How to Choose a Data Mining Suite

Robert A. Nisbet

The choice of a data mining suite is not an easy task. This article provides a brief outline of some considerations that could affect your decision. Contrary to common opinion, the best tool suite for you may not be the most advanced tool, it may not be the one with the most data mining algorithms nor the one that gives the greatest accuracy in prediction. More important than all of these things is identifying the tool suite that is:

- Easy to use,

- Provides acceptable accuracy (even though not the highest accuracy available),

- Able to perform all the common tasks in a data mining project.

Ease of Use

Some traditional (and heavily advertised) data mining tools may provide a rich variety of data processing and modeling capabilities, but require a legion of "priests" to use them. Often, these "priests" of data mining are developed only after many years of practice and travel up the learning curve of the tool's capabilities. Rather than be very procedural (programmed with a scripting language), the user interface to data mining technology should be like the interface to automobile technology. The great success of the automobile is because it brings the benefits of sophisticated engineering technology down to the level of use appropriate to the common man and woman. You don't need to be an expert in internal combustion technology or understand the complex relationships between gear ratios and acceleration to use a car effectively. All you have to do is get behind the wheel, turn on the ignition, step on the gas and steer or use the brakes at the appropriate times and, voilà, you are an expert user of the automobile! Using a data mining tool should be like that. You might be surprised to learn that several modern data mining tool suites approach that ease of use.

Accuracy: How High?

Suppose you could buy a tool for a fraction of the cost of the priestly tool (maybe 20 percent or less), which permitted ordinary business analysts and statisticians to create models that were 80 percent as good as the priest could create. Would you choose to buy the priestly tool, or the 80 percent tool? I think for most companies, the answer is the 80 percent tool. For this case as well, several data mining tool suites available today can provide this functionality. For other purposes, the best tool might be the most accurate tool.

Ability to Perform All Common Data Mining Tasks

Most data miners will tell you that 70-90 percent of the time required to perform a data mining project is spent in data preparation for modeling. Reasons for this include:

- Most data mining algorithms require clean and complete data records as input. No data mining tool in the world can analyze data that does not exist (missing data in some fields).

- Most data in commercial databases was collected from transactional systems to serve

query and reporting purposes, not analytical purposes.

- Most data in commercial databases are rather "dirty." That is, databases often contain inappropriate data, training data, improperly input data or just plain garbage data. Even if the data appears to be clean, historical data records may reflect changes in coding and aggregation rules at various times in the past, which must be reconciled. In addition, data formats may not be consistent across databases used as data sources. Finally, data may require transformation to different ranges or different expressions (letters changed to numbers), or new variables may be needed that are combinations of existing variables. A good data mining suite will provide tools for performing all of these operations. Some data mining suites are better at it than others.

This article reviews five of the most useful and powerful data mining suites available, STATISTICA Data Miner, SPSS Clementine, Affinium Model, Insightful Miner and KXEN. We can use these tools to illustrate how you can evaluate how suitable a data mining tool suite is for your use. Let's cut to the chase right in the beginning.

There is No Best Tool Overall

There is no best tool overall. Are you surprised? Well, competition in the marketplace almost guarantees this. If a particular tool is successful enough to make it into the mainstream of data mining use, it must serve at least a moderate segment of business needs well. Each tool suite has its strengths and weaknesses; each tool suite may be the best for particular needs in particular companies. Each of the five tool suites will be reviewed and classified according to their best uses. From this evaluation, you can gain enough information to take the first step in the choice of the data mining tool suite that is right for you.

The first step is to look at the features and functions of the data mining tool suite. While this only first step in the decision-making process, it may not be the most important consideration for you. Figure1 shows a weighted comparison of the features and functions of the tool suites. You will notice by comparing the relatively moderate cost and the weighted score across all features and functions, that STATISTICA Data Miner is the clear winner. This *does not* mean that this tool is best for you. Your needs may not require (or your budget may not permit) the rich variety of capabilities provided by STATISTICA Data Miner; Insightful Miner (with its great ease of use and affordability) may be just the right tool for you, regardless of its relatively low score in Figure 1. Or, you may want a fully automatic data mining engine that can generate models of the very highest accuracy, to which you are willing to submit data in the suitable format. If so, KXEN is the right tool suite for you, providing the cost is acceptable. Clementine and Affinium Model tool suites provide intermediate solutions between those of KXEN and Insightful Miner, in terms of functionality and cost.

Figure 1

Weighted scores for the ability of five data mining tool suites to perform common data mining tasks. (CLEM = Clementine; STAT = Statistica Data Miner; IM = Insightful Miner; AM = Affinium Model; KXEN = KXEN; MAE = Mean Absolute Error - Typical performance of the five tool suites.)

