

Multivariate Analysis for Predicting Risk of Microbial Contamination of Food

Daria Sorokina
Lujie Chen
Artur Dubrawski

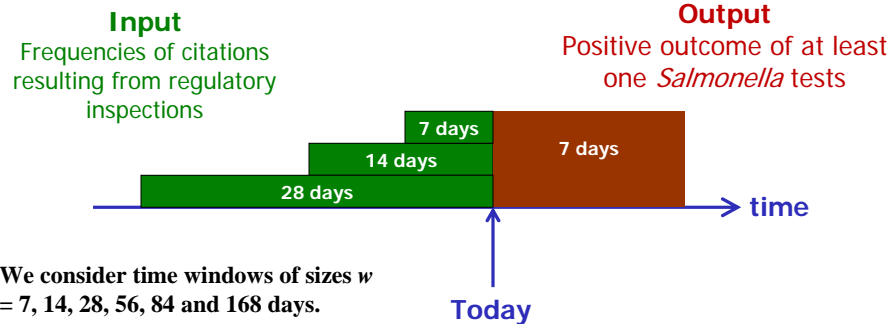
Carnegie Mellon

Auton
Lab

Focus of Our Application: *Salmonella*

- Goal: Estimate risk of positive results of microbial tests in the near future, based on results of regulatory inspections conducted at a food establishment in the recent past

Preparation of Data for Multivariate Analysis



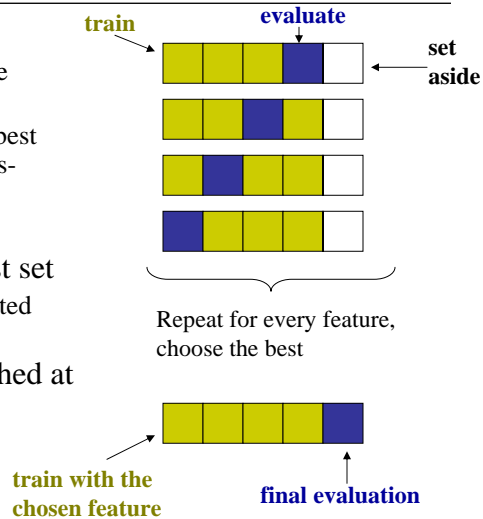
Method of Choice: Logistic Regression

$$p(y = 1 | x_1, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m)}}$$

- Well established
- “White-box”: magnitudes of the regression coefficients characterize the influence of the corresponding features
- Estimated equations can be directly plugged into risk estimation algorithms

Feature Selection

- Wrapper forward selection
 - at each iteration try to add one feature
 - choose the one that provides best results on internal 4-fold cross-validation
- Evaluation on a set-aside test set
 - The whole procedure is repeated several times
- The top performance is reached at ~ 15 features

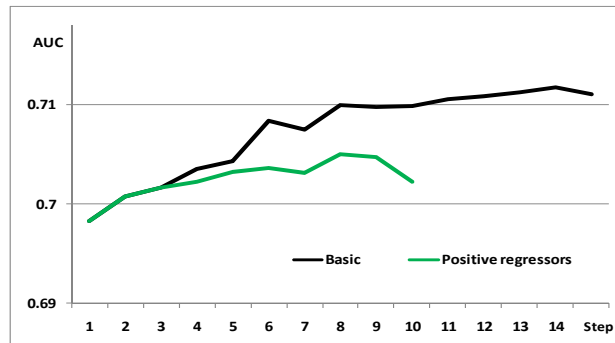


Comparison with Human Expert



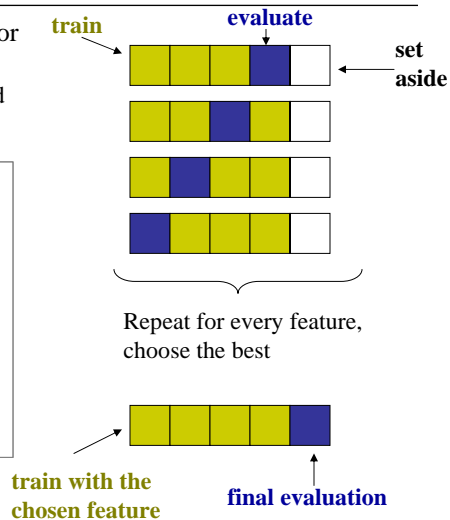
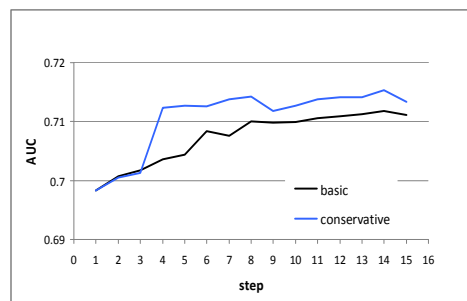
Constraining Logistic Regression

