UDC: 007.52
Original scientific paper

APPLICATION OF KNOWLEDGE DISCOVERY IN DATABASES TO AN ANALYSIS OF SURVEY DATA ON SUBSTANCE ABUSE

Goran Bubaš, Božidar Kliček
University of Zagreb, Faculty of Organization and Informatics Varaždin, Croatia
E-mail: gbubas@foi.hr or bklicek@foi.hr
Željka Čolović Rodik, Zvjezdana Fulir
Center for Prevention and Outpatient Treatment of Addiction, Varaždin, Croatia

Knowledge discovery in databases (KDD) was evaluated for its potential application to the analysis of survey data on substance abuse. Data for 238 psychosocial variables were collected from high school adolescents (N=4198). One criterion variable (marijuana/hashish abuse) and 77 potential predictor variables were selected for linear regression and decision tree analyses. The results obtained by both methods are presented and compared. It is concluded that KDD techniques offer much opportunity for the discovery of new patterns in survey data, model development of substance abuse, and prevention activities.

Keywords: knowledge discovery in databases, data mining, substance abuse, survey.

1. INTRODUCTION

The pervasive use of information technology in business and science results in the dramatic increase of data that is collected and accumulated for various purposes. In addition to the explosive growth in volumes, the structure and complexity of data on many applications is enhanced as well. New computational theories and tools are emerging to assist in extracting useful information from voluminous collections of data, thus constituting a rapidly growing field of *Knowledge Discovery in Databases* (KDD). This interdisciplinary field embraces database theory, pattern recognition, AI, statistics, data visualization, high performance computing and other disciplines [14].

Organizations are facing both the challenge and the imperative to use collected data more effectively and to discover patterns of relations which could be useful for solving problems in business and scientific research. KDD, also known as *data mining* in statistics, is a process of discovering previously unknown patterns in large databases. This could be a very difficult task as the number of variables and volume of data become large and interactions among the variables become complex. The ability to extract valuable information, which is unattainable or difficult to obtain using standard procedures, could enable organizations to provide services and products more adapted to specific customers and to win business in the changing market place [1].

Most of the initial *business* applications of KDD can be found in the fields of marketing, investment, manufacturing, telecommunications, administration, law enforcement and professional sport [16]. Some of the recently reported *scientific* applications of KDD are related to medicine [31], biology [29], geology [13], astronomy [8], and linguistics [10]. It is evident that the scientific challenges of KDD, as well as its potential for practical application, have ignited the interest of both the academic and user communities [27].

This paper is concerned with the potential of KDD for discovering unknown patterns in survey data related to substance abuse. A data set including more than 200 variables and over 4000 records is used to evaluate the applicability of KDD techniques in discovering patterns in survey data which are difficult to uncover by standard methods of statistical analysis. It should be noted that no evidence was found of prior research using KDD in relation to (a) substance abuse, or (b) analysis of psychosocial survey data. This state is expected to change since producers of statistical software packages, used by social scientists in related fields, are putting products with data mining capabilities on the market [11].

2. OBJECTIVES

Effective dealing with substance abuse demands the use of state-of-the-art information technology in both research and prevention activities [5; 9]. This research began with an endeavor, supported by the district municipality, to perform an extensive survey on substance abuse among high school students in one district of northwestern Croatia. After assessing the extent and intensity of substance abuse for various segments of this population, an effort was made to identify predictors of substance abuse behavior using standard statistical procedures. Then, KDD techniques were used to further investigate the relationship between selected subsets of predictor variables with specific consumers of narcotic substances and some of the results are presented in this paper. Finally, the applicability of KDD techniques was assessed for potential use in (a) analysis of survey data, and (b) decision making in substance abuse prevention activities, as part of a research project on *intelligent systems in decision support*.

The objectives of this paper are: (1) an evaluation of the applicability of KDD techniques for analysis of survey data; (2) a comparison of KDD techniques in relation to standard statistical methods; and (3) an uncovering of novel information to improve decision making in substance abuse prevention.

3. KNOWLEDGE DISCOVERY IN DATABASES AND *DATA MINING*

A considerable mental effort is usually required to develop knowledge from complex data, even if the manual analysis and interpretation are based on statistically (pre)processed and organized information. The basic problem of the KDD process is mapping voluminous low-level data into forms which might be (a) *more compact* (i.e. less redundant, more relevant), (b) *more abstract* (for example, a descriptive approximation or model of a process that generated the data), or (c) *more useful* (like a predictive model), all of which should enable a specialist to familiarize himself with the data and serve as an interface between the data and the user [16].

KDD is an iterative process for discovering useful knowledge from data that commonly includes the following phases [see 4 for a detailed comparative example]:

(i) Preparation, selection, cleaning, and transformation of data.
(ii) Data mining to extract patterns.
(iii) Incorporation of prior knowledge to interpret extracted patterns.
(iv) Evaluation of extracted patterns to decide what constitutes "new knowledge".
(v) Validation and consolidation of extracted knowledge.
(vi) Making knowledge available and utilized.

(i) The effective extraction of knowledge from data is feasible if appropriate data sets (i.e. relevant samples, cases, or attributes) are extracted from database(s) and prepared for analysis. Usually, keys have to be reconstructed, encoded values made consistent, structures of data standardized, missing data and errors appropriately represented and handled. *Data warehousing* is a process which sets the stage for effective data mining by collecting and cleaning data, and by providing access to *integrated, detailed and summarized data, historic data and metadata* in renovated data structures [25; 34].

(ii) Data mining is a step of KDD which often involves the repeated iterative application of particular data mining methods in a process directed toward two types of goal: (1) *verification* of user's hypothesis, or (2) *discovery* of new patterns in data. The latter goal is further subdivided into (a) *prediction*, where discovered patterns are used for predicting the future behavior of some entities, and (b) *description*, where the system finds patterns and presents them to users in a comprehensible form [16].

Many different methods and a vast, ever growing number of algorithms have been developed for data mining, but the most commonly used methods in data mining are: artificial neural networks, decision trees, genetic algorithms, nearest neighbor method, and rule induction. Numerous products are on the market that utilize diverse methods and algorithms for analysis [32].

