G. E. Kersten, Z. Mikolajuk and A. Yeh (Eds.), Decision Support Systems for Sustainable Development in Developing Countries, Kluwer Academic, 1999.

18 KNOWLEDGE DISCOVERY IN DATABASES AND DECISION SUPPORT

Anantha Mahadevan, Kumudini Ponnudurai, Gregory E. Kersten and Roland Thomas

1. Introduction

Two types of decision support systems (DSS) are discussed in Chapter 2: model-oriented and data-oriented. The assumption underlying model-oriented support is that accurate models of the decision problem are available or can be constructed prior to decision making. This approach includes optimization, simulation, decision analytic and statistical models. Various statistical methodologies can be used both to construct models and to estimate their parameters. Traditionally, these methods have been applied to relatively small data sets; the resulting models were used to verify given hypotheses, identify relationships among variables, and formulate predictions and scenarios.

Statistical methods require significant expertise in their application and interpretation of results. These methods are difficult to use when there is little explicit knowledge about the issues and when the problems are ill-defined but described by large or very large amount of data. Therefore, these methods are normally used by analysts and researchers rather then directly by decision makers. While one may suspect that valuable knowledge is buried in the data, it was not until recently, that tools for knowledge extraction and presentation in an easily readable form have became available.

Three types of developments made data-oriented DSSs viable and effective. Artificial intelligence (AI) adopted many statistical methods and embedded them in systems that efficiently search for patterns, derive rules, classify data according to some higher level concepts, and construct models from very large databases (Briscoe and Caelli 1996; Kohavi 1998). AI researchers have also developed methods to deal with such

databases, including neural networks, genetic algorithms, and rough sets (Aasheim and Solheim 1996; Bigus 1996; Pawlak, 1995). The second development involves technologies for the construction of multidimensional databases and data warehouses that provide raw materials used by AI systems (Inmon, 1996). Finally, data manipulation and visualization techniques allowed the user search and view databases from different perspectives and to display models in forms that could be understood by decision makers (Fayyad, 1996).

The process of deriving or extracting knowledge from data is known as *knowledge discovery in databases* (KDD). It comprises several distinct steps one of them being *data mining* which specifically deals with model construction. Data mining is used to make generalizations from a given set of data. *Generalization* involves the construction of a model from the data that can be used to describe the population or predict the behavior of its members. While generalization is also one of the key issues in statistics, there is a significant difference between the two. In statistics, the issue of generalization involves the definition of a population to which to generalize. This brings in issues of random sampling. Without careful attention to the sampling method one cannot guarantee statistical generalizability of models.

In data mining, there is no attempt to sample randomly from some identifiable population. In fact, during the first four steps of KDD, the database is viewed more as a population than as a sample. In the later steps, however, the results of the data mining and modeling processes are extended to units outside of the original database. One can regard this as akin to the generalization of mathematical models of processes that are typical of the physical and engineering sciences, where the generalization is non-statistical in nature. Alternatively, one can identify this process with the statistical notion of a finite population generated from a super-population (see, for example, Skinner et al., 1989, p. 14).

In the data mining context, the finite population under study (that is, the database), would be regarded as a random realization from some hypothetical super-population. As an example, this would be a natural conceptualization for a database consisting of scanned data records from stores in a particular chain, for a single month. A key assumption for the extrapolation from such a data mining exercise to future time periods would be that the processes studied, or discovered, will not change structurally over time (an implicit assumption in all economic forecasting). Either way, the generalization of the results relies on two key assumptions that underlie data mining methods:

- a model can be approximated with some relatively simple computational models, at a certain level of precision, and
- the data available in the target database is sufficiently representative of the processes of interest that the generalization is warranted.

It is clear that these assumptions are not very precise. This may be the reason why, in statistics, data mining (also known as data dredging) has traditionally been referred to as a sloppy exploratory analysis with no prior specification of hypotheses (Glymour, Madigan et al., 1997). Recently, however, data mining has become popular because it addresses the needs of business and other organizations in ways that other information systems do not address. Ease of use and intuitive interfaces coupled with often surprising and interesting results have led to a general, if sometimes uncritical, acceptance of data mining.

Several data mining methods based on rule induction are discussed and compared in Chapter 13. In this chapter we discuss the KDD process from the decision making and decision support perspectives. The benefits and shortcomings of data mining are presented with a focus on the potential contributions for decision support.

The KDD process is outlined in Section 2. To illustrate the model construction activity, we have used a popular and accessible decision tree method. In Section 3, the method is introduced and the resulting model discussed and compared with experts' knowledge and with another model obtained using a rule induction method. The results of the comparison give grounds for three experiments discussed in Section 4. A database with over 1,400 records has been used to formulate three descriptive models. Statistical methods are then used to evaluate these models and the results are presented. The potential of KDD for decision support in sustainable development is discussed in Section 4 together with a brief discussion of the future developments in this area.

The difficulties one may encounter in the effective use of the KDD and data mining software do not diminish its usefulness. There have already been many successful applications, some of which are presented in the Appendix.

2. Knowledge discovery

Knowledge discovery in databases (KDD) is the process of identifying significant patterns in data or deriving compact, abstract models from data. It is used to obtain nontrivial information, that is, information that requires search and inferencing rather than straightforward calculation of some predefined quantities (e.g., totals and averages). KDD has evolved from the intersection of machine learning, pattern recognition, statistics, data visualization and high performance computing. These areas of study have been integrated to address problems of interpretation of large databases at a higher-level than provided by management information systems, query facilities and on-line analytic processing.

The KDD process is both interactive and iterative. It includes issues of problem and scope definition, element selection, data standardization, model elicitation (i.e. data mining), and use of recurring patterns in models for developing strategies or procedures (Fayyad 1996; Brachman and Anand 1996). These steps are depicted in Figure 1.

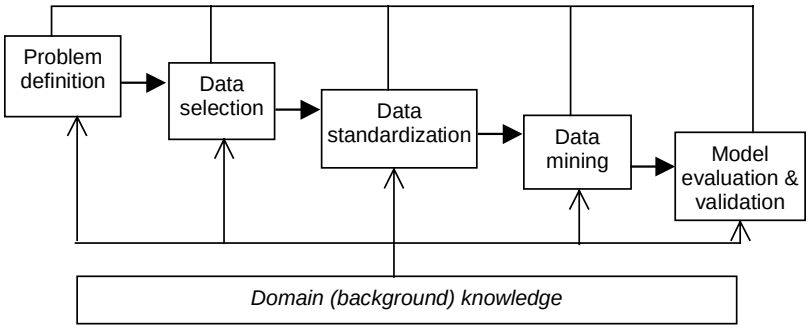

Figure 1. Knowledge discovery process.

Two important features of the KDD process are indicated in Figure 1. First, the process is iterative with possible loops between any two steps. Second, it relies heavily on the knowledge pertinent to the problem, that is, domain (background) knowledge plays a significant role in each step of the process. The process is used to extrapolate models from databases, while incorporating background knowledge throughout.

Domain experts, technology specialists, database documentation, and the end users can provide the domain knowledge necessary for this process. Knowledge about the issues under consideration is used to outline the initial problem and respective data elements. Knowledge is also required for the selection of data relevant to the problem and its standardization. Extrapolated models need to be verified, either through confirmatory statistics or background knowledge.

