

SCIENTIFIC COMPUTING WORLD

KXEN

➤ Shaping the study space



REVIEWS

KXEN: Shaping the study space

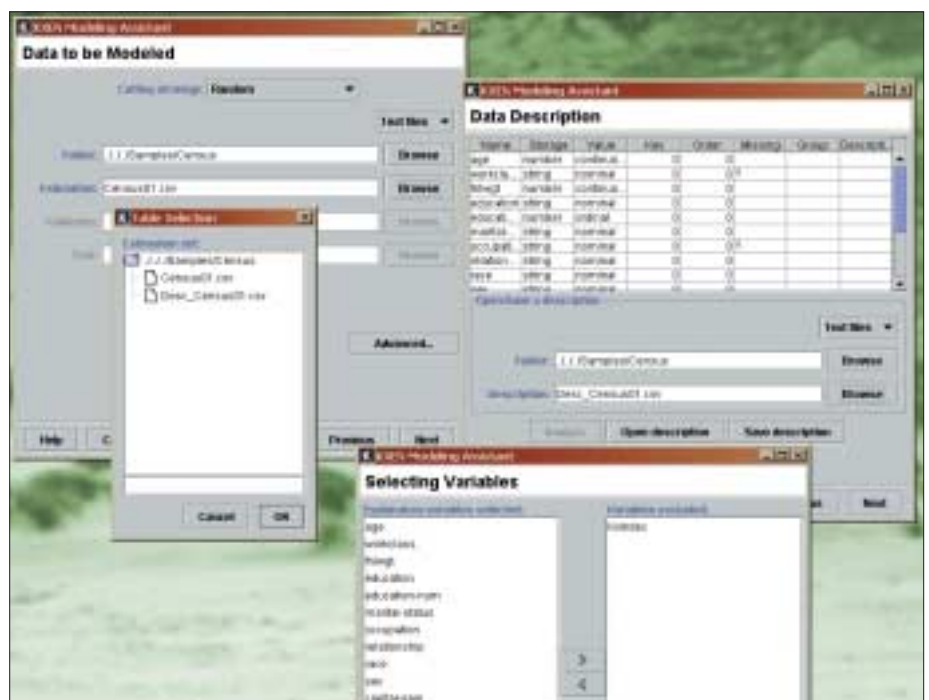
Felix Grant

finds much to applaud in KXEN's 'JCB-style' datamining software

Life, it recently seems, is a road movie. This story took me out and about, pursuing a 'ghost in the machine' through a tour of its application sites. The journey started with me in a distinctly sceptical frame of mind about the product under review; and ended with my conversion to enthusiasm.

Last summer, I wrote about an environmental project: statistical software from several publishers working together, tracking down the source of pollution in a Third World swamp from data on the distributed pollution effects ('Return of the Swamp Thing', *Science Computing World*, October 2002). It prompted a call from Grayson Amies, UK Operations Manager of KXEN Ltd. He believed that his company's software would be of interest in the context of the swamp search. KXEN sounded as though it might be connected with covens and witchcraft (you may, of course, subscribe to the view that all data mining is arguably black magic) but actually abbreviates 'Knowledge eXtraction Engines'. I was clear of the swamp by that time, but we arranged a view of the software anyway.

KXEN Analytic Framework is not an application but a component. It is usually to be found embedded within other software regimes. The engine can be connected to a DBMS system (Oracle, for example, or MS SQL Server) through ODBC drivers. There are also add-in possibilities for other products: I looked at a Clementine installation. The models generated can be saved after development and reloaded for a production (for science, read 'reuse') phase, or (through



the standalone code generator) converted to equivalent C or XML code.

I couldn't, therefore, really get a feel for its operation at the sharp end solely from the core product supplied to me. Evaluating it involved three separate parts: learning to fly the supplied version; testing it against known problems from the archive; and seeing its embedded manifestations at work in existing contexts. I started by writing up the 'innards' as well, the use and balance of algorithms, but abandoned them after a time. Everything I might have said in that department is, to be honest, freely available if you want it from the company or elsewhere on the web. Suffice to say that what is happening under the hood is very different, at least in its operational philosophy, from what we are used to in conventional data mining built around classical techniques or newer approaches such

as neural networks.

