Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification

J. Bala, J. Huang, H. Vafaie, K. DeJong, H. Wechsler
George Mason University Presented by

Sushil J. Louis
Department of Computer Science
University of Nevada
Reno, NV
sushil@cs.unr.edu
`http://gaslab.cs.unr.edu`
`http://www.cs.unr.edu/ sushil`

Outline

- Hybrid Approaches
- Genetic Algorithms
- Decision Trees
- Feature selection
- Facial image data and Satellite images
- Results

Pattern Classification

- Difficult but fundamental task
- Depends heavily on the particular choice of features used by classifier

  Rank features according to some statistic

  Eliminate low-ranked features

$\leadsto$ Ranking is not enough:

  Non-linear interactions between features

  Does not measure effectiveness on the complete task

Feature Selection

- First step in any classification task
- Feature Selection is crucial for classification performance
- Reduces amount of data
- Increases classification accuracy
- Very difficult: Number of features is very large, space of possible subsets is exponential in number of features ($100\text{features} \rightarrow 2^{100}\text{subsets}$).

Optimal Feature Selection

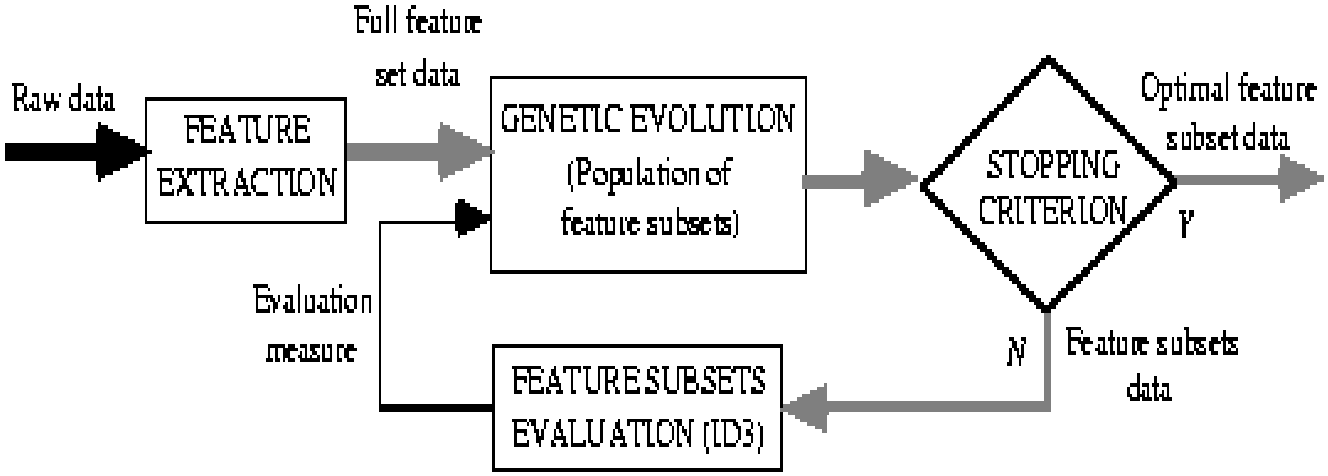
- Minimize number of features
- Improve classification rates

- GA + DT system: GA searches the space of all possible subsets of a large set of candidate discrimination features for an optimal feature subset
  1. For a given subset, ID3 produces a DT
  2. Accuracy on unseen data $\rightarrow$ fitness to GA
  3. GA evolves better subsets