2.1 Problem definition

The first step in the KDD process is to obtain an understanding of the application domain, specify the expected outcomes of the process (user goals and expectations) and define the domain (background) knowledge that may be needed.

Decision support based on model-oriented DSS is restricted to variables and models present in the system. The task of problem definition is not affected by elements outside of the stipulated model. In the absence of a model the initial KDD step requires defining both the scope of the study and problem. This is performed in an iterative manner; the user undertakes subsequent stages several times before arriving at a satisfactory problem definition.

The prerequisite for both model- and data-oriented decision support is the identification of the decision problem. This includes specification of the decision maker's objectives and decision variables (see Chapter 2). DSSs are used to determine the variable values so that the objectives can be achieved.

The objectives of the decision maker guide the support process and provide goals for the KDD process. In other words, the KDD problem definition includes the specification of the goal of the process from the decision maker’s point of view. Two types of goals can be distinguished: *verification* and *discovery* (Fayyad et al., 1996).

The verification goal requires prior specification of hypotheses by the decision maker; contrary to some observations, data mining is not necessarily "hypothesis free" (Glymor, 1997). The hypotheses one needs to formulate include the existence of the relationship between the goal (dependent) variable and other variables (independent) in the database, the type of relationship (for example, causal or functional) and the identification and exclusion of variables that may be disregarded in the verification process. From the decision support viewpoint the goal is to determine the appropriateness of existing DSS models in light of the existing data.

The discovery goal assumes that the process autonomously searches for relationships that define concepts and models. Discovery can be used to obtain descriptive models of an object or process or to formulate predictions of future behavior of some entity. Further, this step involves analyzing the problem to assess if it is appropriate for data mining techniques. If it is, then an assessment is made concerning the avail-

ability of data and the type of techniques to be employed. Discovery, for the purpose of decision support, is oriented toward the search for new models that can be used in a DSS.

These two goals are not mutually exclusive; verification may lead to discovery and discovery requires verification. Their role is to obtain the initial focus for the study whose ultimate goal is to enhance DSSs and provide decision makers with relevant information and knowledge.

2.2 Variable and data selection, and preprocessing

The process of selecting appropriate elements (variables and samples) from the source database (assumed to be the population) relates to restricting the scope of decision support. In this step of the KDD process the variables and the data on which data mining is to be performed are chosen.

Domain knowledge plays a significant role in this step. Dealing with large data sets makes it relatively easy to obtain variables that are highly correlated or appear to be strong predictors of dependent variables. These variables, however, may not be casually related, for example, the observed correlation may be due to chance, or it may be a spurious correlation due to the concomitant variation with some undetermined causal variable.

As an example of another non-causal relationship, consider the problem of plant disease diagnosis. The database may contain variables measuring the symptoms, diseases and treatments of the observed plants. Assume further that a treatment is applied only after a diagnosis has been made. Treatments are highly correlated with diagnoses but, obviously, the causal effect is "from diagnosis to treatment" and not vice versa. If the treatment variable is not removed from the dataset then symptoms need not be considered in order to determine the disease. It may be thus sufficient to select treatments to "predict" a disease. Obviously such a model is of little use to a decision maker who wants to be able to diagnose diseases and only after that decide on a treatment.

Removal of confounding variables and variables that describe implications of goal (dependent) variables is an important component of data selection. It is no less important, however, to retain variables that appear to be insignificant, or that might seem to have little in common with the decision support tasks and goal variables. While this may lead to a discovery of accidental relationships or spurious correlations, it may also provide interesting and previously unknown results.

One of the objectives of the KDD process is to discover new knowledge and not necessarily to reinforce the decision maker's conviction. Data mining methods allow large amounts of data to be processed to identify variables that otherwise might have been ignored. This is one of the reasons for the feedback loops indicated in Figure 1. Preprocessing and data visualization help the decision maker remove variables that should not be considered and keep those that may have some prescriptive or descriptive impact. This is part of the initial exploratory analysis used to provide a preliminary understanding of data.

We mentioned earlier that in KDD the database is treated as a population. Data selection may involve sampling from the database to obtain a sample on which data

mining can be performed. Models are generated (estimated, in statistical parlance) from the sample, known as the training set. Their validity is subsequently verified using the remaining (holdout) records from the database.

Data selected for further processing may be incomplete and require cleaning and standardization. Decisions must be made about what to do with missing data fields. The task of element standardization is influenced by the choice of data mining methods. Further, data transformation may also be required. Processing of transformed data may lead to models with more explanatory power than processing the source data (John, 1997).

2.3 Data mining

The data mining step involves the selection and application of methods to construct models from data. The methods are used to formulate summaries and to identify significant structures and relationships present in data. Data mining is used to develop models that are understandable and that can be intuitively explained. Statistical significance and the validity of model generalization beyond the database are not considered in this step.

Some authors consider query operations, on-line analytic processing, visualization and statistics as data-mining methods (Simoudis, 1996). Though these information processing methods are used in data mining, they do not meet the main objective of data mining which is exploratory model construction.

Three types of models are distinguished in decision science: normative, descriptive, and prescriptive (Bell et al., 1988). *Normative models* provide guidelines for right actions and involve the formulation of principles of alternative evaluation and choice, proposed as rules that decision makers ought to follow. Since KDD is involved with historical data, it may provide information on how decisions were made and what their implications were rather than provide information about principles.

Descriptive models are used to represent actual decision processes and the results of decision implementation. They are formulated on the basis of the observation of decision makers' activities and the analysis of their decisions. One method of constructing such models is through the formulation of hypotheses, and their verification with statistical methods. Data mining, a data-driven approach, can also be used to construct descriptive models.

Prescriptive models focus on providing decision makers with support to make decisions more effectively and to think more constructively. They are used to facilitate the analysis and assessment of information, and to manage decision problem complexity. Prescriptive models can be obtained through the comparison of the expected and achieved outcomes, for example, from the comparison of plant treatments and their results. Data mining methods can be, we believe, used for this purpose.

There are other types of models that are generally not considered in decision science because they do not directly involve the generation of alternatives, their comparison and selection. These are *predictive models* used to determine the future behavior of some entity. These models are important for decision support because they provide information about processes and entities that are outside of the decision maker's control. They are also used to determine parameter values for descriptive and pre-

scriptive models. Data mining has often been used for the construction of predictive models (Fayyad, et al. 1996).

In summary, we distinguish three data mining operations useful for decision support:

1. Predictive modeling - used to generate a model that can be used to determine the future behavior of some entity.
2. Descriptive modeling - used to generate models that provide high-level summary information and that explain relationships between variables in the database.
3. Prescriptive modeling - used to construct models for the determination of decision variable values required to meet objectives specified by the decision maker.

The three types of models can be constructed by the application of one or more data mining tasks. The tasks considered by data mining include:

1. predicting the class to which an object belongs;
2. predicting the dependent variable value given values of independent variables;
3. formulating and describing clusters of similar objects ;
4. describing a group of objects;
5. finding and describing relationships and associations among variables;
6. identifying deviations and changes in a distribution or behavior of objects;
7. identifying variables that control the values of other variables.