Performance

The performance of each tool suite was tested with a neural net (when possible). Prediction accuracies are shown in Figure 2. For data sets with a binary target variable, percent accuracies are listed for the 1s and 0s in the form of 85/56. For the Abalone dataset with an

integer target variable, overall accuracy and mean absolute error (MAE) are listed.

PERFORMANCE	CLEM	STAT	IM	AM	KXEN
Failures dataset Prediction accuracy Model used	69.1/ 83.4% Neural Net	81.2/80.8% Neural Net	74.8/88.7% Neural Net	81.8/96.8% Neural Net	90.6/82.5 SVM
Thrombin dataset Prediction accuracy Model used	30.5/100% Neural Net	1:32% 0:100% Neural Net	31.9/100% Neural Net	31.9/100& Neural Net	75/0% SVM
Abalone dataset Prediction accuracy Model used	MAE = 1.50 Neural Net	MAE = 1.45 Neural Net	MAE 1.68 Neural Net	(Error)	MAE 1.95 SVM

Note: Affinium Model would not perform at all on the abalone data set.

Figure 2: Relative Performance Statistics for Five Data Mining Suites

The relative performance statistics in Figure 2 may not reflect the relative ability of the tool suite to perform on these data sets or on other data sets. Rather generic neural net algorithms were selected for this comparison study. Other algorithms in a given tool suite may perform far better than the neural net on each of these data sets.

Tool Suite Comparisons
SPSS Clementine

Clementine has been around for a long time. The tool suite is very mature and has a very faithful following, particularly in Europe (it was developed in England). Clementine was the first data mining suite to use the graphical programming approach used previously by the scientific programming tools Stella, I-Think, MathCad and MatLab in the 1980s.

Pros:

- Good variety of data mining algorithms.

- Very powerful optimal parameter search routines built into many of the data mining algorithms (automatic trials of different parameter sets).

- Very powerful combination of the type node and quality node for data quality checks and missing value imputation.

- Power meta-learning models can be built, in which the results of one modeling algorithm can be easily streamed as input to another modeling algorithm.

- Powerful (but proprietary) internal scripting language (CLEM) for creating complex variable processing.

- Moderately easy to use.

Cons:

- Relatively little descriptive statistical or parametric statistical analysis capabilities are available directly in the tool (although SPSS nodes can be used for input from and output to the SPSS StatPak.

- Relatively poor descriptive or output graphics forms.

- Model export for scoring outside the tool suite (or to perform deployment calculations at a faster speed than the interpreter based internal tool) must be done via an optional Publisher product ($25,000).

STATISTICA Data Miner

STATISTICA Data Miner (like KXEN) is a tool in a class by itself. This uniqueness is defined

primarily by in terms of the many things it does well, and the completeness in facilitating all tasks of the data mining project. Other tools may be easier to use (e.g. Insightful .Miner) or employ more automation (Affinium Model or KXEN), but no data mining suite available today provides more tools for performing data mining projects. This tool suite is my personal favorite.

Pros:

- Provides the richest combination of parametric statistical and machine learning data mining algorithms

- Relatively easy to use graphical programming user interface.

- Provides tools for all common data mining tasks.

- Highly flexible tools for model output.

- Powerful tools for reduction of dimensionality.

- Scalability (STATISTICA Data Miner can rapidly process larger data sets both in terms of their dimensionality and the overall size than the other products).

- Powerful customization options based on the industry standard VB language.

Cons:

- Lift charts are not easily available for evaluation of neural net models.

- Training in statistical analysis is best for properly interpreting the results of the parametric statistical algorithms.

KXEN

KXEN is one of two tool suites that provide an implementation of a support vector machine (SVM). STATISTICA Data Miner is the other. For this algorithm, the input data must be transformed to the range of -1 to +1 (done automatically by KXEN) and projected into the feature space. An SVM finds the solution of minimum error by identifying the "maximal separable hyperplane" through a cloud of data points in feature-space (a theoretical space with a number of dimensions equal to the number of predictor variables). The error minimization approach of a neural net is an iterative search for the set of predictor variable weights that produces the least amount of error between actual and predicted values. The problem with this kind of a search is that it is quite possible for the search to end at a solution that represents only a "local minimum" for total error across only a region of data points, rather than a "global minimum" total error across all data points. An SVM *induces* the solution of minimum error across all the data points in the feature-space, rather than *deduces* it by an iterative procedure.

Pros:

- KXEN is clearly the most accurate data mining tool available today.

- Various combinations and transforms of existing variables are automatically created and included in the analysis as derived predictor variables.

- This tool is almost fully automatic! For performing analyses on appropriately cleansed and formatted datasets, this tool comes closest to the ideal of the automobile user interface. KXEN is the *only* data mining tool available today that can be so easily embedded into your data processing stream to use as a data mining engine. In fact,

KXEN means "knowledge extraction engine," and it lives up to its name.

Cons:

- A clean data set must be submitted to the Consistent Coder of KXEN in the form of one record per entity to be modeled.