  Stop when performance passes a threshold

System Architecture

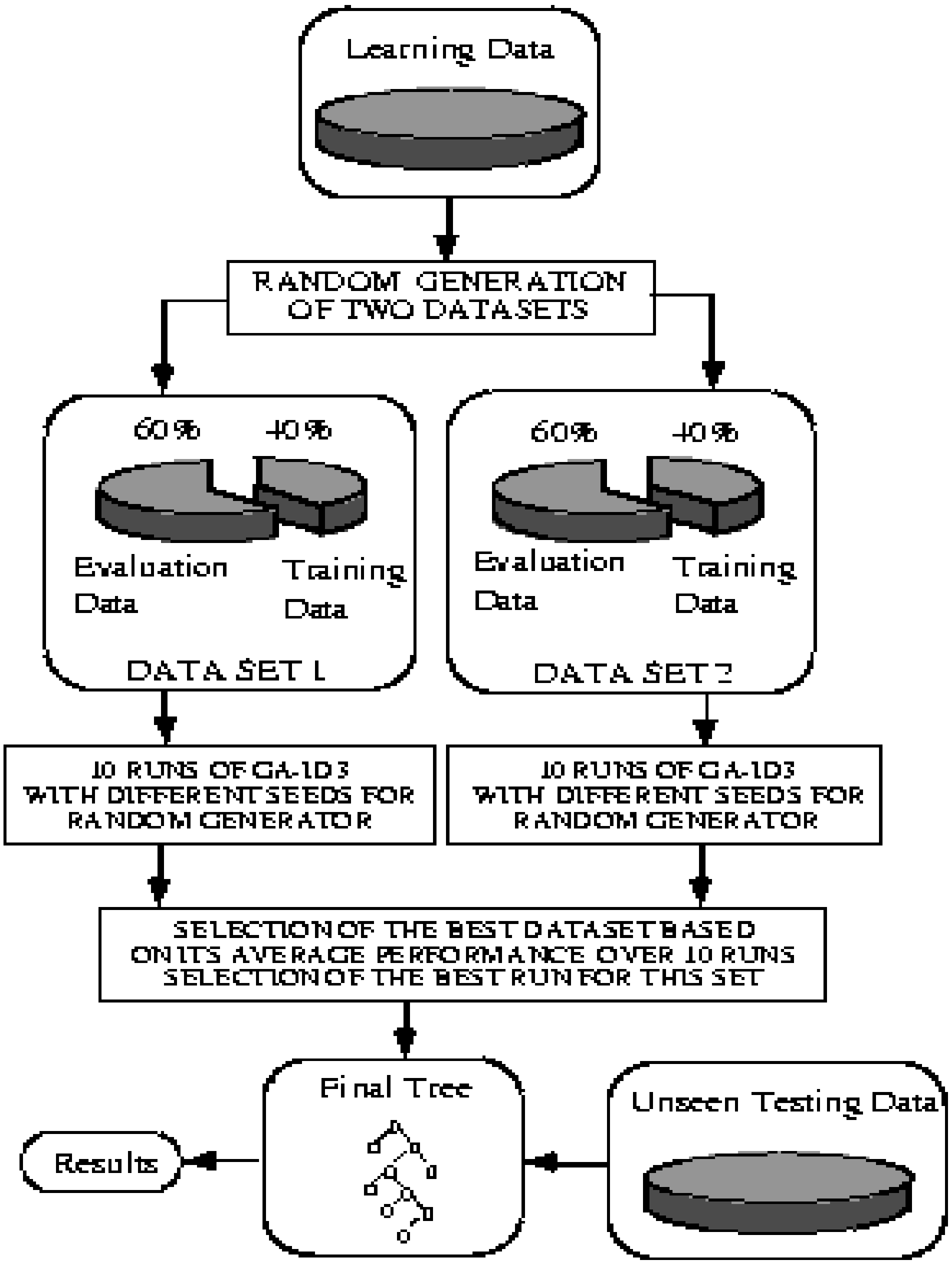
- Initial (large) feature set
- Run GA+ID3
- Recommend best feature set



GA Representation

- Fixed length binary string
- i^{th} position $== 1 \rightarrow i^{th}$ feature is included.
- Genetic Algorithm modifies a population of such strings.
- For each individual in the population
  - Run ID3 on training set containing only these features
  - Evaluate resulting DT on test data
  - Return accuracy to GA as fitness of individual

Methodology



Methodology (cont'd)

- Standard GA, popsize 50, standard parameter values
- C4.5, standard parameter values
- 2-fold Cross-validation
- 40% training 60% testing
- 10 runs of the system with different seeds
- Best tree out of ten is applied to testing data
- Compare
  1. Full feature set
  2. Same number of features but use ranking
- Satellite data (36 features total)
- Facial data (105 features total)

