

# Data Mining - Beyond Algorithms

by Dr Akeel Al-Attar, MD of Attar Software

## The case for data mining

Most organisations can be currently labelled 'data rich', since they are collecting increasing volumes of data about business processes and resources. Typically, these data mountains are used to provide endless 'facts and figures' such as 'there are 60 categories of occupation', '2000 mortgage accounts are in arrears' etc. Such 'facts and figures' do not represent knowledge but if anything can lead to 'information overload'. However, patterns in the data represent knowledge and most organisations nowadays can be labelled 'knowledge poor'. Our definition of data mining is the process of discovering knowledge from data. Data mining enables complex business processes to be understood and re-engineered. This can be achieved through the discovery of patterns in data relating to the past behaviour of a business process. Such patterns can be used to improve the performance of a process by exploiting favourable patterns and avoiding problematic patterns.

Examples of business processes where data mining can be useful are customer response to mailing, lapsed insurance policies and energy consumption. In each of these examples, data mining can reveal what factors affect the outcome of the business event or process and the patterns relating the outcome to these factors. Such patterns increase our understanding of these processes and therefore our ability to predict and affect the outcome.

## Data Mining Technologies

There is a high degree of confusion among the potential users of data mining as to what data mining technologies are. This confusion has been compounded by vendors, of complimentary technologies, positioning their tools as data mining tools. So we have many vendors of query and reporting tools and OLAP (On-Line Analytical processing) tools claiming that their products can be used for data mining. While it is true that one can discover useful patterns in the data using these tools there is a question mark as to who is doing the discovery - the user or the tool! For example, query and reporting tools will interrogate the data and report on any pattern (query) requested by the user. This is a 'manual' and 'validation driven' method of discovery in the sense that unless the user suspects a pattern they will never find it! A marginally better situation is encountered with the OLAP tools, which can be termed 'visualisation driven' since they assist the user in the process of pattern discovery by displaying multi-dimensional data graphically. The class of tools that can genuinely be termed 'data mining tools' are those that support the automatic discovery of patterns in data.

We are going to make one more assertion regarding the difference between data mining and data modelling. Data mining is about discovering understandable patterns (trees, rules or associations) in data. Data modelling is about discovering a model that fits the data, regardless of whether the model is understandable - (e.g. tree or rules) or a black box (e.g. neural network). Based on this assertion, we restrict the main data mining technologies to induction and the discovery of associations and clusters.

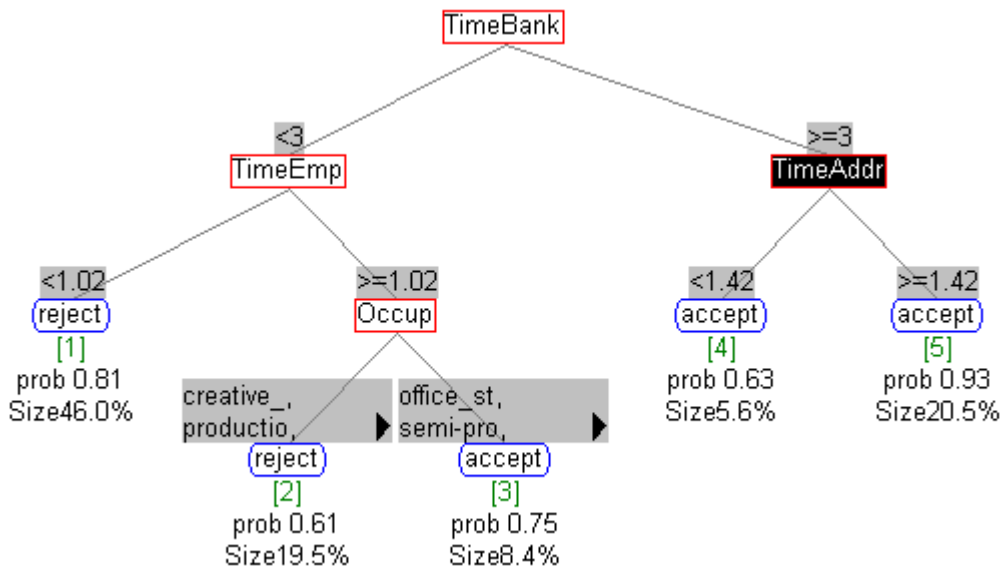
## Rule Induction

Rule or decision tree induction is the most established and effective data mining technologies in use today. It is what can be termed 'goal driven' data mining in that a business goal is defined and rule induction is used to generate patterns that relate to that business goal. The business goal can be the occurrence of an event such as 'response to mail shots' or 'mortgage arrears' or the magnitude of an event such as 'energy use' or 'efficiency'. Rule induction will generate patterns relating the business goal to other data fields (attributes). The resulting patterns are typically generated as a tree with splits on data fields and terminal points (leafs) showing the propensity or magnitude of the business event of interest.