(iii) The application of KDD in science emphasizes the importance of prior knowledge for an interpretation of extracted patterns. It must be accentuated that prior knowledge is more advanced in science than it is commonly in respect to users in the field of business, but the complexity of the problems which are addressed in science can be greater in scale. Prior knowledge is essential for the correct interpretation of discovered patterns, as well as for developing models which are closest to the real world processes they are supposed to represent.

Science users usually know their data in greater detail, and they are better trained to formalize intuitions into procedures and equations. They also have more experience with research methodology and statistical inference. Generally, the scientific applications of KDD should be easier than those for business or other areas [15].

(iv) Prior knowledge is also used to discern 'new knowledge' and to evaluate discovered patterns for their significance. Usually, one or more interpreted pattern structures are chosen to constitute preliminary representative models, out of many

alternative patterns in a set of data with numerous variables that could be derived by data mining methods. The selection process is usually performed iteratively on the basis of reasoning and prior knowledge.

(v) Extracted knowledge should be subjected to validation procedures. Data miners are sometimes thought of as more interested in comprehensibility than accuracy or predictability, with a focus on relatively simple interpretable models involving rules, trees, graphs and so forth. Statistics has much to offer the in evaluation of results acquired by application of data mining methods [19; 20].

Validation ensures that the discovered knowledge is not an artifact produced by an inadequate data set, or by the weakness of employed search procedures. The validation of discovered knowledge is an obligatory step before its application.

(vi) Discovered knowledge may be used to enhance insight in a specific field or support a decision, it can also be incorporated into another system for further use (i.e. in a *pattern recognition*, *DSS* or *expert system*), or it can be documented and reported to interested users [see: 3; 38].

3. METHOD

4.1. Field survey

In a survey on substance abuse 4198 adolescents were polled in a district of northwestern Croatia. The subjects were pupils in their first, third or fourth year of high school (i.e. their ninth, eleventh and twelfth consecutive year of obligatory education in Croatia). The subjects voluntarily and anonymously participated in the survey, which was performed in the second half of the 1997/98 school year.

Each subject's responses were a source of data for a total of 238 psychosocial variables. The survey included questions on substance use (alcohol, tobacco, and narcotics), attitudes related to substance use, consequences of alcohol consumption, social status, social environment, and out of school activities. The content of the questionnaire corresponds to a survey previously completed in 26 European countries [23].

4.2. Preliminary data analysis

Preliminary analysis of the survey data was limited to a frequency analysis of all 238 variables, as well as to an analysis of data on the abuse of specific substances for different school populations. For narcotics, only the abuse of (a) *marijuana or hashish*, and (b*) tranquilizers (sedatives*) were reported with sufficient frequency to use data mining methods for analysis. Abuse of other narcotics, i.e. amphetamines, ecstasy, crack, cocaine, heroin, LSD (or other hallucinogenic substances), and anabolic steroids was reported by limited subsets of subjects which differed in size from 50 to 150 individuals out of a total of 4198 subjects in our survey.

For the purpose of this research, *the reported abuse of marijuana/hashish in the last 12 months* was chosen to be the criterion variable. The frequency of reported abuse of marijuana or hashish is presented in Table 1.

Table 1. Frequency of reported marijuana or hashish abuse by high school children in our survey.

Time period	Reported frequency								
	0 (never)	1-2	3-5	6-9	10-19	20-39	40 or more	No response	TOTAL
a. In your ifetime	3674	229	89	46	40	44	50	26	**4198**
b. In the last 12 months	3461	176	53	37	37	21	26	387	**4198**
c. In the last 30 days	3626	105	34	18	10	4	15	386	**4198**

4.3. Selection of potential predictor variables

Much effort was devoted to the proper selection of potential predictor variables from the total of 237 remaining variables. Correlation, regression, and factor analysis were used, as well as logical reasoning in an iterative process of inclusion and exclusion of potentially useful variables.

Except for the criterion, other variables referring to marijuana or hashish were excluded from further analysis because they were believed to be a part of the criterion. Furthermore, since use of marijuana or hashish is commonly antecedent to the use of other narcotics like heroin and LSD, numerous variables related to other narcotic substances were also excluded from the list of potential predictors. Finally, variables that were for some reason found to be of little or no relevance as potential predictors were excluded as well.

A total of 77 variables was chosen as a list of potential predictors and this primary set was divided into two overlapping subsets: (1) a *drug-independent* subset of potential predictor variables, which are not necessarily related to narcotics, and (2) a discriminative *drug-dependent* subset of variables, which are related to narcotics, with selected social status and activities variables as well.

The first subset of predictors *(drug-independent)* included variables associated with gender, school, family, out of school activities, smoking, alcohol consumption, consequences of alcohol use, and gambling. The second subset of potential predictors *(drug-dependent and others)* included variables associated with: gender, school, smoking, alcohol consumption, use of different alcoholic beverages, consequences of alcohol use, knowledge of narcotic substances, abuse of various narcotics (i.e. tranquilizers, amphetamines, ecstasy, and consuming alcohol in combination with narcotic pills), accessibility of narcotic substances, attitudes toward drug consumption, and substance abuse by friends.

4.4. Linear regression and decision tree method

Data for the criterion and potential predictor variables was recoded prior to the analysis to answer some of the restrictions of the software used to generate decision trees. Two overlapping subsets of potential predictor variables were analyzed by both the stepwise linear regression and the decision tree technique. For the latter, machine learning by neural network was performed, with a learning phase and a test phase each on 50% of the cases, as well as with the visual presentation of a decision tree.

Since the criterion variable was continuous, inference rules were not generated automatically, but deduced manually from the decision tree structure for exemplary purposes.

4. DATA ANALYSIS

5.1. Regression analysis of <u>drug-independent</u> potential predictors

When *stepwise linear regression analysis* was performed on the first subset of 41 potential predictor variables *(drug-independent)*, it brought forth 17 predictors of reported frequency of marijuana or hashish use within the last 12 months (see Table 2.). Multiple R of the predictor variables with the criterion reached 0.456, which explains about 21% of its variance.