The first two tasks can be used for the development of predictive models, tasks three, four and five for the development of descriptive models and tasks six and seven for construction of prescriptive models. Note, however, that, some tasks may serve more than one function. For example, tasks used for description can also be used to specify elements of predictive and prescriptive models. That is, the assignment of tasks to models is indicative rather than exclusive.

Data mining methods have been developed to perform specific tasks. These methods include association rules, cluster analysis, Bayesian classification, decision trees, rough sets, neural networks, and genetic algorithms (Kohavi, Becker et al. 1998; Pawlak, 1995; Srikant and Agrawal 1998; Weiss 1998).

2.4 Model evaluation and interpretation

Effective use of the KDD process requires an appraisal and confirmation of the extracted models. This may be conducted in the form of traditional model testing and through confirmation obtained from the related literature and from domain experts (Glymour et al. 1997). Measures of "interestingness", or measures of intuitive explanation, are often used to sift significant patterns or rules from other output. These measures include strength of rules (for example, number of rule precedents), statistical indicators (for example, goodness of fit) or simplicity values (Fayyad 1996; Mienko et al., 1996). Data visualization is also used to facilitate the evaluation of the significance of extracted information and models.

To use KDD results effectively in decision support, interesting and valid models need to be incorporated into existing knowledge. This may be difficult due to structural differences between the proposed models and existing knowledge. The latter may also be imprecise. The representation embodied in models may differ from the experts' knowledge, requiring the resolution of syntactical conflicts. The meaning of terms and concepts in the database, and those obtained from data mining, may conflict with those of the experts, requiring the resolution of semantical conflicts. Direct involvement of domain experts and decision makers is, therefore, essential in this step. Their verification and interpretations are necessary to convert models into knowledge and to incorporate them in DSSs.

The activities conducted in this step often lead to the repetition of earlier steps of the KDD process.

3. Mining data with decision tree methods

Decision tree methods allow the determination of relationships between sets of variables. They are used to select independent variables to partition data recursively. Classification methods are used for categorical dependent variables and regression methods are used for continuous dependent variables (Breiman, Friedman et al. 1984).

Several approaches to decision tree construction have been proposed including classification and regression trees (CART), chi-square automatic interaction detection (CHAID), and C4.5 (Briscoe, 1996). These approaches differ in the significance measurements used to select independent variables. The CART technique partitions data into two mutually exclusive subsets such that at least one of the resulting subsets has lower dispersion than the previous set of data. CHAID uses chi-square tests to calculate the association between the dependent variable and a chosen descriptive variable. The C4.5 method partitions data into mutually exclusive subsets such that each of the resulting subsets has lower entropy (dispersion) than the previous set of data.

Decision tree methods produce output that is easy to read and interpret. It can be represented as trees or rules; the latter format is especially useful for the development of knowledge-based DSSs. These methods have been implemented in many data mining software packages, for example AnswerTree (SPSS, 1998) and MineSet (SGI, 1998). They can be used for the construction of predictive, normative and descriptive models.

We are interested in KDD and data mining inasmuch they can be used for decision support. Our objective is not to provide an exhaustive analysis of a particular method, nor to present a complex application. Rather, we attempt to outline the opportunities of the application of KDD to obtain models for decision support. We will also discuss some remedies for the difficulties that are inevitably encountered. For this purpose we have selected a popular and easy to understand data mining method and applied it to a small database.

In this section we use the CHAID method to formulate a predictive model for a soybean disease problem; in Section 4 the same method is used to formulate three descriptive models.

3.1 Soybean disease problem

One dimension of expertise is the ability to assess the relative importance and informativeness of symptoms present in a diseased plant and to identify the disease from the symptoms alone. The difficulty in diagnosing some soybean diseases, for example brown spot, alternari spot and frog eye leaf spot, is that they often have similar symptoms. This, according to Mahoney (1996), causes misdiagnosis of diseases even by domain experts. A method that identifies clusters of symptoms for particular diseases may provide experts with additional information and enhance their knowledge.

Michalski (1980) notes that available diagnostic information regarding soybean pathology surpasses by far what a single expert can encompass. This is due to similarity of symptoms and differences in local conditions that may introduce deviations between symptoms for any given disease. Further, symptoms for a particular disease may change slightly over time. The exploration of historical data with methods for identifying deviations and changes may reveal new trends and patterns.

A classification method that can cluster symptoms and diseases and a model that can then be used to predict a disease given symptoms should be useful for decision makers with no deep knowledge of soybean diseases. Knowledge obtained from the KDD process may be used for explanatory and training purposes. Also domain experts and trained pathologists may use data mining tools to validate and possibly extend their knowledge.

A generally accessible small soybean database is available from the MLRepository (1999). The data set consists of 307 cases, with a dependent variable of 19 classes (values). Each class represents a soybean disease and each case in the data set is diagnosed with only one type of soybean disease. Thirty-five descriptors of plant and environmental factors were recorded for a diseased soybean plant. The symptoms were clearly observable conditions obtained with no sophisticated mechanical assistance (Michalski, 1980).

3.2 Decision tree

To illustrate the application of data mining for the soybean disease problem we use the CHAID method implemented in the AnswerTree software (SPSS, 1998). Decision trees generated for all disease classes are too large to be depicted and discussed. For the purpose of this discussion we have selected six of the nineteen classes (values of the dependent variable) for which there are 200 cases in the database.

The CHAID method compares the association between the dependent variable and the independent variable using Pearson's chi-square test. It requires the specification of a minimum significance level value. Independent variables for which significance is below the minimum value are disregarded.

The second parameter that needs be specified is the minimum significance level for merging branches. The CHAID method allows the user to group values (classes) of an independent significant variable, to form groups of values (categories). The procedure adds values to a group until the minimum significance level is achieved. The last required parameter is the minimum number of cases to be considered. If, at a node, the

minimum number is reached the procedure terminates and no further testing is done at this node.

The minimum significance levels for independent variables and for merging branches were both set at 0.05. The minimum number of cases was set at 10.

The decision tree for the soybean problem is given in Figure 2. For illustrative purposes we have truncated the tree to cover the brown-spot disease.

The dependent variable is Diagnosis with six possible values listed in each node. The root (top-level) node comprises all 200 records of the database. The distribution of all records according to the identified diseases, that is values of the dependent variable, is given (for example, there are 40 occurrences of the value *photophthora-rot*).

At each node below the root level a significant predictive variable (calculated using Chi-square tests) is selected. Records are collected for specific value(s) of this predictive variable. For example, the variable Int_Discolor (internal discoloration in the stem) is the first predictor of diagnosis (Chi-square = 200 on five degrees of freedom) and the node for Int_Discolor = 0 is presented below the root level (see Figure 2). Predictive variables are recursively applied to the data, thus the application of a variable on a node is dependent on the variable(s) applied on preceding node(s). This condition is termed the "recursive" nature of decision trees (Weiss and Indurkhya 1998). Hence, variable Plant_Growth depends on Int_Discolor.

Branches indicate the alternative descriptions obtained from one or more values of a predictive variable in a node. In Figure 2 a branch is depicted at the second level for two values of the variable Plant_Growth.

