Predictive Accuracy – Is It Enough?

John Maindonald
Centre for Mathematics and Its Applications, Australian
National University, Canberra ACT 0200, AUSTRALIA.
Email: `john.maindonald@anu.edu.au`

December 6, 2007

Cox (ISI Review 261: 365-367); Cameron's comment

D.R.Cox (ISI Review 261: 365-367): "The discussion is prompted in part by developments in the computer science and philosophical literature *in which a weaker definition of causality tends to be employed*[a], much less cautious than the traditional statistical and epidemiological view summarized in Bradford Hills criteria[b]"

"Cameron: This combination of very large datasets, new questions being asked and a diverse group of participants has led to some interesting work, the invention of some inferior methods and considerable interest from business."

Inferior, in what sense?

From horticulture to involvement with data miners!

[a] Pearl, 1988, 1995; Spirtes, Glymour & Scheines, 1993

[b] Bradford Hill, 1965

A Tentative Answer

It may be enough if there is an adequate and scientifically meaningful, model, with fixed and stochastic parts, that gives meaning to the accuracy measure. Typically this involves:

- Regard to the scientific context.
- Mechanisms that, plausibly, generated the data.
  i.e., model for the data.
- Mechanisms that describe how results will be used.
  Often, a model for the predictive process.

Example – MeBr disinfestation treatments

- Insect kill and fruit damage both increase with temperature and exposure to MeBr (Methyl Bromide).
- Ideally, use the least MeBr that optimizes the trade-off between insect kill & fruit damage.
  (Alas, no-one knows what is optimal!)
  - North America: fumigate at $\geq 20^o$ C
  - NZ: fumigate at 12^o C; use $\geq$ twice as much MeBr
- There was no regard to seasonal variation.
  (Replicates are replicates are replicates?!!)

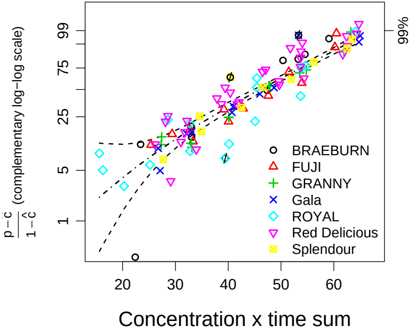
After many years' work . . .

Cultivar	# of 1-day eggs	Year (Replicates)
Pacific Rose	–	1998(3*), 1999(2*)
Braeburn	15,297	1988(2)
Fuji	14,497	1988(2), 1998(2*), 1999(3*)
Granny Smith	13,628	1988(2)
Gala	17,011	1989(2)
Royal Gala	33,160	1988(2)
Red Delicious	16,049	1988(3), 1989(2), 1998(3*), 1999
Splendour	24,182	1989(2)

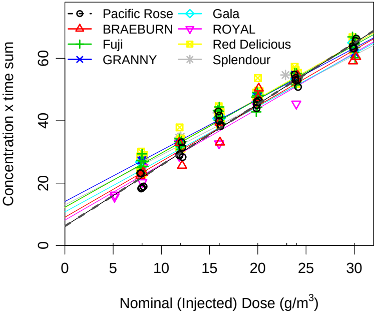
*Trials in 1998 and 1999 examined sorption only.

Each trial gave results for one cultivar at six doses (0, 8, 12, 16, 20, 24, 30g.m^{-2}; except for Royal Gala)

Good news: Insect mortality response to MeBr concentration in the chamber was consistent over varieties, years & trials.



Bad news: MeBr sorption patterns vary between varieties, years & trials.



Graham Williams' 2007 Australia Day Medallion

"…for leadership and mentoring in Data Mining in the Australian Taxation Office and in Australia, and particularly the development and sharing through open source of the [R-based] Rattle system for data mining …"

Web site: `http://www.togaware.com`

Graham's web site is a useful open software resource

"**Freedom:** …We have a choice to either share …to help others in many ways, or to carefully protect our knowledge and understanding, and attempt to become monetarily wealthy. It is, perhaps, a choice between freedom and oppression.

" Software is a creative expression, beautifully crafted, and lovingly nurtured. It is prose that is a pleasure …to read, understand, and learn …

Detection of Tax Fraud – Aust. Financial Rev., 10Jul06

"Traditionally, the ATO concentrated on historical analysis to see if there was a mismatch between income levels & tax paid. But now that it is replacing . . . its legacy systems with brand new software . . . , [it will be possible to do] predictive analysis, which uses past behaviour to determine future actions.

The news for taxpayers may not all be bad!
"The ATO says its new technology will allow it to leave law-abiding taxpayers in peace."

"Predictive analysis" predicts risk, not fraud.
Misleading model coefficients should not be an issue.
The inferential issues are fairly straightforward?