It can be concluded that a large number of the drug-independent predictor variables affect the criterion. However, a list of predictors explains little about the structure of marijuana or hashish use. Furthermore, much more statistical analysis is needed if one needs to highlight specific beneficial configurations of smaller sets of predictor variables with the heaviest influence on various intensity levels of the criterion variable, or to identify specific classes of subjects involved in the survey. From these points of view a decision tree analysis of the same subset of potential predictors may be more informative (see Figure 1).

Table 2. *Drug-independent* predictors of frequency of marijuana/hashish use in last 12 months (N=4198)

Predictor variables	Beta[†]	T[††]	Sig. T[†††]
Occurrences of excessive drinking in last 12 months	.189	11.37	.0000
School absenteeism	.089	5.82	.0000
Frequency of smoking in 30 days	.086	5.51	.0000
Problems with police because of alcohol use	.081	4.91	.0000
Higher year in high school	.069	4.70	.0000
Frequency of playing on gambling machines	.058	3.91	.0001
Entering a conflict or fight because of alcohol	.056	3.35	.0008
Early age of first inhalation of glue or aerosol spray	.054	3.77	.0002
Father not living in same household	.052	3.73	.0002
Higher education of mother	.050	2.95	.0032
Non-relatives living in household	.046	3.33	.0009
Taking tranquilizers (sedatives) prescribed by physician	.042	2.99	.0028
City *(non-rural)* residence	.040	2.78	.0055
Frequency of playing with computer games	.037	2.61	.0090
Higher education of father	.037	2.16	.0309
Mother consumes alcohol excessively	.034	2.46	.0138
Not seeing risk in excessive consumption of alcohol	.030	2.09	.0366

5.2. Decision tree analysis of <u>drug-independent</u> potential predictors

The result of a *decision tree* analysis of the first subset of potential predictors *(drug-independent)* provides information on the *subsets of subjects* in which there are cases of reported use of marijuana or hashish (see Figure 1.). By following the

[†] Beta – regression coefficient; [††] T – value of test of significance; [†††] Sig. T – statistical level of significance.

"branches" of the decision tree which lead from the *"root"* to critical *"leaves"* (i.e. to classes of subjects who reported an on the average higher frequency of marijuana use), a different kind of knowledge can be attained than from the standard regression analysis output. This knowledge explains in more detail the relationship between the more significant predictors and the criterion variable.

The highest average degree of marijuana or hashish abuse was reported by 34 subjects defined by *"leaf" no. 33* (Figure 1.). This "leaf" is determined by the following *profile* of related attributes[†]: *IF* (a) more than 10 occasions of excessive consumption of alcohol in the last 12 months *AND* (b) almost all friends are drinking alcohol heavily at least once a week (R1). This unearthed "rule" can be redefined as a statement: *"the most frequent abuse of marijuana/hashish is related to repeated excessive consumption of alcohol and very frequent heavy use of alcohol by friends as well"*.

Another example of a high average degree of marijuana or hashish abuse is defined by *"leaf" no. 32* with a frequency of 37 subjects (Figure 1.). Succeeding attributes determine its *profile*: *IF* (a) more than 10 occasions of excessive consumption of alcohol in the last 12 months *AND* (b) not all friends are drinking alcohol heavily at least once a week *AND* (c) more than 20 occasions of smoking tobacco during one's lifetime *AND* (d) education level of mother greater than high school (R2). The features of this set of critical subjects constitute a new "rule" which can be defined as: *"abuse of marijuana/hashish is related to repeated excessive consumption of alcohol and an inclination to smoke tobacco for subjects whose mothers received a higher level of education than high school, if all of their friends do not drink heavily at least once a week"*.

By following different paths from the *"root"* of the decision tree to the *"leaves"* with critical subjects, distinct "rules" can be derived, as for the 32 subjects denoted by *leaf no. 28* (Figure 1) to whom the following definition statement applies: *"abuse of marijuana/hashish is related to more than one, and less than 10 occasions of excessive consumption of alcohol in the last 12 months, smoking more than five cigarettes of tobacco per day, the education of the mother to a level higher than primary school, and to the father not living in the same household"* (R3).

By examining which branches of the decision tree in Figure 1 lead to leaves denoting <u>*no instance*</u> *of reported marijuana/hashish abuse* one can also develop the following "rule": *IF* (a) there is no excessive consumption of alcohol in the last 12 months *AND* (b) there is no inclination to smoke tobacco *AND* (c) almost every evening in a week is *not* spent with friends in a bar, disco or at a party *THEN* <u>there is almost no likelihood of marijuana/hashish abuse</u> (R4).

Examples of rule generation from the decision tree structure presented in Figure 1 demonstrate the potential of this technique for a heuristic search of relevant patterns (or *new knowledge*) in the procedure of survey data analysis, as well as for model

[†] Only a limited content of attributes is denoted within the nodes of decision trees in Figure 1 and Figure 2. The questions asked in the survey that correspond to those attributes were numerous, lengthy and with different categories of answers. Since they can be found elsewhere [23], they are not presented in this paper.

