From: AAAI Technical Report WS-94-03. Compilation copyright © 1994, AAAI (www.aaai.org). All rights reserved.

Exploration of Simulation Experiments by Discovery

Willi Klösgen kloesgen@gmd.de
German National Research Center for Computer Science (GMD)
53757 Sankt Augustin, Germany

Abstract: We exemplify in this paper, how a discovery system is applied to the analysis of simulation experiments in practical political planning, and show what kind of new knowledge can be discovered in an application area that differs from others by the high amount of knowledge that the analyst holds already about the process that generates the data. Subgoals like "low classification accuracy", "high homogeneity", "disjoint rules", etc. are introduced into Explora, to select between different statistical tests for each pattern and several search algorithms, allowing the user to adapt the discovery process to the special requirements of the application.

The combination of discovery with simulation is endowed with the main characteristics of both Knowledge Discovery in Databases (KDD) and Automated Scientific Discovery (ASD), i.e. discovery in large databases and experimentation. Analysing a real system with simulation models allows to freely set the experimental conditions. In distinction to the usual KDD assumption of fixed given data, when combining discovery with simulation, additional data can be generated by running new simulations according to the needs of the discovery component.

First, we identify four tasks relevant for exploring simulation experiments which can be supported by discovery methods. Then we describe the application of socio-economic modeling for political planning. Third, we demonstrate for a simple law (financial support of families with children), how the current state of Explora is used for this application. Finally, we discuss some new approaches of Explora to deal with subgoals and outline further work.

Keywords: Knowledge Discovery in Databases, Automated Scientific Discovery, Simulation, Political Planning, EXPLORA

1. Introduction: Machine Discovery and Simulation

Machine Discovery deals with methods and software systems to support knowledge discovery processes. A discovery process aims at finding out new knowledge in an application domain of science or practice by searching in hypotheses spaces. The major directions in Machine Discovery are Knowledge Discovery in Databases (KDD) and Automated Scientific Discovery (ASD). KDD centres around discovery processes in given databases of practice. ASD deals with scientific comprehension by performing and analysing experiments.

A discovery process in ASD may feedback to data generation by claiming further domain information (especially data) to improve the quality and scope of the generated knowledge. Further experiments in a scientific laboratory may be necessary to generate the requested data. In many application areas, experiments with the real system to be investigated are not possible, but computerized models representing the real system are available. Simulation is used to study the behaviour of the real system, approximated by the model. Data about this behaviour is generated by running simulation experiments.

The user of a simulation model can control the experimental conditions and define a simulation *variant* by setting values for the input variables and parameters. Two main questions can be answered by simulation experiments. What happens (what values do the output variables take), if a variant holds given predefined values? And: What has to be done (how must the values of a variant be selected) to achieve a desired output of the model?

A first analysis task in simulation refers to studying the results of a single variant (or factor combination). Predefined factors are given, and the influence of these factors on the output variables of the model have to be analysed and described. For instance, the consequences of a possible decision must be calculated. Then discovery methods can induce relations between input and output variables for the given variant which e.g. can be used as a pragmatical summary of the various detailed relations represented in the model.

The second task is the comparison of two or more variants. Differences between output states belonging to the alternative factor combinations have to be derived. E.g., predictions under alternative conditions must be made, or selections between several already known decision alternatives shall be supported. In this case, discovery methods can derive significant differences between the variants for an output variable in subdomains of input variables.

The next problem is the analysis of a whole space of variants to derive general functional relations between factors and output variables. These relations can clarify the process which is represented by the model. Also they can be used to plan decision alternatives.

A fourth task deals with optimization or goal state experiments. Combinations of factors have to be discovered which produce an optimal output. Often conditions for an aspired output state are given and a variant has to be found which achieves this state.

Assuming that machine discovery can support these simulation tasks, the discovery system Explora (Klösgen 1994) is introduced in applications of German government to support the analyses of simulation results in the area of political planning. In the next section, we give a short overview about this application area.

2. Socio-economic simulation models and political planning

The two main application fields of simulation are physical-technical and socio-economic systems. Socio-economic models are used in government agencies to plan legislation for taxation, transfer subsidies, health care, social security, etc. (Orcutt et al. 1986). Some years ago, we developed the Model Base System MBS (Klösgen et al. 1983) which is widely used in German government to support the application of three types of socio-economic models (macro-, cohort-, and micro-models; compare: Klösgen & Quinke 1985).