My initial scepticism was triggered by the approach taken to data mining, and reinforced by KXEN's briefing material, which emphasised exactly this issue. KXEN's approach of 'fully automated answers ... with a fairly good level of accuracy' is one that instantly fits pragmatic business perceptions, but sits uneasily with those of traditional science. In thinking this way, however, I now realise that I was not examining my own thinking with sufficient rigour.

Rigorous analytic work

There are a number of views taken of data mining in science. At one end of the spectrum lie those who regard it as a mirage, a distraction from rigorous analytic work on the classical model – one academic colleague uses the phrase 'dilettante sledgehammer'. At the

other, some view it as an alternative to conventional analysis. Between those extremes, I see the role of data mining in science as being that of a pathfinder – a way to identify possible avenues of inquiry within an apparently amorphous field of study, for pursuit by more traditional means. For a commercial enterprise, the knowledge that customers who buy nappies often buy eggs as well is a valuable answer in its own right; from a scientific point of view, it is valuable as the beginning of a question that might otherwise not have been asked. Seen in this light, KXEN's 'fully automated ... fairly good' approach makes a lot of sense: narrow down the data-forest to individual question-trees as quickly as possible, freeing up resources for subsequent specifics.

With this in mind, my first port of call was a return to the data (and, so far as possible, team) in the swamp pollution study that triggered this review. How would that study have gone had KXEN been applied to it? Not

Oh, my darling Clementine

There is a KXEN add-in for Clementine, the well-established data mining specialist application from SPSS. This seemed particularly interesting to me, not only as a sort of 'anatomical model' of how the integration with applications is managed, but also as an example of coexistence between KXEN's approach and more traditional data mining techniques. So, with help from inside SPSS, I tracked down a site where the combination was in use.

The installation can't be described as a doddle; but then nor can Clementine itself, and you will obviously not take this route unless you have already decided that Clementine offers much of what you want. If you do not already have a Java 1.3 compatible Virtual Machine, you will have to install one (this is a KXEN requirement; Sun's Java Run Time Environment is provided). From that point on, it's not difficult but you do need to keep your mind on what you are doing. Clementine's user interface needs to be updated with (supplied) CEMI files; then the KXEN modules (four of them: K2R, K2S, KSVM Linear and KSVM Non-Linear) must be added into Clementine as nodes, through the palette manager.

Once the initial set-up is done, if you know Clementine you can just take it away. Insertion, execution and interrogation runs as you would expect. This was a very rewarding part of the review process; in fact, I was sufficiently impressed to use this as a testbed for rerunning and crosschecking a sample of swamp project comparisons.



as easy to test as it sounds. To make a fair comparison, the data (which was assembled slowly, with analyses conducted alongside collection) had to be approached in the same sequence of partial sets as the original. The audit trail of the study contained the necessary detail for reconstruction, but not in a format that anticipated reconstruction of an exact replica. Then again, the introduction of a new software component raised issues of best resource deployment. To cut a long story short, we ran several different simulations of the original field study, with different types and degrees of KXEN involvement, to see what the outcomes might be.

We applied the KXEN software to each stage of the original investigation, to see what effect it would have had if available for that stage. Then we tried again, re-simulating the study from various key stages through to the end. The results were complex. Full replacement of all software on the original study with KXEN speeded up some parts of the study, slowed down others. The result was a much longer time-line though fewer resource demands; but this is not a likely or sensible scenario. More likely would be addition of KXEN to the existing regime, or its considered replacement of one or more existing tools within that regime

Subsequent refinement

The big gain over conventional analytic tools was in the 'fully automated' bit. Not that we could run it in a turnkey mode, but a set of rule-based operating schedules achieved the same effect. This is not a program which can be picked up and instantly used by any intelligent five year old; but for routine tasks it is certainly more rapidly mastered by people with less experience or statistical knowledge, than most alternatives. A group of 17-year old students, sufficiently interested in environmental issues to take part but with no science or statistics training, were able to take data and instruction sheet in hand and turn out useful results for subsequent refinement. The original study had few lay, non-expert members involved in data analysis; but that is because the opportunity for them was not there.