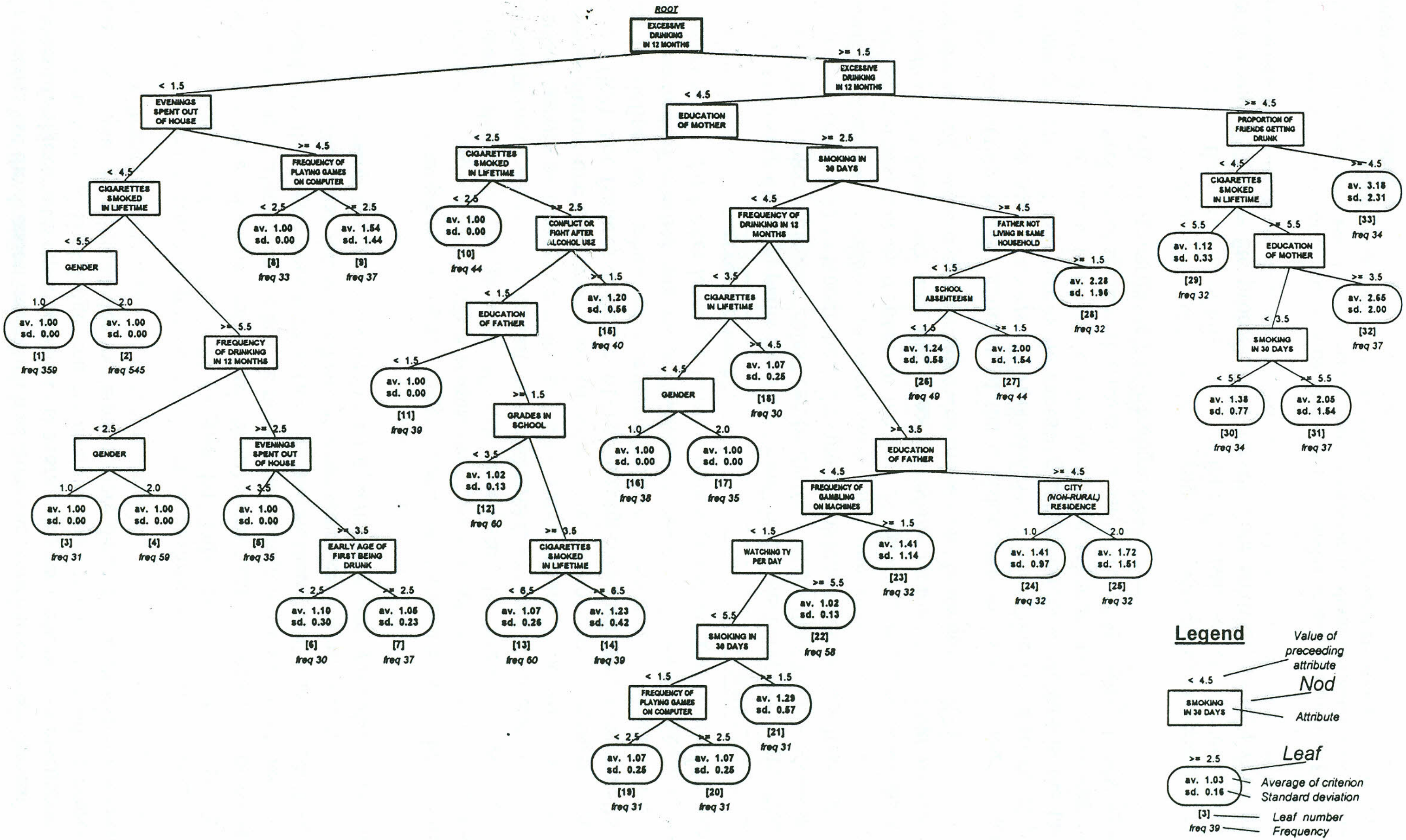

Figure 1. Results of decision tree analysis for *drug-independant* potential predictors of marijuana/hashish abuse

development of real world processes of substance abuse. Different pattern structures for different sets of *predictor variables* (as in Figure 2) or different *criteria* (i.e. abuse of alcohol, tranquilizers, and other narcotics) can be obtained in minutes, and explained without much exertion.

However, more simplistic and obvious decision trees could be obtained *if* either (a) different algorithms are used for structure generation and display, *or* (b) insignificant branches/paths are pruned, *or* (c) the minimum number of cases in terminal nodes (i.e. *"leaves"*) is increased. Our exemplary decision trees (in Figure 1 and Figure 2) are presented as they were produced by the software used for data analysis, and have a considerably "overgrown" structure as far as lucidity and functionality are concerned.

5.3. Regression analysis of drug-dependent and other potential predictors

In a further statistical analysis, *stepwise linear regression* was performed on the second subset of 53 potential predictor variables *(drug-dependent and other variables)* and it brought forth another combination of 23 predictors of *reported frequency of marijuana or hashish abuse within the last 12 months* (see Table 3.). Multiple R of these predictor variables with the criterion reached 0.619, which explains about 38% of its variance. When *drug-dependent and other* potential predictor variables were included in regression analysis many more predictors were found to affect the criterion.

It must be noted that one of the most significant predictor variables, *"proportion of friends who smoke marijuana or hashish"*, has been intentionally excluded from regression analysis (it correlated 0.50 with the criterion). This was done because it dominated in the structure of the decision tree derived from the second subset of potential predictors *(drug-dependent and other)*, and it is thought of largely as a part of the criterion.

5.4. Decision tree analysis of drug-dependent and other potential predictors

The results of the decision tree analysis of the second subset of potential predictor variables *(drug-dependent and other)* are presented in Figure 2. From the structure of this decision tree, the following rules can be unfolded:

- *Leaf 31*: Very probable marijuana/hashish use in the last 12 months: *IF* (a) 10 or more occasions of consumption of alcohol in the last 12 months *AND* (b) use of ecstasy, or first use of ecstasy at an early age (R5).
- *Leaf 29*: Probable use of marijuana/hashish in the last 12 months: *IF* (a) 10 or more occasions of consumption of alcohol in the last 12 months *AND* (b) no reported use of ecstasy *AND* (c) there is a proportion of friends who take tranquilizers or sedatives without prescription *AND* (d) third or fourth year of high school *AND* (e) no disapproval of people who have used ecstasy once or twice (R6).
- *Leaf 13*: Probable use of marijuana/hashish in the last 12 months: *IF* (a) more than one, and less than 10 occasions of consumption of alcohol in the last 12

months *AND* (b) smoking of more than 11 cigarettes per day *AND* (c) awareness of amphetamines like kalicor or encephabol (R7).

