The Role of Applications in the Science of Machine Learning

Foster Provost
New York University

Thanks to: John Aronis, Bruce Buchanan, Scott Clearwater, Andrea Danyluk, Tom Fawcett, Shawndra Hill, David Jensen, Ronny Kohavi, Andrew McCallum, Sofus Macskassy, Jen Neville, Claudia Perlich, Jeff Simonoff, Ben Taskar, Gary Weiss, and many others

ICML-2003
Washington, DC

Roles of applications

• Help to stimulate researchers
• Attract (good) graduate students
• Help to convince funding sources
• Provide tests of prior research results
  – of relevance
  – of necessity
  – of efficacy
• Highlight insufficiencies of the state of the art
  → pointing to important areas for future research

“In the meantime, while the creative power of pure reason is at work, the outer world again comes into play, forces upon us new questions from actual experience, opens up new branches [of the science]”

David Hilbert, 1900
A spectrum of applicability

Applications

<table>
<thead>
<tr>
<th>Generality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied Research</td>
</tr>
<tr>
<td>Experimental Research using synthetic data</td>
</tr>
</tbody>
</table>

- “I make no mockery of honest ad hockery”
  attributed to John Tukey

- NB: research using benchmark (e.g., UCI) data sets is **not** applied research!

---

**Data set** | **Winner AUR** | **Winner Acc** | **Result**
---|---|---|---
Nurse | none | none | Indistinguishable
Obedient | none | none | Indistinguishable
Letter-F | C1 | C1 | C4 domination
Letter-A | C4 | C4 | C4 crossover
Ionosphere | C1 | C4 | C4 domination
DNA | C4 | C4 | C4 domination
Covtype | C4 | C4 | C4 domination
Telcom | C4 | C4 | C4 domination
Horse | C4 | C4 | C4 domination
Car Eval | none | none | C4 crossover
Spam | C4 | C4 | C4 domination
Chess | C4 | C4 | C4 domination
CA1 Bons | C4 | C4 | C4 domination
Amazon | C4 | C4 | C4 domination
Credit | C4 | C4 | C4 domination
Adult | C4 | C4 | Mixed
Census | C4 | C4 | C4 domination
More | C4 | C4 | C4 domination
Dow Jones | C4 | C4 | C4 domination

C4-LR

High separability

16-2

Low separability

1-8

Perlich et al., JMLR 2003
Ref: Lim, Chih-Wei H. (2001)
A spectrum of applicability

• “I make no mockery of honest ad hockery”
  John Tukey

• NB: research using benchmark (e.g., UCI) data sets is not applied research!

Example applications of ML

Business
• Credit scoring
• Customer retention
• Drug design
• Financial modeling
• Fraud detection
• Games (“video”)
• Information retrieval
• Information monitoring
• Operations support
  – diagnosis/troubleshooting
  – monitoring
  – quality control
• Robotics
• Spam filtering
• Targeted marketing
• Web personalization
  • and more …

Science/medicine/government
• Astronomic sky surveys
• Bioinformatics
• Biosurveillance
• Brain imaging
• Counterterrorism
• Crystallography
• Law enforcement
• Medical diagnosis
• Natural language processing
  – and other AI areas
• Planetary exploration
• Planetary image analysis
  – e.g., volcanoes on Venus
• Public health research
  – e.g., predicting carcinogenicity
• Weather forecasting
  • and more …
Example: data mining for fraud detection

Can we identify which accounts have been compromised so as to take corrective action?