The size of the tree is controlled by the user. If desired, the tree can be constructed so that it represents most records in the training set. This would increase the applicability of the model to those records, but reduces the chance that the model would generalize to the population. This situation is referred to as “over-fitting” the model to the training data. Analysis can be performed using only one dependent variable at a time.

The accuracy of classification of the model was 86.5%. For this model accuracy rate was calculated using the same observations that were used to build the model. When the model is tested for accuracy, the most probable outcome of a terminal node is assigned the observation that meets specified conditions. This accuracy rate may not be a reasonable indication of the performance of the model in classifying new cases, as the model was tested on the same cases that were used to build it.

3.3 Rule-based model

Decision trees can be represented with rules. Each path leading from the terminal node to the root node is a rule. For each rule the likelihood of observing a particular class of the dependent variable is calculated (it is a percentage of all the observations at a terminal node).

Four rules corresponding to the four numbered paths in the tree depicted in Figure 2 are given in Table 1.

Only rule 2 presented in Table 1 implies a single disease. The remaining three rules are not discriminatory because their conclusions indicate two or three possible diseases with a different degree of prevalence. This is the consequence of different diseases having similar symptoms.

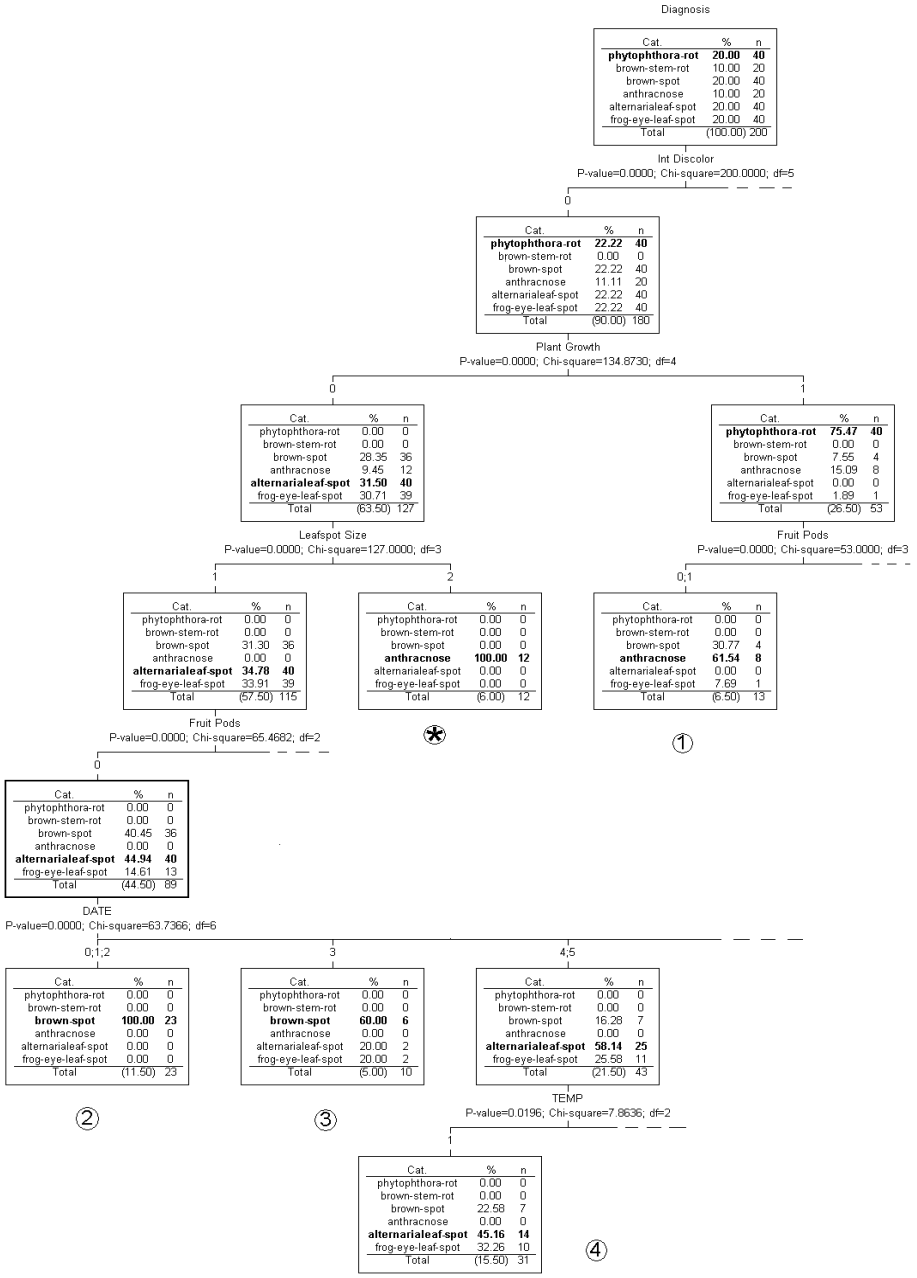

Figure 2. Soybean disease decision tree

We have also obtained rules that have little value. Consider the rule obtained from the path of which the terminal node is indicated in Figure 2 with an asterisk:

"If there is no internal discoloration in the stem, the plant growth is normal and the leaf spot size descriptor is not applicable, then the disease is anthracnose (100%)."

This rule is misleading because it suggests that regular (healthy) conditions indicate anthracnose disease.

The example of a misleading rule indicates the need for a preliminary analysis of observed models even before experts and decision makers became involved in the model evaluation and validation step of the KDD process. This analysis may also involve removal of rules with many precedents if there are other rules that contain only a subset of these precedents and have the same consequents. Rules with many precedents are highly specialized and considered "less interesting".

Table 1. Rule set generated with the AnswerTree CHAID method

Rule	Symptoms	Disease (%)
1	There is no internal discoloration in the stem AND the plant growth is abnormal AND the fruit pods are normal or diseased	Anthracnose (61.54%) Brown spot (30.77%)
2	There is no internal discoloration in the stem AND the plant growth is normal AND leaf spot size is above 1/8 inch AND the fruit pods are normal AND disease occurred in April, May or June	Brown-spot (100%)
3	There is no internal discoloration in the stem AND the plant growth is normal AND leaf spot size is above 1/8 inch AND the fruit pods are normal AND disease occurred in July	Brown-spot (60%) Alternaria leaf-spot (20%) Frog-eye-leaf-spot (20%)
4	There is no internal discoloration in the stem AND the plant growth is normal AND leaf spot size is above 1/8 inch AND the fruit pods are normal AND disease occurred in August or September AND temperature is normal	Brown-spot (22.6%) Alternaria leaf-spot (45.2%) Frog-eye leaf spot (32.2%).

To analyze the model and assess its potential usefulness one can repeat CHAID for different values of the method parameters and compare the results. A model is robust if, for the given dataset, changes in parameter values do not lead to significant changes in the model. Such a model is of greater value for decision makers because it is less sensitive to occasional records or outliers in the database.