In this paper, we concentrate on the application of discovery for microanalytic models (Klösgen 1986). Especially for these models, the volume of data is so large that a "manual" analysis of simulation results is necessarily very restrained and discovery approaches can offer a more complete and systematic analysis approach. The "manual" analysis is supported in MBS by standard queries with reports and some analyses (e.g. cross tabulations).

Microanalytic models operate on large sample data. This application area uses a sample of about 50,000 households in Germany. Each household record has some 200 socio-economic input variables and a weight variable indicating the representativity of the household within total population in Germany. Output variables are calculated by a model, depend on the parameters belonging to a variant, and are repetetive over a sequence of variants.

The microanalytic models represent various tax and transfer laws (income tax, subsidies for households with children, subsidies for students, etc.). In a single simulation experiment, the model representing the law is applied for a given variant of the law to each household. A variant includes a value for each of the parameters of the law. Output variables are derived by the model calculating output values for each household using input values and parameters. The main questions which are studied with these models are the determination of the total cost of a variant and the winners and losers of a planned legislative measure.

3. Connecting Explora to simulation

Explora (Klösgen 1993) is a discovery system for Apple-Macintosh™ which is freely available by "anonymous ftp"[1]. For this application, the system has to be linked to the Model Base System

[1] Connect to *ftp.gmd.de* (or 129.26.8.90) and transfer *Explora.sit.hqx* from the directory *gmd/explora. README* informs about the installation. The "end user" version of Explora is available for practical applications on medium sized data bases (up to 100,000 records). It is distributed as a stand alone program in object code and stripped of the access to LISP, not requiring a Macintosh Common LISP (MCL)™ license.

MBS which is running on mainframes and was just transferred to an UNIX environment. The first applications are based on a loose coupling of the two systems. At first, the households records of a MBS-microsimulation including the results of several variants for one simulation model are selected and file transferred to the Macintosh environment.

Explora relies on an internal data approach which is based on an inverted list data organization to achieve efficient discovery (Klösgen 1994). This variable-oriented data structure gives the preferences to the efficiency requests of discovery, whereas the original MBS record-oriented data organization follows simulation requests. In a second step, the Explora user calls an MBS-import menu option to initiate the conversion of the microdata into the Explora data structures. Explora manages and accesses data via a hierarchy of applications, segments, and variables. Here, an Explora application is associated to a simulation model and a segment to a variant. An input segment stores the input variables of a model which are variant-independent to avoid data redundancy in Explora. Variables can be imported separately and further variants can be added.

Before starting discovery in simulation results, the Explora user chooses goals and subgoals which determine patterns, their statistical tests, and search strategies. Then variants, (in)dependent, and weight variables are selected in the Explora windows (Klösgen 1994).

Extensions were necessary to adapt Explora to weighted cases. Further extensions refer to the introduction of subgoals (see section 6). The statistical tests applied in the verification methods of patterns to calculate the significance of an instance of a pattern (hypothesis) have to consider weights. Several statistical approaches are possible. Two extreme solutions are offered: The weighted value of a variable can be seen as a mean value in a group of cases (the weight indicates the number of cases in the group) or as a random realization of one case in this group. Although the order of magnitude of the statistical significance of a hypothesis differs considerably for these two alternatives, the results produced by the refinement algorithm of Explora (Gebhardt 1991) are very similar. This is due to the fact that refinement in Explora is invariant to linear transformations of the evidence value of a hypothesis.

Another extension refers to the bit-vector approach used in Explora to calculate the logical operations and the number of cases in a group which is defined by a logical expression. Now, to sum the number of cases of a group it is not sufficient to count the number of bits which are set to 1, but one has to sum the weights for these cases. When substituting the simple function which does this calculation, run time is considerably increased, because this is the most called function during the search process in the "inner loop".

So, the run times presented for some application cases of Explora in (Klösgen 1994) would rise enormously, because of the large data amount and the more time consuming computation. Fortunately, we can profit by the technological progress in hardware. Already substituting a Macintosh IIx by a Macintosh Quadra allows to do these discovery runs still in dialog. Also large databases can still be handled by extending the external and internal storage capacity; e.g. 36 MB of main memory are used for these applications of Explora.