S e x	A g e	Time Addr	ResStat	occup	Time Emp	Time Bank	House Exp	Decision
M	50	0.5	owner	unemploye	0	0	00145	reject
M	19	10	rent	labourer	0.8	0	00140	reject
F	52	15	owner	creative_	5.5	14	00000	accept
M	22	2.5	rent	creative_	2.6	0	00000	accept
M	29	13	owner	driver	0.5	0	00228	reject
F	16	0.3	owner	unemploye	0	01	00160	reject
M	23	11	owner	professio	0.5	01	00100	accept
F	27	3	owner	manager	2.8	01	00280	reject
F	19	5.4	owner	guard_etc	0.3	0	00080	reject
F	27	0.3	owner	manager	0.1	01	00272	reject
M	34	4	rent	guard_etc	8.5	07	00195	accept
M	20	1.3	rent	labourer	0.1	0	00140	reject
M	34	1.3	owner	guard_etc	0.1	0	00440	reject

As an example of tree induction data mining consider this data table which represents captured data on the process of loan authorisation. The table captures a number of data items relating to each loan applicant (sex, age, time at address, residence status, occupation, time in employment, time with the bank and monthly house expenses) as well as the decision made by the underwriters (accept or reject).

The objective of applying rule induction data mining to this table is to discover patterns relating the decisions made by the loan underwriters to the details of the application.



Such patterns can reveal the decision making process of the underwriters and their consistency, as shown in this tree. It reveals that the time with the bank is the attribute (data field) considered most important with a critical threshold of 3 years. For applicants that have been with the bank over 3 years the time in employment is considered the next most important factor, and so on. The tree below reveals 5 patterns (leaves) each with an outcome (accept or reject) and a probability (0 - 1). High probability figures represent consistent decision making.

The majority of data miners who use tree induction will most probably use as in automatic algorithm which can generate a tree once the outcome and the attributes are defined. Whilst this is a reasonable first cut for generating patterns from data, the real power of tree induction can be gained using the interactive (incremental) tree induction mode. This mode allows the user to impart his/her knowledge of the business process to the induction algorithm. In interactive induction, the algorithm stops at every split in the tree (starting at the root) and displays to the user the list of attributes available for activating a split, with these attributes being ranked by the criteria of the induction engine for selecting attributes (significance, entropy or a combination of both). The user is also presented with the best split of attribute values (threshold or groups of values) according to the algorithm. The user is then free to select the top

ranking attribute (and value split) according to the algorithm or select any other attribute in the ranked list. This allows the user to override the automatic selection of attributes based on the user's background knowledge of the process. For example, the user may feel that the relationship between the outcome and the best attribute is a spurious one, or that the best attribute is one that the user has no control over and should be replaced by one that can be controlled.

Interactive induction can also be seen as bridging the gap between OLAP based manual data segmentation / exploration and algorithm assisted segmentation.

## The Discovery of Associations

This is the second most common data mining technology and involves the discovery of associations between the various data fields. One popular application of this technology is the discovery of associations between business events or transactions. For example discovering that 90% of customers that buy product A will also buy product B (basket analysis) or that in 80% of cases when fault 1 is encountered then fault 7 is also encountered. If the sequence of events is important then another data mining technology for discovering sequences can be used.

