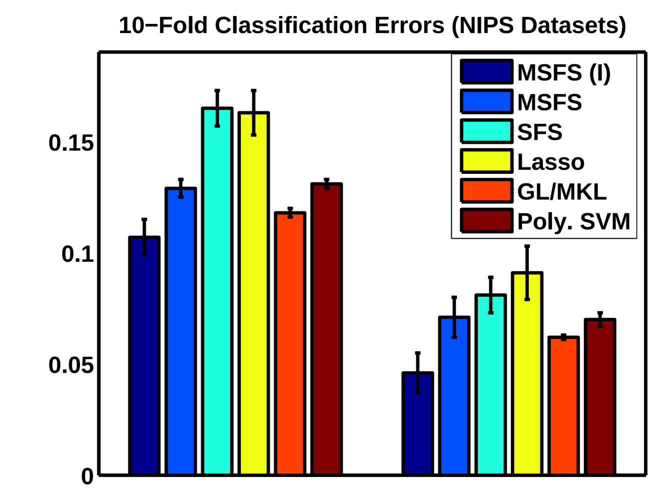
Experimental Results (Contd.)

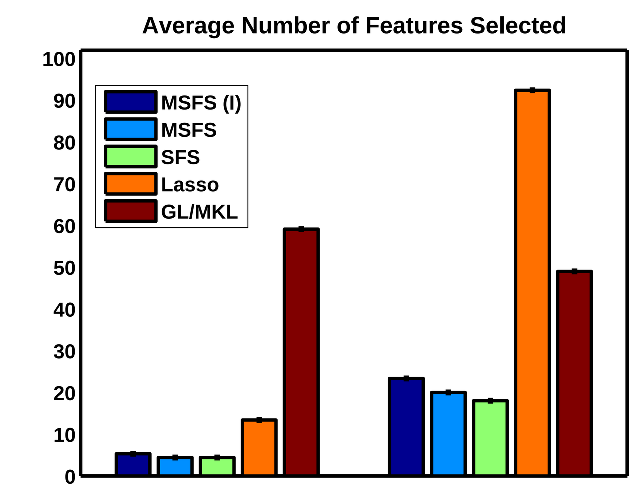
- All datasets (5 shown below) were augmented with extra static feature classes (groups) for PCA and "squares" of original features.





- NIPS 2003 data (2 shown below) did not contain feature classes to begin with so we artificially created classes by clustering the features. (Details in the paper.)





- Time complexity of MSFS
  - Orders of magnitude faster than "batch" algorithms. (Plots are in the paper.)

Summary

- **MSFS extends streaming feature selection to the case of multiple feature classes (groups)**
  - Straightforward extension of SFS to incorporate group structure.
  - It also allows dynamic generation of feature groups (classes).
  - Extremely computationally efficient compared to competing methods.
- Performs significantly better than state-of-the-art "batch" methods in terms of predictive accuracy and run time.

References

[1] P. S. Dhillon, D. Foster, and L. Ungar. Efficient feature selection in the presence of multiple feature classes. In *ICDM*, pages 779–784, 2008.

[2] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *JRSS: Series B*, 68(1):49–67, 2006.

[3] J. Zhou, D. P. Foster, R. A. Stine, and L. H. Ungar. Streamwise feature selection. *JMLR*, 7:1861–1885, 2006.