UCRL-PRES-206158

Feature Selection in Scientific Applications

Erick Cantú-Paz, Shawn Newsam, and Chandrika Kamath

Center for Applied Scientific Computing

Lawrence Livermore National Laboratory

cantupaz@llnl.gov

www.llnl.gov/casc/sapphire

Summary

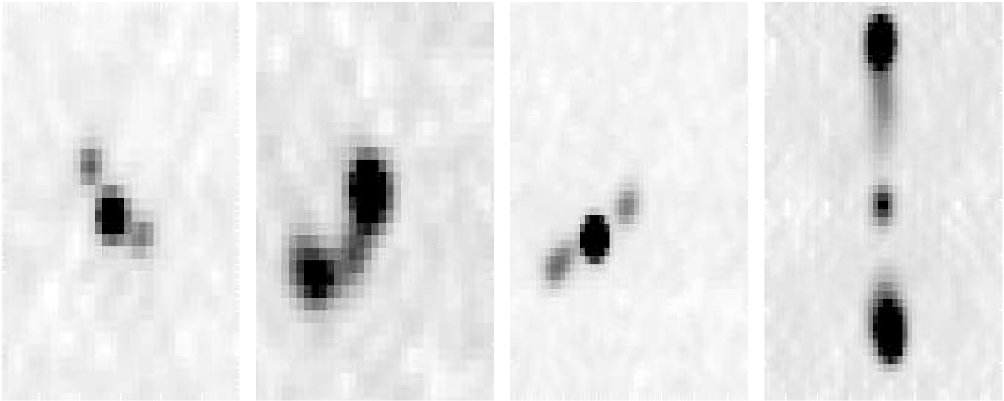
- We applied feature selection methods to three applications:
  - classification of galaxies
  - retrieving interesting objects from image databases
  - detection of human settlements in remote sensing images
- We want to build models that discriminate between objects of different classes
- Model builders are sensitive to irrelevant and redundant features
- Domain information and exploratory analysis have limitations

Feature Selection Algorithms

- Filters
  - KL distance between histograms
  - Chi-Square from contingency tables
  - Stump using Gini index
  - PCA filter
- Wrappers
  - Sequential Forward Selection (SFS)
  - Sequential Backward Elimination (SBE)
- Accuracy estimated by 10-fold crossvalidation and Naive Bayes

FIRST Astronomical Survey

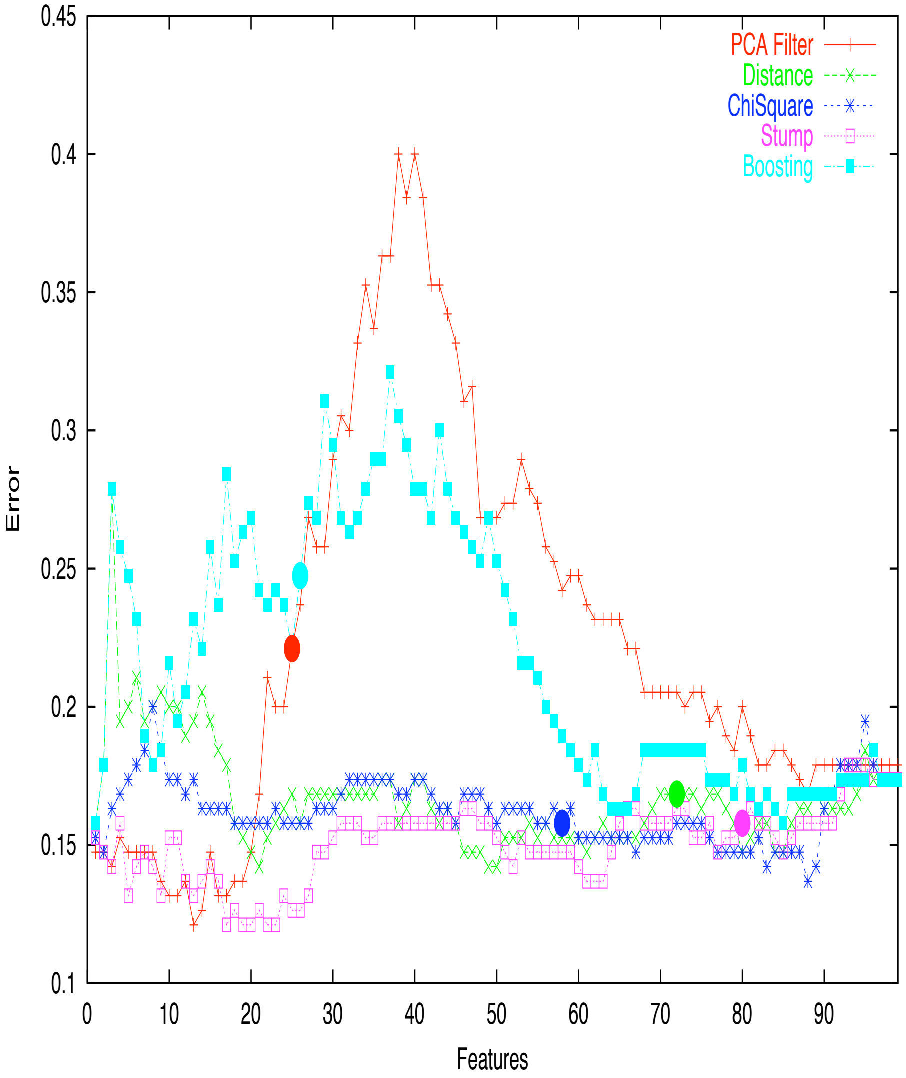
- Astronomers interested in identifying bent-double galaxies
- Objective is to maximize accuracy of classification
- 99 features were extracted
  - Several measures of “bentness” and symmetry
  - Unclear which one(s) are preferable