Replacing the analytic products with KXEN but leaving the visualisation software in place, there is still a time penalty – but almost all the experts became redundant. The original project did not have to pay its experts; they gave their time free. But rounding up a team of specialised volunteers whose expertise is available all at the same time is not an easy matter – just as much of a resourcing challenge as finding a budget to pay them.

The best result came, not from replacing other tools with KXEN, but from allying them with it in a new mix. Allowing it first bite at the data, then comparing the results with visualised evidence, gave potent pointers towards 'strong' and 'weak' datasets, for subsequent detailed analysis in other ways. We estimate that, had we applied KXEN in this way from the outset, we could have cut time to completion by between 20 per cent and 50 per cent while simultaneously reducing investment in specialist expertise by something like the same amount.

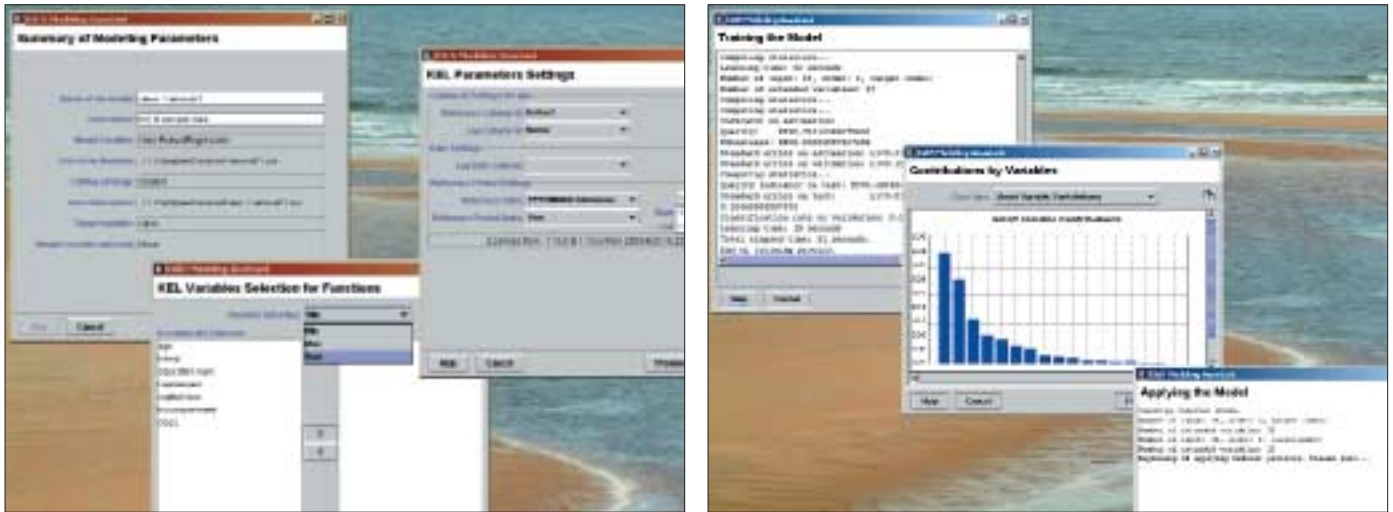
Similar comparisons on other environmental studies gave results that varied according to

'For routine tasks it is certainly more rapidly mastered by people with less experience or statistical knowledge, than most alternatives'

the type of study. The biggest payoffs came in settings where there were no existing hypotheses from which to work (a sudden case of die-off among a particular species, with no apparent environmental change, for instance).

The undoubted power of most data mining tools is not easily brought to bear in effective ways without experience. I don't suggest that KXEN is 'data mining for dummies', but its inbuilt heuristics and largely preset tuning do increase the hit rate for less experienced workers – or for experienced workers who have very little time.

However carefully done, post hoc comparisons are always questionable. Next step, then, was to apply the software to new problems. In the case of a factory malfunction, results were encouraging, producing a solution without resort to any of my usual toolkit. Then I went out on the tour of patient KXEN users on the ground, taking with me the dataset from a habitat encroachment study about to be started by a colleague. In the interests of impartiality, I prefer to track down my own contexts rather than having them provided by a software provider.



