ELSEVIER

Information & Management 37 (2000) 271–281

INFORMATION & MANAGEMENT

www.elsevier.com/locate/dsw

Briefings

Methodological and practical aspects of data mining

A. Feelders[a,*], H. Daniels[a,b], M. Holsheimer[c]

[a]*Department of Economics and Business Administration, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands*

[b]*Rotterdam School of Management, Institute of Advanced Management Studies, PO Box 1738, 3000 DR Rotterdam, The Netherlands*

[c]*Data Distilleries, Kruislaan 419, 1098 VA Amsterdam, The Netherlands*

Received 15 October 1998; accepted 5 September 1999

Abstract

We describe the different stages in the data mining process and discuss some pitfalls and guidelines to circumvent them. Despite the predominant attention on analysis, data selection and pre-processing are the most time-consuming activities, and have a substantial influence on ultimate success. Successful data mining projects require the involvement of expertise in data mining, company data, and the subject area concerned. Despite the attractive suggestion of 'fully automatic' data analysis, knowledge of the processes behind the data remains indispensable in avoiding the many pitfalls of data mining.

Keywords: Data mining; Knowledge discovery in databases; Data quality

1. Introduction

Data mining is receiving more and more attention from the business community, as witnessed by frequent publications in the popular IT-press, and the growing number of tools appearing on the market. The commercial interest in data mining is mainly due to increasing awareness of companies that the vast amounts of data collected on customers and their behavior contain valuable information. If the hidden information can be made explicit, it can be used to improve vital business processes. Such developments are accompanied by the construction of data warehouses and data marts. These are integrated databases that are specifically created for the purpose of analysis rather than to support daily business transactions.

Many publications on data mining discuss the construction or application of algorithms to extract knowledge from data. The emphasis is generally on the analysis phase. When a data mining project is performed in an organizational setting, one discovers that there are other important activities in the process. These activities are often more time consuming and have an equally large influence on the ultimate success of the project.

Data mining is a multi-disciplinary field, that is at the intersection of statistics, machine learning, database management, and data visualization. A natural question comes to mind: to what extent does it provide a new perspective on data analysis? This question has

* Corresponding author. Tel.: +31-134668201; fax: +31-134663377.
E-mail address: a.j.feelders@kub.nl (A. Feelders)

0378-7206/00/$ – see front matter
PII: S0378-7206(99)00051-8

received some attention within the community. A popular answer is that data mining is concerned with the extraction of knowledge from *really large* data sets. In our view, this is not the complete answer. Company databases indeed are often quite large, especially if one considers data on customer transactions. One should however take into account the fact that:

- Once the data mining question is specified accurately, only a small part of this large and heterogeneous database is of interest.
- Even if the remaining dataset is large, a sample often suffices to construct accurate models.

If not necessarily in the size of the dataset, where does the contribution of the data mining perspective lie? Four aspects are of particular interest:

1. There is a growing need for valid methods that cover the *whole* process (also called Knowledge Discovery in Databases or KDD), from problem formulation to the implementation of actions and monitoring of models. Methods are needed to identify the important steps, and indicate the required expertise and tools. Such methods are required to improve the quality and controllability of the process.
2. If it is going to be used on a daily basis within organizations, then a better integration with existing information systems infrastructures is required. It is, for example, important to couple analysis tools with Data Warehouses and to integrate data mining functionality with end-user software, such as marketing campaign schedulers.
3. From a statistical viewpoint it is often of dubious value because of the absence of a *study design*. Since the data were not collected with any set of analysis questions in mind, they were not sampled from a pre-defined population, and data quality may be insufficient for analysis requirements. These anomalies in data sets require a study of problems related with analysis of 'non-random' samples, data pollution, and missing data.
4. Ease of interpretation is often understood to be a defining characteristic of data mining techniques. The demand for explainable models leads to a preference for techniques such as rule induction, classification trees, and, more recently, bayesian networks. Furthermore, explainable models encourage the explicit involvement of domain experts in the analysis process.

2. Required expertise

Succesful data mining projects require a *collaborative* effort in a number of areas of expertise.