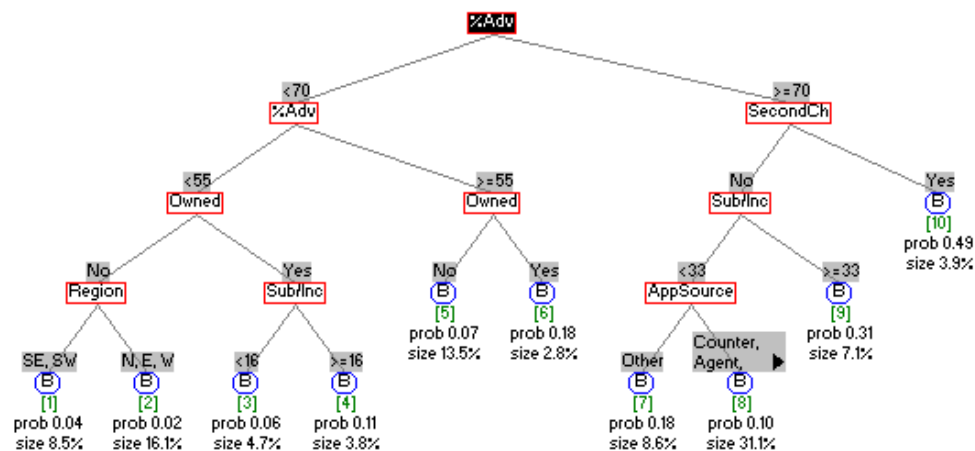
A second application of associations discovery data mining is the discovery of associations between the fields of case data. Case data is data that can be structured as a flat table of cases. Records of mortgage applications is an example of case data. In such data, associations can be found between data fields; for example that 75% of all applicants that are over 45 and in managerial occupations are also earning over £40,000. Such associations can be used as a way of discovering clusters in the data. Note that this differs from rule induction on case data in that no outcome needs to be defined for the discovery process.

## Case Studies

A number of case studies are described in the following sections which detail the background to each case studies, the data mining approach used and the benefits gained by the organisations concerned.

### Case Study 1: Mortgage Lending

This case study comes from a UK Mortgage Lender which had a mortgage portfolio in which 9.8% of all accounts were in arrears (over 3 months in arrears) and 4.1% of all accounts were in severe arrears (over 6 months in arrears). The objectives of the data mining project was to discover patterns relating the propensity of arrears to the mortgage application data. Such patterns can be applied at the front end applications processing to reduce the level of arrears and can result in better management of accounts that go into arrears.



Rule induction data mining was used with the outcome of the analysis being the arrears status healthy, moderate arrears or severe arrears. The attributes of the data mining analysis were the mortgage application data such as age, income, occupation, term, loan amount, region etc. Two separate data mining analysis were carried out; one to discover the patterns of arrears and the second to discover the patterns of severe arrears. Rule induction generated the following tree for arrears with splits on the attributes %Adv (% of loan to property value), SecondCh (second charge on property), Owned (is the property already owned), sub/Inc (subscription to income), AppSource (application source) and Region. The tree reveals 10 profiles with a propensity for arrears ranging from 0.02 to 0.49.

The trees discovered from the arrears data was used in three ways:

- To generate policy rules which were introduced at the application processing stage to reduce arrears.
- To formulate a marketing strategy targeting low risk profiles.
- To focus arrears management resources on accounts most likely to end up in severe arrears.

### ***Case Study 2: Life Underwriting***

This case study is from the Hibernian in Ireland who like other Life Insurers were facing the challenges of reducing costs, maintaining market share and meeting market demands. In order to meet these challenges, Hibernian decided to re-engineer its Life Underwriting Process in order to speed up the process and reduce its costs.

The first phase of re-engineering involved the implementation of a rule based underwriting system which was used to automate the processing of Life Proposals at the point of application. The system involved capturing underwriting knowledge which resulted in 51% of proposals being underwritten automatically with the remaining cases being referred to head office for manual underwriting.

While the automated underwriting system proved very successful, Hibernian looked for ways of increasing the percentage of cases that can be processed by the system. Attempts were made to capture more advanced underwriting rules, however this proved to be very difficult. Data mining was then considered as an alternative for generating additional knowledge. The basic premise was that out of the 49% of cases being referred to Head Office a significant number were underwritten with no or a very small additional premium (less than the cost of the manual underwriting!). It was therefore decided to apply rule induction analysis to cases being referred to Head Office with the amount of additional underwriting premium being used as an outcome. Rule induction analysis generated patterns of low additional premiums of the following format :

**If AGE > 30 & AGE < 41 and HEIGHT-WEIGHT = NORMAL Then PREMIUM LOADING = 1%**

The generated patterns for low additional premiums were qualified and checked for risk by the actuaries at Hibernian before being added as additional underwriting rules to the automated underwriting system. The result was to increase the rate of automated underwriting to 78%.

This case study illustrates how data mining helped Hibernian in Ireland develop new ways of processing life proposals and these methods now underpin a cost effective new business process.

### ***Case Study 3 : Gas Processing Plant***

This project was carried out for an oil company and was based in a remote US oil field location. The process investigated was a very large gas processing plant which produces two useful products from the gas from the wells, natural gas liquids (NGL) and miscible injectant (MI). NGL is mixed with crude oil and transported for refining, and MI is used to improve the viscosity of oil in the fields to improve crude oil recovery.