In the next sections, we refer to one comparatively simple law which regulates the financial support of households with children. We address here only one aspect of this law, i.e. the payment of subsidies for children (a negative income tax regulation and an amount free of tax are further aspects of equalization of burdens for families with children). These simple regulations were chosen to limit the substantial complexity of the examples.

4. Analysis of a single variant

In this chapter, we analyse a single variant. The following questions shall be answered:

- Who gets subsidies (or no subsidies)?
- Who gets how much subsidies?
- To which groups the total amount of subsidies is distributed, i.e. which are the dominating key groups?

The first question is important additionally to the second one, because for (political) consciousness of people it sometimes plays a role to get a support, independent from the level of the support. Of course, also other dichotomizations than 4.1 are possible and useful.

In the examples of section 4, we use a database with simulation results for one selected variant of the law regulating subsidies for the 25,000 households with children (within the representative sample of 50,000 households). The dependent variable quotes the calculated amount of money paid to a household with children. To the independent variables belong taxable income, number of children, family status, and many others.

Discovery can be applied to a vast spectrum of applications. One dimension which can be used to classify this spectrum is the amount of knowledge about the application domain which the analyst holds already. On the one end of this dimension lies an application where the user has no knowledge at all about the process that generates the data. This simulation example lies on the other end of this dimension. The process that generates the data is known already in detail. This process is given by the law, resp. the detailed representation of the law in the model. The three questions of this chapter are now treated to demonstrate what kind of new knowledge can be derived in this situation by discovery methods.

4.1 Who gets subsidies?

The parameters set in the definition of the variant treated as an example in this section can be transformed by elemental calculations (without using the simulation model) and compiled in the form of table 1. This table shows the boundaries of the taxable income up to which subsidies to families with children are paid. The boundaries depend on the number of children and the family status. This table gives already one possible answer to the posed question which can be stated e.g. in the form of 20 rules, for instance:

If 1 child and single and taxable income < 18800, then subsidies.
If 1 child and single and taxable income >= 18800, then no subsidies.

Table 1
Variant 94; boundaries of taxable income for subsidies

	single	married
1 child	18800	29800
2 children	30200	41600
3 children	53600	65000
4 children	77000	88400
5 and more children	98400	111800

These rules are not known within the discovery system, because the system exploits only the simulation results. To the discovery patterns which can be used to treat the above questions, belong several types of rules and classification trees. It is necessary to further detail the discovery goals to fix an appropriate pattern. In this situation, one analysis subgoal (comp. section 6) is, to uncover the dominant structure of the problem and to derive conditions on the independent variables and their interactions that drive the dependency relation. The goal is not, to produce an accurate classifier, because the user knows already this classifier (by table 1).

Therefore, we use a probabilistic rule pattern for this question. The results discovered by Explora are given in table 2. The 20 rules of table 1 are reduced to 2 rules in table 2 (resp. 4 or 10 rules on the next levels of detail). This wanted generalization is due to the statistical test used for the probabilistic rule pattern. No exact rules are discovered, but it is sufficient, if the group percentage differs significantly from population percentage (see section 6). Explora here applies an exhaustive search strategy and by treating overlappings due to multicorrelations between independent variables in a refinement phase, it is achieved that heterogeneities in subgroups are discovered. Also, the refinement algorithm succeeds in identifying the correct independent variables by suppressing overlappings (compare tables 2 and 3). The probabilistic rule pattern and the refinement algorithm of Explora are described in more detail in (Klösgen 94 and Klösgen 92).

Other rule patterns (strong rules) and statistical tests for rules (see section 6) were applied in Explora, e.g. to analyse the capability of a discovery system to re-identify the law which in this simple case is already given by table 1. The available tests include evaluations which are used in classification trees (Breiman et al. 1984) such as an information theoretic criterion (compare also: Quinlan 1986), the CART and the Gini criterion (Breiman et al. 1984), and the CN2 criterion (Clark & Niblett 1989). Because these criteria are oriented to more classification accuracy, the amount of generalization of the rules of table 1 is different. Special problems in re-identification are due to missing simulation results for some rare events (e.g. children >3, single, taxable income > 60,000). Therefore, to derive exactly the boundaries given in table 1, it would be necessary to install a feedback from discovery to simulation. Using methods of ASD (Shrager and Langley 1990), a discovery component generates a simulation experiment which derives the requested data.

