

The Wide Scale Deployment of Active Data Mining Solutions

A White Paper by Attar Software

Data Mining is a process

There is increased interest in a process or methodology for data mining. It is argued that such a formalised process will widen the exploitation of data mining as an enabling technology for solving business problems. It will allow people with varying expertise in data mining and from different business sectors to carry out successful data mining projects with a high degree of consistency.

There are a number of initiatives for the development of a formal/documented data mining process both in Europe and North America. It is reassuring to the data mining community that the processes emerging from all of these initiatives reveal a large degree of similarity. There is widespread agreement on the main steps (stages) involved in such a process and any differences relate only to the detailed tasks within each stage. A summary of the major stages of a data mining process is:

- **Goal definition:** This involves defining the goal or objective for the data mining project. This should be a business goal or objective which normally relates to a business event such as arrears in mortgage repayment, customer attrition (churn), energy consumption in a process, etc. This stage also involves the design of how the discovered patterns would be utilised as part of the overall business solution.
- **Data selection:** This is the process of identifying the data needed for the data mining project and the sources of this data.
- **Data preparation:** This involves cleansing the data, joining/merging data sources and the derivation of new columns (fields) in the data through aggregation, calculations or text manipulation of existing data fields. The end result is normally a flat table ready for the application of the data mining itself (i.e. the discovery algorithms to generate patterns). Such a table is normally split into two data sets; one set for pattern discovery and one set for pattern verification.
- **Data exploration:** This involves the exploration of the prepared data to get a better feel prior to pattern discovery and also to validate the results of the data preparation. Typically, this involves examining the statistics (minimum, maximum, average, etc.) and the frequency distribution of individual data fields. It also involves field versus field graphs to understand the dependency between fields.
- **Pattern Discovery:** This is the stage of applying the pattern discovery algorithm to generate patterns. The process of pattern discovery is most effective when applied as an exploration process assisted by the discovery algorithm. This allows business users to interact with and to impart their business knowledge to the discovery process. In the case of inducing a tree, users can at any point in the tree construction examine / explore the data filtering to that path, examine the recommendation of the algorithm regarding the next data field to use for the next branch then use their business judgement to decide on the data field for branching. The pattern discovery stage also involves analysing the ability of the discovered patterns to predict the propensity of the business event, and for verification against an independent data set.
- **Pattern deployment:** This stage involves the application of the discovered patterns to solve the business goal of the data mining project. This can take many forms: Patterns presentation: The description of the patterns (or the graphical tree display) and their associated data statistics are included in a document or presentation. This requires the data mining tool to generate text reports and WMF (Windows Meta File) representations of the graphical decision tree.

- **Business intelligence:** The discovered patterns are used as queries against a data base to derive business intelligence reports. This requires the data mining tool to generate SQL representations of the decision tree.
- **Data Scoring & Labelling:** The discovered patterns are used to score and/or label each data record in the database with the propensity and the label of the pattern it belongs to. This can be done directly by the data mining tool or through generation of SQL or C representation of the decision tree
- **Decision Support Systems:** The discovered patterns are used to make components of a decision support system. This can be achieved by embedding the data mining tool as a decision making component, or a C module generated by the data mining tool.
- **Alarm monitoring:** The discovered patterns are used as 'norms' for a business process. Monitoring these patterns will enable deviations from normal conditions to be detected at the earliest possible time. This can be achieved by embedding the data mining tool as a monitoring component, or through using SQL generated by the data mining tool.
- **Pattern Validity monitoring:** As a business process changes over time, the validity of patterns discovered from historic data will deteriorate. It is therefore important to detect these changes at the earliest possible time by monitoring patterns with new data. Significant changes to the patterns will point to the need to discover new patterns from more recent data.

The wide scale deployment of data mining solutions

A repeatable data mining process will help ensure the success of a data mining project. However, a successful data mining project also needs developers with the following skills:

- Deep knowledge of the data and its history.
- Insight into the specific business area.
- Proficiency in the use of the data mining tool.

