Astronomical Data Mining and Databases

Clive Page, University of Leicester,

2001 January 29

1 Astrogrid Objectives

The main Astrogrid objectives are:

1. To integrate existing and future data archives into a "virtual observatory".
2. To develop facilities for data mining.

When we started work on Astrogrid nearly a year ago, I assumed that some standard database management system, whether relational or object-oriented, would form the infrastructure which would support both of these objectives quite well. But now I have come to doubt whether a standard DBMS is very suitable, particularly for data mining.

2 What do we mean by Data Mining?

- Essentially: finding information buried in a mass of data.
- An alternative term is: Knowledge Discovery from Databases (KDD)
- Elements include:
  - Exploratory data analysis or Data Prospecting
  - Data visualisation.

In the era of massive datasets and standard processing pipelines, it is all too easy to reject everything not predicted in advance.

Data mining should be **contrived serendipity**.

3 Data Mining Functions

- Regression analysis (cross-correlation) to find association rules.

  Examples from the history of astronomy include: the discovery of the HR diagram; the Hubble recession law.

- Deviation detection: finding outliers from distribution, or exceptions to general rule.

  Examples: discovery of pulsars while studying interplanetary scintillation; finding quasars as stars with strange spectra.

- Sequence analysis, such as auto-correlations and period searching in time-series.

  Examples: pulsar searches; SETI@home.

- Clustering and classification algorithms.

  Examples: type I and type II supernovae.

- Similarity searches.

  Examples: finding novae, SNRs, or minor planets using a blink photometer.

- And others...

4 Data Mining Challenges

4.1 Data Quality

Commercial experience is that:
20 – 30% of time spend understanding data characteristics, and
50 – 70% of time spent on data cleaning.

Attention must be paid to: bad data points, duplicated records, null values, errors of observation, upper-limits.

Data Warehousing

Commercial databases are dynamic. Data mining requires creation of a Data Warehouse which containing static, carefully cleaned data, all in compatible form.

4.2 Performance

If you try to read 1 terabyte of data at 10 MB/sec (the speed of a typical Fast Ethernet) it takes 1.2 days.

Possible solutions to the performance problem:

- Sampling – suitable for some work, e.g. regressions.
- Indexing – to select subsets of interest.
- Parallelism in hardware: multi-processor systems, Beowulf clusters, RAID-1 disc systems, etc. Computational Grid concepts clearly very relevant here.
- Efficient data formats, e.g. binary not text, column or row ordered to match algorithm requirements, related data clustered together.

4.3 Heterogeneous Data Formats

Astronomical data may be held in any of:

- FITS tables and images

- XML structures (soon)
- DBMS formats: Oracle, Sybase, PostgreSQL, etc. – no metadata
- Other astronomical package formats: HDS, MIDAS, IRAF, STSDAS, QPOE...
- Plain text files or binary files

Possible solutions:

- Convert required datasets to some common format (what?)
- Convert on the fly using middleware.

4.4 Access to remote data

Will often be needed especially for multi-wavelength data mining.

Two options:

- Access over the network. The SuperJanet4 backbone will reach 20 Gbits/sec by 2002, which is 200 times bandwidth of Fast Ethernet. But latency may still be a problem.
- Copy required datasets to local node. Fortunately data mining mostly requires either sections of tabular datasets, or parts of images, which may be small enough for this to be feasible.

5 Software

5.1 Existing Astronomical Packages

Data mining elements already exist in Starlink software collection, MIDAS, IRAF, FTOOLS, etc.

But: they use a wide variety of data formats and user interfaces.

5.2 Commercial DBMS

Many sites already use Oracle, Sybase, Informix, O2, etc.

- Good basic data management, but optimised for transactions not for fast access to static data.
- Indexing usually limited to hash tables and B-trees in one dimension. With few exceptions, most DBMS lack multi-dimensional indexing, bit-map indexes, inverted lists.
- Proprietary data formats.

- SQL (or OQL) are pretty useless for data mining. A query to find all rows with values of some parameter more than 3 sigma from the mean takes some very complex SQL, which is unlikely to be executed efficiently.
- There is essentially no metadata support in relational DBMS.

5.3 Statistical and Visualisation packages

Some data mining functions exist in many current packages, for example: Genstat, IRIS Explorer, IDL, Maple, Mathcad, Mathematica, Matlab, SAS, S-plus, SPSS, SCILAB, Statistica.

But:

- All have their own formats and user interfaces.
- None provides more than a small part of the required functionality.
- They were designed for megabyte datasets, not terabyte ones.
- The very fact that there are so many packages with large overlaps of functionality suggests strongly that there is a large market, but that no one package is judged to be very good, otherwise it would surely have come to dominate that market.

5.4 Data Mining Packages

Many are appearing, but all seem to come from relatively small companies, and are designed for use on commercial data warehouses.

One interesting package: Kensington Enterprise Data Mining System is a spinoff from the Department of Computing at Imperial College. This claims to support "e-science" as well as "e-commerce".

6 Conclusions

- Commercial DBMS do not appear to be the solution.
- Some existing software packages show promise and should be investigated.
- Much new software has to be developed if we are to satisfy the data mining needs of astronomers.