Table 2

```
Microsimulation: Subsidies for children
Analysis of a single variant: V94
Problem: Who gets (no) money?
Pattern: probabilistic rules; refinement of results
Low classification accuracy, low overlappings, exhaustive search

Households: West Germany, with children, 94

55% of these households get no money. These are:
     95% of TaxableIncome > 40000

     Exceptions (5%) with money:
        Children > 2, TaxableIncome 40000 - 80000 (96%)

          in detail:
             Children = 3, TaxableIncome 40000 -  62000   (99%)
             Children = 4, TaxableIncome 40000 -  86000   (99%)
             Children > 4, TaxableIncome 40000 - 116000  (100%)

45% get money. These are:
     83% of TaxableIncome < 40000

     Exceptions (15%) without money:

        Children = 1, TaxableIncome 29000 - 40000 (98%)

          in detail:
             single,  Children = 1, TaxableIncome > 18000 (99%)
             married, Children = 1, TaxableIncome > 29000 (99%)
             single,  Children = 2, TaxableIncome > 29000 (99%)
```

These results clarify that the variable TaxableIncome is the most important one for getting money (dichotomy; amount of money: compare 4.2) assuming this variant. Nearly all households with children and a taxable income higher than 40,000 get no money. The next important variable is then the number of children which characterizes the exceptions or counterpoints. Family status is important only for a further refinement.

A diagram representing the groups, their sizes and inclusion relationships can support this textual form and illustrate the results more evidently. The information in the tables 1 and 2 overlaps. On the one hand, table 2 generalizes some unimportant cases (singles with more than 2 children and high income do not appear, because this group is too unimportant considering its size). On the other hand, table 2 contains more information than table 1. This is due to the analysis of the joined distributions (e.g. income and children) represented in the simulation results. E.g., to derive the percentages for subgroups using the weights of the cases, the analysis of the simulation results is necessary.

Benefit of discovery: Discovery can derive additional interesting knowledge about the variant: A system of rules (given law) is generalized considering the importance of the rules, and distributional information is also used to add information about the size of the subgroups. The practical relevance of this additional value of discovery depends on the complexity of the law. For more complex tax and transfer laws, it is not possible to list the various subcases of the law in such a simple way as in table 1. The probabilistic rule pattern can then be used to uncover the dominant structure of the dependency.

Another advantage of using Explora can be demonstrated with the results of table 3. The refinement algorithm of Explora eliminates overlappings and identifies the determinative variables. An option is available to uncover such overlappings. Because of correlations between the independent variables, other groups which have a high overlapping with the "primary" groups are also affected by the law. The discovery of these secondary effects is important for the political discussion, especially in the case when special political target groups (e.g. civil servants, employers, employees, pensioners) are affected.

Table 3

```
Microsimulation: Subsidies for children
Analysis of a single variant: V94
Problem: Who gets (no) money?
Pattern: probabilistic rules, overlappings

Households: West Germany, with children, 94

55% of these households get no money. These are:
     80% of spouse is employee
     89% of spouse is civil servant
     63% of civil servant, spouse not employed
     60% of civil servant, single
     ...
45% get money. These are:
     80% of employers
     68% of singles
     52% of spouse not employed
      ...
```

4.2 Who gets how much subsidies?

This question refers to functional dependencies between output and input variables. In some discovery systems (e.g. Forty Niner, Zytkow 1993), patterns are available to discover the existence of a functional dependency and the equation connecting the variables. Other patterns to treat this problem include regression trees (Breiman et al. 1984) and mean-patterns.

Regression trees or mean patterns identify groups which are homogeneous or have significantly larger or smaller means (relating to the dependent variable). If the functional dependencies are known, one can derive such subgroups by interpreting the equations. Regression trees and mean patterns derive such conclusions directly without to determine the equations.

The functional dependencies which characterize the variant used as an example in section 4 are also known by the law regulations and can be compiled in the form of table 1. However, each entry of this new table now holds a linear (stairs-) function. E.g., singles with 1 child get 135 DM monthly subsidies up to 11,000 DM yearly taxable income. The subsidies are decreased by 10 DM in steps of 600 DM taxable income. Law regulations incorporate typical nonhomogeneities: different relations hold between variables in different parts of the population (here: a system of 10 functions).

