Ein-Dor, P. (ed.), *Artificial Intelligence in Economics and Management: An Edited Proceedings on the Fourth International Workshop*, Boston, Kluwer Academic Publishers, pp. 193-203, 1996.

Multistrategy Conceptual Analysis of Economic Data

Kenneth A. Kaufman and Ryszard S. Michalski *
Machine Learning and Inference Laboratory,
George Mason University,
Fairfax, Virginia, 22030, USA

* Also with Department of Computer Science, Department of Systems Engineering, and the Institute of Computer Science, Polish Academy of Sciences.

{kaufman, michalsk}@aic.gmu.edu
Phone: (+1) 703-993-1542
Fax: (+1) 703-993-3729

Multistrategy Conceptual Analysis of Economic Data

KENNETH A. KAUFMAN AND RYSZARD S. MICHALSKI
George Mason University, Fairfax VA, 22030, USA

{kaufman, michalsk}@aic.gmu.edu

Abstract

The goal of the multistrategy tool, INLEN, is to serve as an intelligent assistant for discovering knowledge in large databases. INLEN has been applied to, and is well-suited for the exploration of databases consisting of economic and demographic facts and statistics. Preliminary experiments on several data sets have focused on discerning and comparing various patterns in the status and development of countries in different regions of the world. These experiments have provided some interesting and often unexpected results, and serve as an example of one way in which such data can be explored. This paper describes in brief the INLEN methodology, presents examples of its learning and discovery operators, and demonstrates its application to economic domains.

Keywords: Multistrategy Learning, Knowledge Discovery in Databases, Knowledge Transmutations, Economic and Demographic Databases

1. Introduction

As the amount of electronically available information grows explosively, it is becoming both critically important and increasingly difficult to analyze the data to derive desired knowledge from them. Traditional tools for data analysis employ mostly statistical concepts and methods. These methods can be particularly useful for such tasks as detecting statistical trends, correlations between attributes, data distributions, etc. They are, however, limited in the types of knowledge and regularities they can derive from data.

For example, a statistical analysis can detect a correlation between given factors, but cannot produce a conceptual explanation why such a correlation exists, nor can it formulate any specific quantitative and/or qualitative law(s) responsible for this correlation. A statistical technique can determine a central tendency and variability of some properties, or fit a curve to a set of datapoints, but it cannot explain them in terms of causal dependencies or qualitative relationships. Attributes that define a similarity and the measures of similarity involved must be given in advance. In short, these techniques require that an interpretation of the findings, i.e., a "conceptual" analysis of data, be performed by a human analyst. As the quantity of available data increases, the complexity of such an analysis can easily outstrip human capabilities.

This paper gives a brief description of INLEN, a methodology and multistrategy system for intelligent conceptual data analysis, and presents examples of its application to the domain of economics. INLEN has been designed to overcome some of the limitations of

statistical data analysis by applying advanced methods of machine learning. It integrates database, knowledge base and machine learning technologies into an intelligent system for data analysis and knowledge discovery. It offers a data analyst a powerful tool for discovering patterns of non-statistical nature, determining logic-style data descriptions, and can producing justifications or explanations of the discovered patterns (Michalski et al, 1992).

For example, it can describe the data in terms of logic-style decision rules that link several conditions to a desirable outcome. Individual conditions may contain not only attributes present in the data, but also *constructed* attributes. Such attributes are generated in the process of *constructive induction*, a process that splits the search for patterns or general descriptions into two phases – the first to search for the "best" representation space (the space of attributes or descriptive terms), and the second to determine the "best" hypothesis in the found space. New attributes are generated by using hints from an expert, or by constructing single- or multi-level functions of the original attributes.

Symbolic learning programs have been developed that can determine rules for distinguishing between many classes of items, find conceptually useful ways to group objects, apply knowledge in order to predict missing values in a data set, select representative subsets of a large data set best suited for a particular learning task, etc. The various learning and discovery programs in INLEN are accessed in the form of *knowledge generation operators* (KGOs) that can be applied in sequence, with each KGO capable of taking advantage of its predecessor's findings. Because of the symbolic nature of these operators and their adaptability to different problems, the INLEN methodology is well-suited for many economic domains in which databases contain a large number of records and attributes, and in which the results of analysis must be understood by non-experts.