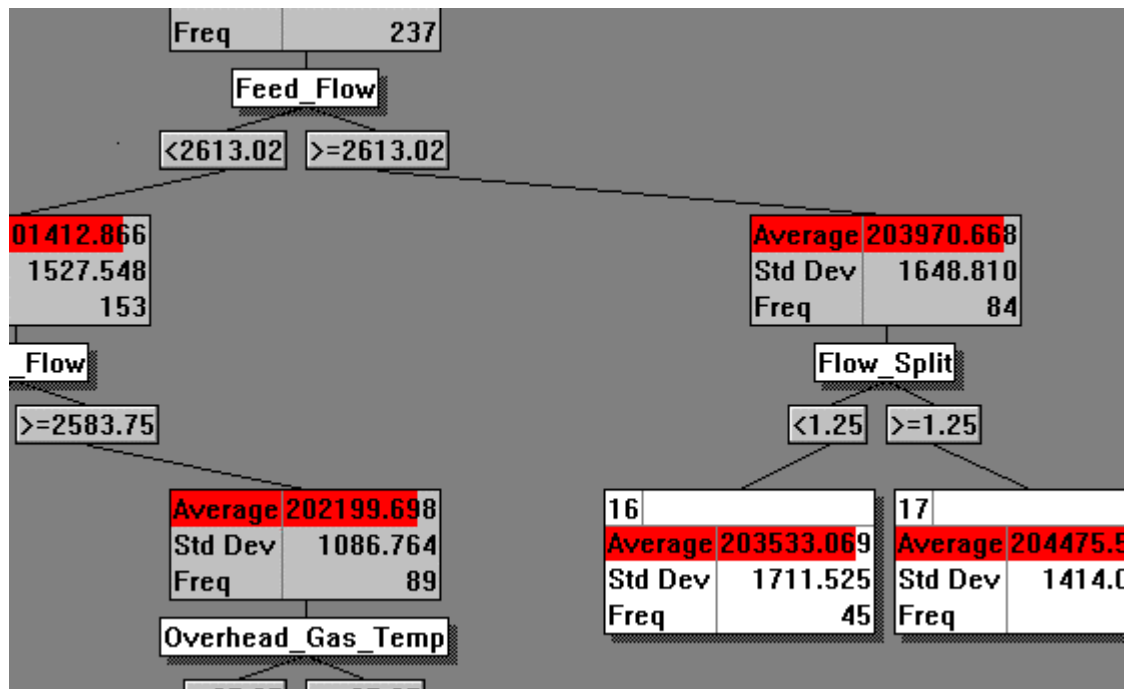
The aim of the study was to use data mining techniques to analyse historical process data to find opportunities to increase the production rates, and hence increase the revenue generated by the process. Approximately 2000 data measurements for the process are captured every minute.

Rule induction data mining was used to discover patterns in the data. The business goal for data mining was the revenue from the Gas Process Plant, while the attributes of the analysis fell into two categories:

Disturbances, such as wind speed and ambient temperature, which have an impact on the way the process is operated and performs, but which have to be accepted by the operators.

Control set points, these can be altered by the process operators or automatically by the control systems, and include temperature and pressure set points, flow ratios, control valve positions, etc.

An important part of the process is where the incoming feed gas is pre cooled with heat exchangers in two parallel process streams. The Oil company has always believed that there is an opportunity to improve process performance by altering the split of flows, however, it was not sure in which way to split the flow and what the impact will be on the revenue. Therefore flow split was put forward as an attribute for data mining.



This is the tree generated by rule induction. It reveals patterns relating the revenue from the process to the disturbances and control settings of the process. In particular the impact of the flow\_split is revealed with a critical ratio of 1.25 : 1.

The benefits derived from the generated patterns include the identification of opportunities to improve process revenue considerably (by up to 4%). Mostly these involve altering control set points, such as altering the flow splits. Some of the discovered knowledge can be implemented without any further work and for no extra cost (e.g. Altering flow splits). In other cases, it is necessary to provide the operators with timely advice about the best combination of settings for a given circumstance. This can be achieved cost effectively by delivering the rules generated as part of an expert system.