Again, the subgoal of this discovery problem is, to uncover the dominant interactions of variables and not, to re-identify the already known system of equations. The statistical test used for a pattern and the search approach must be chosen in accordance with this subgoal. A mean pattern of Explora (Klösgen 1992) treats this question. Subgroups are discovered with an over- (under-) proportional mean.

The results of table 4 were derived using a statistical mean test as evaluation for the significance of a subgroup. More homogeneous subgroups, but also more (and smaller) groups are discovered, if the variance criterion of regression trees (Breiman et al. 1984) is used.

Table 4

```
Microsimulation: Subsidies for children
Analysis of a single variant: V94
Problem: Who gets how much money?
Pattern: comparison of means
Medium homogeneity, low overlapping, exhaustive search

Households: West Germany, with children, 94

These households get on av. 148 DM. More get:
    Children   >= 5, TaxableIncome < 60000             1652
    Children   >= 5, TaxableIncome < 96000             1597
    Children    = 4, TaxableIncome < 48000             1052
    Children    = 4, TaxableIncome < 72000              994
    Children    = 3, TaxableIncome < 42000              650
    Children    = 2, TaxableIncome < 22000              335

Less subsidies for children get:
    TaxableIncome >= 40000                               15
    TaxableIncome >= 30000                               33
    Children    = 1, TaxableIncome >= 22000               4
    Children    = 1                                      36
    single, Children = 2, TaxableIncome 24000 - 30000    71
```

Another factor related to discovery subgoals is the search approach. Classification and regression trees use an unrevocable search approach which is one-step optimal. Inherent to this approach is also the partition of the population into disjoint subgroups. As already in case of the dichotomy pattern (4.1), this leads to a more accurate re-identification suffering however by large trees with many subgroups (findings).

A beam search is applied by the CN2 method (Clark & Niblett 1989). Overlappings, which may occur due to the beam search, are avoided by eliminating all cases which are covered by a found rule for the next rule identification step.

In Explora, an exhaustive search approach can be applied to identify an optimal set of findings, if the size of the search space can be restricted (otherwise also Explora uses a stepwise approach). The search space can be restricted in these applications by limiting the order of conjunctions, e.g. by including only conjunctions of maximally 3 selectors. The overlappings of the subgroups which are necessarily available in an exhaustive search are reduced in a refinement phase (compare also: Gebhardt 1994). A more detailed comparison of the various search approaches and evaluation criteria is beyond the scope of this paper.

Problems in table 4 still occur with some cumulatings of continuous variables. So, it would be preferable, if the refinement algorithm could discard e.g. the subgroup "Children >= 5, TaxableIncome < 96000" and identify the subgroup "Children >= 5, TaxableIncome 60000 - 96000" to stress the degression against "Children >= 5, TaxableIncome < 60000" more directly.

Benefit of discovery: The goal of problem 4.2 is, to get a complete overview about the values a dependent variable takes in a population by listing a small number of homogeneous (referring to the dependent variable), large subgroups. Of course, there are trade-offs between the subgoals (high degree of completeness, small number of subgroups, large size of subgoups, high degree of homogeneity). Preferences for these subgoals must be given by the user which can be used to select appropriate evaluation criteria and search approaches.

For this application, the additional value of applying discovery approaches is determined by the achievement of these subgoals. The discovery results have to be compared with the elementary alternative which consists of using the model or the known system of functional equations to calculate the values for some typical individual cases. The additional value is the higher the more independent variables are involved.

4.3 Which groups get a large part of the total subsidies?

This question relates to the breakdown of total subsidies to key subgroups (compare the KeFir system, Piatetsky-Shapiro & Matheus 1993) and is particularly important in today's political environment which is often characterized by a need to save transfers and to increase taxes. Here the size of the groups is still more dominant than in 4.2. Subgroups with a high average value that undergo a deduction may be too small to amount to the aspired savings.

In the statistical test of the mean pattern (4.2), the size of a subgroup is already considered by evaluating the significance of an outstanding mean as proportional to the square root of the weighted size of the subgroup. This exponent (0.5) can be further increased to strengthen the evidence of large subgroups. An exponent equal to 1 corresponds to the share of the cumulated value of the dependent variable for the subgroup related to the cumulated value in the population. But this cumulated value alone is not appropriate as an evidence criterion for a subgroup. The total population would have the highest value.