The following sections describe briefly INLEN's methodology and its applications to the analysis of several databases consisting of world economic and demographic facts. These applications are illustrated with examples of the system's discoveries based on the patterns in the data. The last sections summarize the development status of INLEN, the advantages and current limitations of the proposed methodology as applied to problems in economic domains, and an outline of the plans for future research

2. The INLEN methodology

The INLEN system integrates a database, knowledge base and machine learning technologies into a unified knowledge discovery environment (Fig. 1).

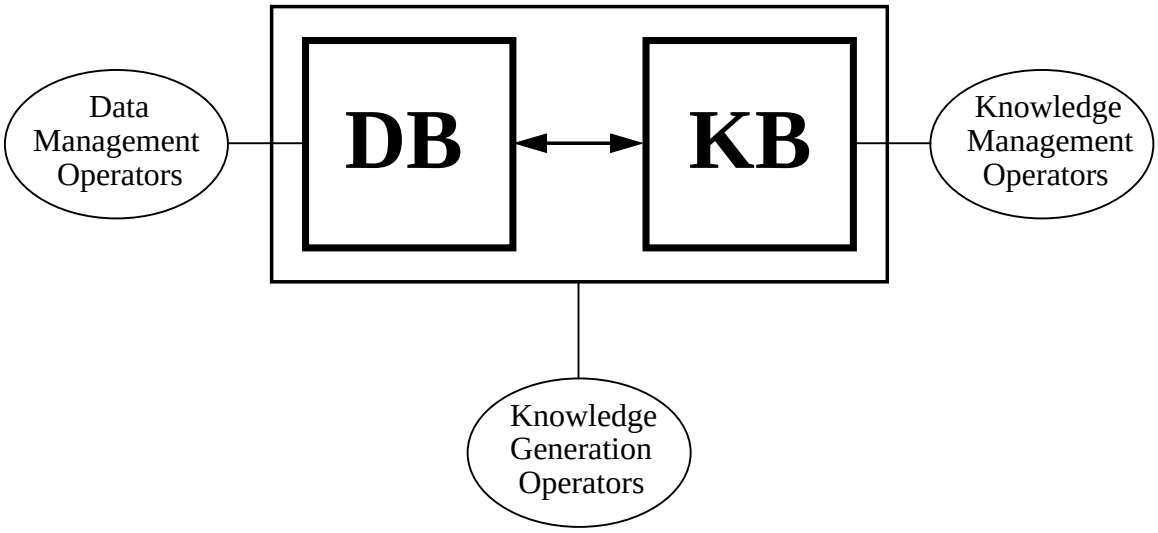

Figure 1. The functional architecture of INLEN.

A relational database is used for storing data, and can be modified by the user through *data management operators.* A knowledge base stores and maintains rules, equations, decision procedures, representative examples, and concept hierarchies that are employed in the process of data analysis and knowledge discovery. The knowledge base includes domains of the attributes, problem background knowledge, discovered knowledge, and any knowledge entered or modified by the user. The contents of the knowledge base can be modified by an expert through *knowledge management operators*.

The central engine of the system is a set of operators, called *knowledge generation operators*, that invoke learning, discovery and inference programs to perform data analysis tasks. These operators take input from the database and/or knowledge base, and produce outputs that enhance the data and/or knowledge bases. They are adept at diverse tasks, such as learning symbolic rules to differentiate among several classes of data items, conceptually dividing a data set into two or more groups (conceptual clustering), modifying the representation space into one more suitable for a particular learning task, testing a set of knowledge for consistency with respect to new data, and predicting values for missing data.

The idea of such a multi-operator approach to knowledge discovery was formulated by Michalski in the 1980s. The first such effort, from which INLEN derived much of its conceptual architecture, was the QUIN system (**Qu**ery and **In**ference), a combined database management and data analysis environment (Michalski, Baskin and Spackman, 1982; Michalski and Baskin, 1983; Spackman, 1983).