A typical frauded account

<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Day</th>
<th>Duration</th>
<th>From</th>
<th>To</th>
<th>Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/01/95 10:05:01</td>
<td>Mon</td>
<td>13 mins</td>
<td>Brooklyn, NY</td>
<td>Stamford, CT</td>
<td></td>
</tr>
<tr>
<td>1/05/95 14:53:27</td>
<td>Fri</td>
<td>5 mins</td>
<td>Brooklyn, NY</td>
<td>Greenwich, CT</td>
<td></td>
</tr>
<tr>
<td>1/08/95 09:42:01</td>
<td>Mon</td>
<td>3 mins</td>
<td>Bronx, NY</td>
<td>White Plains, NY</td>
<td></td>
</tr>
<tr>
<td>1/08/95 15:01:24</td>
<td>Mon</td>
<td>9 mins</td>
<td>Brooklyn, NY</td>
<td>Brooklyn, NY</td>
<td></td>
</tr>
<tr>
<td>1/09/95 15:06:09</td>
<td>Tue</td>
<td>5 mins</td>
<td>Manhattan, NY</td>
<td>Stamford, CT</td>
<td></td>
</tr>
<tr>
<td>1/09/95 16:28:50</td>
<td>Tue</td>
<td>53 sec</td>
<td>Brooklyn, NY</td>
<td>Brooklyn, NY</td>
<td></td>
</tr>
<tr>
<td>1/10/95 01:45:36</td>
<td>Wed</td>
<td>35 sec</td>
<td>Boston, MA</td>
<td>Chelsea, MA</td>
<td>YES</td>
</tr>
<tr>
<td>1/10/95 01:46:29</td>
<td>Wed</td>
<td>34 sec</td>
<td>Boston, MA</td>
<td>Yonkers, NY</td>
<td>YES</td>
</tr>
<tr>
<td>1/10/95 01:50:54</td>
<td>Wed</td>
<td>39 sec</td>
<td>Boston, MA</td>
<td>Chelsea, MA</td>
<td>YES</td>
</tr>
<tr>
<td>1/10/95 11:23:28</td>
<td>Wed</td>
<td>24 sec</td>
<td>White Plains, NY</td>
<td>Congers, NY</td>
<td></td>
</tr>
<tr>
<td>1/11/95 22:00:28</td>
<td>Thu</td>
<td>37 sec</td>
<td>Boston, MA</td>
<td>EastBoston, MA</td>
<td>YES</td>
</tr>
<tr>
<td>1/11/95 22:04:01</td>
<td>Thu</td>
<td>37 sec</td>
<td>Boston, MA</td>
<td>EastBoston, MA</td>
<td>YES</td>
</tr>
</tbody>
</table>

Example: fraud detection

Profile customer behavior

(Fawcett & Provost 1996, 1997)

Example fraud indicator: calls from the Bronx at night

Example profiles:
- Customer makes an average of 2 ± 2 calls at night
- Customer makes 12 ± 5 calls during business hours
- Customer makes 0 ± 0 calls from the Bronx at night

Data Mining

Fraud indicators

Profiler construction

Evidence combining

Detector

Profiling Templates

Behavior Profilers

Example profiles:
- Customer makes an average of 2 ± 2 calls at night
- Customer makes 12 ± 5 calls during business hours
- Customer makes 0 ± 0 calls from the Bronx at night
Is this behavior evidence of fraud?

It depends on the customer.
Example: fraud detection
Profile customer behavior

(Fawcett & Provost 1996, 1997)

Call Data

Data Mining

Fraud indicators

Profiler construction

Profiling Templates

Evidence combining

Detector

Example profiles:
• Customer makes an average of 2 ± 2 calls at night
• Customer makes 12 ± 5 calls during business hours
• Customer makes 0 ± 0 calls from the South Bronx at night

Example fraud indicator: calls from the South Bronx at night

Example: fraud detection
Evaluation of fraud detection methods

Accuracy

Cost

Cost ($) vs. Accuracy (%) for different fraud detection methods:
- Alarm on None
- Collisions & Velocities
- High Usage
- Best Individual
- SOTA
- DC-1 Detector

Accuracy vs. Cost graph:
- Highest accuracy at lowest cost:
  - SOTA
  - DC-1 Detector

Accuracy vs. Cost graph:
- Lowest accuracy at highest cost:
  - Alarm on None
  - Collisions & Velocities
  - High Usage
  - Best Individual
Today I'll discuss two groups of issues, and associated research problems

• Costs
  – many different costs to consider
  – see (Turney, 2000) for an excellent overview

• Relational data
  – networked data
  – multi-table RDBs

What “new” research issues emerge?

Costs & machine learning
  • e.g., Workshop @ ICML 2000

• Different misclassification costs
  – false-positive vs. false-negative errors

• Misclassification error cost varies case-by-case
  – e.g., (Zadrozny & Elkan ICML 2001, etc.) (Bennett, SIGIR 2003)

• Highly unbalanced class distribution
  – Workshops @ AAAI 2000, ICML 2003

• Don’t know target costs & class distributions
  – (Provost & Fawcett, KDD-97, MLJ 2001) (Latinne et al., ICML 2001)

• Procuring training data costly
  • budget-sensitive learning?
    – e.g., what proportion of positives and negatives for training?
Given an example procurement budget, what class distribution should be used?