Finding users of a product is usually straightforward; but with an embedded component it becomes more difficult. Surprisingly few KXEN users, I discovered, are aware that they are such. This is a striking compliment to the software, but a headache for someone seeking it. Nevertheless, I identified a number of sites willing to let me play with their facilities.

KXEN's existing market seems to be in business, not science; but there is no inherent reason why. Data is data. In a large marketing operation, I fed habitat data to a machine used to a diet of loyalty card analyses. There was some need to pad or trim the data set, ensuring that the variables matched (in position and type) those that the machine expected; and, of course, to interpret the output, which referred to customers and products rather than organisms and influences. But, beneath the superficial, an analysis of data was being done without regard to purpose. This analysis was repeated with other commercial concerns, and the accumulated results taken home for examination.

Interest follows annoyance

One thing which became clear in all of these commercial contexts was that very little statistical or data mining expertise was available at the point of analysis or decision. The set-up had been made by 'somebody from IT', and was now run as a black box by administrative staff who sifted the output in a sort of triage system – reject, pass onward to somebody else for consideration, or act upon. It seemed very effective.

In my own case, I took all the output back to my habitats colleague for comparison with his own first efforts. His initial response was annoyance, quickly moving to interest. About 85 per cent of what I brought back with me, he had already arrived at on his own. Of the remaining 15 per cent, which represented

possible avenues of inquiry that he had missed, roughly 12 per cent looked promising for his purposes while 3 per cent were spurious (similar proportions applied within the 85 per cent overlap, as well). To a large extent, though, that 15 per cent improvement and 3 per cent sink rate both miss the point. If you eliminate finding the KXEN installations in the first place, travelling to them, negotiating with their owners, and so on, I had spent perhaps five hours of actual work on this, only minutes of computer time, and no expert knowledge. He, by contrast, had worked many hours over a period of six weeks, applying his specialist subject and statistical know how – which, of course, accounts for the annoyance.

He had made use of another, very good set of data mining tools found within his institution's standard analytical package. Does this mean that KXEN knocks that product of the shelf? No, it doesn't, for reasons which lie within that 3 per cent margin of spurious results. It does, however, argue very convincingly that KXEN's operational philosophy earns its own place on that shelf alongside the established toolset.

Another person looking separately at KXEN in a science context, whom I am not free to name or quote, raised the question of whether this is a trend which will make data mining and statistical specialists redundant? Their conclusion was that it would not, but that the role of those specialists will change. I agree with that conclusion, but for different reasons. An old story from my childhood, when many of my relations were in the building trade, has a man complaining that 'one worker in a mechanical digger makes redundant a hundred with shovels'; to which the riposte was that 'one with a shovel makes redundant a hundred with teaspoons'. The mechanical digger is too valuable to ignore and there remains, as any archaeologist will tell you, a valuable place for the careful use of shovel,

trowel and even teaspoon. A tool like KXEN strikes me as being akin to a JCB – it carves its way though an awful lot of data, in a very short time, freeing up the specialists to spend more time applying their expertise to the subsequent phases. It also increases the degree to which non-specialists can become valuable contributors. Generalists like me make their living by working in the crevices of science, where big systems can't reach; I reckon that will increasingly be true of specialists as well. There is an apt military analogy: data mining could be described as a means by which to 'shape the study space'. My judgment would be that, as an addition to the repertoire of investigative science, KXEN greatly enhances the speed and efficiency with which the first, broad brush shaping can be done.

What's going on?

KXEN doesn't approach data through the array of model-building tools conventionally applied to data mining; it looks instead to the classes of functions which may underlie them. It owes much of its theoretical foundation to Vladimir Vapnik, a Russian mathematician who spent his early working life in the USSR's Institute of Control Sciences before moving to the US and AT&T/Bell. He is behind the development of the Support Vector Machine (SVM), of an expected risk minimisation theory for empirical data, and of the 'VC dimension', which describes mapping complexity in a family of mathematical functions.

The VC dimension can predict the consistency of models and their stability in application. He has applied this to, amongst other things, study of the reasons for neural network success. SVM theory is a learning machine approach to control of the VC dimension in a model.