27 July 1998

Technique Selection in Machine Learning Applications

Ronny Kohavi

Silicon Graphics, Inc.

ronnyk@sgi.com

ICML 98 Workshop:

Methodology of Applying Machine Learning

Outline

- Learning is Impossible without assumptions.
- Some useful assumptions.
- Choosing an accurate algorithm vs. test–driving one.
- Desired properties.
- The big picture: technique selection is a small part of the Knowledge Discovery process.
- Summary

Learning is Impossible without Assumptions

- Watanabe's Ugly Duckling Theorem.
- Mitchell's version spaces.
- Schaffer's conservation law.
- Wolpert's no free lunch theorem.

Researchers keep discovering that generalization is impossible without assumptions.

No Free Lunch Theorem

Theorem: For any two algorithms *A* and *B*, there exist datasets for which algorithm *A* will outperform algorithm *B* in prediction accuracy on unseen instances.

Proof: Take any Boolean concept.
If *A* outperforms *B* on unseen instances, reverse the labels and *B* will outperform *A*.

Extension: For discrete spaces, the number of concepts for which *A* will outperform *B* in prediction accuracy is equal to the number for which *B* will outperform *A*.

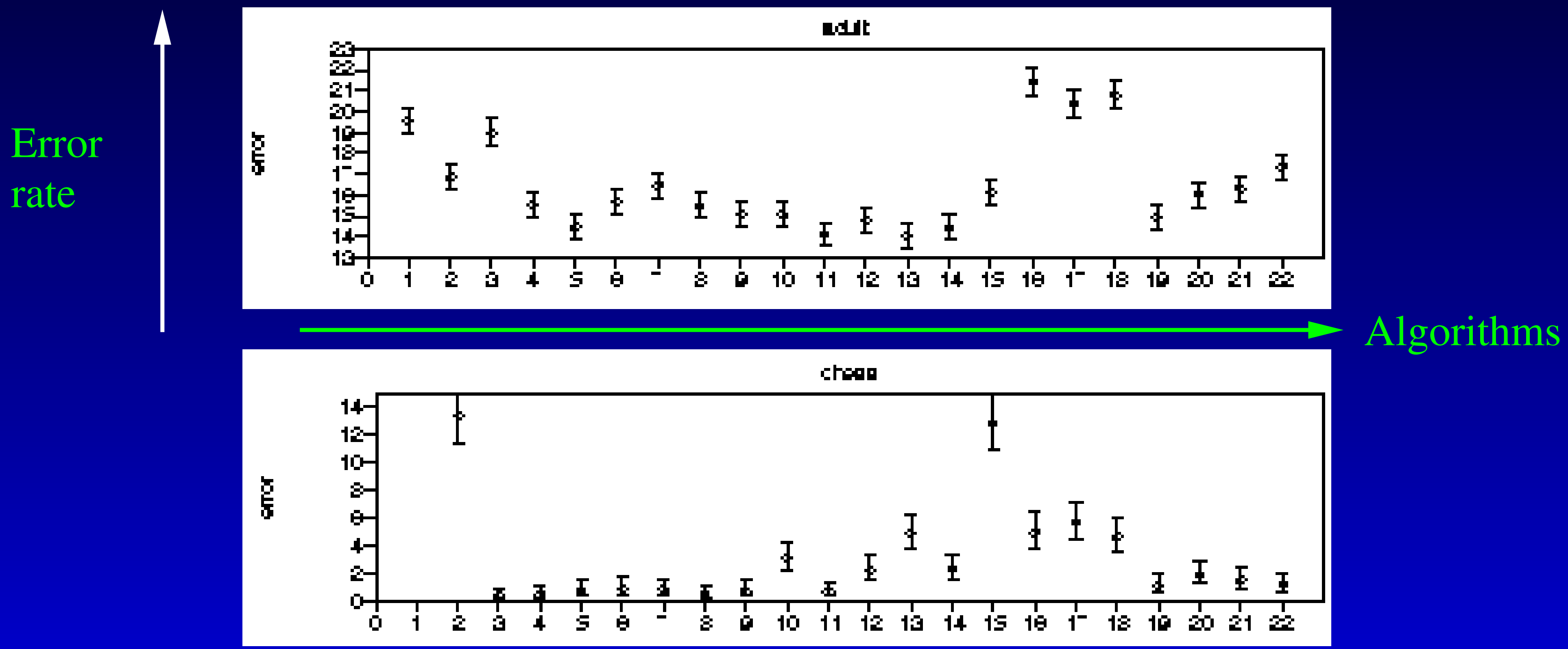
Observations from NFL

- A simple majority learner that predicts the most frequent label will outperform any fancy algorithm on as many concepts as the fancy one outperforms the majority.
- Observations from NFL Learning curves must sometimes decrease in accuracy.
- Meta–level techniques that choose algorithms based on holdout, cross–validation, or bootstrap are still subject to the theorem.

Why is there a Machine Learning Field then?

Does the NFL Hold in Practice?

The NFL is relevant to real world problems (Kohavi, Sommerfield, Dougherty, 1997)



There was no clear winner, but several algorithms performed better on average.

Useful Assumption – Smoothness