Athletes as Data Providers – Aust. Financial Rev., 8Jul06

Over 25 years, athletes have been measured more than tailors' dummies. . . . From about 2002, . . . AIS began to view athletes not just as physical, mental & moving specimens but as providers of data that could be integrated . . . [to] provide a new language about performance. The result is an $AU8.7 million data mining initiative . . . .

Not so fast . . . There's huge scope for confounding, regression coefficients that go in the wrong direction, etc. Note however:

- Statistician involvement will doubtless continue. (c.f., the dataset `ais` in the *DAAG* package.)
- Teamwork between statisticians, "data miners" & other specialists has dividends both for the immediate project & for the future work (new insights, etc.) of all team members.

Are Airbags Effective

Meyer & Finney (MF)[a] studied US data, for 1997-2002, from police-reported car crashes in which there was a harmful event (people or property). The debate is ongoing[b].

If a key factor is omitted, confounding is spectacular.

- Round 1: (MF) Without allowance for both seatbelts & speed of impact, summary tables are misleading. With such allowance, airbags appear useless, possibly dangerous.
- Round 2a: (Farmer) There are other problems. Here is another, maybe better way, to assess the evidence.
- Round 2b: (Meyer) Airbags are dangerous, full stop!

[a] Who wants airbags?. Chance 18:3-16, 2005

[b] Farmer, Chance 19:15-22, 2006; Meyer, Chance 19:23-24, 2006

Are Airbags Effective (N = No airbag; A = airbag)

No adjustment for other factors

Sbelt	N_dead (total)	A_dead (total)	xtra_dead
none	39676 (5484922)	25919 (6648610)	-22175

Adjust for seatbelt use

Sbelt	N_dead (total)	A_dead (total)	xtra_dead
none	24067 (1366089)	13760 (885635)	−1842
belted	15609 (4118833)	12159 (5762975)	−9681
			−11703

Adjust for seatbelt use & speed of impact

Excess risks are

1-9km/h (N, A), 10-24 (), 25-39 (), 40-54 (), 55+ ()

(140, 0), (1282, -611), (-2097, -2048), (2045, -612), (1320, 1220)

Total Extra Dead = 618

Motivations

D.R.Cox (ISI Review 261: 365-367): "The discussion is prompted in part by developments in the computer science and philosophical literature *in which a weaker definition of causality tends to be employed*[a]*, much less cautious than the traditional statistical and epidemiological view summarized in Bradford Hills criteria*[b]"

"... computer science ... a weaker notion of causality ..."

"Data mining" fits readily into this context.

- Statisticians commonly seek a "good" model, expecting that good models will do well on any sensible criterion.
- Data miners may make predictive accuracy the priority.

  Training/test set and source/target issues are then crucial!

[a] Pearl, 1988, 1995; Spirtes, Glymour & Scheines, 1993

[b] Bradford Hill, 1965

Is Prediction, then, a Royal Road?

"Royal", implying avoidance of statistical niceties?

- Much data mining literature suggests that it is.
- Breiman's 2003 paper[a] seems to support that view.
  "Stop fussing about niceties, be a little less precious!"?

JM's view: Prediction is, often, rather important. But it makes no sense until the form of the model is clear.

e.g., AIC can reasonably compare predictive accuracies between normal & lognormal errors, but not between 2 & 3 levels of random effects?

[a] Breiman, L. 2001.Statistical modeling: the two cultures (with discussion). Statistical Science 16: 199- 231.

Data Mining a/c Data Miners (Wikipedia article)

“. . . the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc.

“. . . has links with . . . core fields such as computer science and adds value to . . . seminal computational techniques from statistics, information retrieval, machine learning and pattern recognition.

“Example: . . . If a clothing store records the purchases of customers, a data mining system could identify those customers who favour silk shirts over cotton ones.

“The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction.

“Cross validation is a common approach to evaluating the fitness of a model . . . – data is divided into a training subset and a test subset to . . . build and then test the model.”

http://en.wikipedia.org/wiki/Data_mining

Methods Popular with Data Miners

Data mining might alternatively be characterized by the methods used. Common methods are:

- Trees and ensembles (or “forests” of trees).
- Neural nets.
- Support Vector Machines.
- Regression.
- Various clustering approaches.

Inference almost inevitably assumes a model that has independent and identically distributed (iid) errors.

Comments

- Automation, most likely with implicit iid assumptions, is fine if it is effective.
- Some methods are in vogue that do not readily lend themselves to classical inferential approaches, using Bayes &/or maximum likelihood.
- Accuracy is commonly assessed using a random training/test split, or using cross-validation, and thus relates to prediction for the population from which the data were derived.
- Predictive accuracy is a worthy aim. But what if the interest is in estimation of population parmeters?
  [c.f., HRT studies (CHD risk); US accident data (airbag effects)]

The Source/Target Distinction

Source versus target	*Are data available from target?*
1: Identical (or nearly so)	Yes; data from source suffice
2: Source & Target differ	Yes, we have data from target
3: Source & Target differ	No. But a model-based estimate of predictive accuracy is available. (cf: multi-level models; time series)
4: Source & Target differ	No; must make an informed guess.