The above skills may be combined in one person or may require more than one person. However, even in the largest of organisations there is a relatively small number of such specialists/teams with the above skills. To maximise the returns on data mining, the role of these specialists in a data mining project should be to prepare a specific 'Data Mining Business Scenario'. Once such a scenario is prepared it can be deployed on a much wider scale to a large user community - inside or outside the organisation. A Data Mining Business Scenario can also be called a Data Mining Solution.

Preparing a Data Mining Business Scenario involves all the steps of the data mining process; goal definition, data selection, data preparation and transformation, data exploration, pattern discovery and pattern deployment. The business scenario can be deployed to a wide user base. As an example, consider the business scenario of mortgage arrears in the portfolio of a financial institution:

Goal definition

Identify the profiles of mortgage accounts with a high or low propensity to default on mortgage payments. Define default as 3 or more months in arrears. The patterns discovered will be issued to branch managers to help them with the processing of mortgage applications. The patterns will also be issued to marketing managers to help them in their targeted marketing and in the definition of new products/mortgage packages. Finally, the patterns will be used by the Credit Manager to monitor the changes in the mortgage portfolio over time.

Data selection

Identify the source data as the Mortgage Applications data base and the monthly payments database. Furthermore, focus on historic mortgage applications, for example, those made in 1996 and 1997 and all payments records from 1996 until present date.

Data Preparation

- Extract mortgage application records from 1996 and 1997.
- Extract payments records from 1996 until present date.
- Join mortgage application and payment tables.
- Derive the new fields age (from DOB), total income and Loan/property value ratio.

Data Exploration

- Explore the frequency distribution of data fields.
- Explore the correlation between data fields.
- Plot the goal (arrears status) against other fields.

Pattern Discovery

- Induce a decision tree profiling arrears.
- Get the domain expert to validate the tree.
- Verify the patterns against test data sets.

Pattern Deployment

- Generate the patterns (tree) in WMF format. Import the graphical WMF file into MS-Word and print out the tree as a chart. Issue the tree print out to branch and marketing managers.
- Generate the patterns as SQL which is used to generate regular reports for the Credit Manager on the proportion of new business matching each of the discovered patterns.

The Deployment of Active Data Mining Solutions

The methods of deployment of patterns listed in the previous section can be described as passive deployment. This is because the solutions deployed can only utilise the patterns previously discovered. In Active Mining Deployment, the users are empowered to discover and explore new patterns within the business scenario (solution) delivered to them. For example, in the area of mortgage arrears described above, the business scenario can be prepared as described, but the Credit Manager and the users in the branches and marketing department can be given the ability to:

- Interactively develop new tree patterns in line with their business expertise/requirements.
- Develop new tree patterns from new data as it becomes available.
- Monitor the impact of new data on existing patterns.
- Within the same business scenario, change the outcome field and develop tree patterns for a new goal (outcome). For example, instead of profiling the arrears propensity, the user may be interested in the profile of mortgages with high loan amounts.
- Explore data throughout the tree patterns.

The active deployment of data mining, turns data mining into vertical business applications for wide scale use by business people who otherwise would not have the skills to develop a data mining process.

Technologies for the Deployment of Active Data Mining Solutions

There are a number of software technologies required in order to realise the benefits of the active deployment of data mining. These data mining software components allow the creation of vertical applications with embedded active data mining for use by business users.

- ***Embedded Data Transformation Engine:*** This component allows the data transformation process designed by the creators of the business scenario to be run against newly available data without any technical intervention by the business user.