2.1. Subject area expertise

Scenarios, in which the subject area expert provides the data analyst with a dataset and a question, expecting the data analyst to return 'the answer', are doomed to fail. The same is true for situations where the subject area expert does not have a specific question and ask the analyst to come up with some interesting results. Data mining is not some 'syntactical exercise', but requires knowledge of the *processes behind the data*, in order to:

- Determine useful questions for analysis;
- Select potentially relevant data to answer these questions;
- Help with the construction of useful features from the raw data; and
- Interpret (intermediate) results of the analysis, suggesting possible courses of action.

2.2. Data expertise

Knowledge of available data within — and possibly outside — the organization is of primary importance for the selection and pre-processing of data. The data expert knows exactly where the data can be found, and how different data sources can be combined. Peculiarities of data conversions that took place years ago can have substantial influence on the interpretation of results; for example:

> A large insurance company takes over a small competitor. The insurance policy databases are joined, but the start-date of the policies of the small company are set equal to the conversion date, because only the most recent mutation date was recorded by the small company. Lacking the knowledge of this conversion, one might believe there was an enormous 'sales peak' in the year of conversion.

2.3. Data analysis expertise

Slogans such as 'data mining for the masses' suggest that any business user can mine her or his own data, and no particular data analysis skills or experience are required, provided that the right tools are used. This suggestion is, however, misleading. In fact, the analysis of data residing in company databases often requires sound statistical judgment, and may have to take into account phenomena such as selection bias and population drift [8].

Data mining expertise is required to recognize that a particular information requirement may be fulfilled by data mining. Furthermore, it is crucial to decide which data mining algorithm or tool is most suited to address the question. In selecting the appropriate algorithm one should also take into account non-technical aspects. Once an algorithm and tool have been selected, the information requirement of the user has to be translated into its terms, and usually the data has to be pre-processed before analysis can occur. The data mining expert also supports the interpretation of results, and translates model results back to the language of the domain expert.

3. Stages of the data mining process

Data mining is an explorative and iterative process:

- During data analysis, new knowledge is discovered and new hypotheses are formulated. These may lead to a focussing of the data mining question or to considering alternative questions.
- During the process one may jump between the different stages; for example from the analysis to the data pre-processing stage.

There is a need for a sound method that describes the important stages and feedbacks of the process [4]. The method should ensure the quality and control of the process. In Fig. 1, the important stages are depicted.

3.1. Problem formulation

In this stage, information requirements to be fulfilled with data mining are identified. In general these are questions concerning patterns and relations that may be present in the database. Examples of typical data mining questions are: 'How can we characterize customers that spend a lot on audio equipment?' and 'Which groups of applicants tend to have problems in repaying their loans?' A specific question such as 'Is there a relation between income and audio equipment expenditure?' could be tested using classical statistical methods. Data mining comes into its own when many possible relations between a large number of attributes have to be evaluated. This allows for the discovery of relations that are totally unexpected.

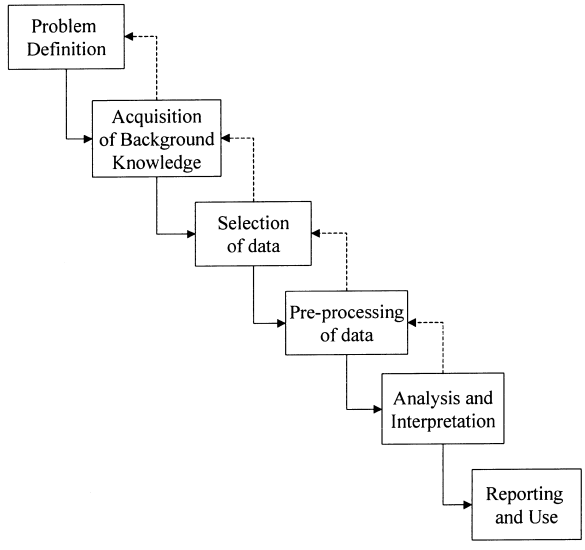

Fig. 1. The major steps in the data mining process.

In general, the initial question in a data mining project is rather vague, for example because it contains terms that have to be made operational. What do we mean by 'a client': do we mean an individual person or a household? Of course one may allow for different possibilities to be explored during analysis, but still these issues have to be addressed before data selection and pre-processing takes place.