The knowledge generation operators in INLEN represent various knowledge transmutations as are catalogued in Michalski's Inferential Theory of Learning (1993). For instance, learning classification rules from examples involves an inductive generalization transmutation; conceptual clustering is an act of inductive agglomeration;

constructive induction can consist of generation, selection or abstraction operators; and knowledge-based prediction of missing values incorporates a similization transmutation.

An important feature of INLEN's knowledge generation operators is that many of them allow the user to define parameters and tailor their performance to a specific learning task. In this way, the user (data analyst) can specify from among the many possible outputs consistent with the data which type of knowledge is likely to be most useful. Two examples below illustrate the choices of parameters for using the knowledge generation operators.

One of the backbone abilities of INLEN is rule learning from examples, performed using the AQ15c learning engine (Wnek et al, 1995). The user can accept the default parameters or instruct the program to bias its rule selection toward highly specific characterizations, short rules for discrimination, a set of maximally simple rules, rules that tend to incorporate or avoid specific attributes if at all possible, or the user can create and specify a different set of preference criteria.

Another knowledge generation operator is based on the PROMISE program (Baim, 1982). It examines the set of attributes in the database and ranks them according to their applicability to a given classification problem. The user can then choose to learn from the subset of the data consisting of only the attributes likely to be more relevant to the learning task. This operator can be used in two modes. In the first, each attribute's overall discrimination ability is calculated using an information gain metric similar to that used by the C4.5 (Quinlan, 1990) family of algorithms in selecting attributes to serve as the roots of the decision trees it generates. The second mode focuses on the attributes most likely to produce concise rules; it selects attributes that contain some values that discriminate very well between classes, even if most of the other values provide little information about how to classify the examples.

3. Application to economic and demographic domains

INLEN has been applied to learning and discovery problems in such domains as engineering design, disease diagnosis, intelligence gathering and economic analysis. The following example illustrates the role an intelligent agent can play in the discovery of knowledge from data:

> *The United States government maintains records of the import and export of goods from various countries of the world. The different products and raw materials are divided and subdivided into different categories. In the early 1980s the data showed a sharp decline in the import of trucks from Japan, while there was a corresponding increase in imports from Japan in the auto parts category. It took several years before analysts noticed that fact and concluded that Japan was shipping the chassis and truck beds separately to the US, where they would be subsequently assembled, thereby avoiding a high US tariff on imported trucks that was directed primarily at Europe and had been on the books since World War II. When United States analysts inferred this, the US and Japan commenced trade negotiations pertaining to the import of trucks.*

How much sooner would that trend have been noticed had a discovery program been applied to the data and pointed out the opposite changes in two related categories to an analyst? How much revenue did the undiscovered truth cost the US before they could

finally work out a new agreement with Japan? Noticing economic trends and patterns is a difficult task, as humans can easily get overwhelmed by the amount of data and figures.

Based on such motivations, the analysis of economic and demographic data has become one of the focus domains for INLEN development and testing. These experiments have involved two similar data sets – one provided by the World Bank consisting of information on 95 attributes in 171 countries for the period of 1965 to 1990, and one extracted from the 1993 World Factbook (CIA, 1993) containing several databases of information on 190 countries.

For example, one experiment focused on distinguishing between development patterns in Eastern Europe and East Asia (Kaufman, 1994). A conceptual clustering operator determined that one way of distinguishing between the typical Eastern European country and the typical Far Eastern country was through examining the country's change in the percentage of its population in the labor force between 1980 and 1990. Most of the European countries had a labor force change below a threshold determined for the region by the learning program, while most of the Asian countries had changes above their region's threshold.