To eliminate this problem, Explora introduces further conditions which must be valid for an "interesting" subgroup. A first criterion relates the share of the subgroup in the total population measured in (weighted) cases to the above share by a factor (this is equivalent to the comparison of means) and a second condition sets a floor for the minimum share.

A cumulating pattern in Explora offers three parameters which relate to subgoals which can be stressed within this pattern: to get large groups (exponent), groups with high or small mean (factor for the means), and high cumulating (minimum share). The results in table 5 were discovered with an exponent 1, factor 1.2, and minimum share of 30%, resp. 10%.

Benefit of discovery: The results of table 5 can not be derived directly by knowing the regulations of the law, because they rely on distributional information which is only available in the simulation results. Explora searches systematically in a space of subgroups built by all combinations of (selected) independent variables including intervals for continuous variables and taxonomies for nominal variables and applies a refinement algorithm. In general, this systematic approach produces better results than a manual approach relying on user defined cross tabulations.

Table 5

```
Microsimulation: Subsidies for children
Analysis of a single variant: V94
Problem: Distribution of total subsidies to key groups
Pattern: cumulating
Medium cumulating, low overlapping, exhaustive search

Households: West Germany, with children, 94

More than 30% of total subsidies for the groups:
    Children >= 2                   48% of households get 84%
    Children >= 3                                 14% get 54%
    Children >= 3, TaxableIncome < 18000           5% get 31%
    Children >= 3, TaxableIncome < 42000           9% get 50%
    Children = 3                                  10% get 30%
    TaxableIncome < 48000                         62% get 98%
    TaxableIncome < 0                             18% get 42%

Less than 10% of total subsidies for the groups:
    TaxableIncome >= 36000                        52% get  8%
    Children = 1, married                         41% get  8%
    Children = 1, single                          11% get  8%
    Children = 1, TaxableIncome > 0               43% get  7%
    Children = 2, TaxableIncome > 18000           26% get 10%
```

5. Analysis of two or *k* variants

In this chapter, we introduce a second variant V93 with is characterized by the following regulations: The subsidies are oriented to the total income of a household (and not to the taxable income), the parameters of the linear functions for the calculation of subsidies have other values (e.g. the subsidies are decreased only down to a minimum and not down to 0), and there are additional amounts free of tax for households with children.

We treat here only the case of comparing two variants. Again, the following types of questions are interesting:

- Which groups hold a larger share of supported households under the first variant?
- Who gets significantly more money under the first variant?
- Which groups get a larger share of total subsidies under the first variant?

To save space, we show only one result. In general, the profit of discovery methods is larger than in analysing one variant, because it is more difficult to assess the overlapping effects of both variants by elementary calculations.

We assume two populations for the statistical approach of the comparison patterns. In special cases, it would be possible also to refer to one population and to analyse e.g. the difference of a dependent variable for the two variants. But in general, the weights can be different for the variants.

By investigating the inverse share, we get results (table 6) on deviations in the composition of the supported households. With these variants, these are the key groups with no support under V94 and a (strong) support under V93.

Table 6

```
Microsimulation: Subsidies for children
Comparison of 2 variants: V93 versus V94
Problem: Different structures of supported households
Pattern: dichotomy (inverse), 2 populations
Low separation, low overlappings, exhaustive search

Households: West Germany, with children; comparison V93 vs. V94

Under V93 more supported families belong to the groups:
    TaxableIncome >= 72000                   18% vs. 0%
    TaxableIncome 48000 - 72000              20% vs. 3%
    Children = 1, married                    41% vs. 2%
```

6. Methodical extensions

This application relies on the basic Explora approaches (Klösgen 1992, 1994) of embedding different patterns into a general search algorithm and refining a first brute force search by an evaluation of further interestingness criteria. Due to space limitations, we can here describe the recent methodical extensions of these approaches only very concisely.

A pattern is determined by the system by considering the analysis goals specified by the user (analysis of one variant of the law or comparison of 2 or more alternative regulations of the law, dichotomous, discrete or continuous problem, etc.). Extending the previous Explora approch, additionally also subgoal specifications of the user are examined by the system to determine the statistical test in the verification method of the selected pattern. As one example for these subgoal interpretations, we discuss here the rule pattern (goals: one variant, dichotomous problem) and the subgoal "low classification accuracy".