It is important to determine *in advance* how the results will be used. We distinguish between three basic cases [7]:

1. Description/insight: the goal is to obtain an intelligible description of interesting segments or groups; for example, customers.
2. Prediction: discovered relations are used to make predictions about situations outside the database.
3. Intervention: results may lead to active intervention in the system being modeled.

Table 1
Example data for credit scoring (? = value unknown)

Age	Income	Zip code	...	Accepted	Defaulted
23	40000	10285	...	Y	N
35	35000	90057	...	Y	Y
21	30000	90054	...	N	?
...	...	...	...	...	

Of course, one must be aware of possible biases in the data; one has to consider whether the data is representative with respect to the question to be answered. Biased data give erroneous 'insights', and impedes generalization to cases outside the database. One also has to consider the role of causality. Our expectation that a particular policy has the desired effect is based on the assumption that the correlation found corresponds to a causal relation. Justification of this assumption cannot be obtained by data analysis alone but requires knowledge of the application domain.

Consider a bank that intends to analyse data on personal loans in order to create profiles of applicants with a relatively high risk of defaulting. Table 1 shows a number of records of the data available for creating such profiles.

The goal of the analysis is to *predict* whether or not *new applicants* will default on the loan. One can now make two important observations

1. The records for which we know the outcome (default or not), do not represent the population of new applicants, since we only know the outcome for applicants that were *accepted* in the past.
2. In order to make predictions, causal relations are not required. The model may use the attribute 'Zip Code' to predict the risk associated with an applicant, even though we may doubt that it *causes* the clients repayment behavior.

3.2. Identification of background knowledge

Databases are usually not created for the purpose of analysis, but rather to support vital business processes. The implication for analysis is the virtual absence of *statistical design*. Therefore, it is important to identify possible biases and selection effects that could limit the *generalization* of patterns to cases not recorded in the database. One frequently occurring bias often leads to 'knowledge *rediscovery*' rather than knowledge discovery. A company may have held certain 'campaigns' to sell a product to a particular market segment, say a publisher has held a campaign to interest student in a particular magazine. Upon analysis of his client database he may now find that students are particularly interested in this magazine, with the possible conclusion that his marketing effort should be directed at this interesting group. But in a sense he only finds these patterns because he put them there himself. Knowledge of past campaigns for this product is required to properly interpret the results.

Another important type of domain knowledge is about causal relations in the application domain. Some techniques allow for the incorporation of prior knowledge, most notably Bayesian Networks [9]. In other techniques, for example rule induction, the user would have to guide the analysis in order to take causal relations into account. Thus, if we analyse data on traffic-accidents with the purpose of finding situations with a substantially increased risk of a fatal accident, we may find a relation between 'pavement type' (asphalt or brick-pavement) and risk. Common sense may suggest that this correlation is due to other underlying factors. An influence diagram may look like Fig. 2. Here the arrows represent genuine influences and the dotted line may be a spurious correlation. The spurious correlation could emerge because city areas tend to have a lower maximum speed and more brick-paved roads. It would however not be a good idea to replace asphalt with brick-pavement for safety reasons.

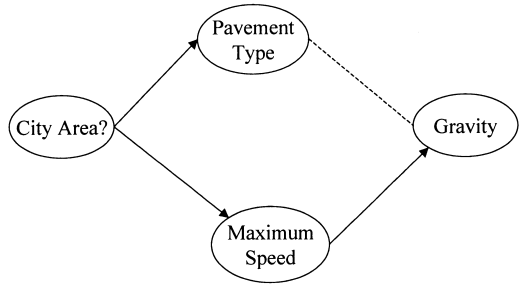

Fig. 2. Influence relations in traffic accident domain (dotted line represents 'spurious' correlation).

3.3. Selection of data

One should next decide what data may be relevant to answer the question. This selection should be made with an 'open mind' since the strong point of data mining is to 'let the data speak for itself' rather than restricting the analysis to a pre-specified hypothesis. In the bank example, we may consider several data sources, for example

1. Application data, such as Age, Income, Marital Status, Zip Code, Occupation, etc.
2. Data on possession and use of other products at the bank, for example earlier loans.
3. External data from a Central Credit Bureau (availability differs by country), in order to obtain the *credit history* of the applicant.