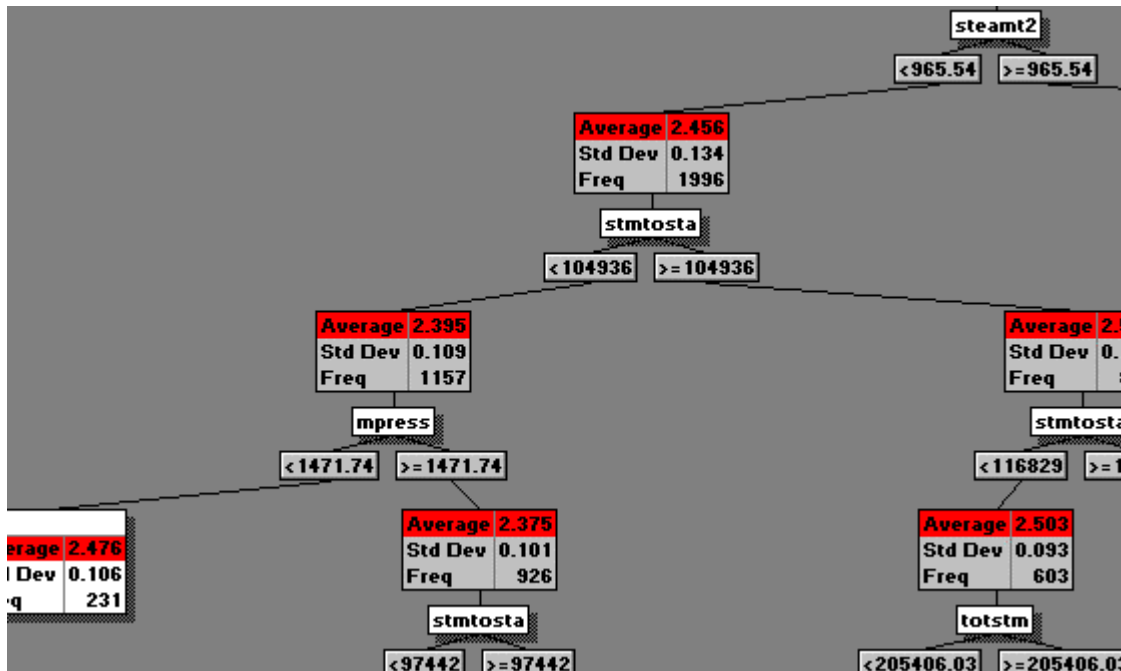
The company is in the process of carrying out another rule induction analysis with product quality as an outcome. The discovered patterns will allow the plant to be operated at the maximum revenue possible with acceptable product quality.

#### **Case Study 4 - Energy Usage in a Power Station**

ICI Thornton Power station, produces steam for a range of processes on the site, and generates electricity in a mix of primary pass out and secondary condensing turbines. Total power output is approximately 50 MW (i.e. a small power station). Fuel and water costs amount to about £5 million a year (depending on site steam demands).

The objective of the data mining project was to identify opportunities to reduce power station operating costs. Costs include the cost of fuel (gas and oil to fire the boilers) and water (to make up for losses). Electricity and steam are sold and represent a revenue.

Rule induction data mining was used for the project with the outcome (goal) of the analysis being the net cost of steam per unit of steam supplied to the site (i.e. the cost of the product). The attributes fall into two categories; disturbances such as ambient temperature and the site steam demand over which the operators have no control, and control settings such as pressure and bled steam rates.



Here is a section of the induced tree revealing the variation of steam cost with attribute values. It identifies the main contributors to efficient operations as manifold pressure (i.e. pressure at primary turbine inlet), steam flow to the secondary turbines and the total site steam flow.

The benefits derived from the generated patterns include the identification of opportunities to improve process revenue considerably (by up to 5%). Mostly these involve altering control set points which can be implemented without any further work and for no extra cost. Implementing some of the opportunities identified needed additional controls and instrumentation. The pay back for the additional controls would be a few months.

### The Current Issues in Data Mining

With real case studies of organisations deploying data mining as a catalyst for enhancing and re-engineering their business processes, data mining is now entering mainstream IT as a mature and tested technology. With this phase of evolution data mining has moved beyond the debate on algorithms and into the debate on usability. There are three main issues which should be considered by any organisation considering the introduction of data mining; methodology, ease of use and performance / scalability.

#### Methodology

For data mining to gain wide acceptance, it is important to have a step by step methodology for a data mining project. This ensures that the benefits reported by seasoned data miners are repeatable by other people in various business sectors. This can help dispel the belief that data mining is a kind of 'black art' which can only be practised by specialist. Such a methodology is beginning to emerge and there is certainly wide agreement on the main steps of such a methodology. These are :

### **Problem analysis**

This step involves analysing a business problem to assess if it is suitable for tackling using data mining. If it is, then an assessment has to be made of the availability of data, the data mining technology to be used (induction, associations, etc.) and how the results of data mining will be deployed as part of the overall solution

### **Data Preparation**

This step involves extracting the data and transforming it to the format required for the data mining algorithm. This involves, aggregation, table joins, deriving new fields, data cleansing, etc.