• What does the manual tell us to do?

  “if the sample size is fixed, a balanced sample will usually produce more accurate predictions than an unbalanced (one)” (SAS, 2001)

• General recommendations (based on C4.5):
  – For accuracy use natural distribution
    • but could do better
    • in 9 of 26 cases get burnt (with confidence)
    • (balanced sample not recommended!)
  – For AUC use balanced distribution
    • but could do better
    • in 7 of 26 cases get burnt (with confidence)

Relative performance of different class distributions is remarkably consistent across training-set sizes

(Weiss & Provost, JAIR 2003)

(Weiss & Provost, JAIR 2003)
Budget-sensitive progressive sampling

% Minority Class Examples vs. Portion of budget spent

- Beam_top
- Beam_bottom

10% (best)
Budget-sensitive progressive sampling

% Minority Class Examples

Portion of budget spent

beam_top

beam_bottom

10% (best)

Budget-sensitive progressive sampling

% Minority Class Examples

Portion of budget spent

beam_top

beam_bottom

10% (best)
Budget-sensitive progressive sampling

(Weiss & Provost, JAIR 2003)
What “new” research issues emerge?

Costs & machine learning
- e.g., Workshop @ ICML 2000
- Different misclassification costs
  - false-positive vs. false-negative errors
- Misclassification error cost varies case-by-case
  - e.g., (Zadrozny & Elkan ICML 2001, etc.) (Bennett, SIGIR 2003)
- Highly unbalanced class distribution
  - Workshops @ AAAI 2000, ICML 2003
- Don’t know target costs & class distributions
  - (Provost & Fawcett, KDD-97, MLJ 2001) (Latinne et al., ICML 2001)
- Procuring training data costly
  - budget-sensitive learning
  - e.g., what proportion of positives and negatives for training?

What “new” research issues emerge?

- Learning from relational data
  Recently:
  - SRL Workshops @ AAAI 2000, IJCAI 2003
  - MRDM Workshops @ KDD 2002, KDD 2003
  - Networked data
  - Multirelational databases
Mining networked data

• Classifying fraudulent entities “by the company they keep”?
Mining networked data

• Classifying fraudulent entities “by the company they keep”?

Can we generalize this idea?

Dialed-digit detector (Fawcett & Provost, 1997)
Communities of Interest (Cortes et al. 2001)
A relational-neighbor classifier?

Is there theoretical justification?  

Thanks to (McPherson, et al., 2001)

- *Birds of a feather, flock together*  
  attributed to Robert Burton (1577-1640)

- *(People) love those who are like themselves*  
  -- Aristotle, *Rhetoric* and *Nichomachean Ethics*

- *Similarity begets friendship*  
  -- Plato, *Phaedrus*

- *Hanging out with a bad crowd will get you into trouble*  
  -- Mom

Relational neighbors and beyond

- **Homophily**
  - fundamental concept underlying social theories (e.g., Blau 1977)
  - one of the first features noticed by analysts of network structure
    - antecedents to SNA research from 1920’s (Freeman 1996)
    - fundamental basis for links of many types in social networks
      *Patterns of homophily:*
        - remarkably robust across widely varying types of relations
        - tend to get stronger as more relationships exist
        - likely to be present for non-person entities?

  → **Simple relational-neighbor classifier may be effective**

- **A baseline against which we should compare relational learning algorithms?**

- (More complex structure can be learned)
  - e.g., see Daphne’s talk on Sunday
Classifying firms “by the company they keep”...

Classifying movies by the company they keep...

Prior results: predicting movie success (AUC)
Relational Trees: 0.82
Relational Bayes: 0.85
IMDB data, using attributes on related entities (e.g., the most prevalent genre of a movie’s studio).