Ideally, one would like to make use of a data warehouse [18] or data mart to select potentially relevant data. In current practice, such an integrated data source is rarely available, so data from different sources and organizational units may have to be combined. This may lead to many difficulties in coupling different databases, inconsistent data models, and so on. If these issues have not been resolved prior to the project, they will take up large part of its time.

3.4. Pre-processing the data

Even if a data warehouse with all the potentially relevant data is available, it will often be necessary to pre-process the data before they can be analysed. This usually takes substantial project time, especially when many aggregations are required.

3.4.1. Deriving new attributes

It is often possible to 'help' the data mining system by adding new attributes that are derived from existing attributes in the mining table. Different income- and expense items of a loan applicant are listed separately in the initial mining table. Domain knowledge suggests that the *difference* between income and expenses is highly predictive of the repayment behavior of the applicant. Therefore, it makes sense to add a new attribute, that is derived from the individual income- and expense items. This may be especially helpful if the algorithm would not be able to figure out this relation by itself, for example in case of a classification tree algorithm. Even when the algorithm could in principle figure this out by itself, for example in case of a neural network, it is still beneficial to put as much domain knowledge as possible in the construction of new attributes.

3.4.2. Aggregation

The *mining-entity* is the entity in the data model that corresponds to the unit of observation at which analysis takes place. If we are looking for profiles of fatal traffic accidents, then 'traffic accident' is the mining-entity. Much pre-processing is due to the existence of 1 : N relations between the mining-entity and other entities in the database. Most data mining algorithms require that all data concerning one instance of the mining entity are stored in one record; analysis takes place on one big table where each record corresponds to one instance of the mining entity. Consequently, a data structure such as depicted in Fig. 3 has to be 'flattened'.

The user has to make a number of non-trivial choices. For the sake of simplicity, assume that either one or two objects (vehicle, pedestrian, animal or fixed object, like a lamppost) are involved in a traffic accident, how are we going to aggregate the 'objects involved'? Suppose we guess that the type of the object may be relevant. We could then create attributes 'type-object-1' and 'type-object-2' but there really is no order in the objects, so 'type-object-1 = car & type-object-2 = truck' represents the same situation as 'type-object-1 = truck & type-object-2 = car'. Therefore, it would be better to create only one attribute that has, as values, all possible combinations of objects, irrespective of their order in the database. If we furthermore suspect that 'nationality' may be of interest, we could aggregate it in the same manner to the 'traffic accident' level. Now we have lost the information about the nationality of a specific object.

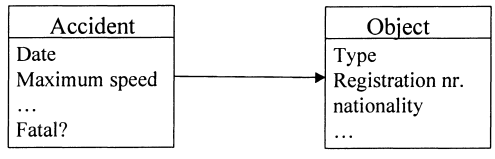

Fig. 3. Part of traffic accident data model (arrow indicates 1 : N relation).

Could this be important? All choices can have a substantial influence on the outcome. Domain knowledge and common sense are still required to determine the appropriate aggregations. Inductive logic programming (ILP) [1] approaches to data mining allow for learning from multiple relational tables, rather than one big table with a fixed number of attributes. Nevertheless, particular aggregations considered to be of interest may now have to be hand coded by the user as so-called *background knowledge*. Still, an ILP algorithm could handle multiple objects being involved in one accident in the above example, and is more flexible for learning from multiple relational tables.

3.5. Analysis and Interpretation

It is beyond the scope of this paper to describe the full range of data mining algorithms; the interested reader is referred to [5,10] for a comprehensive overview.

All three types of expertise are required during the analysis phase. *Knowledge of the application domain* is required to interpret (intermediate) results, and indicate which should be further explored. *Data expertise* is required to explain strange patterns that may be due to data pollution or other causes such as data conversions. *Data mining expertise* is required for the 'technical' interpretation of results; that is, the translation of results to the language of the domain- and data expert. Ideas for further analysis are translated into the formalism of the data mining algorithm. The way of working proposed is best supported by interactive algorithms that yield patterns that are easy to understand, for example rule induction or classification tree algorithms.