Models are often verified using holdout data. The initial model is constructed from a sample drawn from the data base and assessed through the analysis of the remaining data. The two datasets are known, in data mining, as training and verification data sets respectively. The soybean dataset is small and therefore we used the entire set to construct the decision tree and rules. The use of training and verification data, and the estimation of the performance of the rules are discussed in Chapter 13.

3.4 Model comparison

Models obtained from data mining need be evaluated and verified (see Section 2.4). There are several possibilities, including the use of a different data mining method, comparison of the models with other models available in the literature, and the assessment of the model by the domain experts.

To compare rules obtained from the CHAID method we selected rules generated with AQ11, an inductive program developed by Michalski (1980) and applied to a different soybean database with 290 records. The observations were gathered from questionnaires completed by plant pathologists. The predictor and dependent variables available in these two data sets were identical. Rules that define brown spot disease obtained from the three sources are given in Table 2.

Table 2. Expert knowledge, and AQ11 and AnswerTree results for brown spot disease.

Expert knowledge	Inductive program AQ11	AnswerTree CHAID
Leaf condition is abnormal	Leaf condition is abnormal	
Leaf spot halos are present		
Leaf spot water soaked margins are absent		
Leaf spot size is above 1/8 inch	Leaf spot size is above 1/8 inch	Leaf spot size is above 1/8 inch
Disease occurred in May, August or September		Disease occurred in May, August or September
Precipitation is above normal	Precipitation is above normal	
	Crop is repeated in the same field for more than a year	
	Not the whole field is infected with the disease	
	Leaf is not malformed and roots are normal	
	No yellow leaf spot halos	
	No leaf spot water-soaked margins	
		No internal discoloration in the stem
		Fruit pod is normal

The three rules given in Table 2 contain some common precedents. There is one precedent present in all rules. Several precedents are present in two but not all rules. The rule generated with AQ11 is most specialized and has seven precedents. CHAID generated the rule with the least number of precedents (4), thus one may assume that this rule is most general. Interestingly, the expert's knowledge does not contain negative precedents that indicate that a particular symptom is absent for the brown spot disease. Both programs generated such precedents.

No rule from one source contradicts any other rule and none is a generalization of another. The question is if one should use all the rules or amend the expert's knowledge with precedents that appear in the generated rules but that are absent in the

statement of experts. We believe that the answer to this question should be left to domain experts.

4. Model verification

In the preceding section we compared two models obtained from the application of data mining methods with expert knowledge. While such a comparison is useful, additional and detailed analysis is often required especially if the decision problem is complex or domain knowledge is not available. The data mining literature suggests model verification by means of sub-samples of the database (Kohavi, 1995). In the *k* fold cross-validation method, the database is partitioned into *k*, ($k \geq 2$) subsets. The model is constructed in *k* iterations, where in each iteration a different subset is used for calculating the accuracy of the model constructed using the remaining $k-1$ subsets. Another approach found in the literature is to sample the database *k* times and average the results (Weiss and Indurkhya, 1998).

In this section other approaches to model verification are presented. For this purpose we use the database containing over 1,400 transcripts of negotiations conduced via the INSPIRE system and two questionnaires filled out by the users. The system and negotiations are discussed in more detail in Chapter 11.

For the purpose of model verification three models are constructed and analyzed. As in Section 4, the CHAID AnswerTree method (SPSS, 1998) is used with a minimum significance level of 5% for considering significant predictors and for merging branches. The decision trees are constructed using all available data.

4.1 Model confirmation

INSPIRE negotiations are anonymous and bilateral. Each user negotiates with their counterpart for up to three weeks. Upon completion of the negotiation users are requested to fill in a post-negotiation questionnaire. Several questions pertain to the user's perception of their counterpart. One of the questions asks about the user's willingness to work in future with their counterpart. We selected this variable (denoted workwopp) as the dependent variable. Forty-four variables were considered as possible independent variables. In this model we seek an explanation rather than prediction; thus the model is descriptive and not predictive.

The decision tree generated with the AnswerTree CHAID method (SPSS, 1998) is depicted in Figure 3. The following four variables have a relationship with the dependent variable (variable name in parantheses):

1. Whether the counterpart was honest or deceptive (opphones).
2. Whether the counterpart was cooperative or selfish (oppcoop).
3. How much control the user had during negotiations (control).
4. Whether an agreement was reached (agr).

There have been no similar experiments prior to the negotiations conducted via the INSPIRE system. The users are allowed to preserve their anonymity, they can be from

any country and region, and they can use several decision support techniques. We do not have access to any expert knowledge that may be used to verify the above model. Although cross-validation may possibly improve the results, the problem of further verification remains.

To verify the model, logistic regression with the four descriptive variables and one dependent variable was conducted. An overall likelihood ratio test of the logistic regression model yields a likelihood ratio statistic of 200.0 on four degrees of freedom, with a significance level, *p*, less than 0.0001. This provides evidence that the four variables included in the model make a strong contribution to describing the user's willingness to work with his/her opponent in future. A standard goodness-of-fit test indicates that the model fits the data well ($p = 0.19$). In other words, there is no reason to suspect that important explanatory variables have been omitted (Agresti, 1996).

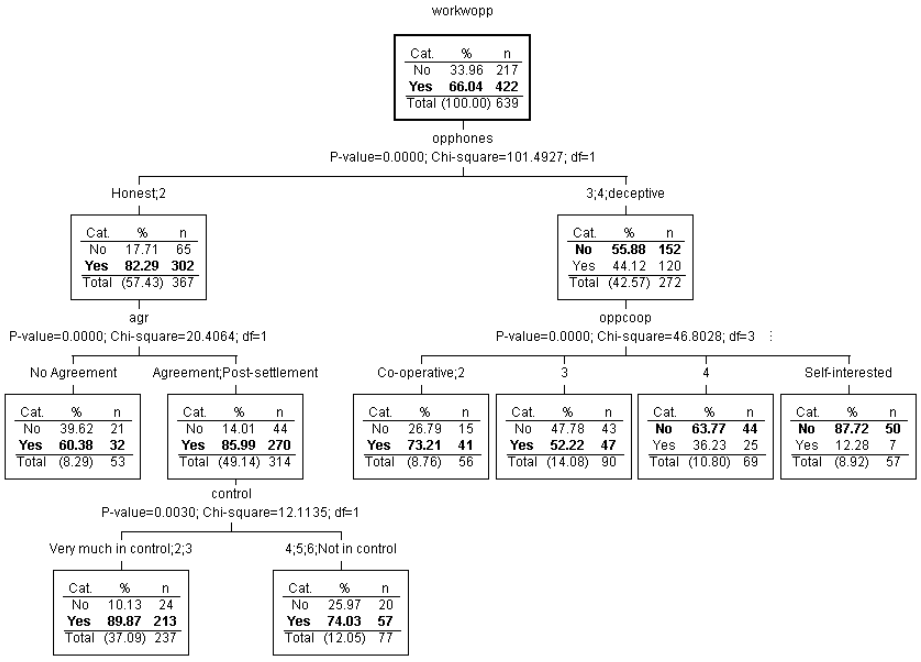

Figure 3. Decision tree for the users' desire to work with counterpart.