- Food safety analysts are primarily interested in understanding and monitoring the risk-increasing factors
- Constraint: allow the algorithm to use only positive coefficients



Constraining Logistic Regression

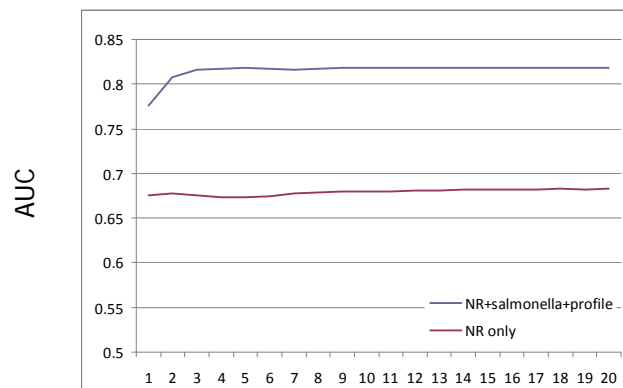
- Constraint: use conservative metric for feature selection
 - Penalize average score by 2 standard deviations



Additional Sets of Features

- Salmonella history
 - past occurrences of Salmonella
- Plant profile features
 - types of meat processed
 - production volumes
- Geographical features
 - coordinates
 - regions

Improvement from Extra Features



Analyzing Model Complexity

- New features might introduce interactions
 - e.g., different NRs are important in different regions
- Logistic regression does not handle interactions by default
 - Complex black-box models do
 - But we can't use them in this application
- If we know specific interactions, we can use them in LR
 - introduce new components
 - build several LR models
- Need to detect interactions

Interaction Detection with Additive Groves

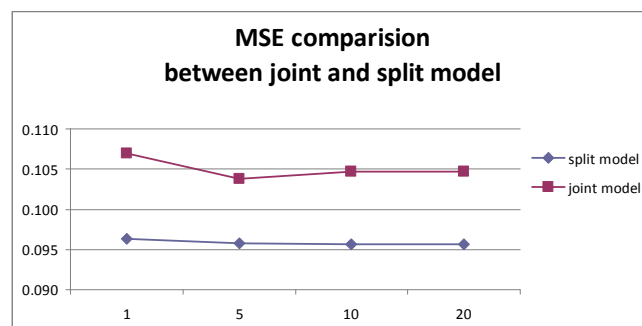
- Powerful predictive black-box algorithm
 - Third place at ICDM'09 Data Mining Contest
- Based on
 - Additive Models
 - Bagged Decision Trees
- Useful for interaction detection
 - Captures interactions
 - Easy to restrict specific interactions
- Available at
 - www.cs.cmu.edu/~daria/TreeExtra.htm

Interaction Detection Results

- Detected 5 interactions
- 4 of them included slaughter_chicken variable

- Decision – split the data based on slaughter_chicken value
 - Build two LR models: one for plants that slaughter chickens one for plants that do not

Two LR Models vs. Joint Model



Different Sets of Features

Chicken slaughter present

past_Salmonella_w84
Meat_Processing
Citation_xxx_w56
region_Mid_Atlantic
past_Salmonella_w28
Citation_xxx_w168
region_West_North_Central
region_West_South_Central
Citation_xxx_w28
Citation_xxx_w7

Chicken slaughter absent

past_Salmonella_w168
slaughter_Cattle
aggr.Citation_xxx_w84
slaughter_Turkey
Citation_xxx_w168
past_Salmonella_w14
Citation_xxx_w168
aggr. Citation_xxx_w84
Meat_Slaughter
Citation_xxx_w56

Summary

- Logistic regression is a useful white-box algorithm for this application
 - Improves over human experts
 - Easy to analyze
- We analyzed several ways to improve it:
 - Forward feature selection with conservative metric
 - Choosing only positive coefficients
 - Introducing interactions discovered by a black-box model