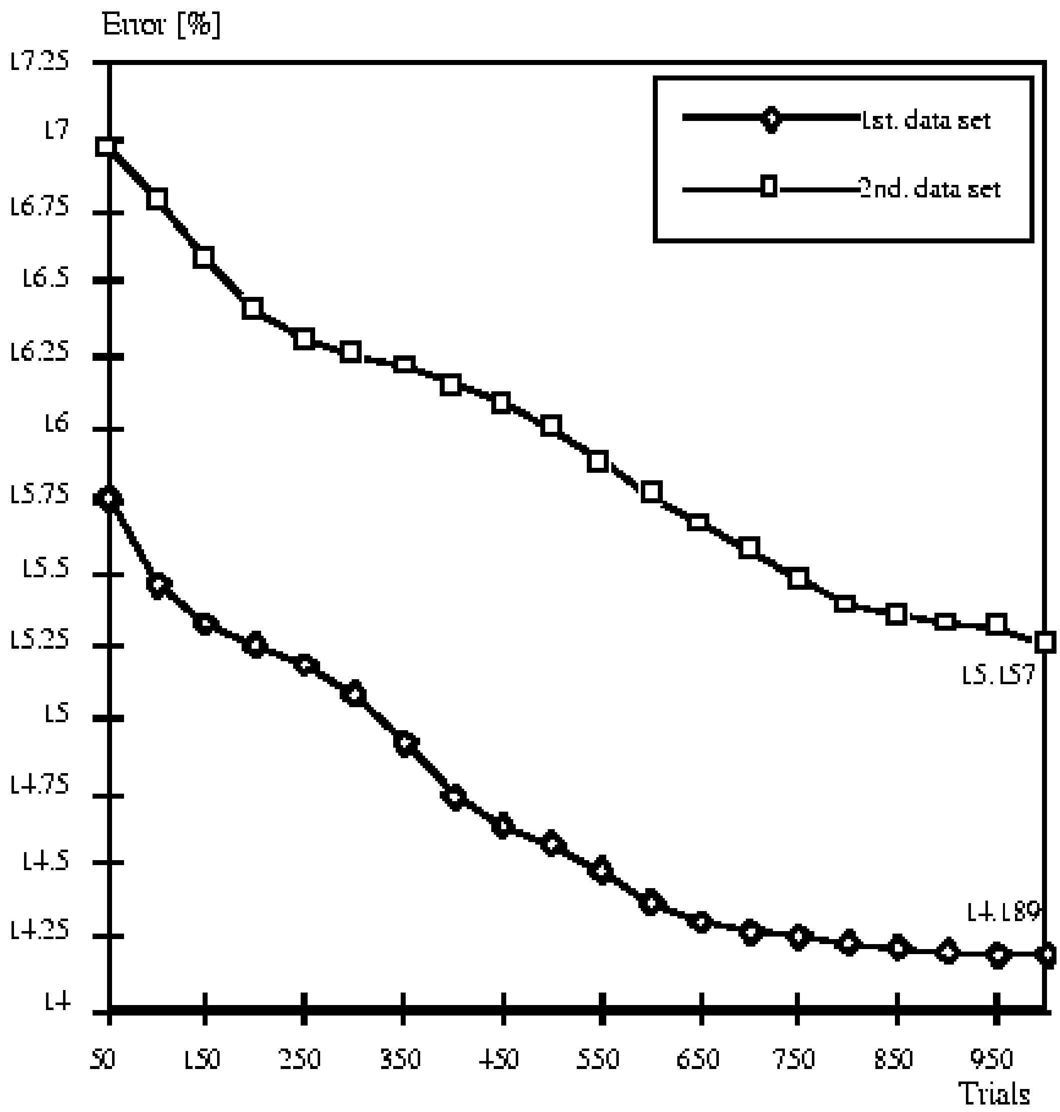
Satellite Data

- Predict central pixel classification based on 3x3 neighborhood multi-spectral values
- Nasa Landsat images
- 6 classes
- 4 spectral bands, 9 pixel values $\rightarrow$ 36 features
- Feature/pixel values are $0 - 255$
- Learning/training set $=$ 4435 examples, test data $=$ 2000 examples