Logistic regression provides additional information. Each variable in the specified model has the following interpretation (under the assumption that the values of other variables are constant):

1. opphones: The odds of a user agreeing to work with their counterpart are 3 times greater for those who believe that their counterpart is honest, than for those who believe that their counterpart is deceptive.

2. oppcoop: The odds of a user agreeing to work with their counterpart is twice as great for those who believe that their counterpart is cooperative, than for those who believe that their counterpart is selfish.

3. control: The odds of a user agreeing to work with their counterpart are 7 times greater for those who believe that they are in control during negotiations, than for those who believe that are not in control during negotiations.
4. agr: The odds of a user agreeing to work with their counterpart are 0.27 times greater for those who reach an agreement, than for those who do not reach an agreement.

4.2 The case of two models

Negotiators are often satisfied with inefficient (not Pareto-optimal) compromises. The second example considers the dependent variable opt (whether the final agreement was efficient). There were twenty-two variables initially considered as possible predictive variables.

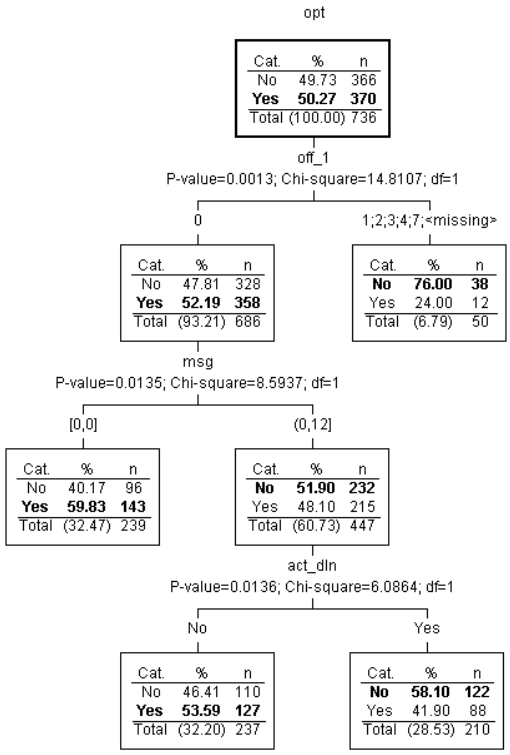

Figure 4. Decision tree for the agreement efficiency problem.

From the decision tree model depicted in Figure 4, it follows that the following three variables may be significant predictors for the dependent variable:

1. Number of offers conveyed by user on the second-last day of negotiations (off_1).

2. Number of messages sent by user to their counterpart (msg).
3. Whether there is activity forty hours prior to the deadline (act_dln).

We conducted, as in Section 5.1, a logistic regression. In this model, however, only two variables (off_1 and act_dln) are significant, at the 5% significance level. Variable msg is not significant at this level.

The msg variable is discrete, $0 \leq msg < +\infty$. The decision tree (see Figure 4) indicates that of the possible values in the variable, only two categories have significant association with the dependent variable (opt).

A second iteration of logistic regression was conducted, with the msg variable recorded into two categories: 1) no messages were conveyed by the user during negotiations, and 2) at least one message was conveyed by the user during negotiations. The descriptive variables with significant logistic parameters are off_1, act_dln, and msgbin (dichotomized msg variable).

An overall likelihood ratio test of the logistic regression model yields a likelihood ratio statistic of 24.6 with three degrees of freedom, ($p < 0.0001$). This provides evidence that the three variables included in the model make a strong contribution to explaining the probability of reaching an efficient agreement. A standard goodness-of-fit test indicates that the model fits the data well ($p = 0.17$). In other words, there is no reason to suspect that important explanatory variables have been omitted.

The interpretation of the logistic regression for each significant predictor is (keeping other variables constant):

1. off_1: An additional offer sent on the second-last day of negotiations reduces the odds of reaching an efficient agreement by .5 times.
2. act_dln: The odds of reaching an efficient agreement are 0.25 times lower when there is activity forty-eight hours prior to the deadline.
3. msgbin: The odds of reaching an efficient agreement are .54 times higher when no messages are sent by the user.

The presence of significant categories in msg might not have been considered unless there was prior knowledge that combining categories would increase the contribution that the variable makes to estimating the probability of success. The use of CHAID, therefore, indicates the possibility of increasing explanatory power when only significant categories are considered.

4.3 Interactions

INSPIRE users have access to data visualization and decision support tools. One of them is the history graph that depicts the flow of offers and counteroffers between a user and their counterpart (an example is depicted in Chapter 11, Figure 6).

In the third example we study the contributing factors to the use of the negotiation history graph. The dependent variable is graphuse (Whether the user viewed the history graph at any time during negotiations). There were twenty-three variables that might be related with the dependent variable. The decision tree for this problem is given in Figure 5.

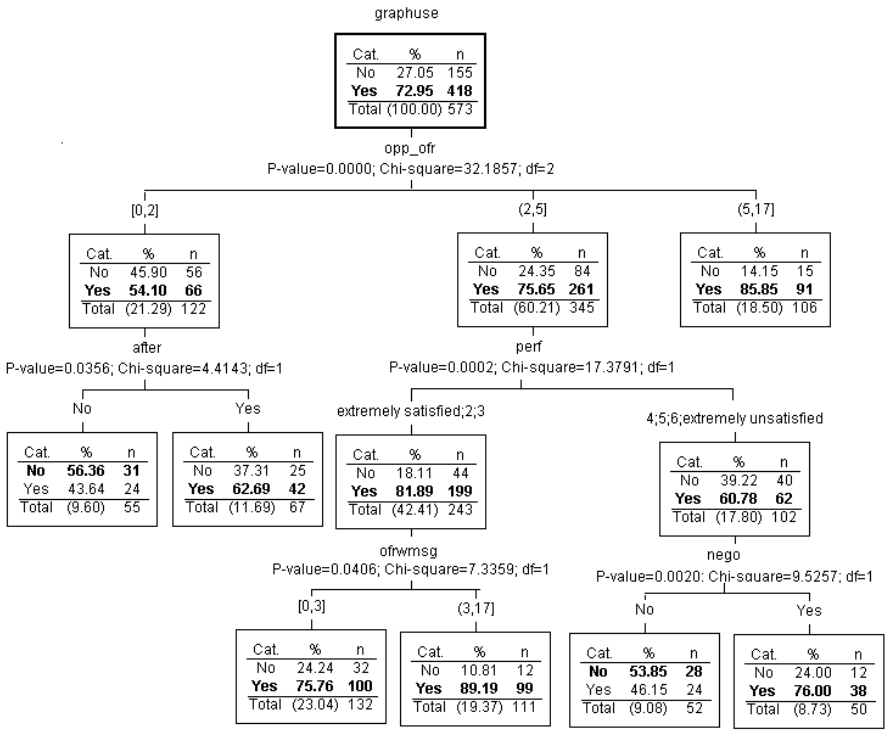

Figure 5. Decision tree representation of history graph problem.

From the initial set of variables the following five are related to the dependent variable:

1. Number of offers received from counterpart during negotiations (ofr).
2. Number of offers conveyed that included a message (ofrwmsg).
3. Whether the user would increase use of Internet after using INSPIRE (after).
4. Whether the user would use a similar system (to INSPIRE) in real life (nego).
5. The user's perception of their performance during negotiations (perf).

