Data Mining Methodological Weaknesses & Suggested Fixes

John Maindonald

November 30, 2006

Four Motivations

- Report for ANU Administration
  - How do data miners explain themselves?
  - What is the practice; how is it done?
- Refereeing experience.
- Teaching a DM course.
- Frustration with the superficiality of data mining texts. (Math3346)[1]

Key themes

- There are good & bad approaches to inference.
- Effective inference blends computing power with analytical insight and skill.
- Two (or more) cultures?

NB: Any use of data to reach a conclusion is an inference.

[1] http://datamining.anu.edu.au/student/math3346_2006.html

Inference I

Data –
from where?
how collected?

Analyse data –
Draw conclusions

Use of
analysis
results.

- What justifies drawing the arrow?

  Contrast audacious archers (feeble justications suffice) with savvy sleuths (who assess the hazards).

Two sets of terminology – populations or processes

- Sample from source population; sample from target.
- Data from source process; data from target process.

Source of data

Target for
use of results

Inference II

Naïve view

- Find data that seem relevant.
- Analyze data (trees, neural net, latest DM gismo, . . . ?)
- Write a report.

More scientifically – consider the why and how!

- Why are we doing it?
  - Among the many reasons, none justify mindless flailing!
- Identify and collect the relevant data.
- Use methods whose properties are known and understood.
  - Finally gold must be distinguished from dirt.
  - In new territory, the user must do his/her own evaluation.
  - Analyses with many features are new territory for everyone.
- Use a presentation that conveys the message effectively.

Ideas that Underpin Inference from a Given Dataset

- Populations and Samples;
- Source vs target population;
- Modeling issues (algorithms are not enough);
- Prediction & predictive accuracy
  But what is the relevant measure of predictive accuracy?
- Detecting pattern, cf also Exploratory Data Analysis
  Interestingness has to be modeled!

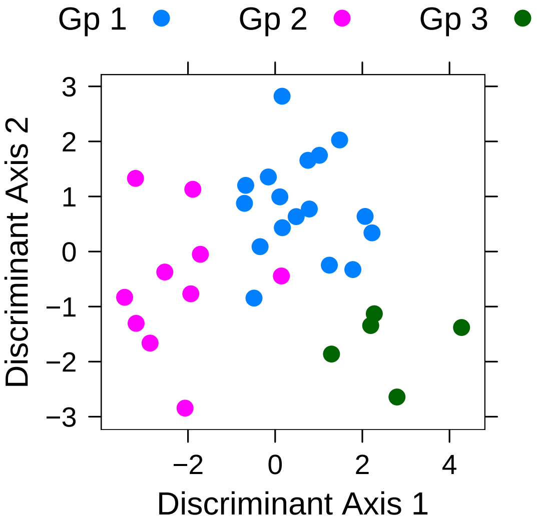
General Observations

- 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!

What is the appropriate measure of predictive accuracy?

Spurious Appararent Pattern (Interestingness?)



Method

Data were random – 3 groups, 32 points in total. Select the "best" 15 features, from 500. Plot first two linear discriminant scores.

Source vs target population

Are source (from which data were collected)
and target (to which results will be applied) the same?

Usually, No!

Examples

Source of training data	*Target*
Victorian pre-election polls	Election results
Historical credit scoring & loan default data	Current loan applicants
Christmas 2005 sales	Christmas 2006 sales
NSW country towns	Victorian country towns
2005 successful applicants	All 2006 applicants
Expression array experimental data	All such experiments (or amounts of RNA?)

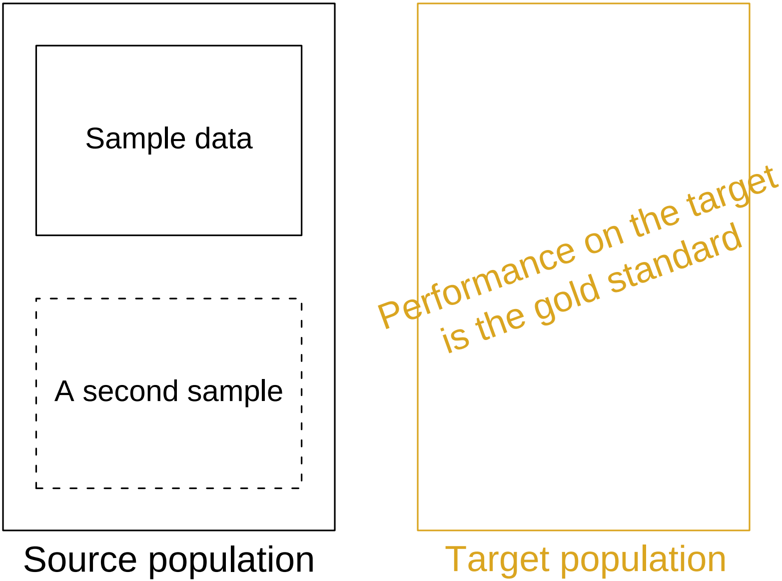
Weak and Strong Testing

- Test data must be independent of training data; else the accuracy measure will be flawed.
- Use of training/test data from the source population, and cross-validation, provide weak accuracy measures. (Section 1: may be better than nothing, if correct!)
- Strong accuracy is accuracy for an intended practical use; test data must be from the target population.