- Only 195 labeled instances available

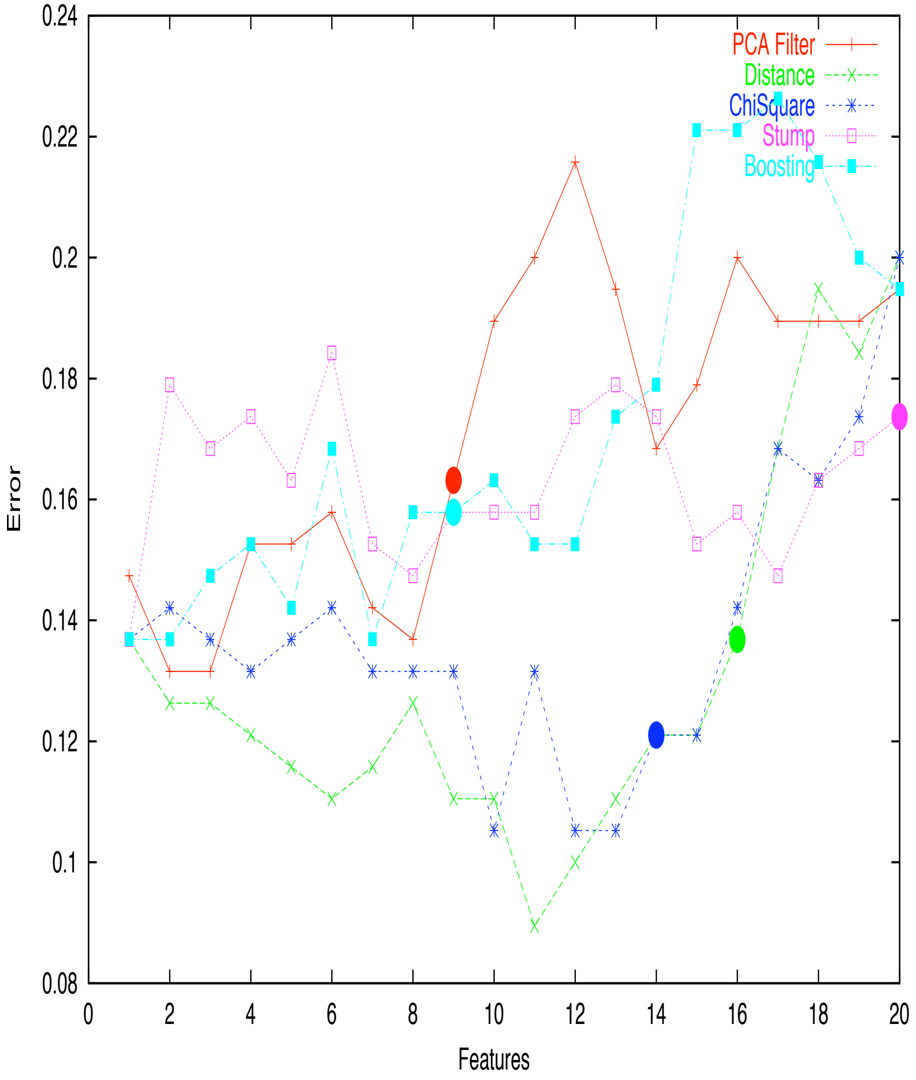
FIRST Results I

- Using all 99 available features as input



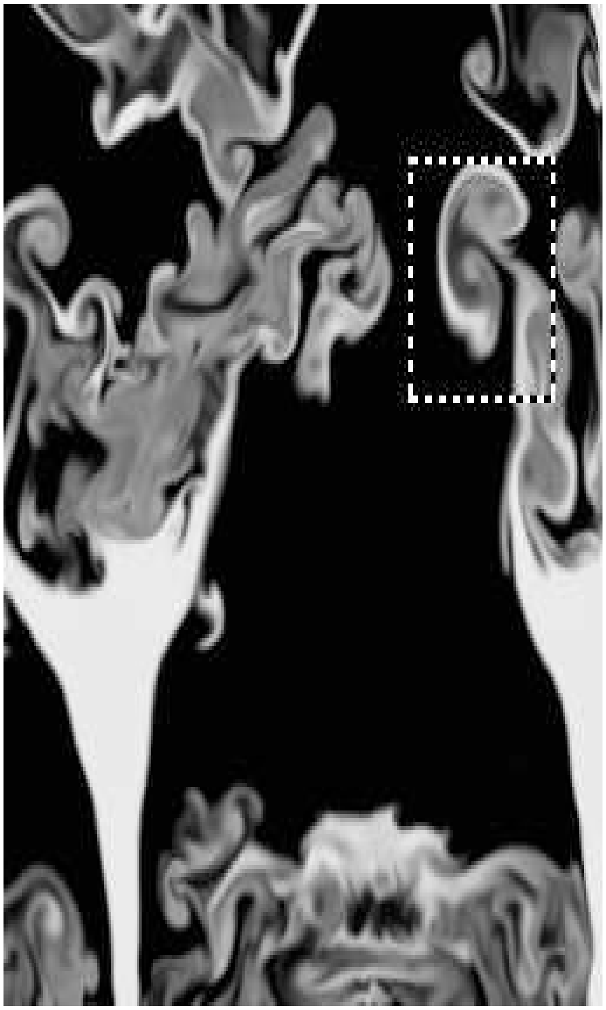
FIRST Results II