- Typically, the user gives a qualitative specification of a subgoal (e.g. low, medium, high classification accuracy) and the system transforms this qualitative direction into a quantitative numerical parameter (e.g. an exponent in a formula). By tuning a slider (e.g. "still higher classification accuracy"), the user can modify his previous specification, if he is not yet satisfied with the discovery results presented by the system. Of course, he can also set a special value of the parameter, if he knows about its theoretical foundation.

The statistical evaluation of a rule is done in a two dimensional p-q space, with p as the probability of the rule, i.e. p = P(RHS | LHS), and q as the relative frequency of the conditional part (concept), i.e. q = P(LHS). Let p_0 = P(RHS) the relative frequency of the conclusion which is fixed since we regard all rules with a given right hand side (RHS). The admissible section in that p-q space (due to constraints) and the isolines of the following evalation functions cannot be discussed here due to space limitations.

The following evaluation functions are associated in Explora to the three qualitative values:

(6.1) $\sqrt{q}\,(p - p_0)$ (high accuracy)

(6.2) $(q\,/\,(1-q))\,(p - p_0)^2$ (medium accuracy)

(6.3) $q\,(p - p_0)$ (low accuracy)

If the user wants a still higher accuracy, the exponent in 6.1 is further decreased. In the example belonging to the rules of table 2, accurate classification rules (table 1) can be re-identified with an exponent 1/4 (fourth root of q in 6.1).

6.1 is equivalent to the binomial test (used previously as only criterium for probabilistic rules in Explora). One can show (proof omitted here), that 6.2 is equivalent to the criteria of CART, Gini index, chi-2 test for 2X2 contingency tables, and INFERULE (Uthurusamy et al. 1991) criterium for 2 classes (all these criteria are equivalent). 6.3 is equivalent to a criterium that was introduced by Piatetsky -Shapiro (1991) as the simplest criterium satisfying some basic principles.

Equivalence is defined related to allowed transformations on the evaluation functions. The above equivalence refers to possible multiplications with constant factors (constant with respect to p and q). This kind of equivalence is important for the refinement algorithm of Explora which is invariant to such multiplications.

Another equivalence can be defined with respect to maximum preserving transformations. In case of disjoint rules (another subgoal is related to overlapping of rules), maximum preserving transformations of evaluations are allowed, because in this mode, the simplest strategy is to iteratively select the rule with the maximal evaluation, discard all overlapping rules, and select the next maximal rule. This strategy is applied in a similar way by the CN2-algorithm. Subgoals related to search strategies can not be discussed here in more detail due to space limitations.

7. Further simulation tasks

The preceding examples relate to the first two simulation tasks introduced in section 1 as a potential application area of discovery methods. The other two simulation tasks require a closer coupling of simulation and discovery components. Then however, the computing efforts will increase considerably, so that these tasks will not run in dialog with the presently used systems or will need another hardware base.

The next task would be to derive for parts of the population the dependencies between parameters in a factor space and output variables. A three dimensional factor space can e.g. be analysed for the above example. The dimensions of this space are given by the parameters "subsidies amount per household", "income where reduction of subsidies starts" and "slope of reduction". A dependency between the output variable "total subsidies" and the three factors can be discovered in parts of the population (e.g. parts defined by combinations of "family status" and "number of children"). Discovery methods which find equations (e.g. Zytkow 1993) can be applied which may feedback to simulation when generating new points in the factor space and run the simulation model for these factors to derive the output variable. Because components of these equations are known by studying the kind of relationship, an equation finder should already rely on these components.

The derived functions can also be used for goal state experiments. Given e.g. a predetermined total amount of subsidies and its aspired distribution to some population parts, factor combinations have to be derived using these equations which approximately result in the desired goal state. The model can then be run iteratively to optimize this approximation. The consequences of the identified solution, or in case of several possible solutions, the differences between the corresponding variants, are analysed as described in sections 4 and 5.

Identification of a law by analysing examples is a primary goal in Machine Learning. To exactly reconstruct laws in such practical areas like tax and transfer legislation from data seems to be a tremendous challenge for Machine Learning approaches, because of the heterogeneities and various exceptions in these laws. Therefore, this application area is an interesting domain also from a theoretical point of view. For practical applications, the re-identification problem is interesting to serve as a test area for discovery methods and systems, to study the performance and e.g. to confirm that these methods identify the correct variables.