Even though the two subsets of potential predictor variables (i.e. *drug-independent* and *drug-dependent and other*) overlap for a number of attributes, there is very little similarity in the decision trees in Figure 1 and Figure 2, as well as little similarity in the rules revealed by their structure. Exclusion of important attributes (like *"occurrence of alcohol consumption in the last 12 months"* and *"occurrence of excessive alcohol consumption in the last 12 months"*) would considerably alter and distinguish these models. Our selection of two subsets of potential predictors demonstrates: (1) the existence of parallel feature structures within survey data sets, and (2) a potential for the comparison of the relative importance of different attributes when generating concurrent interpretable decision tree structures for equal criteria or real world processes.

Table 3. *Drug-dependent and other* predictors of frequency of marijuana/hashish use in last 12 months (N=4198)

Predictor variables	**Beta**	**T[††]**	**Sig.[†††]**
Frequency of abuse of ecstasy	.191	12.07	.0000
Freq. of inhalation of glue or aerosol spray in last 12 months	.168	11.29	.0000
Occurrences of excessive drinking in last 12 months	.123	7.94	.0000
Frequency of abuse of amphetamines	.115	7.26	.0000
Occurrences of consuming alcohol in combination with narcotic pills	.115	7.49	.0000
Approval of people who have used ecstasy once or twice	.069	5.29	.0000
Cognizance of amphetamine or related narcotics	.060	4.23	.0000
Disagreement with the statement that "people who drink lose control in a displeasing manner"	.057	4.47	.0000
Frequency of smoking in a 30 day period	.050	3.58	.0004
City *(non-rural)* residence	.049	3.90	.0001
Problems with police because of alcohol use	.049	3.56	.0004
Proportion of friends who take tranquilizers (sedatives) without prescription	.046	2.88	.0040
Higher year in high school	.046	3.44	.0006
School absenteeism	.046	3.36	.0008
Early age of first inhalation of glue or aerosol spray	.044	3.20	.0014
Proportion of friends who use substances which can be inhaled (i.e. glue)	.037	2.37	.0177
Awareness of methadone	.035	2.43	.0014
Gender (male)	.033	2.59	.0097
Frequency of drinking wine in last 30 days	.031	2.26	.0240
Frequency of reading non-school books for fun	.030	2.41	.0159
Not taking tranquilizers (sedatives) prescribed by physician	.030	2.34	.0195
Early age of first attempt to use tranquilizers or sedatives without prescription	.029	2.12	.0340
Frequency of playing with computer games	.027	2.14	.0324

[†] Beta – regression coefficient; [††] T – value of test of significance; [†††] Sig. T – statistical level of significance.

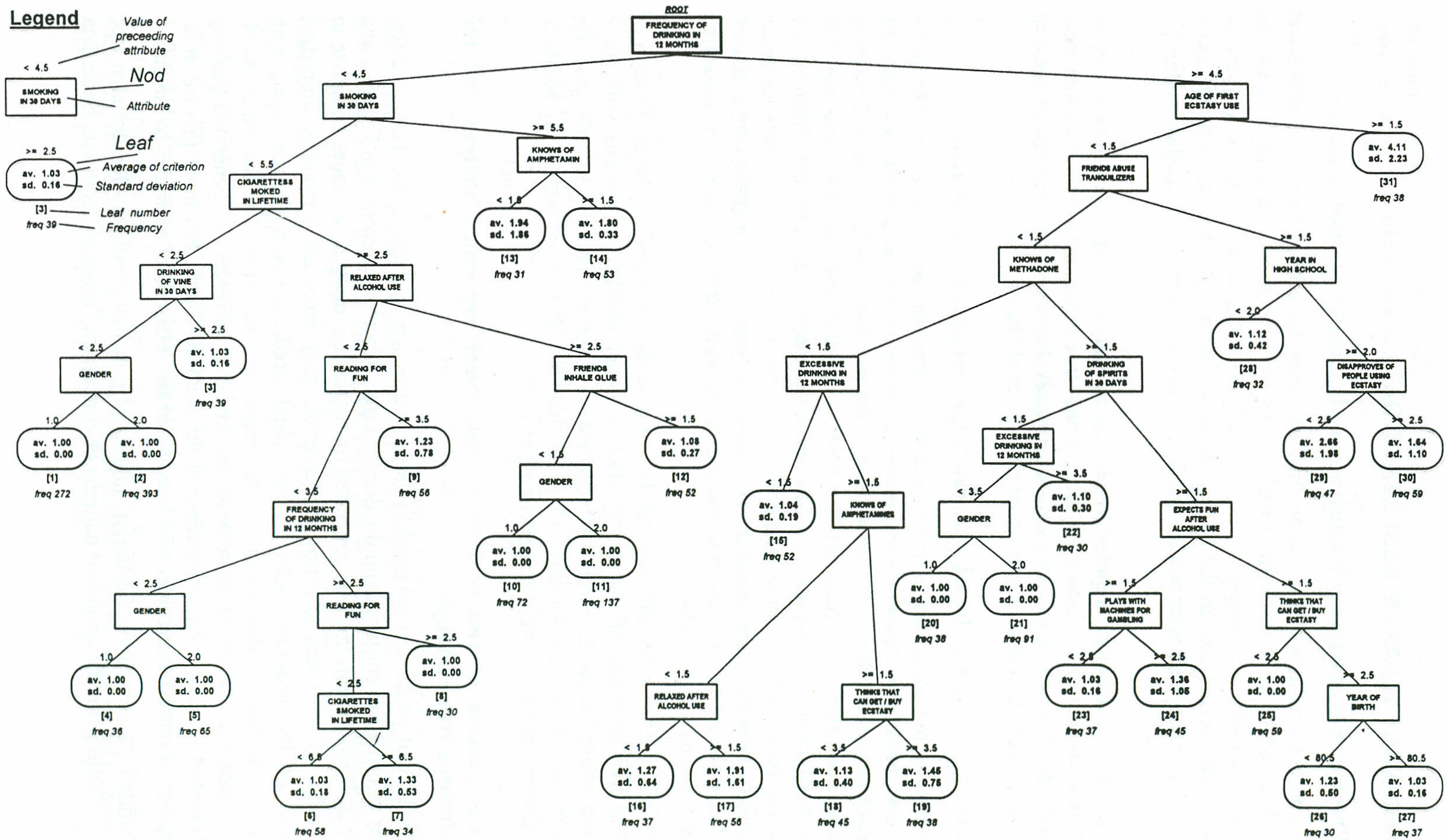

Figure 2. Results of decision tree analysis for *drug-dependant and other* potential predictors of marijuana/hashish abuse