Statisticians have made smoothness assumptions for years:
- For real–valued attributes
- With high probability, an infinitesimal change will not change the label.

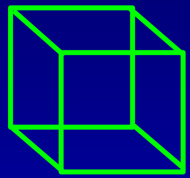
Fix and Hodges (1951):
- Statistical consistency for nearest neighbors. As the training set grows, the accuracy approaches the Bayes optimal. (Asymptopia.)

Gordon and Olshen (1984):
- ditto for Decision Trees for some algorithms.

Useful Assumption – Few Attributes

Feature selection methods assume that a small number of attributes suffices.

Bellman's curse of dimensionality implies that in high dimensions everything is "far"



Test yourself:
- 20 dimensional space.
- Each attribute is real valued in the range 0 to 1.
- 100,000 instances uniformly distributed.

What is the expected distance to nearest neighbor?

Non–interesting assumptions

For natural datasets, the following are not very interesting:

- Dissimilar: great for parity.
- 1R: single attribute.
- Some PAC–motivated spaces.
  Nice theorems can be proved about some hypothesis space, but that does not make them natural.

How to Choose an Accurate Algorithm in Advance

Several papers exist on rules of thumb for technique selection.

- Carla Brodley (selective Superiority; 1993).
- Ross Quinlan (sequential/parallel, 1994).
- Peter Adriaans (1996)

Problem: usually works fine for artificial concepts. Much harder in real life with natural concepts.

Proposed method: Test Drive

Since the theory of choosing an algorithm is weak, my recommendation is to test–drive different algorithms: TRY THEM.

Run several algorithms and measure the accuracy/error/loss and other important properties (discussed soon).

Caveat: Choose from a small set of algorithms.

Created by Donghoon Shin, Art Center College of Design.

Properties: Accuracy/Loss

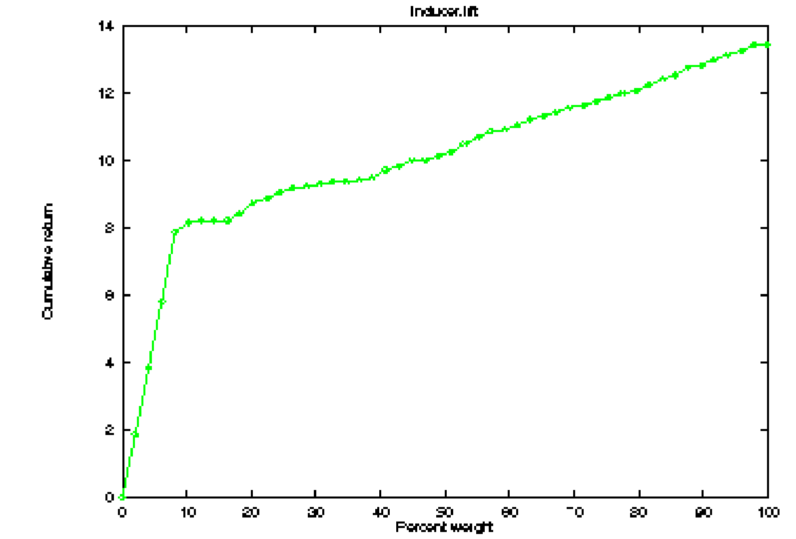
- Accuracy/error are often the ***wrong*** measure.
- Confusion matrices:

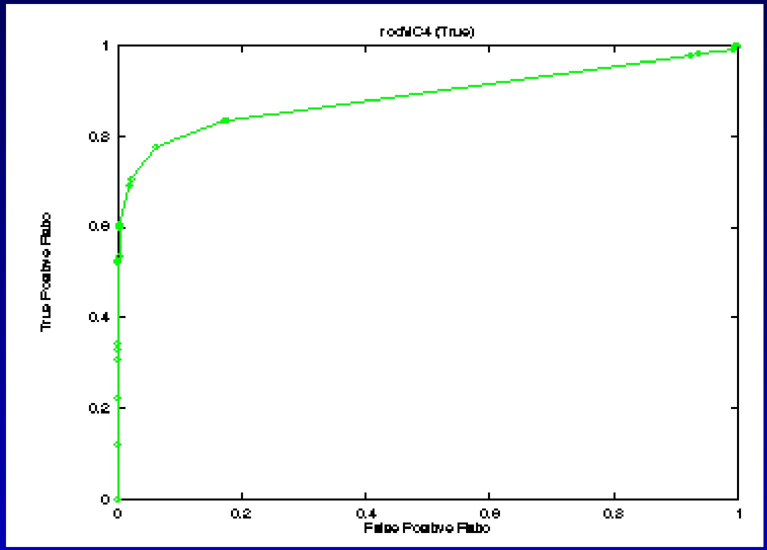
		Predicted		
		?	Yes	No
Actual	Yes	10	0	50
	No	10	5	0

- Measure the loss/utility, not simple error. Real life problems have associated costs with false positive and false negative classifications.
- Support unknown predictions.

Properties: Lift Curve/ROC curve

- Lift curves show how good the classifier is at predicting probabilities for a given class. Great for mailing campaigns.





  - Used to test stability and power of probabilistic predictions.
  - The two graphs are isomorphic.

Properties: Comprehensibility

- In many cases, especially in business settings, comprehensibility is crucial.
  - Can you explain how the classifier predicts (as opposed to how it was built)?
  - Can the model be visualized in a way that is comprehensible to the (business) user?
- Is the model compact?
  - Usually compactness helps comprehensibility.

Properties: Training/Class Time

- How long does it take to train the model?
  Neural network are slow to train.
  Nearest neighbor are trivial to train.

- How long does it take to classify?
  Nearest neighbor are slow to classify.
  Neural network are fast to classify.

One can define a utility function with the properties and pick the one with the highest estimated utility (Fayyad, Piatetsky–Shapiro & Smyth).

Data Mining (Knowledge Discovery)

Generalize to the future

Knowledge discovery is iterative. As you uncover "nuggets" in the data, you learn to ask better questions

