Combining clustering of variables and random forests for high-dimensional supervised classification

M. Chavent\textsuperscript{1,2}  R. Genuer\textsuperscript{1}  J. Saracco\textsuperscript{1,2}

\textsuperscript{1} Bordeaux University, \textsuperscript{2} INRIA Bordeaux Sud-Ouest

August 28

COMPSTAT 2012
Introduction

Main goal:
- dimension reduction for high-dimensional data
- supervised classification framework

Proposed methodology:
1. clustering of variables (using \textit{ClustOfVar})
2. selection of the most important synthetic variables (using \textit{VSURF} a variable selection procedure based on random forests)

Main idea:
- eliminate redundancy before applying a variable selection method
Framework

\( \mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) i.i.d. random vectors with same distribution as \((X, Y)\)

\( X \in \mathbb{R}^p \): (quantitative) variables

\( Y \in \{-1, +1\} \): (dichotomous) response

Typically, \( n << p \)

Examples: gene expression data, fMRI data...
ClustOfVar (Chavent et.al. 2012)

Generality:
- lumps together strongly related variables into clusters of variables
- computes synthetic variables associated to each cluster
- useful for case studies and dimension reduction

Tools:
- homogeneity criterion based on squared correlations (hence, synthetic variable = 1st principal component of PCA applied to the cluster variables)
- hierarchical clustering algorithm (Ward like)
- k-means type partitioning algorithm
VSURF (Genuer et.al. 2010)

VSURF = Variable Selection Using Random Forests

Based on Random Forests (Breiman 2001)
- nonparametric statistical learning method
- aggregation of a collection of classification trees
- trees constructed on bootstrap samples with randomly drawn variables

Interesting outputs of RF
- Out-Of-Bag (OOB) error: prediction error estimation
- Variable Importance (VI) score: helps to determine which variables explain the most the response

⇒ VSURF uses both OOB error and VI for variable selection
1. Introduction

2. Toy example

3. Real data
Toys data (Weston et.al. 2003)

Original dataset: 6 true variables and 194 noise variables

- two independent groups of 3 true variables, related with $Y$
- within each group, true variables are mediumly correlated with each other
- noise variables, independent with $Y$

Modified dataset:
for each 6 true variables and 6 noise variables: addition of 10 highly correlated variables

$\Rightarrow$ 12 groups of 11 highly correlated variables, and 188 noise variables

$n = 200 \quad p = 320$
Comparison of ClustOfVar and ACP using RF and LDA

Test error rate (in %)

Number of synthetic variables
ClusOfVar combined with VSURF for toys data

Test error rate (in %)

Number of synthetic variables

CoV−RF
CoV−VSURF
CoV−VSURF−vars
RF with all vars
RF with true vars
Prostate data (Stephenson et.al. 2005)

- **Gene expression data** in a cancer study
- 79 treated patients:
  - 37 recurrent primary prostate tumor
  - 42 non-recurrent
- 7884 gene expressions

\[ n = 79 \quad p = 7884 \]
Prostate data (Stephenson et.al. 2005)

- Gene expression data in a cancer study
- 79 treated patients:
  - 37 recurrent primary prostate tumor
  - 42 non-recurrent
- 7884 gene expressions
  \[ n = 79 \quad p = 7884 \]

Clustering of the 7884 variables:

1. kmeansvar with 120 clusters of variables
2. hclustvar on the 120 corresponding synthetic variables
Comparison of ClustOfVar and ACP using RF and LDA

Test error rate (in %)

Number of synthetic variables

CoV−RF
ACP−RF
CoV−LDA
ACP−LDA
ClustOfVar combined with VSURF for Prostate data

Test error rate (in %)

Number of synthetic variables

CoV−RF
CoV−VSURF
CoV−VSURF−vars
RF with all vars
Results:

- Choice of the number of synthetic variables using OOB error rates $\Rightarrow$ **13 synthetic variables** for Prostate
- VSURF then selects 4 synthetic variables corresponding to **516 original variables** (gene expression)
  $\Rightarrow$ we discard **93.5%** of the variables.

Remarks:

- Eliminate redundancy before variable selection seems to bring some **benefits for prediction**.
- R packages: ClustOfVar (available), VSURF (in construction)

Perspective:

- **516 variables** may still be too large.
  $\Rightarrow$ Select variables within each selected cluster ?
  Select variables directly among the 516 variables ?
Références


ClusOfVar combined with VSURF for Prostate data

OOB error rate (in %)
Number of synthetic variables
CoV−RF
CoV−VSURF
Comparison with ClustOfVar combined with LDA

Test error rate (in %)

Number of synthetic variables

CoV–RF
CoV–VSURF
CoV–LDA
CoV–LDA–varselect
Comparison of the two variable selection procedure

Number of selected variables vs. Number of synthetic variables for CoV-VSURF and CoV-LDA-varselect.
ClustOfVar (Chavent et.al. 2012)

Homogeneity measure $\mathcal{H}$ of a partition $P = (C_1, \ldots, C_K)$:

$$\mathcal{H}(P) = \sum_{k=1}^{K} H(C_k)$$

with

$$H(C_k) = \sum_{X^j \in C_k} r^2(X^j, c_k)$$

$c_k$ is a synthetic variable

$r^2$ is squared correlation
ClustOfVar (Chavent et.al. 2012)

Homogeneity measure $\mathcal{H}$ of a partition $P = (C_1, \ldots, C_K)$:

$$\mathcal{H}(P) = \sum_{k=1}^{K} H(C_k)$$

with

$$H(C_k) = \sum_{X_j \in C_k} r^2(X_j, c_k)$$

$c_k$ is a synthetic variable

$r^2$ is squared correlation

Synthetic variable of a cluster:

$$c_k = \arg \max_{u \in \mathbb{R}^n} \left\{ \sum_{X_j \in C_k} r^2(X_j, u) \right\}$$

$\Rightarrow c_k = 1$st principal component of PCA applied to the cluster
Classification tree

Tree: piecewise constant predictor obtained by dyadic recursive partitioning of $\mathbb{R}^p$

At each step of the partitioning process, we seek for the "best" split of the data from $\mathcal{L}_n$


Figure: Classification tree
Random Forests

- Introduction
- Toy example
- Real data

```
\begin{align*}
\mathcal{L}_n & \xrightarrow{\text{Bootstrap}} \mathcal{L}_{n}^{\Theta_1} \\
\mathcal{L}_{n}^{\Theta_1} & \xrightarrow{\text{RI Tree}} \hat{h}(., \Theta_1, \Theta'_1) \\
\mathcal{L}_n & \xrightarrow{\text{Agregation}} \hat{h}_{RF-RI}(.) \\
\mathcal{L}_{n}^{\Theta'_1} & \xrightarrow{\text{}} \hat{h}(., \Theta_1, \Theta'_1) \\
\mathcal{L}_n & \xrightarrow{\text{}} \hat{h}(., \Theta_q, \Theta'_q)
\end{align*}
```