6. INTERPRETATION OF RESULTS OF DATA ANALYSIS

6.1. Predictors of marijuana/hashish abuse obtained by regression analysis

Numerous predictors of marijuana/hashish abuse have been listed in Table 2 and Table 3. It can be concluded that variables that are both *independent* of drug use, as well as those that are *dependent* on drug use affect the chosen criterion, i.e. occurrence of marijuana/hashish use in the last 12 months. There is a great correspondence in these lists of predictors with predictors of substance abuse reported in other studies [2; 6; 12; 18; 22; 36].

The lists of predictors acquired by regression analysis can be utilized to direct the selection of potential predictors for the decision tree technique. Using decision trees on hundreds of potential predictors simultaneously, however, may be impractical, and in addition many specific but important patterns could be neglected.

6.2. Use of decision tree based rules related to marijuana/hashish abuse

The rules that are generated from the decision trees in our examples (or derived by other data mining methods) could be used as some of the precursors to a complex model of substance abuse, as well as part of the predictive tools for identification of the *risk prone* (by using rules R1-R3, R5-R7) and the *risk free* (by using rule R4) segments of adolescent populations. Moreover, such knowledge extracted from survey data could both lead the course, and constitute some of the content of substance abuse prevention. However, an important antecedent to practical application of extracted knowledge should be its evaluation and the testing of models that have been generated by using this new knowledge.

To increase the reliability and validity of models developed by the application of decision tree analyses of our survey data, a larger number of records with more criterion related subjects should be obtained in a survey, and the sample used for research should be representative of more than one district or region. In addition, validation procedures should be performed as well.

6.3. Comparison of results derived from linear regression analysis and the decision tree technique

The lists of predictors obtained by linear regression in Table 2 and Table 3 differ in their information content from the decision tree structures presented in Figure 1 and Figure 2. Even though the lists generated by regression analysis are more complete in the number of attributes affecting the criterion, they represent a trivially simplified structure of the relations between individual predictors and the criterion, with quantification largely dependent on the regression equation. On the other hand, decision tree analysis of the same potential predictor variables has generated patterns which turned out to be less embracing of different attributes, but more precise in its description of pattern structures which represent the relations of attributes to the data.

Evidently, a selection of different subsets of potential predictors for decision tree analysis could generate abundant patterns or concurrent "models" from one basic data set with a large number of variables. Even though these co-existing representations may complicate a search for *one* consolidated model of real world processes from which the data were collected, nevertheless they provide many opportunities for the deliverance of concealed knowledge in the iterative process of its elicitation and verification.

The potential to generate complex rules from decision trees can guide exploratory research activities when large numbers of variables constitute the collected data sets. More than a mere description of the processes can be obtained through models that incite a more detailed research design, simulation, or experimental appraisal.

Since the accurate use of the decision tree technique may require a sufficiently large number of records, numerous subjects or instances of measurement may need to be included in the research data used for analysis. Moreover, when a criterion variable is selected one must assure that enough cases can be related to it and that a suitable distribution of data values is present within the criterion variable. If cases related to the criterion are not numerous, the subdivision of those cases by the nodes of the decision tree could affect rule generation by disabling the reliable generation of complex rules for *critical* "leaves" or terminal nodes which could provide useful information about processes that are inspected (i.e. those terminal nodes would probably contain a statistically insufficient numbers of cases). This problem is enlarged when a proportion of the records is used as a learning sample and the rest as a test sample, such as was the case with our analysis.

7. THE POTENTIAL OF KDD IN SURVEY RESEARCH AND SUBSTANCE ABUSE PREVENTION AND TREATMENT

Researchers in the fields of psychology and sociology often use surveys, questionnaires, tests, inventories, interviews, and other methods to collect numerical and categorical data from hundreds of subjects. Moreover, their acquaintance with the research methodology and statistical methods is usually very close. Still, they often find the process of uncovering patterns of relations among numerous psychosocial variables by standard methods of data analysis to be a complex and exhaustive task. Our demonstration of the potential of the *data mining* technique for the analysis of survey data on substance abuse illustrates that researchers could profit considerably from the use of such methods. Previously collected research data could be *reanalyzed* and future research designs *adapted* so that the format of the acquired input data is more appropriate to the data mining methods that would be used for data analysis. It must be noted that the sound knowledge of statistics and research design considerably improves the capability for proper evaluation and verification of knowledge discovered by the application of data mining methods.

Furthermore, the potentials of KDD can be associated with recent trends in the field of substance abuse related activities, where increased interest has evolved toward a prevention-oriented, proactive approach, as opposed to the treatment-oriented, reactive approach [12]. The cost of health care and the social loss associated with addiction, with ever limited resources for prevention activities, gives motivation for a more effective identification of those most "at risk" of drug abuse and who need to be covered by prevention activities. Presently, more advanced models are sought after in relation to substance abuse, prevention and treatment, as well as for evaluation, especially since some recent studies have disclosed *insignificant outcomes* of some prevention programs [7, 26; 35], as well as the potential to increase the effectiveness and reduce the cost of expensive treatment [28]. As demonstrated in this paper, *data mining methods* could enhance the knowledge of substance abuse related processes and contribute to the development of more pertinent models.

It must also be emphasized that lately much attention has been given to the development of information systems for substance abuse treatment activities [30; 33]. *KDD techniques* have much to offer such systems by the application of *data warehousing* to database design and maintenance, *data mining* to the analysis of collected data, as well as by improving the means by which knowledge is made available for screening purposes, evaluation of treatment, and decision making.