The non–trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

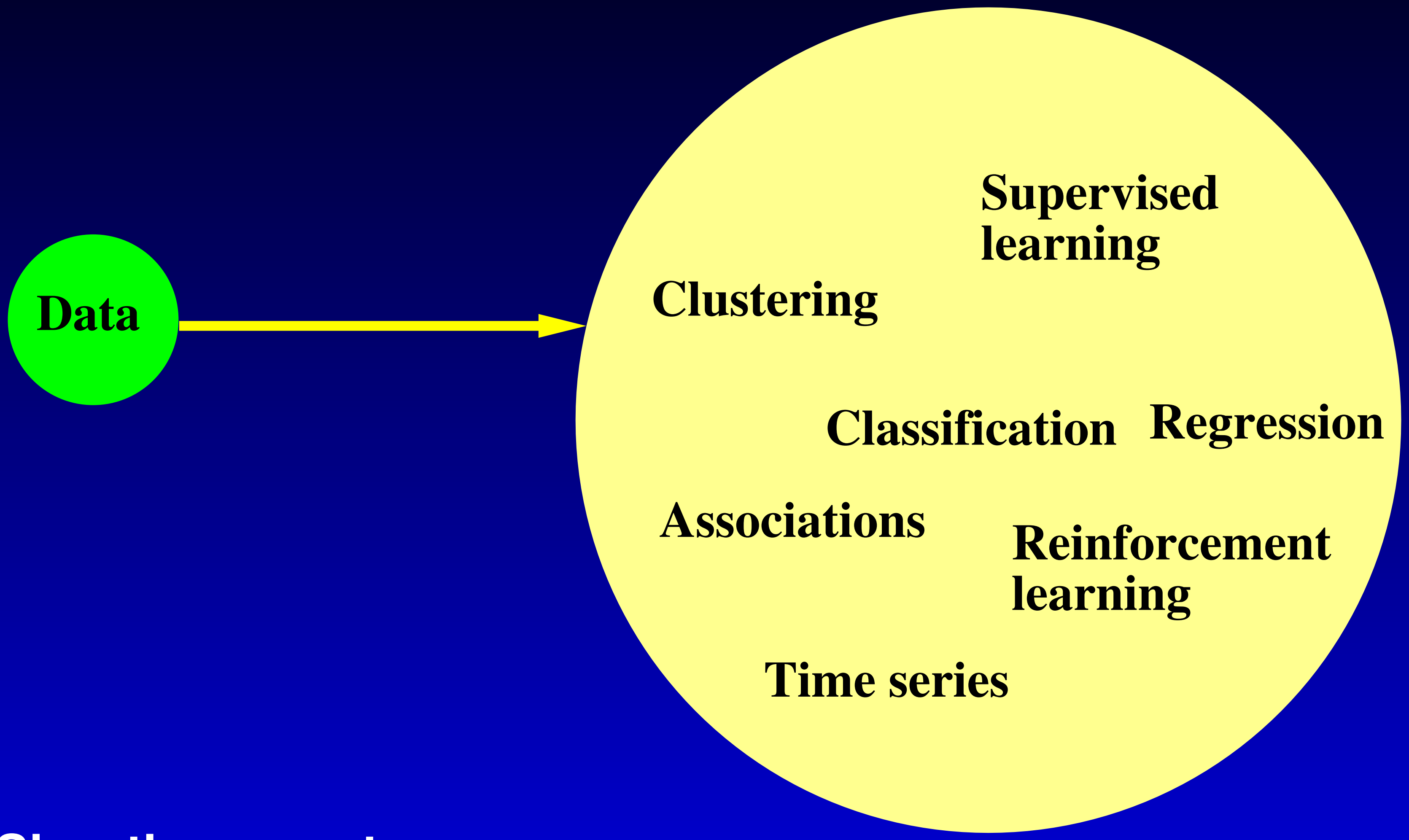
–– Fayyad, Piatetsky–Shapiro, Smyth [1996]

Not something we already know

Process leads to human insight.

For our task. Actionable

The BIG Picture: ML View



Size=time spent

The BIG Picture: Actual View

Summary

- Technique selection should *not* be based solely on accuracy. Also take into account:
  - Loss matrix/lift curve/ROC curve
  - Comprehensibility
  - Training/test times

- The goal in most real–world situations is to get insight, not just predict.
  Visualize and study the models, for they provide insight.

- There is no best car or a best graph. Test drive techniques on your specific dataset.