Results

FULL FEAURE SET 36 Features		REDUCED FEATURE SET BY THE GA-ID3 SYSTEM 17 Features		REDUCED FEATURE SET BY FEATURE RANKING 17 Best Features Chosen	
Error Rate	Tree Complexity	Error Rate	Tree Complexity	Error Rate	Tree Complexity
18.5%	99 nodes	16.9%	72 nodes	20.6%	83 nodes

Results



Error [%]
17.25
17
16.75
16.5
16.25
16
15.75
15.5
15.25
15
14.75
14.5
14.25
14
1st. data set
2nd. data set
15.157
14.189
50
150
250
350
450
550
650
750
850
950
Trials

Facial Data

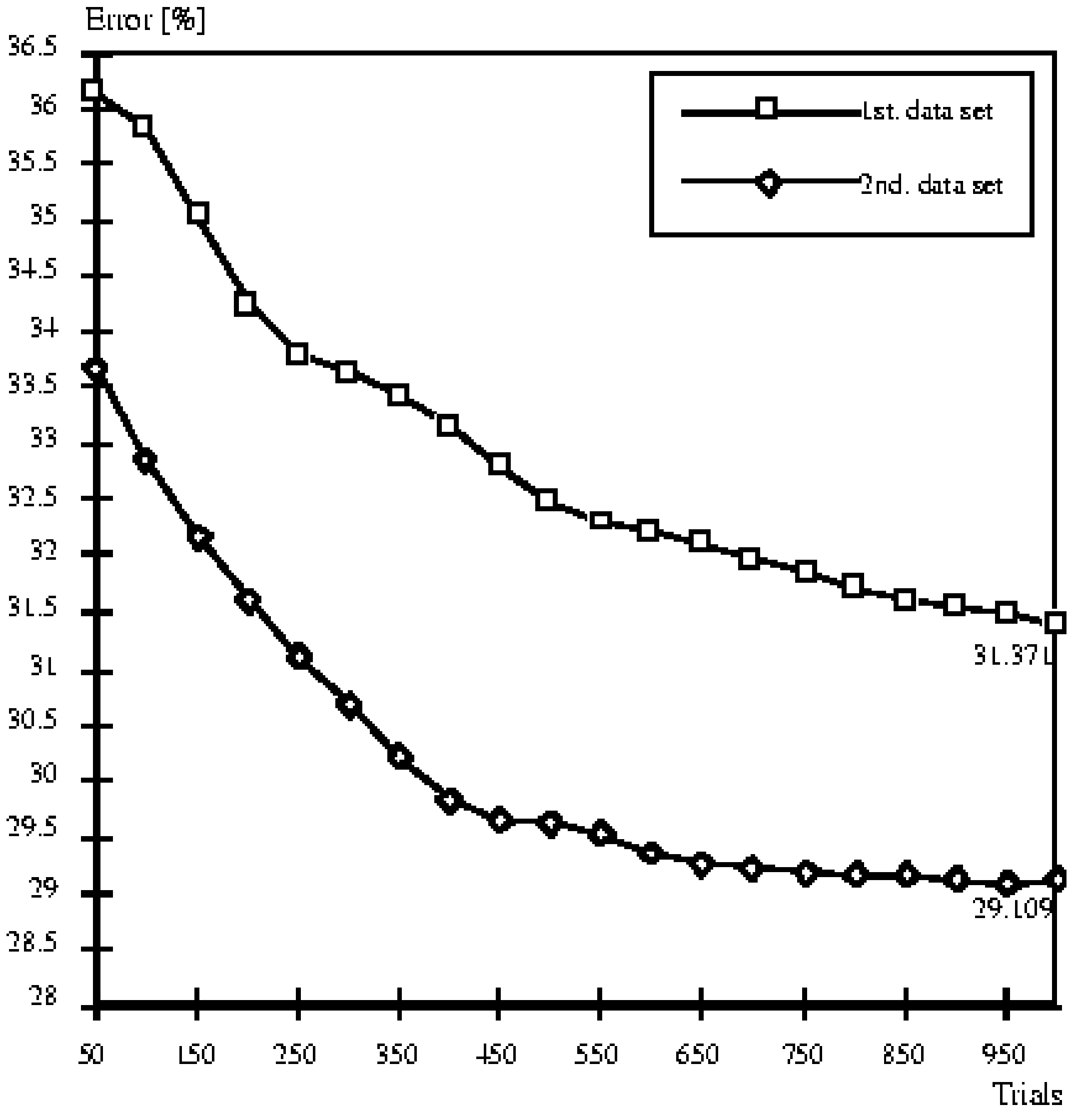
- Detecting eyes very important for face recognition
- 102 eyes, 102 nose, 102 other facial region examples for training
- 52 eyes, 52 nose, 52 other facial region for testing
- 105 features for each example
- Feature values between $0 - 255$
- Three classes