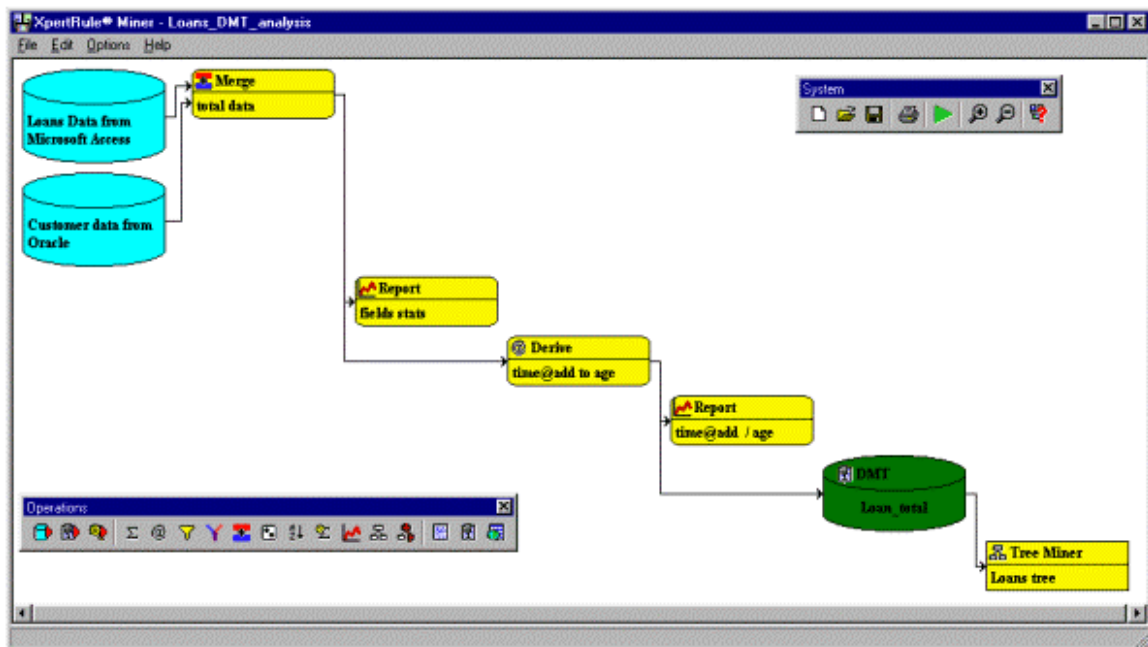
- **Embedded Pattern Discovery:** This allows the pattern discovery components to be embedded within a vertical end user application. For example, the mining of mortgage arrears can be embedded as part of a Customer Relationship Management System.
- **Graphical and interactive Pattern Discovery:** An essential part of active data mining is to allow the business user to interact with the data mining algorithm to ensure that business oriented patterns are discovered. Extensive pattern visualisation and exploration features are also an important aspect of active data mining.
- **Scalability, Architecture and performance:** Embedding data mining as a component within a vertical application will in no way reduce the need for the data mining component to work within the IT infrastructure and to be capable of handling large data volumes. The interactive nature of embedded data mining makes it even more important to have high performance pattern generation and exploration.

XpertRule Miner for the Active Deployment of Data Mining

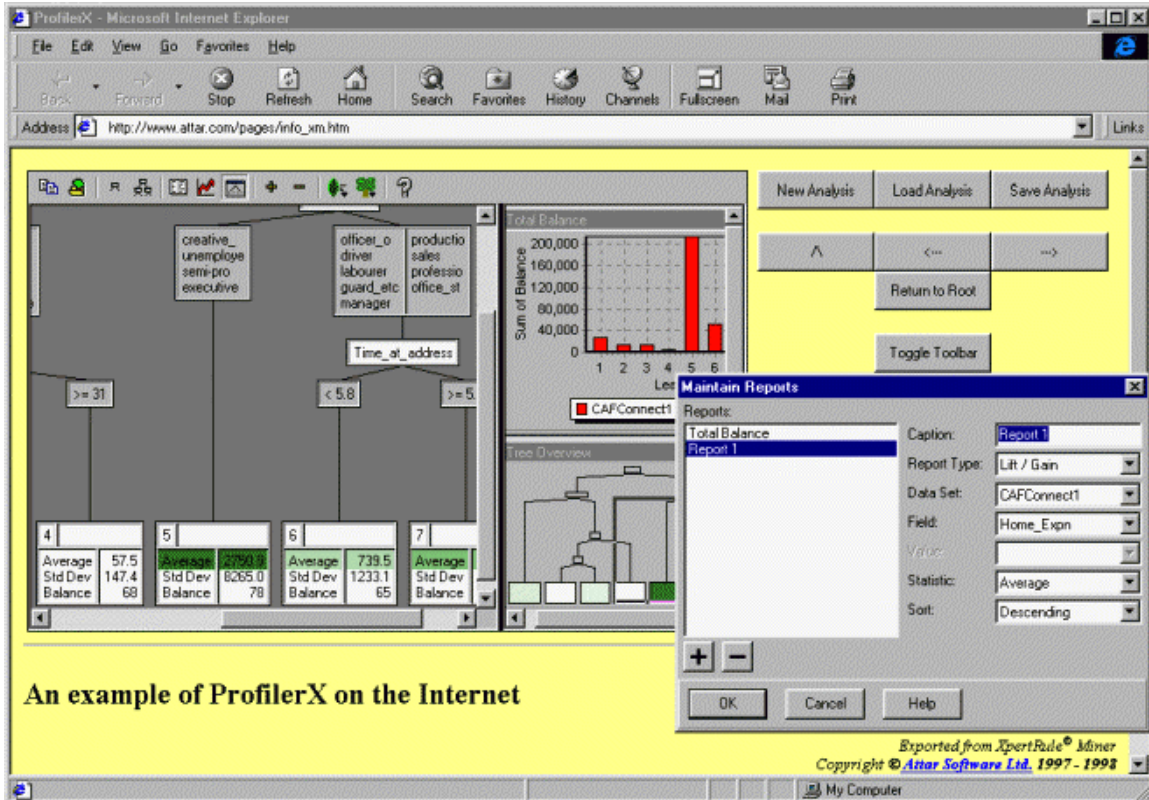
XpertRule Miner is designed from the ground up to enable the active deployment of data mining. It achieves this through the following features:

- **Graphical Support for the full Data Mining Process**

XpertRule Miner provides a graphical environment for supporting all the stages of the data mining process. The click, drag and drop environment allows non programmers to carry out complex data preparation, mining and deployment processes. It is an ideal environment for the development and testing of data mining business scenarios.

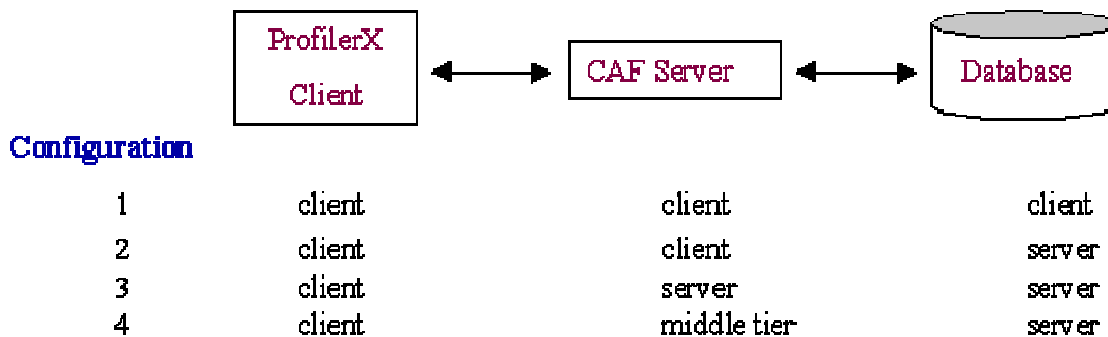


- **Embedded Data Transformation Engine:** Any data transformation process developed within XpertRule Miner can be executed by its Data Engine which is a component that can be embedded in any business application.
- **Embedded ActiveX Data Mining Component:** The ProfilerX tree induction component is delivered as an ActiveX component which is highly graphical, interactive and can be embedded within other business applications. The component exposes methods and objects which enables it to be seamlessly embedded within other applications - such as Customer Relationship Management (CRM) Systems.



- **Flexible architecture, High performance and scalability**

XpertRule Miner provides one of the most flexible deployment architecture as illustrated here:



The ActiveX tree induction client allows data mining to be embedded within other applications or deployed over the Internet/Intranet. The CAF (Contingency AND Frequency) servers can be deployed on the client, middle tier or server and are scalable and highly performant. These CAFs can exploit the high performance available from parallel processing database servers by the firing of intelligent query streams at the server. Alternatively, the CAF server can cache data from any ODBC compliant database into a highly tokenised format which is optimised for high performance mining on very large data tables. Using this caching technique allows an ODBC data source of millions of rows to be mined in minutes on average specification machines (e.g. 300 MHz Pentium with 64MB of RAM).