Many data mining questions can be formulated as finding subregions of the attribute space for which the value of the target variable is significantly larger than the global average of the target variable [6]. Consider the analysis of accepted loans: an interactive rule induction algorithm may yield the intermediate results of Fig. 4. The leftmost node indicates that of all 8955 people in the mining table, 3.1% defaulted. Directly to the right of this node, three groups with a significantly higher defaulting percentage are displayed. The domain expert is particularly intrigued by the group of 'room renters' and would like to know whether all room renters are an increased risk, or whether there are exceptions within this group. On the next level the mining algorithm discovers that room renters above 34 years of age are in fact very good risks. In this way the mining table may be explored interactively for interesting groups, allowing for a substantial amount of control of the subject area expert. Apart from this interactive analysis, one can also let the mining algorithm search several levels deep for groups with a substantial increased or decreased risk of defaulting (see [11]).

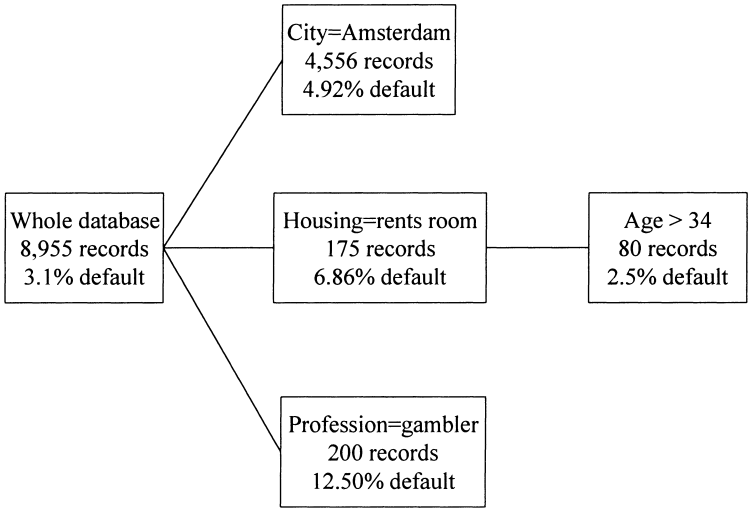

Fig. 4. Interactive rule induction.

3.6. Use of results

The results of data mining may range from its use as input for a (complex) decision process to its full integration into an end-user application.

Patterns such as 'The probability of a fatal accident increases significantly if road illumination is absent, and the maximum speed is 80 km/h' may eventually lead to measures to improve safety. This involves a complex decision process in which many other factors and constraints play an important role.

Data mining results can also be used in a knowledge-based system or decision support system. An important motivation for research on machine learning — one of the pillars of data mining technology — was to avoid the 'knowledge acquisition bottleneck' in the development of knowledge-based systems. Human experts find it difficult to articulate their own decision process, but if experts are confronted with a particular case, they are able to indicate the correct decision. The idea of machine learning was to create a number of examples, and to use induction algorithms to generate a knowledge-base, in the form of general decision-rules. This way the time-consuming 'manual' knowledge-acquisition process is side-stepped.

Finally, results may also be used in other tools; for example in direct mailing selection. The customer profiles derived with a data mining algorithm can, for example, be used by a marketing campaign scheduler to make the best selection from the customer database. Data mining functionality is embedded in an end-user program as part of a more complex process. Needless to say, this form of use is only viable for applications that are well-understood and regularly require analysis of new data.

4. Model interpretability

Ease of model interpretation is an important requirement. The widespread use of classification trees and rule induction algorithms in data mining applications and — tools aids in interpretation of results. Often there is a trade-off between ease of model interpretation and predictive accuracy, and the goal of the modeling task determines which quality measure is considered more important. Ease of interpretation has two major advantages:

1. The 'end product', that is the final model, is easy to understand.
2. The different model 'versions' created during the iterative data mining are easy to understand.

Although these advantages are clearly related, it is beneficial to consider them separately. The advantage mentioned under (2) becomes clear when we look at 'horror stories' about black-box approaches such as neural networks. A practical example illustrates the drawbacks of a black-box modeling approach.

Consider a direct marketing bank that receives both written and telephonic loan applications. The data of written applications are always entered into the company database, whether the loan is accepted or rejected. For telephonic applications, however, the data of the applicant are not always entered. If the bank employee quickly finds out that the applicant is not accepted, the conversation is usually ended without recording data. This allows the employee to help more clients, but clearly yields an incomplete database. If the goal of the modeling task were to predict whether an applicant is accepted or rejected, then the type of application (written or telephonic) appears to be highly informative! This apparent relation is however due to the 'selection mechanism'. When we use a black-box model, the use of this highly suspicious correlation may not be noticed by the modelers, due to a lack of insight in the model. If a decision tree were used instead, it would be clear immediately that 'type of application' appears to be highly informative, arousing the suspicion of domain experts. Further investigation of the data entry process would then lead to the source of this false correlation. Since data mining has a highly explorative nature, ease of interpretation is important in order to facilitate discovery and interpretation of interesting or suspicious relations in the data.