8. STATISTICS AND KDD IN THE ANALYSIS OF SURVEY DATA

The early application of KDD brought forth significant results, but the field is now abounding in problems that arise from the difficulty of model inference from a limited number of observations and managing uncertainty in the results derived through the application of knowledge extraction methods [14].

Statistics has a wealth of technical procedures to offer data miners confronted by such questions, and several simple methodological morals to serve them in their efforts [19]: uncertainty should be used and revealed, and not concealed in the applications and in the use of the results of analyses; the errors of search procedures should be calibrated; the error probabilities of hypothesis tests should not be taken as the error probabilities of search procedures; and it should be proven that estimation and search procedures used in data mining are consistent under conditions thought to realistically appear in applications.

If data mining experts take into account the statistical component of the classification problems for which they use data mining methods, they can further improve the classification performance of their systems [17]. It must be accentuated that the use of KDD in the analysis of survey data does not presume that one can circumvent the standard statistical problems of inference and uncertainty. Naturally, the know-how of statistics should be utilized as well.

When using data mining searches one should consider (a) how many discovered patterns are real and not products of chance fluctuations in databases, (b) how to make valid probability estimates of those patterns, and (c) how many of them are non-trivial, interesting and valuable [21]. Statistics can greatly contribute to the advancement of data mining results in the validation phase of KDD. It is opportune that "new" knowledge and "models" created out of survey data by data mining undergo statistical and empirical verification.

However, it should also be noted that data mining, and the problems this field addresses, offers many opportunities for advancement in statistics, and that both statisticians and data miners can profit by the examination of each other's methods and by using appropriate combinations of them to solve diverse real world problems [24]. One must recognize that *data mining* greatly corresponds to statistical *exploratory data analysis* (EDA) and that *decision trees* are analoguous to *classification trees* in statistics [37].

For an analysis of survey data with numerous variables and a sufficient number of instances or cases, one should combine both statistics and KDD in a beneficial way to surpass the limitations of each class of method in the analysis of a specific data set.

9. CONCLUSION

Data collected in a poll of substance abuse among high school adolescents was used to assess the potential of KDD in an analysis of survey data. We have demonstrated that KDD can be effectively applied to the process of discovering new patterns of knowledge in survey data if a sufficient number of cases constitute the data set. With large numbers of variables or attributes within data sets, KDD methods may offer a convenient solution for discovering the structural relations of the criterion and other variables, as well as for prediction and model development. However, statistical methods may be more appropriate in other areas, i.e. for the verification of discovered knowledge and validation of models developed from that extracted knowledge.

Numerous predictors of marijuana/hashish abuse have been revealed in this research paper, with a demonstration of how decision tree analysis can complement a structure to the lists of predictors generated by linear regression. More in-depth research is to be conducted on this issue, and the structure of abuse of other narcotic substances and alcohol shall be examined as well.

Our results demonstrate that both the diagnosis and the prevention of substance abuse has much to gain from the application of advanced techniques of data analysis. Useful pattern structures may be uncovered from collected data by data mining methods without excessive effort, but, nevertheless, a conscientious and thorough validation of the new knowledge or models that are being developed should be an obligatory step in this type of research and its practical activity.

REFERENCES

[1] J. H. Ahn, K.J. Ezawa. Decision support for real-time telemarketing operations through Bayesian network learning. *Decision Support Systems*, Vol. 21, 1997, pp. 17-21.

[2] S. L. Albrecht, C. Amey, M. K. Miller. Patterns of substance abuse among rural back adolescents. *Journal of Drug Issues*, Vol. 26, No. 4, 1996, pp. 751-781.

[3] F. Ö. G. Ali, W. A. Wallace. Bridging the gap between business objectives and parameters of data mining algorithms. *Decision Support Systems*, Vol. 21, 1997, pp. 3-15.

[4] S. S. Anand, A. R. Patrick, J. G. Huges, D. A. Bell. A data mining methodology for cross-sales. *Knowledge-Based Systems*, Vol. 10, 1998, pp. 449-461.

[5] S. T. Berg, L. M. Blotzer. New "weapon" enters war on drugs. *OR/MS Today*, Vol. 21, No. 3, 1994, pp. 40-45.

[6] G. Bhattacharya. Drug use among Asian-Indian adolescents: identifying protective/risk factors. *Adolescence*, Vol. 33, No. 129, 1998, pp. 169-184.

[7] J. H. Brown, I. G. G. Kreft. Zero effects of drug prevention programs: issues and solutions. *Evaluation Review*, Vol. 22, No. 1, 1998, pp. 3-14.

[8] M. C. Burl, L. Asker, P. Smyth, U. Fayyad, P. Perona, L. Crumpler, J. Aubele. Learning to recognize volcanoes on Venus. *Machine Learning*, Vol. 30, 1998, pp. 165-194.

[9] J. P. Caulkins. Dealing with the country's drug problem. *OR/MS Today*, Vol. 22, No. 1, 1995, pp. 32-40.

[10] W. Daelemans, P. Berck, S. Gillis. Data mining as a method for linguistic analysis - Dutch diminutives. *Folia Linguistica*, Vol. 31, 1997, pp. 57-75.

[11] *Data Mining News*. Intelligent Data Analysis Group *(IDAG)*, Vol. 1, No. 16, 1998.

[12] K. T. Doherty, L. B. Szalay. Statistical risk versus psychological vulnerabiliy: why are men at greater risk for substance abuse than women? *Journal of Alcohol and Drug Education*, Vol. 42, No. 1, 1996, pp. 57-77.

[13] P. W. Eklund, S. D. Kirkby, A. Salim. Data mining and soil salinity analysis. *International Journal of Geographical Information Science*, Vol. 12, 1998, pp. 247-268.

[14] U. Fayyad. Diving into databases. *Database Programming & Design*, Vol. 11, No. 3, 1998, pp. 24-31.

[15] U. Fayyad, D. Haussler, P. Stolorz. Mining scientific data. *Communications of the ACM*, Vol. 39, No. 11, 1996, pp. 51-57.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, Vol. 17, No. 3, 1996, pp. 37-54.