From the application of logistic regression we find that all descriptors, except for the variable after, are significant at the 5% significance level. In the previous example (Section 5.2), the msg variable was found to be significant if dichotomized**.** In considering the after variable, there is no further refinement possible because it is a binary variable. This raises concern with respect to the choice of this variable in CHAID.

Decision tree methods are used to construct models that represent interactive relationships among descriptive variables. Logistic regression measures the significance of the contribution a descriptor makes towards estimating a probability — interactions

among variables are not automatically included in a logistic regression model, though they can be separately coded and included as additional descriptive variables.

Interactions can be efficiently explored with loglinear modeling (Agresti, 1996). We observed interactions between the dependent variable and the 3-way interaction between after∗nego∗ofrwmsg, and between the dependent variable and perf (a∗b denotes interaction between variables a and b). Each *m*-way interaction ($m \geq 2$) implies the existence of lower order interactions. The variable perf has no interaction with any other variables, and so, it has a direct interaction with graphuse.

These initial results indicate a 3-way interaction between variables after, nego, and ofrwmsg. Such a high-order interaction is difficult to interpret. An approach to this problem is to divide the dataset according to a variable that appears in an *m*-way interaction. In this example, the dataset is divided into two subsets using the binary variable after. Loglinear analysis is used to construct further models on each of the constructed subsets.

For the subset of data where all observations take a value of after = 0, it is found that the dependent variable has an interaction with the generating class nego∗ofrwmsg and a direct interaction with variable perf. For the subset of data where all observations take on the value of after = 1, it is found that graphuse has direct interaction with variables ofrwmsg and nego. This step reveals that division of the initial data set has reduced the complexity of the existing interactions (from third-order to second-order, in one subset). A two-way interaction still represents complex interaction, and, therefore, further division is suggested. Correspondingly, the subset consisting of observations with values of after = 0 is partitioned using the values in variable nego. The subset with after = 1 is held constant since there is no higher order interactions for that data.

From the original set of data, there are now three subsets (after = 0 and nego = 0; after = 0 and nego = 1; and after = 1). For the subset of observations where after = 0 and nego = 0, it is found that the dependent variable has an interaction with variable perf alone. For the subset of observations where after = 0 and nego = 1, it is found that the dependent variable has an interaction with variable ofrwmsg alone.

Conducting logistic regression on each subset of data (using only corresponding descriptors) yields significant contributions of the descriptors to the estimation of success: the user viewed the history graph at least once during negotiations. These models (and corresponding *p*-values at alpha of 5%) are presented in Table 8. This sequential division of the database to yield simple logistic regression is interesting in that it can be regarded as a secondary CHAID-type process. Presumably, some optimal compromise between the database division and simple model representation exist. This general issue is of importance for the use of KDD but it is beyond the scope of this chapter.

Table 8. Logistic regression models for each of three subsets of data

Conditions of after and nego	Descriptors
after = 0 AND nego = 0	perf (p = 0.0121)
after = 0 AND nego = 1	ofrwmsg (p = 0.0121)
after = 1	ofrwmsg (p = 0.0001), nego (p = 0.0224)

5. Discussion

In the previous sections we presented several potential difficulties with the use of the KDD for model construction and validation. The four examples are based on relatively small databases with problems characterized by less than fifty variables. Large and very large databases and problems with thousands of potentially relevant variables will only multiply the difficulties of model validation and interpretation.

The problems of applying KDD methods to real-life situations need to be addressed. At present, this requires direct involvement of experts and the use of statistical methods. This does not mean, however, that KDD is not useful and that it provides no new insights. It also does not imply that the role of statistical methods in model formulation and verification remains the same irrespective of the use of data mining methods. Statistical methods, as we have attempted to show in this chapter, complement and enhance data mining methods. The latter allow for easy and efficient model construction, while the former are used to assess, verify and interpret the results.

The situation with KDD systems is similar to the earlier generations of DSSs that could not be used directly by decision makers. Instead, they were used by analysts and MS/OR specialists. Hence, one strategy for the effective utilization of KDD software is by back-office analysts who have an understanding of the domain as well as the software and methods. These experts may provide decision makers with interesting and relevant models or embed them in the DSSs that the decision makers use directly. Another strategy requires a significant extension of the capabilities of the existing software for KDD so that it can provide integrated support for all steps of the KDD process, including model validation.

Systems for KDD are considered intelligent (Limb and Meggs 1994). This intelligence, however, is limited to the use of AI mechanisms in the data mining software. At present the effort is primarily on data mining with other KDD steps only partially supported. There is little integration between the support provided during the different steps of the process.

Knowledge discovery from databases is a novel approach for model construction and validation. It requires dedicated software, powerful computing and communication facilities and access to databases. Many developing countries face shortages of computers and software, however, efforts of their governments and international organizations help to alleviate these problems. In many chapters of this book large complex DSSs have been developed and implemented in developing countries. Several examples of GISs, that have requirements similar to KDD, are discussed in chapters 3 and 5. Voluminous data for both developed and developing countries is collected from satelites and by other means (NASA, for example, collects data at the rate of 50 gigabytes per hour). Investments in communication technologies will allow for efficient access to data and to KDD methods.

Application of model-based DSSs in developing countries have been criticized because they often do not account for local indigenous knowledge and ignore local practices. Their incorporation into systems can be achieved through direct studies; an example of this approach is presented in chapters 10 and 13. It is also possible, however, to extract information about local procedures and traditions from historical data describing communities and regions.

Annex. KDD and sustainable development

1. Forest management

The analysis of ecosystems, including studies of biodiversity, forest density, wildlife population and water pollution, is aided through images collected via remote sensing. Traditionally, statistical methods, such as linear discriminant analysis, were used to analyze images. The use of these methods was cumbersome because of the very large amount of data. Flinkman et al. (1998) applied the rough set method (Pawlak, 1991) to analyze interactions between land uses, vegetation types, forest density, and other biotic, abiotic and anthropogenic conditions in the Siberian forest. The objective of their study was to identify key attributes for the formulation of sustainable forest management policies.

Precise and robust classification of land cover is required to conduct environmental assessment and to formulate plans and management policies. Friedl and Brodley (1997) used three decision tree methods (univariate, multivariate and hybrid methods) to obtain models for land cover classification. The models were compared with models obtained from two statistical methods (maximum likelihood and linear discriminant analysis) and used cross-validation for model comparison. The study was conducted on a global database, a database for North America, and one describing a forest in California. According to the authors, the decision tree models have accuracy comparable to statistical models. The advantages of the data mining tools include their good performance in the presence of disparate and missing data, and the ease of model construction given the complexity of the data.

2. Precision agriculture

Farming requires efficient management of resources (such as water and fertilizer) that needs to account for changes in the environment. Dong (1998) used remotely-sensed data to construct descriptive models of agricultural practices. The association rules analysis was used for this purpose. The author states that harvests from different fields can be mapped to each season and varying yields analyzed according to seasons and locations to determine the relationship between seasonality and location on crop yield. Farm managers and extension workers may use this information to allocate resources in order to maximize yield.