There may be several reasons why the final model(s) should be easy to interpret. The goal of modeling may not be pure prediction, but rather to gain an understanding of say particular groups of customers, in order to develop marketing strategies. Thus, the description of a customer segment may only be the beginning of further policy development. Clearly, a black-box predictive model is almost useless in such a situation. In addition, it may be required that model 'decisions' (predictions) are explainable

either to the model user or the customer. Providing a sensible explanation to the customer may be a legal obligation, in case the model prediction entails a decision 'to the disadvantage' of the customer (for example rejection of a loan application or insurance policy; see Section 6). For 'scoring' models, predictive accuracy is however also very important. Therefore, explainability may only become a concern after the predictive accuracy has been optimized. The objective then is to provide *satisfactory* explanations, and to comply with any legal obligations, mainly through *post-processing* of a highly complex model.

5. Missing data

Data quality is a point of major concern in any information system, and also in construction of data warehouses, and subsequent analyses ranging from simple queries to OLAP and data mining [14,19]. Although all aspects of data quality are relevant to data mining, we confine discussion to the issue of completeness. If many data values are missing, the quality of information and models decreases proportionally. Consider the marketing department of a bank that wants to compute the average age of customers with a specific combination of products. If 10% of the age values are missing, the uncertainty of the correct average increases, and consequently the quality of information decreases. One would like to solve this problem at the source, that is where data are collected and entered into the operational systems. With organizations becoming more and more aware of the value of high-quality data, there is an increasing attention for data quality issues. Nevertheless, sometimes other goals will prevail.

There has been a fair amount of work on handling missing data in the field of statistics [13]. In the past various ad-hoc procedures were used to *eliminate* the missing data; for example, by simply ignoring records with missing values, or by filling-in (*imputation*) single values, such as means. The problem of these procedures is that they often require ad-hoc adjustments depending on the specific analysis one intends to perform. Consequently, they are not suited to obtain generally satisfactory answers to missing data problems.

The description of the expectation maximization (EM) algorithm in Ref. [3], has provided a breakthrough in the sense that it gives a unifying view for analysis with missing data. A disadvantage of the algorithm is that it must be implemented differently for each type of model.

Multiple imputation [16,17] is a simulation-based approach where *a number* of complete data sets are created by filling in alternative values for the missing data. The completed data sets are subsequently analyzed using standard data mining methods, and the results of the individual analyses are combined in the appropriate way. Thus, one set of imputations can be used for many different analyses. The hard part of this exercise is to generate the imputations that may require computationally intensive and complex algorithms, such as *Markov Chain Monte Carlo* and *Data Augmentation* algorithms [17].

Multiple imputation was originally implemented in systems to handle the problem of missing data of *public-use* databases, where the database constructor and the ultimate user are distinct entities [16]. The situation is somewhat analogous for data warehouses in large organizations. A multitude of different analyses is performed by different users, so it would be beneficial to solve the missing-data problem *once*, and allow the end-user to use his or her preferred complete-data analysis software. A complicating factor is that data warehouses are often more dynamic than public survey data, since they are updated regularly to reflect the changes. For the purpose of data mining, one often takes a 'snapshot' of a relevant part of the data warehouse, to avoid changes during the project. Multiple imputation may then be used to replace missing values.

6. Legal aspects

In data mining projects where personal data are used, it is important for management to be aware of the legislation concerning privacy issues. For example, in the Netherlands the National Consumers' Association recently stated, that personal data of Dutch citizens are stored in more than 100 different locations. Therefore, the code of law on privacy will be renewed and reinforced in the near future. The Dutch law on privacy protection ('Wet Persoons Registratie') dates

from 1980 and it became operational in 1989. It is based on the principle of protection of personal privacy that is recorded in the written constitution of the Netherlands. Legislation in other European countries on privacy issues is similar, because the guidelines for the implementation have been embedded in the 1981 treaty of Strassbourg.

