Automatic Sample-by-sample Model Selection
Between Two Off-the-shelf Classifiers

Steve P. Chadwick

University of Texas
at Dallas
Model Selection by Predicting the Better Classifier

Idea:
- Two classifiers, "primary" and "secondary"
- Use confidence to predict which one is expected to perform best

Pima Indian Diabetes

Primary classifier: Fisher LD

Fisher LD classifies over 70% of the data before half the total error is accumulated.

Secondary classifier: 1-Nearest Neighbor

1-Neigh classifies about 50% of the data when half the total error is accumulated.
Primary classifier: Fisher LD

Fisher LD classifies over 60% of the data before half the total error is accumulated.

Secondary classifier: Nearest Unlike Nei.

NUN classifies about 60% of the data before half the total error is accumulated.
Confidence measure profiles

(Ljubljana Breast Cancer)

Fisher LD using $1/(w'x+s)$

1-Nearest Neighbor using 1-Neighbor distance

MSE using $Q$

1-Nearest Neighbor using distance from centers

1-Neighbor using nearest unlike neighbor ratio
Differential error
---------------------
(Ljubljana Breast Cancer)
30% class A, 70% class B

Differential error
---------------------
(Pima Indian Diabetes)
35% class A, 65% class B

Differential error
---------------------
(Synthetic Data)
50% class A, 50% class B
Obstacles

- About 25% of the training data contributes to calculating the selection LD when combining linear discriminant and nearest neighbor classifiers.

![Selection LD and data in q-space](image)

- The different confidence measures have different ranges, which makes them difficult to compare with each other.