RN Classifier:

<table>
<thead>
<tr>
<th>RN link type</th>
<th>AUC</th>
<th>stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>0.766</td>
<td>0.003</td>
</tr>
<tr>
<td>director</td>
<td>0.658</td>
<td>0.007</td>
</tr>
<tr>
<td>producer</td>
<td>0.850</td>
<td>0.005</td>
</tr>
<tr>
<td>production company</td>
<td>0.862</td>
<td>0.003</td>
</tr>
</tbody>
</table>
Iterative/collective classification using belief propagation

Idea: belief about class membership can be propagated through a network of linked examples, to help to classify unlabeled/uncertain examples.

- Probabilistic belief propagation (e.g., Pearl, 1988)
- Relaxation labeling (e.g., Chakrabarti et al., SIGMOD 1998)
- PRMs (e.g., Koller et al. AAAI-98, IJCAI-99, ICML-01, IJCAI-01, UAI-02, etc.)
- Mining viral marketing (Domingos & Richardson, KDD-2001)

• Simple belief-propagation version of RN?
  - Useful when:
    • few labeled data
    • linked to unlabeled data
  - A baseline against which we should compare relational learning algorithms?
    • (Macskassy, MRDM -2003 next week)

Classifying research papers by the company they keep...
Mini-summary:
- people
- companies
- movies
- research papers
- …

Simple relational classifiers should be used as baselines in relational learning studies

Classifying web pages by the company they keep...

WebKB Comparison on Washington data: RN vs. RPT/RBC

cf. (Gaines, ICML 1989)

(Macskassy, MMMM 2003 next week)
What “new” research issues emerge?

Today I’ll discuss two groups of issues, and associated research problems

• Costs
  – many different costs to consider
  – see (Turney, 2000) for an excellent overview

• Relational data
  – networked data
  – multi-table RDBs
How should training examples be constructed?

one-to-many relationship

must aggregate values

(see C. Perlich’s talk on Tuesday at KDD-2003)
Future “new” research areas?

- Comprehensibility!
  (Pazzani, *IEEE Intelligent Systems*, 2000)
- Drift
  - concept (e.g., Lane & Brodley, KDD-1998; Bartlett et al. MLJ 2000)
  - population (e.g., Bartlett, CoLT-92; Long, CoLT-98)
- Activity monitoring
  - timeliness of detection vital
    - seldom taken into account by ML studies
    - (Weiss & Hirsh, KDD-98)
      - network monitoring
    - (Fawcett & Provost KDD-99)
  - biosurveillance …

ML for biosurveillance

(Wong, Moore, Cooper, Wagner, *J. Urban Health* 2003)

Sunday Afternoon: (Wong, Moore, Cooper, Wagner, *ICML-2003*)
Summary

Applications:
- help to stimulate researchers
  - and attract graduate students
- help to convince funding sources
- provide tests of prior research results
  - of relevance
  - of necessity
  - of efficacy
- highlight insufficiencies of the state of the art
  - pointing to important areas for future research

“In the meantime, while the creative power of pure reason is at work, the outer world again comes into play, forces upon us new questions from actual experience, opens up new branches [of the science]”

David Hilbert, 1900

On writing, reviewing, editing ...

- Keep complexity of applications in mind
  - must ignore certain things
  - but can’t “assume” them away
- Keep “spectrum” in mind
  - different sorts of contributions
- Have a clear focus
  - don’t try to sneak your new X algorithm through a back door
    - where X is a well-known problem
- Argue for generality
- Don’t forget the basics
  - what are the claimed contributions?
    - novel? important?
  - how are they supported?
  - how does the work fit in the context of prior work?
  - be scholarly
Potential contributions?

- In-depth descriptions of (new) practically important problems
  - that can’t be solved with existing methods
- A solid case study
  - Feedback on the (lack of) utility of prior results
  - Areas of weakness in our knowledge
- A new benchmark data set
- Good novel algorithms/methods
  - maybe ad hoc
  - potential starting point for future studies

Summary

Applications:
- help to stimulate researchers
  - and attract graduate students
- help to convince funding sources
- provide tests of prior research results
  - of relevance
  - of necessity
  - of efficacy
- highlight insufficiencies of the state of the art
  \(\rightarrow\) pointing to important areas for future research

“In the meantime, while the creative power of pure reason is at work, \textit{the outer world} again comes into play, forces upon us new questions from actual experience, opens up new branches \textit{of the science}!”

David Hilbert, 1900
References 1


References 2

References 3


References 4

References 5


References 6