FEATURES	DESCRIPTION
From x_1 to x_{35}	Obtained by averaging a 4x4 window which has 50% (2 pixels) overlap in each shift.
From x_{36} to x_{70}	They represent the standard deviation of a 4x4 window.
From x_{71} to x_{105}	They represent the entropy for a 4x4 window.

Results

FULL FEATURE SET *105 Features*		REDUCED FEATURE SET BY THE GA-ID3 SYSTEM *41 Features*		REDUCED FEATURE SET BY FEATURE RANKING *41 Best Features Chosen*	
Error Rate	Tree Complexity	Error Rate	Tree Complexity	Error Rate	Tree Complexity
38.4%	8 nodes	27.5%	13 nodes	38.5%	11 nodes

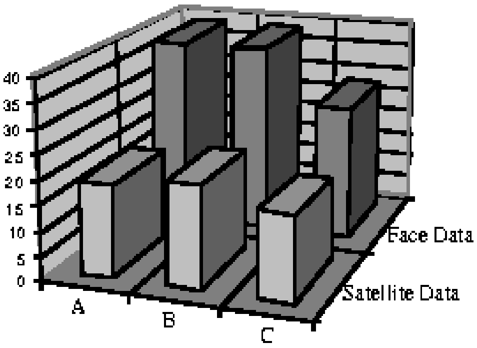
Face Results



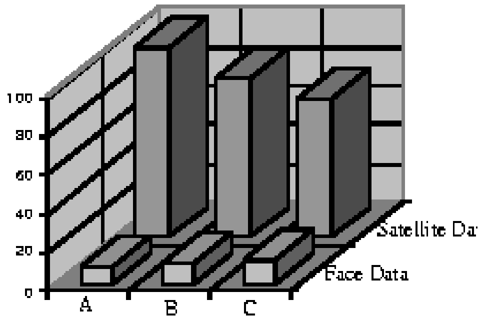
Results

- Better error rates
- Reduced tree complexity for Satellite
- 60% reduction in facial features, 52% in satellite

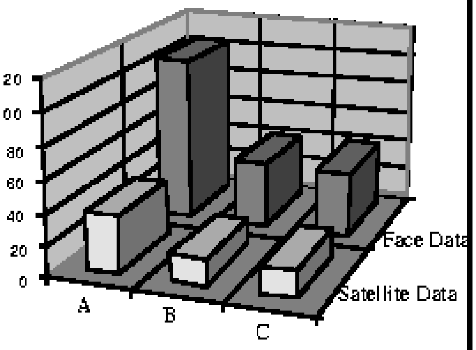
Results



Recognition Error Rates in Percentages



Description Complexities Expressed as the Number of Nodes



Number of Features in Sets

A: Full Feature Set
B: Reduced by Independent Ranking
C: Reduced by the GA-ID3 System

Conclusions and Future Work

- These preliminary results are encouraging
- Larger more complex problems in image processing
- Use information other than classification accuracy as GA fitness