Conclusions

The current state of a combined simulation and discovery system offers to the user the choice between 18 problems determined by three goals: number of variants to be analysed (1, 2, k>2), qualitative (dichotomy or other discretization) or quantitative (continuous) dependency, and significant subgroups or distribution into key groups. According to the 18 possible combinations of these goals, a pattern is selected in Explora for a discovery process.

Subgoals determine the criterion which will be applied by the verification method of the pattern to evaluate the evidence of a hypothesis. Some subgoals relate to the homogeneity, size, number, and strength of the subgroups to be discovered. Other subgoals refer to the accuracy of the findings by fixing e.g. the granularity of the intervals built for continuous variables or other aspects of the language used to construct subgroups, and the search strategy (exhaustive or stepwise). The degree of overlapping or the disjointness, and the focus on separation (e.g. minimal classification error) or key structure are fixed by further subgoals.

First practical applications for legislative planning show that discovery methods can constitute a valuable approach also in an area where the analyst has already a lot of knowledge on the domain. The systematic search in hypotheses spaces which does not skip any important hypotheses ensures better results than those achieved by "manually" analysing large simulation datasets e.g. by standard reports and cross-tabulations. Therefore, discovery methods can contribute to generalize a system of heterogeneous regulations, give a representative overview about the values a multidimensional continuous function takes, identify key groups and decover secondary groups by uncovering overlappings

Further advances for these applications can be realized, if discovery will be applied to goal state experiments to find "optimal" solutions. Also, from a theoretical point of view, legislative planning seems to be an interesting problem for Machine Learning and Discovery.

References

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning 3,* pp. 261-283.

Gebhardt, F. (1991). Choosing among Competing Generalizations. *Knowledge Acquisition* 3, pp. 361-380.

Gebhardt, F. (1994). Discovering interesting statements from a database. *Applied Stochastic Models and Data Analysis* 10 (1).

Klösgen, W., Schwarz, W., and Honermeier, A. (1983). *Modellbank-System MBS (User Manual).* Arbeitspapiere der GMD, 32. GMD, Sankt Augustin.

Klösgen, W. and Quinke, H. (1985). Sozioökonomische Simulation und Planung: Entwicklungsstand und Computerunterstützung. *Informatik-Spektrum 8,* pp. 328-336.

Klösgen, W. (1986). Software implementation of microanalytic simulation models - state of the art and outlook. In: Orcutt, G., Merz, J., Quinke, H. (eds.) *Microanalytic simulation models to support social and financial policy;* pp. 475-491 . North Holland, Amsterdam.

Klösgen, W. (1992). Problems for Knowledge Discovery in Databases and their Treatment in the Statistics Interpreter EXPLORA. *International Journal for Intelligent Systems* vol 7(7), pp. 649-673.

Klösgen, W. (1993) *Explora: A support system for Discovery in Databases, Version 1.1, User Manual.* GMD, Sankt Augustin.

Klösgen, W. (1994). Efficient Discovery of Interesting Statements in Databases. To appear in: *Journal of Intelligent Information Systems.*

Orcutt, G., Merz, J., Quinke, H. (eds.) (1986). *Microanalytic simulation models to support social and financial policy.* North Holland, Amsterdam.

Piatetsky-Shapiro, G. (1991). Discovery, Analysis, and Presentation of Strong Rules. In Piatetsky-Shapiro, G., and Frawley, W.J. (Eds.), *Knowledge Discovery in Databases.* MIT Press, Cambridge, MA.

Piatetsky-Shapiro, G. and Matheus, C. J. (1993). KeFir: Key Findings Reporter, *1993 AAAI Knowledge Discovery Workshop, Poster Session Presentation.*

Quinlan, R. (1986). Induction of decision trees. *Machine Learning 1,* pp. 81-106.

Shrager, J. and Langley, P. (eds.) (1990). *Computational Methods of Scientific Discovery and Theory Formation.* Morgan Kaufmann, San Mateo, CA.

Uthurusamy, R., Fayyad, U.M., and Spangler, S. (1991). Learning Useful Rules from Inconclusive Data. In Piatetsky-Shapiro, G., and Frawley, W.J. (Eds.), *Knowledge Discovery in Databases.* MIT Press, Cambridge, MA.

Zytkow, J. and Zembowicz, R. (1993). Database Exploration in Search of Regularities. *Journal of Intelligent Information Systems* 2, pp. 39-81.