Burrowing into the bran tub

Felix Grant assesses statistical information mining techniques

Intelligence organs of a fairly small, but technologically proficient, country recently managed to clone the email and document bases of a fairly large, but unsuspecting, company with an active and diverse research operation. Not the research results bases themselves; just the email and document archives. The immediate benefits of the theft were obvious, mundane, and irrelevant here. More interesting from a scientific computing viewpoint (if no more morally or legally defensible) were the spinoffs from applying statistical information mining techniques to the combined database contents.

As a direct result of the operation other companies, seen as friendly to the small country's interests, received repackaged and anonymised material suggesting productive new lines of enquiry. Patents resulting from these have already been filed; others are in the pipeline. The large company that was the target of the operation, significantly, remains unaware of the opportunities that exist within components of its own activities, which have never been brought together.

Or so, at least, I am told. I have no way to verify the truth of this story, of course. But the young man who tells it to me, sitting in the café at the National Gallery, his hands continually roving across the keyboard and digitiser pad of a notebook as he talks, has never sold me a lemon yet. And, true or not, it illustrates a truth: that much knowledge is locked away in information stores assembled for one set of reasons and never reexamined in other ways.

Less melodramatically, and less dubiously, information openly published on the internet forms a huge field within which to prospect potential information seams – 'The low user entry barrier of the web has resulted in massive amounts of unstructured and weakly structured data referring to objects, concepts, user interests and communities,' to quote the Digital Enterprise Research Institute at...
Galway. As SAS’s David Smith points out (see ‘Keep taking the tablets’), tweets and blog entries can contain pointers to early identification of potentially vital phenomena. This is an aspect of what is known in the industry as pharmacovigilance, which ‘can be defined as a set of practices aiming at the detection, understanding and assessment of risks related to the use of drugs in a population, and the prevention of consequential adverse effects [of] in a narrower sense ... postmarket surveillance’.

A leader in making such text searching accessible to smaller, nontraditional users and demonstrating the value of placing intelligent defaults in their hands, is the Data Miner Recipes (DMR) tool in Statsoft’s Statistica. It provides a clear cut, step-by-step path through a data mining project from initial cleaning and preparation of the data through to building and evaluating a model. It doesn’t quite pass my ‘10 year old test’ (in which I ask a child to try and use a software tool for a practical purpose), but that is mainly a cognitive difficulty in grasping the ideas involved. Over the past month, on the other hand, I have seen a dozen 14-year-olds embrace my copy with enthusiasm and derive meaningful results from experimental data in a few mouse clicks. Once the data file and variables are selected, the project can carry the user through to finished models without much interference – though any degree of sophisticated control is possible at every stage. There are other approaches, but the DMR wizard removes most of the barriers to initial adoption and familiarity.

Extracting connections from text bases is, of course, neither the only way to skin a data set nor separate from other approaches. I’ve recently been watching a team of farmers without formal statistical training use DMR to interrogate a combination of numeric and text data. Seeking exploitable patterns in the behaviour and influence of natural pollination vectors, they join a wide range of observational and administrative records using time as a common key and then let DMR do the rest.

Agriculture is a rich recipient of the benefits accruing from information mining approaches. A quick dip into the literature on trait selection over the past year showed them to be behind six of the first seven results: adipocytes[2] and growth[3] in meat stock, milk production[4], neuroendocrine correlation in poultry[5], protein interactions in yeast[6], and crop breeding[7]. This is not surprising, since information mining is one of the staple of new methods that accompany the explosive blossoming of genetics. As the editor-in-chief of the Federation of American Societies for Experimental Biology’s FASEB Journal observed a few months ago (see ‘Science in the information age’), information mining rides a wave of new approaches which, in aggregate, represent a radical shift away from traditional hypothesis-based science. Or, to borrow a Dutch colleague’s picturesque metaphor, ‘We no longer ask Sinterklaas and Zwarte Piet for a specific named present at the beginning of the research season and wait patiently to see whether we get it at the end; we go and burrow through their lucky dip bran tub to see what they’ve got.’

This is one of those areas where traditional hypothesis-based approaches may never uncover a linkage, because there is nothing to suggest the hypothesis in the first place. Or, alternatively, one function of information mining can be seen as an enhancement of intuition, dramatically accelerating the rate at which hypotheses can be generated.

While the benefits of information mining are real and valuable at the small user end of the research spectrum, they grow exponentially with scale and are significant building blocks for science and for national or regional economies. At that scale, methods development has to be explored outside the user or provider framework. In Europe, the EPCC (Edinburgh Parallel Computing Centre)
Centre) manages the ADMIRE (Advanced Data Mining and Integration Research for Europe) project, which 'aims to deliver a consistent and easy-to-use technology for extracting ... meaningful information by data mining combinations of data from multiple heterogeneous and distributed resources... which will give users and developers the power to cope with complexity and heterogeneity of services, data and processes'.

ADMIRe emphasises the need to accommodate the increasing size of information stores, sophistication of extraction requirements, and complexity of the resulting systems, responding with a unified conceptual structure based on integration of separate expertise types in relation to a defined component library structure. The outputs are ...a framework, an architecture and a set of use cases that illustrate how they can be used to improve DMI [data management integration]. Alongside conceptual development and implementation, it has applications on which its methods can be demonstrated in practice as well as a growing number of studies that call on its capabilities.

Keeping a grip on the bigger picture

Algorithms that work on complex networks with hundreds to thousands of vertices often disintegrate on real networks with millions (or more) of vertices. For example, betweenness centrality is not robust to noisy data (biased sampling of the actual network, missing friendship edges, etc.) They require niche computing systems that can offer irregular and random access to large global address spaces. ... the newest breed of supercomputers have hardware set up not just for speed, but also to better tackle large networks of seemingly random data. ... Applications include anywhere complex webs of information can be found: from internet security and power grid stability to complex biological networks."

David A Bader, Georgia Institute of Technology[3]

Labelling the building blocks

Liangxi Han and others[6] (conclusions of an ADMIRE supported study) state: "...we have developed a new data mining framework to facilitate the automatic annotation of gene expression patterns of mouse embryos. There are several important features of our framework: (1) the combination of statistical pattern recognition with image processing techniques can help to reduce the cost for processing large amounts of data and improve the efficiency. We have adopted the image processing method to standardise and denoise images. Wavelet transform and Fisher Ratio techniques have been chosen for feature generation and feature extraction. The classifiers are constructed using LDA. (2) For enhancing the extensibility of our framework, we formulate our multi-class problem into a two-class problem and design our classifiers with a binary status: 'yes' or 'no'. One classifier only identifies one anatomical component. Classifiers for each gene expression are independent on each other. If new anatomical component need to be annotated, we do not have to train previous classifiers again. The classifiers can be assembled and deployed into the system based on user requirements. (3) We have evaluated our proposed framework by using images with multi-gene expression patterns and the preliminary result shows our framework works well for the automatic annotation of gene expression patterns of mouse embryos."

References and sources

For a full list of the sources and references cited in this article, please visit www.scientific-computing.com/features/referencesaug10.php