3. Oil wells

Oil drilling has significant impact on the surrounding eco-system. Selection of the type of mud where drilling is undertaken has, in turn, impact on drilling efficiency and costs, bore-hole instability and cleaning, and the cost of oil extraction. Project managers in an oil field in Oso, Nigeria encountered problems (for example, high incidents of stuck pipe and filter cake build up) and cost overruns when their decisions to select mud types were based on experience, the available resources and personal preferences. To address these problems they used data mining techniques to choose between alternative mud types (Dear III, 1995).

A structured approach based on a decision tree method was devised in order to determine the relationships between the attributes of mud type, drilling (including costs) and eco-system. The analysis revealed that mud types that were recommended by experience often led to higher costs, while types proposed with data mining and previously not considered, allowed lowering of the long-term costs (Dear III, 1995).

4 Chemical pollutants

Concentration of chemicals in the environment is regulated to minimize their impact on the ecosystem. This requires frequent assessment of chemical levels. Ranking systems, based on subsets of aggregated factors, have been introduced for this purpose. Aggregation, however, often does not include interactions among these factors which are required to determine acceptable levels of exposure. According to Eisenberg (1998) some of the ranking systems are highly complex and cannot be understood by decision makers and, therefore, are considered as “black boxes”. They provide estimates of chemical exposure but not information about the uncertainties and sensitivities associated with these estimates.

Eisenberg (1998) developed a methodology based on the CART decision tree method for the assessment of chemical exposure levels. It is used to identify properties of chemicals, conditions of the environment, and the relationships among chemical properties that are most important for classifying an area according to its exposure level. The CART method was used to derive classification rules of chemicals and identify factors affecting chemical exposure levels. These rules are also used for the analysis of changes in factors and corresponding changes in the classification (chemical exposure levels).

Acknowledgements

We thank David Cray for his comments and suggestions. This work has been supported by the Social Science and Humanities Research Council of Canada and the Natural Sciences and Engineering Research Council of Canada.

References

Aasheim, O. T. and H. G. Solheim (1996). "Rough Sets as a Framework for Data Mining", Norwegian University of Science and Technology**:** 148.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, New York: Wiley.

Becker, B., R. Kohavi, et al. (1997). "Visualizing the Simple Bayesian Classifier". *KDD 1997 Workshop on Issues in the Integration of Data Mining and Data Visualization.*

Bell, D. E. H. Raiffa and A. Tversky (1988). *Decision Making. Descriptive, Normative and Prescriptive Interactions*, Cambridge, MA: Cambridge Univ. Press.

Bigus, J. P. (1996). *Data Mining with Neural Networks*, McGraw-Hill.

Brachman, R. and T. Anand (1996). "The Process of Knowledge Discovery in Databases: A Human-centered Approach", in U. Fayyad et al., *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, 37-58.

Breiman, L., J. H. Friedman, et al. (1984). *Classification and Regression Trees*, Belmont, Wadsworth International Group.

Briscoe, G. and T. Caelli (1996). *A Compendium of Machine Learning, Vol.1*. Norwood, Ablex.

Dear III, S. F., R. D. Beasley, et al. (1995). "Use of a Decison Tree to Select the Mud System for the Oso Field, Nigeria." *Journal of Petroleum Technology*, 47(10), 909-912.

Dong, J. (1998). "Mining Association Rules from Imagery Data", Computer Science Dep. North Dakota State University, Fargo, ND.

Eisenberg, J. N. S. and T. E. McKone (1998). "Decision Tree Methods for the Classification of Chemical Pollutants: Incorporation of Across-Chemical Variability and Within-chemical Uncertainty." *Environmental Science and Technology*, 32(21), 3396-3404.

Fayyad, U. M. (1996). "Data Mining and Knowledge Discovery: Making Sense Out of Data" *IEEE Expert*, 11(5), 20-25.

Fayyad, U. M., G. Piatetsky-Shapiro, et al. (1996). "From Data Mining to Knowledge discovery: An Overview", U. M. Fayyad, G. Piatetsky-Shapiro et al. (Eds.), *Advances in Knowledge Discovery and Data Mining*, Menlo Park, MIT Press.

Flinkman, M., W. Michalowski et al. (1998). "Identification of biodiversity and Other Forest Attributes for Sustainable Forest Management: Siberian Forest Case Study", IR-98-106, International Institute for Applied System Analysis, Laxenburg, Austria.

Friedl, M. A. and C. E. Broadly (1997). "Decision Tree Classification of Land Cover from Remotely Sensed Data", *Remote Sensing of Environment*, 62, 399-409.

Glymour, C., D. Madigan, et al. (1997). "Statistical Themes and Lessons for Data Mining" *Data Mining and Knowledge Discovery*, 1(1): 11-28.

Inmon, W. (1996). *Building the Data Warehouse*, New York: Wiley.

John, G. H. (1997). "Enhancements to the Data Mining Process", Computer Science Department, School of Engineering, Stanford University**:** 194pp.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *International Joint Conference on Artificial Intelligence (IJCAI)*.

Kohavi, R. (1998). "Crossing the chasm: From academic machine learning to commercial data mining". Mountain View, SGI, http://reality.sgi.com/ronnyk/chasm.pdf.

Kohavi, R., B. Becker, et al. (1998). "Improving simple bayes". Mountain View, Data Mining and Visualization group, Silicon Graphics, Inc.

Limb, P. R. and G. J. Meggs (1994). "Data Mining - tools and techniques" *BT Technology* 12(4): 32-41.

Mahoney, J. J. (1996). "Combining Symbolic and Connectionist Learning Methods to Refine Certainty-Factor Rule-Base", Department of Computer Sciences, The University of Texas. [on-line] http://www.cs.utexas.edu/users/ml/abstracts.html

Michalski, R. S. and R. L. Chilausky (1980). "Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology." *International Journal of Man-Machine Studies*, 12, 63-87.

Mienko, R, et al. (1996). "Discovery-oriented Induction of Decision Rules", Cahier du LAMSADE, No. 141, Paris.

MLRepository (1999). Machine Learning Database Repository, University of California-Irvine, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Pawlak, Z., I. Grzymala-Busse, et al. (1995). "Rough Sets", *Communications of the ICM*, 38(11), 89-95.

Sasisekharan, R., V. Seshadari, et al. (1996). "Data Mining and Forecasting in Large-scale Telecommunication Networks", *IEEE Expert*, 11(1), 37-43.

SGI (1998). "MineSet 2.5 tutorial", Silicon Graphics Inc.

Simoudis, E. (1996). "Reality Check for data Mining" *IEEE Expert* 11(5), 26-33.

Skinner, C. J., D. Holt and T. M. F. Smith (1989). *Analysis of Complex Surveyes*, Chichester" Wiley.

SPSS (1998). "SPSS in Data Mining". Chicago, SPSS Inc., http://www.spss.com.

Srikant, R. and R. Agrawal (1998). "Mining quantitative association rules in large relational tables", IBM Almaden Research Center.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Reading, MA: Addison Wesley.

Weiss, S. M. and N. Indurkhya (1998). *Predictive Data Mining: A Practical Guide*, San Francisco, Morgan Kaufman.