Commentary

- Better weak accuracy performance may not imply better strong accuracy performance! See Hand's paper.
- Consider fortification, i.e., add elements of strength?
- Strong (or even fortified) testing has been unusual in the DM literature, notwithstanding its practical importance.

Target Population Performance vs the Gold Stamdard



Different Relationships Between Source & Target

Source versus target	*Are data available from target?*
1: Identical (or nearly so)	Yes; data from source suffice
2: Source & Target differ	Yes
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.

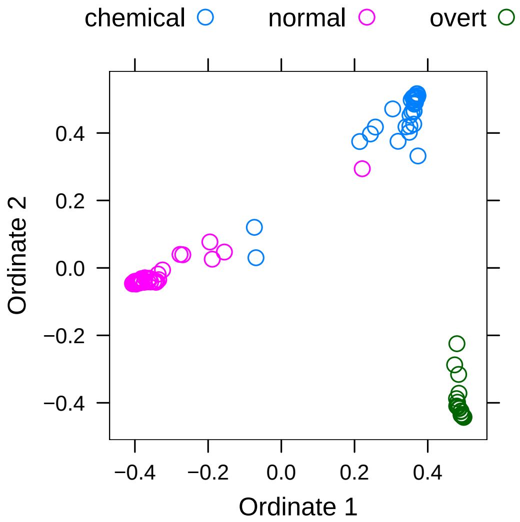
When algorithms are evaluated or compared ...

- 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 comparison with random data.
  (It can be a useful reality check.)
- Try each algorithm with simulated data.
  (Under what circumstances does it perform well/badly?)
- Give a 2-D or 3-D view that identifies "difficult" points.
  - Note 1: Is 2-D adequate? Should it be 3-D, 4-D, ...?
  - Note 2: Graphs for the training data are, strictly, flawed.

Even if done well, most papers compare weak accuracies.

Be up front; admit the weakness!

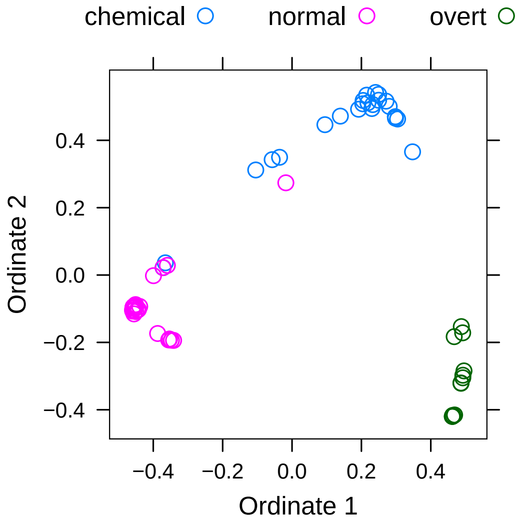
I: A Mildly Biased (but Nonetheless Useful) Plot



2D View I

Data are 3 measures on 145 diabetics. Use 1-proximity, from random forest, as pairwise distance.
NB: Distances are for training data; hence mild bias.

II: An Unbiased Plot, for 50% of Data



2D View II

50% of data, plus the 3 "doubtful" points, were set aside for testing. Plot is for these test data; hence unbiased.

Why plot the data?

- Which are the difficult points?
- Some points may be mislabeled (faulty medical diagnosis?)
- Improvement of classification accuracy is a useful goal only if misclassified points are in principle classifiable.

What if points are not well represented in 2-D?

Alternatives include identification of points that are outliers on a posterior odds (of group membership) scale.

Take-home message: There are other issues than predictive accuracy.

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 have fixed and random parts, right? Models are needed that allow a complex error structure.)
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[2] – is essential

The good news is that we now have, for many applications, marvellous software that will take care of the calculations (but large datasets may require a super-grunty computer!)

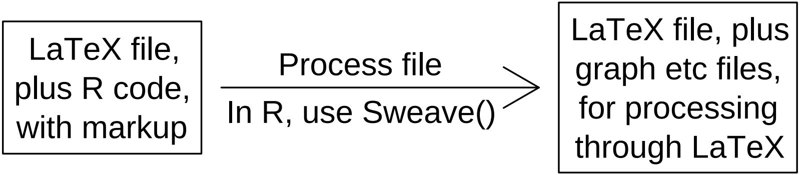
[2] (also p-values may not be a high priority!)

Reproducible Reports – the Gold Standard

Give a file that, when processed:

- reproduces all computations whose results are given;
- combines those results with the text of the paper to reproduce the entire paper. This includes
  - results in the text, tables, graphs, and other output;
  - any of the computer code that appears in the paper.

One possibility – Use R's `Sweave()` function



NB: The markup information is used to generate all needed `includegraphics` etc LaTeX commands.

In Summary

- Source and Target
  – flawed, weak and strong measures;
- Complex structures of variation (errors);
- Tell it with graphs.
- In reporting evaluations/comparisons
  - Tell all algorithmic steps, in careful detail;
  - Report reproducibly (`Sweave`, etc.)