- Using 20 features that depend on three blobs



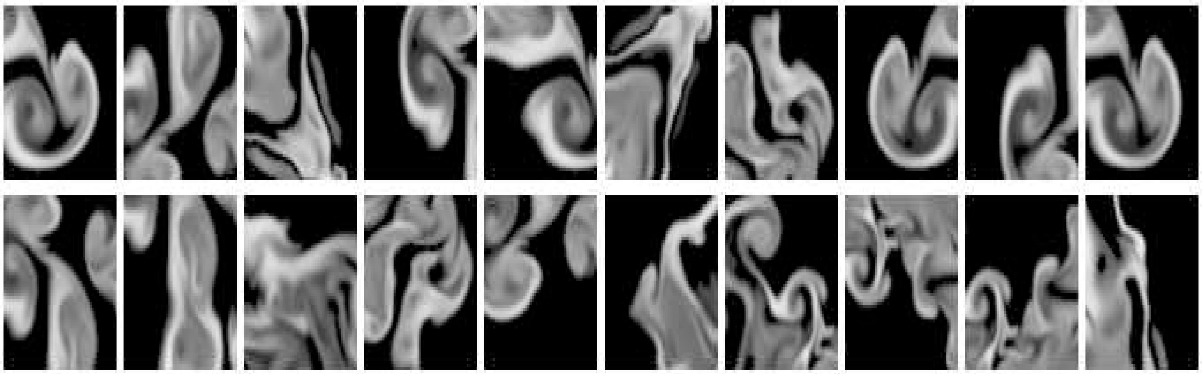
Similarity Based Object Retrieval

- Find objects similar to a query
- Objects are described by texture and shape features



