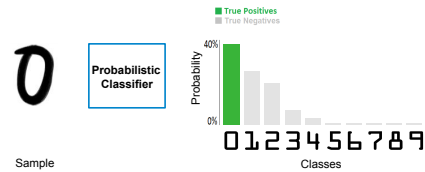


Visual Methods for Analyzing Probabilistic Classification Data

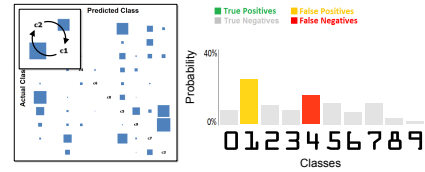
Bilal Alsalakh, Allan Hanbury, Silvia Miksch, Andreas Rauber, Helwig Hauser

Presented by Yinggal Dong
Information Visualization
UBC Computer Science
Nov. 2014

Probabilistic Multi-class Classification



Probabilistic Multi-class Classification



Analyzing Probabilistic Classification Data

Tasks

(A) Analyze probability distributions (of samples in a class)

Both overall distribution and sub-distributions by classification results

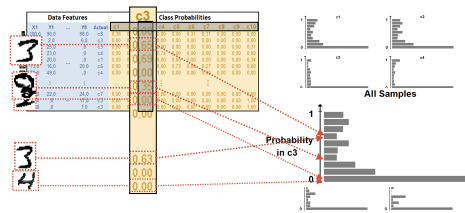
(B) Select samples

by probabilities, classification results, data features, certain class confusions, ...

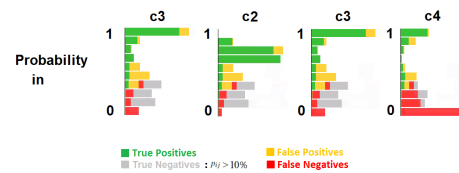
(C) Analyze separability (of *correct* and *incorrect* classifications)

can certain *FPs* / *FNs* be separated from *TPs* / *TNs* by *data features*?

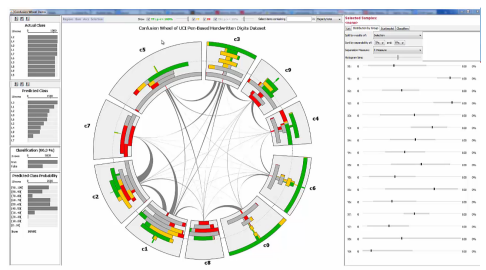
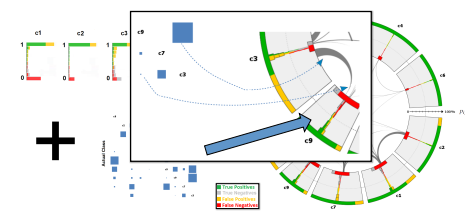
Task A: Analyze Probability Distributions



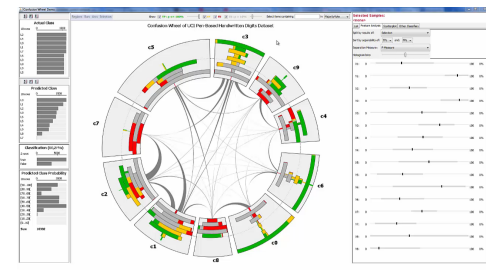
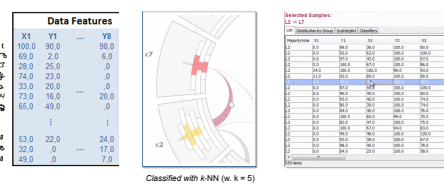
Probability Sub-Histograms



Augmenting Class Confusion



Task B: Interactive Selection

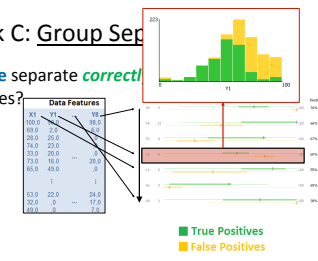


Task C: Group Separation

Can a *data feature* separate *correct* classified samples?

Sort features by separability:

- p -value

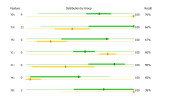


Tasks

(A) Analyze probability distribution



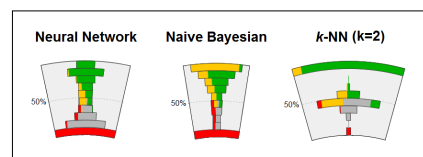
(B) Select samples



(C) Analyze separability

Usage Scenarios

Analyzing Classifier Behavior



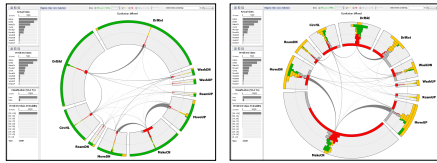
Based on Task A (analyzing probability distribution)

Visual Inspection



Based on Task B (interactive sample selection)

Visual Inspection



TPs > 90% are filtered

Defining Post-Classification Rules

Rectify certain **false negatives**

$$R_{s_j} : q_1(s) \wedge \dots \wedge q_k(s) \wedge P(s) \neq c_j \Rightarrow P(s) \leftarrow c_j$$



Rectify certain **false positives**

$$R_{s_j} : q_1(s) \wedge \dots \wedge q_k(s) \wedge P(s) = c_j \Rightarrow P(s) \leftarrow c_j : P(s, c_j) = 2$$



Based on Task C (separability analysis)

Limitations

Visual Complexity

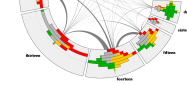
Too much information
Needs extensive training

Scalability in # of classes



Issues with the radial layout

Arcs instead of straight bars
Histograms with no base lines



Conclusion

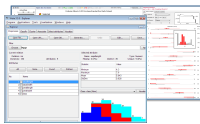
Classification probabilities

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-views.
- Scale
 - # Classes up to 20. # samples up to tens of thousands.

Thanks!