Other possibilities, where source & target differ

- Train (1) a model that is optimal for the source data and (2) a model that underfits.

  In day to day use, run them side by side.
- Seek out comparable historical "source" data, for which matching historical target data are available.

Maybe a weak/strong testing jargon is useful?

- Unless test data are independent of training data; the accuracy measure is flawed.
- Use of training/test data from source population, and cross-validation, provide weak accuracy measures.
- Strong accuracy is accuracy for an intended practical use; test data must be from the target population.

Commentary

- Better weak accuracy may not imply better strong accuracy[a]!
- Consider fortification, i.e., add elements of strength?
- Strong (or even fortified) testing has been unusual in the DM literature, notwithstanding its practical importance.
- Strong testing is a choice of model issue!

NB: Distinguish the above from the weak/strong learners jargon!

[a] See Hand, D. J. 2006. “Classifier Technology and the Illusion of Progress”

When algorithms are evaluated or compared . . .

- What training/test data were used?
- Describe algorithmic steps in precise detail.
- Include precise details of any tuning or variable selection or transformation steps.
  (For cross-validation; were these repeated at each fold?)
- Expose code to public display and scrutiny.
- Try the algorithm with random data.
  (This can be a useful reality check. If a pattern appears & seems to check out statistically, be worried!)
- . . . to be continued . . .

When algorithms are evaluated or compared (cont.) . . .

- Try each algorithm with simulated data.
  - Under what circumstances does it perform well/badly?
  - Are the statistical properties what they are claimed to be?
    (A common error is that the CV[a] does not account for all steps.)
- Give a 2-D or 3-D views that identify points that the different algorithms classify differently.
  - Note 1: Is 2-D adequate? Should it be 3-D, 4-D, ...?
  - Note 2: Graphs for the training data are, strictly, flawed.

The above may still compare only "weak" accuracies.

Performance may, in practice, be different!

[a] cross-validation

Towards strong accuracy measures

1. Use training/test data that cross the source/target split. cf Eamonn Keogh's collection.
2. Relatively sophisticated modeling can be essential – cf time series, multi-level models, spatial models, . . . Models with a complex error structure may be needed for conceptualization, even if not for analysis.
3. NB also simulated data – use a model to generate data. Simulation allows scenarios that are unlike the past.

For 2 & 3, mastery of the statistical issues – ideas, not necessarily the mathematics[a] – is essential

The good news is that we now have, for many applications, marvellous software that will take care of the calculations.

[a] Also, *p*-values may not be a high priority!

AIC and Friends, for Complex Error Structures

linear iid models: Parameter variances are multiples of s^2.

multi-level models: Variance estimates for different parameters &/or statistics of interest are different functions of the components of variance. Focus for optimization?

e.g. balanced hierarchical 3-level model:

$$\mathrm{var}[\mathrm{y_{ijk}}] = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$$

Means over *n* low-level units:

$$\mathrm{var}[\bar{\mathrm{y}}_{\mathrm{ij.}}] = \sigma_1^2 + \sigma_2^2 + \frac{\sigma_3^2}{\mathrm{n}}$$

Are all parameters equal?

c.f., Vaida & Blanchard: Conditional Akaike information for mixed-effects models. Biometrika 92: 351370, 2005.
(model fitted with random clusters; within cluster estimates)

Challenges for Statisticians

- Engage with these “alternative” streams of statistical development. Maybe appropriate some of their ideas.
- Why do random forests often do so well?
  Can more structured models benefit from the same ideas?
- Get a better handle on predictive accuracy measures (especially for for models with a complex error structure).
- High dimensional data (e.g., expression arrays) & Variable/feature selection are huge challenges.
- Make analysis as automatic as possible, but not more automatic.[a]
- Automate using first-class analysis tools. (We need R^{++}).
  Get greater involvement from computer scientists?

[a] This misquotes a saying that is attributed to Einstein!

In Summary

- Check accuracy claims with great care:
  - Are they (in some weak sense) justified?
  - Do they generalize to other datasets?
  - Are they relevant to likely practical use of the method?
- Complex structures of variation (errors);
- Tell it with graphs.
- In reporting evaluations/comparisons
  - Tell all algorithmic steps, in careful detail;
  - Report reproducibly (`Sweave`, etc.)

Reference

This talk is based in part on

Maindonald, J. (2006). Data Mining Methodological Weaknesses and Suggested Fixes. In Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney. CRPIT, 61: 9-16. `http://crpit.com/Vol61.html`

Note also:

Maindonald, J. H. 2005. Data, science and new computing technology. New Zealand Journal of Science 62: 126-128. `http://www.maths.anu.edu.au/~johnm/science/rethink.pdf`