In the Netherlands the National Audit Office is responsible for the enforcement of the act. Both private institutions and governmental departments are regularly audited. There are a few concessive clauses added to the code of law that enable some governmental departments to fight fraude and crime. Guarding the privacy of citizens has become a lot more difficult since the collection and distribution of mass data can be done so easily and at very low cost. Firms may buy additional data obtained from so-called list-brokers and merge these with their own company databases. Data mining techniques can be applied to combined databases and may derive new delicate information, of which the subjects are unaware (cf. [12]). In 1995 the European Parliament and the European Council started to formulate new guidelines for the protection of privacy of European citizens. In the Netherlands it became operational in October 1998 and the act is known as 'Wet Bescherming Persoonsgegevens'. The main principles are listed below:

1. The act covers the use, collection and distribution of personal data.
2. It is applicable to all automatized data processing.
3. Recording of personal data, other than for communication purposes like address data, is only allowed if the subject has explicitly consented.
4. The subject has the right of inspection and correction of personal data.
5. Institutions that collect personal data are obliged to state clearly the goal of data collection and processing.
6. In case the data of the subject will be used for data mining the subject has to be informed about the right of removing the data without any cost.
7. Data may not be passed to a third party unless the subject has explicitly consented.
8. The holder of data has the duty to provide for all necessary measures to protect personal data.

The new law will have important consequences for data mining projects, where personal data are involved. For example, it will not be possible to enrich customer data with demographic data bought from list-brokers without permission of the subjects. Also, unequal treatment of customers is not allowed unless it can be explained using reasonable arguments. For example, insurance companies may only charge higher premiums if the insured has an exceptional record of claims. They may not differentiate premiums on the basis of general risk analyses with data mining.

7. Tools

Many of the early data mining tools were almost exclusively concerned with the analysis stage. They are usually derived from algorithms developed in the research community, for example C4.5 [15] and CART [2], with a user-friendly GUI. An interactive GUI is *not* just a superficial 'gimmick'; it encourages the involvement of the subject area expert, and improves the efficiency of analysis. For frequent use in business this functionality is however insufficient. Also, many early systems require loading the entire data table in main memory before processing can take place, allowing the analysis only of relatively small data sets.

Selection and pre-processing of data should be supported extensively by the tool. This includes a coupling with DBMSs in which the company data reside, and extensive possibilities to manipulate and combine data tables. For the analysis stage, the system should contain a number of algorithms that jointly cover the types of problems most frequently encountered. Since data mining projects often require a number of analysis sessions, spread over a couple of months, it is essential that the system takes care of administrative tasks such as logging of models, data tables, annotations to models, and so on. Adequate support should also be provided for the use of results. This ranges from reporting facilities to the possibility to export models. One step further is the integration of data mining functionality in application software, such as a marketing campaign scheduler. In this way data mining functionality is integrated in a more comprehensive business process (see Fig. 5).

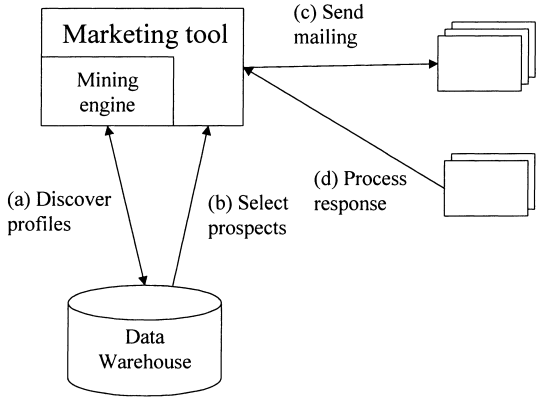

Fig. 5. Data mining embedded in marketing campaign scheduler.

8. Conclusions

Data mining or knowledge discovery in databases (KDD) is an exploratory and iterative process that consists of a number of stages. Data selection and data pre-processing are the most time-consuming activity, especially in the absence of a data warehouse. Data mining tools should therefore provide extensive support for data manipulation and combination. They should also provide easy access to DBMSs in which the source data reside.