- There are no data preparation tools to help you put the data in this form (although many data preparation steps required by parametric statistical or neural nets/decision trees are not necessary with SVMs).

- No coincidence (or "confusion") matrix is available for binary output, from which precision and recall values can be calculated. You can create one, if you can determine the correct threshold to use on the decimal output to convert it to binary predicted values.

Insightful Miner

This tool suite may be the best one available for a company that would like to use ordinary business analysts to do relatively simple data mining projects. Insightful Miner is a good data mining tool for use with S-PLUS systems, because the entire library of S-PLUS functions are available for use with it! In addition, it provides a rich assortment of data mining and statistical data mining algorithms (but not nearly as rich as does STATISTICA Data Miner). For almost all common steps in the data mining project, Insightful Miner clearly gives the best bang for the buck.

Pros:

- Excellent tools for data import/export, data exploration and data cleansing tasks, and reduction of dimensionality prior to modeling.

- Even though it does not employ a graphical programming interface, it is relatively easy to use by non data miners.

- The most complete general purpose data mining suite available, and it is relatively inexpensive.

Cons:

- A relatively low level of automation.

- No scripting interface for coding of complex problems.

- Recoding must be done via an expression language in the create columns node.

- No model exporting capabilities.

Affinium Model

This tool is the easiest to use response modeling product on the market, even easier than STATISTICA Data Miner. This is the best package for use by the non-data miner/statistician, for whom the lack of a rich statistical and graphical backbone is not a problem. The automatic operation of the modeling engine shields the user from many data mining operations that must be manually performed by users of other packages, including choice of algorithms. The user has only to choose the level of analysis from quick to extensive, and the tool automatically creates models from a small to a large number of algorithms and parameter settings, while saving the current best model. Four different modeling applications: Response Modeler, Cross-Seller, Customer Segmenter and

Customer Valuator are actually very similar in function and differ only the terms used in creating the model. This seems like just a repackaging of the same thing, but that is part of the appeal of this tool. In the mind of the non-data miner, perception is most of the problem in using data mining tools for different purposes. The modeling button brings up a list of modeling options (quick-intermediate-extensive) that cycles through an increasingly large number of modeling algorithms and associated parameters to find the optimum model for each data set. Optionally, you can select one particular model type, or no model. Following that choice, the modeling process in Affinium Model is automatic.

Pros:

- The menu items are arranged from left to right showing modeling application, data import, modeling, model reports, scoring, scoring reports and variable name editing.

- The data is imported into an internal spreadsheet, like in STATISTICA Data Miner, but the only manipulation of the data is permitted through the Edit button (only for variable name changes) and via a drop-down menu with a data quality option (for missing data reports and imputation).

- New variables can be derived in the spreadsheet with a rich set of macro functions.

- Interpretation of the model results is very intuitive. The user has the choice of viewing a brief report, a detailed report, a lift curve or a variable sensitivity report.

Cons:

- No data exploration tools.

- The biggest potential drawback with this product is the almost complete lack of data preparation functions. Input data must be properly prepared in other tools before import into Affinium Model.

- There is no evidence that data is standardized (conversion of the ranges of all variables to a common scale) before submission to the modeling algorithm.

How can you tell which tool would be best for you? One way is to match the tool to the data scenarios for which you plan to use it. The following are several common data scenarios that you might encounter in your business.

Data Scenarios

The choice of the proper data mining suite for your use may depend on the data environment in which you would like to use it. Here are some data scenarios and choices of appropriate tools to use.

Scenario #1. If the company has access to (or is willing to hire) people with statistical expertise, the best tool will be one that statisticians understand and can use effectively:

- STATISTICA Data Miner
- SPSS Clementine (in conjunction with SPSS Stat package)

Scenario #2. If data preparation must done by hand inside the data mining package, then the best tools would include:

- STATISTICA Data Miner
- Insightful Miner

- SPSS-Clementine
- Affinium Model would not be a good choice here, because relatively few data preparation operations are supported in the tool.

Scenario #3. If the company wants to do data mining modeling with lower-level business analysts, then the best tool will have a relatively high degree of automation:

- Affinium Model
- KXEN
- Insightful Miner

Scenario #4. If the company has it own in-house analytical tools that require some enhancement to provide data mining capability, then the best data mining tool will be one that is easily embedded into their existing systems:

- KXEN
For example, NCR's Warehouse Miner could be coupled with KXEN to provide access to very fast data mining solutions within the Teradata database system. KXEN can be used to find the optimum support vector for a solution, in which some of the important variables in the solution may include non-intuitive constructs created by the KXEN Consistent-Coder prior to modeling.
- Affinium Model

Robert A. Nisbet, Ph.D., is an independent data mining consultant with over 35 years experience in analysis and modeling in science and business. You can contact him at Bob@rnisbet.com or (805) 685-0053.

For more information on related topics, visit the following channels:

- Data Mining