### **Data Exploration**

This step precedes the actual pattern discovery stage. It involves visualisation driven exploration and its aim is to give the user a good feel for the data and to reveal any errors in the data preparation / extraction.

### **Pattern Generation**

This step involves using rule induction (automatic or interactive) and associations discovery algorithms to generate patterns. This step also involves validating and interpreting the discovered patterns

### **Pattern Deployment**

This step involves deploying the discovered patterns as designed in the problem analysis stage. Patterns are typically used in decision support systems, to produce reports / guidelines, or to filter data for further processing.

### **Pattern Monitoring**

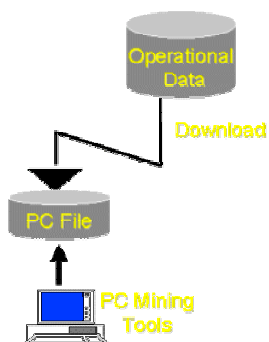
A main premise for the deployment of data mining results is that the future resembles the past and hence historic patterns can be applied to future situations. This strategy is safe only if the historic patterns are regularly monitored against new data to detect shifts in these patterns at the earliest possible time.

### ***Ease of Use***

Data mining tools are increasingly used by computer literate business users. This requires these tools to be no more difficult to use than a spreadsheet program. Furthermore the data mining tool needs to support all the steps of a data mining methodology. Finally, because of the nature of data mining, the tool has to support extensive data and patterns reporting and visualisation.

### ***Performance and Scalability***

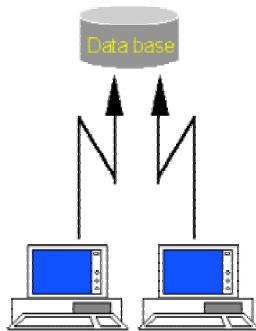
With the decreasing costs of data processing and storage comes the data rich organisation. It is now common place for small and medium sized organisations to hold gigabytes of data relating to a business process. It is therefore essential that data mining tools can deliver acceptable performance on large volumes of data regardless of the computing platform / architecture being used. There are a number of computing architectures for data mining



### **Client based data mining**

In this architecture the data to be mined is downloaded (extracted) and stored on the client machine (Windows 95 or NT). All the data preparation and mining is carried out on the client.

Until recently this approach was limited to mining tens of thousands of records (in acceptable times of under an hour). Recent advances has made it possible for millions of records to be mined on a client in tens of minutes and Attar Software's Profiler is an example of a data mining tool which such a capability.

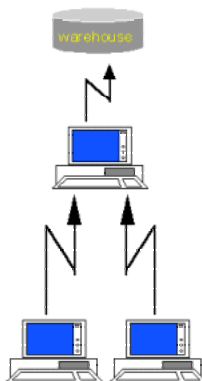


### Two tier client server data mining

In this architecture the data is extracted and stored on a server but is mined from the client machine(s). There two distinct flavours of this architecture:

**Standard platform server (NT or Unix).** In this scenario the server is capable of running a data mining engine which is invoked by the data mining front-end running on a client machine. This is the most common architecture for 2 tier data mining and can handle the mining of millions of records on a high performance Unix or NT hardware.

**High Performance Dedicated server** .In this scenario the server is a dedicated machine which can not run a data mining engine (for example Teradata). Most data mining tools can only address this situation by extracting (copying) data from the dedicated server to another standard platform server. This approach is not liked by users since it involves much data and hardware duplication. Recent advances have been made whereby the data in dedicated servers is mined in situ by the data mining client firing intelligent queries at the server (Profiler from Attar Software is an example of a tool with such capability).



### Three tier client server data mining

This architecture typically involves a dedicated high performance server (such as Teradata from NCR), a standard platform (Unix but increasingly NT) middle tier and a number of data mining clients. Again there are two distinct scenarios

**Thin middle tier** This scenario is only feasible if the middle tier can have running on it a data mining engine which is invoked by the client, but which can mine the data in situ on the server. Again this involves the data mining engine on the middle-tier firing intelligent queries at the server.

**Fat middle tier** This scenario involves the middle tier loading the data where it is mined on the middle-tier and the results are passed on to the data mining client.