Similarity Search

- Compute distance between vector of features describing query and image tiles

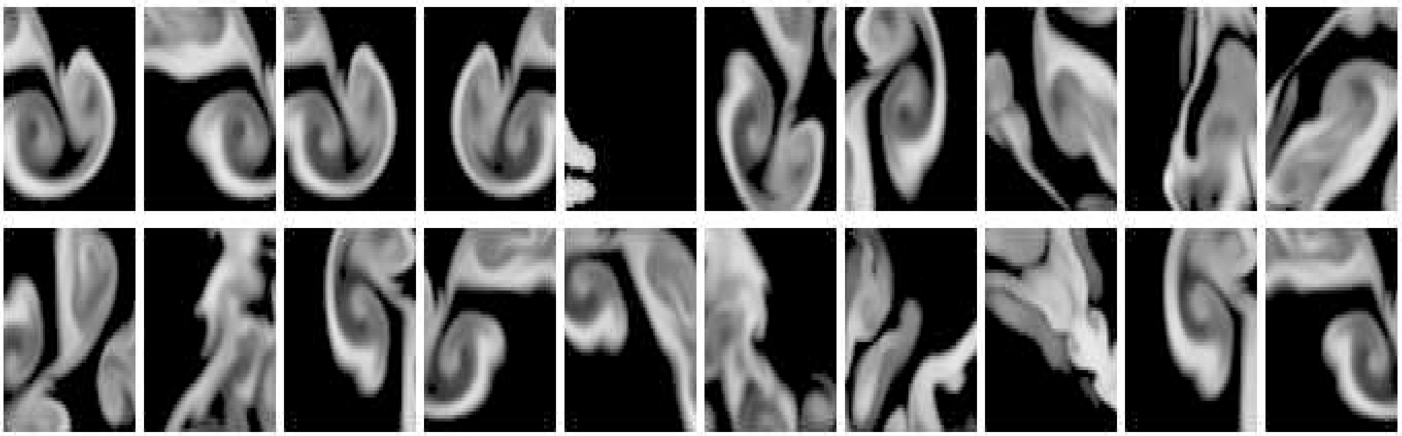


Relevance Feedback

- Improve results based on feedback from user
- Learn a model that captures what user finds interesting
- Use the model to find other interesting objects
- Objectives are to maximize accuracy and identify useful features
- Texture features usually ranked high

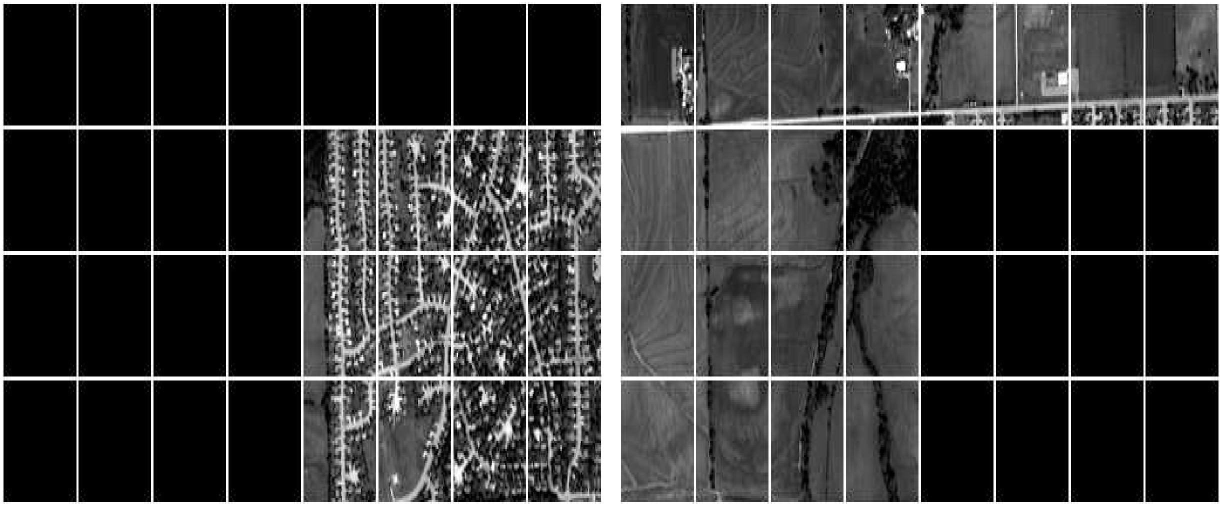
Using Feedback

- Rank features, build models on increasingly larger subsets
- The model with highest accuracy is used on the entire database



Detecting Human Settlements

- Four-band IKONOS satellite images
- Extracted total of 496 features from 7419 64×64 tiles
- Unclear which bands or which features maximize accuracy



Human Settlements Results

- Minimum error rates for each filter considering feature types independently and in combination

Method	Pow. Sp	GLCM	Wavelet	Gabor	All
No filter	29.0	27.5	28.8	33.2	41.8
PCA	28.0	27.1	28.2	27.8	28.8
KL	27.1	26.0	26.7	26.1	26.0
χ^2	27.0	26.0	26.5	26.1	26.0
Stump	28.2	26.6	27.8	26.6	26.9

Conclusions

- Identifying relevant features can save computational resources
  - Human settlements problem can be solved with two bands
- Unnecessary features degrade performance of classifiers
  - Even with domain knowledge, choosing features is hard
- Feature selection provides insights
- Simple methods work well in many cases