[17] J. H. Friedman. On bias, variance, 0/1 – loss, and the course-of-dimensionality. *Data Mining and Knowledge Discovery*, Vol. 1, 1997, pp. 55-77.

[18] E. R. Galaif, C. P. Chou, S. Sussman, C. W. Dent. Depression, suicidal ideation, and substance use among continuation high school students. *Journal of Youth and Adolescence*, Vol. 27, No. 3, 1998, pp. 275-299, 1998.

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

[20] C. Glymour, D. Madigan, D. Pregibon, P. Smyth, P. Statistical inference and data mining. *Communications of the ACM*, Vol. 39, No. 11, 1996, pp. 35-41.

[21] D. J. Hand. Statistics and more? *The American Statistician*, Vol. 52, 1998, pp. 112-118.

[22] P. A. Harrison, J. A. Fulkerson, T. J. Beebe. DSM-IV substance use disorder criteria for adolescents: a critical examination on a statewide school survey. *The American Journal of Psychiatry*, Vol. 155, No. 4, 1998, pp. 486-492.

[23] B. Hibel, B. Andersson, T. Bjarnason, A. Kokkevi, M. Morgan, A. Narusk, *The 1995 ESPAD Report: Alcohol and Other Drug Use Among Students In 26 European Countries*. The Swedish Council for Information on Alcohol and Other Drugs, CAN, Stockholm, 1997.

[24] J. R. M. Hosking, E. P. D. Pednault, M. Sudan. *A Statistical Perspective on Data Mining*. Research report RC 20856 (92315), IBM Research Division, T.J. Watson Research Center, Yorktown Heights, NY, 1997.

[25] W. H. Inmon. The data warehouse and data mining. *Communications of the ACM*, Vol. 39, No. 11, 1996, pp. 49-50.

[26] I. G. G. Kreft. An illustration of item homogeneity scaling and multilevel techniques in the evaluation of drug prevention programs. *Evaluation Review*, Vol. 22, No. 1, 1998, pp. 46-77.

[27] H. Y. Lee, H. Lu, H. Motoda. Knowledge discovery and data mining. *Knowledge-Based Systems*, Vol. 10, 1998, pp. 401-402.

[28] C. G. Leukefeld, T. K. Logan, S. S. Martin, R. T. Purvis, D. Farabee. A health service use framework for drug abusing offenders. *The American Behavioral Scientist*, Vol. 41, No. 8, 1998, pp. 1123-1135.

[29] T. C. Lin, M. Pourahmadi. Nonparametric and non-linear models and data mining in time series - a case-study on the Canadian lynx data. *Applied Statistics - Journal of the Royal Statistical Society*, Series C., Vol. 47, Part 2, 1998, pp. 187-201.

[30] D. McCarty, T. G. McGuire, H. J. Harwood, T. Field. Using state information systems for drug abuse services research. *The American Behavioral Scientist*, Vol. 41, 1998, pp. 1090-1106.

[31] J. M. McDonald, S. Brossette, S. A. Moser. Pathology information systems - data mining leads to knowledge discovery. *Archives of Pathology & Laboratory Medicine*, Vol. 122, 1998, pp. 409-411.

[32] L. Morrill. Tools for mining. *Database Programming & Design*, Vol. 11, No. 4, 1998, pp. 62-64.

[33] A. C. Ogborne, K. Braun, B. R. Rush. Developing an integrated information system for specialized addiction treatement agencies. *The Journal of Behavioral Health Services & Research*, Vol. 25, No. 1, 1998, pp. 100-107.

[34] *An introduction to data mining: discovering hidden value in your data warehouse*. White paper, Pilot Software, 1998. Web address: http://www.pilotsw.com/dmpaper/dmindex.htm

[35] D. Rindskopf, L. Saxe. Zero effects in substance abuse programs: avoiding false positives and false negatives in the evaluation of community-based programs. *Evaluation Review*, Vol. 22, No. 1, 1998, pp. 78-94.

[36] L. M. Scheier, G. J. Botwin. Relations of social skills, personal competence, and adolescent alcohol use: a developmental study. *Journal of Early Adolescence*, Vol. 18, No. 1, 1998, pp. 77-114.

[37] StatSoft, Inc. (1997). *Electronic Statistics Textbook*. StatSoft, Inc., Tulsa, OK, 1997. Web address: http://www.statsoft.com/textbook/stathome.html

[38] E. Thomsen. Smart decision-support systems. *Database Programming & Design*, Vol. 11, No. 6, 1998, pp. 59-61.

Received: 8 July 1998
Accepted: 26 October 1998

Goran Bubaš
Božidar Kliček
Željka Čolović Rodik
Zvjezdana Fulir

PRIMJENA TEHNIKA OTKRIVANJA ZNANJA U BAZAMA PODATAKA ZA ANALIZU ANKETNIH PODATAKA O ZLOPORABI OPOJNIH SUPSTANCI

Sažetak

Otkrivanje znanja u bazama podataka (knowledge discovery in databases, KDD) razmatra se s obzirom na potencijale za analizu anketnih podataka o zloporabi opojnih tvari. Anketom su prikupljeni podaci za 238 varijabli od 4198 adolescenata, učenika srednjih škola. Kao jedina kriterijska varijabla izabrana je zloporaba marihuane ili hašiša, dok je 77 varijabli izabrano za prediktore. Analize podataka izvršene su uporabom dviju metoda: linearnom regresijom i generiranjem stabla odlučivanja. Nakon prezentiranja rezultata analiza podataka razmatrane su i uspoređene obje korištene metode. Zaključeno je da KDD tehnike imaju znatne mogućnosti za otkrivanje novih struktura i veza između varijabli sadržanih u anketnim podacima, za razvoj modela zloporabe opojnih tvari te za aktivnosti prevencije. Također je naglašena važnost statističke provjere otkrivenog znanja primjenom KDD tehnika.

Ključne riječi: otkrivanje znanja u bazama podataka, rudarenje podataka, zloporaba opojnih tvari, anketiranje.