The commitment of a subject area expert, data mining expert as well as a data expert to the project is critical for its success. Despite the attractive suggestion of 'fully automatic' data analysis, knowledge of the processes behind the data remains indispensable to avoid the many pitfalls of data mining.

Although company databases are usually quite large, proper formulation of the analysis question and an adequate sampling scheme often allows the database to be reduced to manageable size. It is typical for data mining projects that the data have not been collected for the purpose of analysis, but rather to support daily business processes. This may lead to low-quality data, as well as biases in the data that may reduce the applicability of discovered patterns.

References

[1] I. Bratko, S. Muggleton, Applications of inductive logic programming, Communications of the ACM 38 (11), 1995, pp. 65–70.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM algorithm, Journal of the Royal Statistical Society B 39 (1977) pp. 1–38.

[4] U. Fayyad, D. Madigan, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine 17(3) (1996) pp. 37–54.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

[6] J.H. Friedman, N.I. Fisher, Bump hunting in high-dimensional data, Statistics and Computing 9 (2), 1999, pp. 123–143.

[7] C. Glymour, D. Madigan, D. Pregibon, P. Smyth, Statistical themes and lessons for data mining, Data Mining and Knowledge Discovery 1, 1997, pp. 11–28.

[8] D.J. Hand, Data mining: statistics and more? The American Statistician 52 (2), 1998, pp. 112–118.

[9] D. Heckerman, Bayesian Networks for Knowledge Discovery, in: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 273–305.

[10] M. Holsheimer, A. Siebes, Data Mining: The Search for Knowledge in Databases, Technical Report, CS-R9406, CWI, 1994.

[11] M. Holsheimer, M. Kersten, A. Siebes, Data Surveyor: Searching the Nuggets in Parallel, in: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 447–467.

[12] K.C. Laudon, Markets and privacy, Communications of the ACM 39 (9), 1996, pp. 92–104.

[13] R. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, 1987.

[14] D.E. O'Leary, The impact of data accuracy on system learning, Journal of the Management Information Systems 9 (4), 1993, pp. 83–98.

[15] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[16] D.B. Rubin, Multiple imputation after 18+ years, Journal of the American Statistical Association 91 (434), 1996, pp. 473–489.

[17] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, 1997.

[18] A. Subramanian, L.D. Smith, A.C. Nelson, J.F. Campbell, D.A. Bird, Strategic planning for data warehousing, Information and Management 33, 1997, pp. 99–113.

[19] R.Y. Wang, D.M. Strong, Beyond accuracy: what data quality means to data consumers, Journal of Management Information Systems 12 (4), 1996, pp. 5–34.

A. Feelders is an Assistant professor at the Department of Economics and Business Administration of Tilburg University in the Netherlands. He received his Ph.D. in Artificial Intelligence from the same university, where he currently participates in the Data Mining research program. He worked as a consultant for a Dutch Data Mining company, where he was involved in many projects for

banks and insurance companies. His current research interests include the application of data mining in finance and marketing. His articles appeared in *Computer Science in Economics and Management* and *IEEE Transactions on Systems, Man and Cybernetics*. He is a member of the editorial board of the *International Journal of Intelligent Systems in Accounting, Finance, and Management*.

H. Daniels is a Professor in Knowledge Management at the Erasmus University Rotterdam and an Associate Professor in Computer Science at the Department of Economics at Tilburg University. He received a M.Sc. in Mathematics at the Technical University of Eindhoven and a Ph.D. in Physics from Groningen University. He also worked as a project manager at the National Dutch Aerospace Laboratory. He published many articles in international refereed journals, among which the *International Journal of Intelligent Systems in Accounting, Finance, and Management,* the *Journal of Economic Dynamics and Control* and *Computer Science in Economics and Management*. His current research interest is mainly in Knowledge Management and Data mining. He is a member of the editorial board of the journal *Computational Economics*.

M. Holsheimer is President of Data Distilleries. Previously Holsheimer spent several years at CWI, the Dutch Research Center for Mathematics and Computer Science. In 1993 he was asked to start the data mining research at CWI, one of the first European centers to start data mining research, and now a leading institute in this area. Since the second half of the 1990s major banks and insurance companies in the Netherlands expressed their need for data mining software and consultancy. Together with Martin Kersten, and Arno Siebes, Holsheimer founded Data Distilleries in the summer of 1995.