Tasks

(A) Analyze probability distributions (of samples in a class)
Both overall distribution and sub-distributions by classification results

(B) Select samples
By probabilities, classification results, data features, certain class confusions, ...

(C) Analyze separability (of correct and incorrect classifications)
Can certain FP/FN be separated from TP/TN by data features?

Usage Scenarios

(A) Analyze probability distributions
(B) Select samples
(C) Analyze separability

Analyzing Classifier Behavior
Neural Network
Naive Bayesian
k-NN (k=2)
Based on Task A (analyzing probability distribution)

Visual Inspection
Based on Task B (interactive sample selection)
Visual Inspection

TPs > 90% are filtered

Defining Post-Classification Rules

Rectify certain false negatives

\[ \mathbb{P}(y|X) \cdot \mathbb{P}(X|y) = \mathbb{P}(X) \]

Rectify certain false positives

\[ \mathbb{P}(y|X) \cdot \mathbb{P}(X|y) = \mathbb{P}(X) \]

Based on Task C (separability analysis)

Limitations

Visual Complexity
Too much information
Needs extensive training

Scalability in # of classes
Issues with the radial layout
Arcs instead of straight bars
Histograms with no base lines

Conclusion

Classification probabilities
Rich of information
Explain classifier behavior

Interactive exploration
Reveals several insights
Guides new improvements

http://www.cvast.tuwien.ac.at/ConfusionAnalysis

Summary

- What?
  - High-dimensional probabilistic classification data

- Why?
  - Task: Analyze probability distribution, select samples, analyze separability

- How?
  - Stacked histogram (bar), wheel layout, boxplot, multiple-view

- Scale
  - # classes up to 20, # samples up to tens of thousands.

Thanks!