Based on this, the rule learning operator (based on the AQ15c inductive learning program) was then called upon first in characteristic mode to categorize the differences between the Asian-like countries (those above their regional thresholds) and the European-like countries (those below their regional thresholds), and then in discriminant rule-optimizing mode to condense those characterizations into the following simple decision rules:

Country is Asian-Like if:

A.1 Change in Labor Force Participation • slight_gain, *(9 countries)*
or
B.1 Working Age Population • 64%,
2 Life Expectancy is in 60s. *(2 countries)*

Country is European-Like if:

A.1 Life Expectancy is not in 60s,
2 Change in Labor Force Participation is near 0 or decreasing, *(7 countries)*
or
B.1 Percentage of Labor Force in Industry • 40. *(1 country)*

The rules show that the features aside from change in labor force participation instrumental in distinguishing between the European-style and Asian-style development patterns include life expectancy, working age population and degree of industrialization. In both the Asian- and European-Like cases, the first rule accounted for most of the countries fitting the class, while the second one described the remainder.

In another experiment, the AQ15c module of INLEN, determined characteristic descriptions of various regions of the world based on some typical "core" countries in each region, as suggested by a domain expert (Kaufman, 1994). An example of such characterization of three core countries of Southeastern Europe is below:

Conditions characterizing Southeastern European Countries:

		Pos	Neg	Sup
1	Urban_Population_Percentage = 50% to 70%	3	5	37%
2	GDP Agricultural percentage • 10%	3	5	37%
3	Infant Mortality Rate • 10	3	6	33%
4	GNP Per Capita • $3500	3	7	30%
5	Life Expectancy = 70 to 73 years	3	8	27%
6	Percentage of Labor in Industry = 30% to 45%	3	8	27%
7	Percentage of Labor in Agriculture • 15%	3	9	25%
8	Fertility Rate = 5/3 to 3	3	15	18%
9	Death Rate = 8 to 12	3	16	15%
10	Population Growth Rate • 1%	3	17	15%

where Pos denotes positive examples (i.e., countries from Southeastern Europe), Neg denotes negative examples (countries from other regions), and Sup denotes a relative support, according to the expression: *Supp = Pos / (Pos + Neg) x 100%* Sup values provide the user with an idea of how much support that condition alone would provide toward the hypothesis in question. In this case, an urban population percentage between 50% and 70% provides much stronger evidence that the country in question was from Southeastern Europe than the fact that its population growth rate was less than 1%.

In this experiment some strong trends were discovered. For instance, Western European core countries had high per-capita gross national products (GNPs), low infant mortality rates and low percentages of workers in agricultural fields. Eastern European countries showed low life expectancies and low per-capita GNPs, but while those in the northern part of the region were also characterized by high death rate, high primary school education and moderate-to-high infant mortality, those in the southern part of the region exhibited low urbanization, slow population growth and relatively highly agriculturally based economies (this characterization is shown above). Southern European countries showed low allocation of resources to education, moderate per capita GNPs and low fertility rates. The developed countries of East Asia all had very low death rates.

To see how well these characterizations fit the other countries in each region, the non-core countries in the examined regions were then tested against the discovered patterns using the rule testing operator developed from the ATEST program (Reinke, 1984). This operator tests a rule set for consistency and completeness. Among the findings was the tendency for island countries to deviate from the mainland norms for their region more than other mainland countries, even when they might appear sociopolitically closer to the regional norms. For instance, development patterns in Switzerland closely matched those found elsewhere in Western Europe (INLEN reported an 80% match with the region's characteriztion), while Ireland's were significantly further from the typical regional pattern (only a 40% match).

Other findings suggested similarities between the development patterns of South American countries and many of the countries of East Asia. The results also confirmed the suggestion that sociocultural influences are often stronger than geographical proximity in determining a country's economic and demographic indicators. For example, Italy shows a pattern consistent with that of Western Europe, rather than Mediterranean Europe.

Sometimes more than one rule is necessary to fully describe a class of data. In such cases, the examples united under a single rule tend to have a conceptual similarity to one another. For instance, in characterizing the regions of the world based on the 1990 World Bank data, INLEN found a characterization pertaining to most of the countries of South America, another one encompassing Bolivia, Colombia, Suriname and Uruguay, and two more representing the exceptional cases of the Guyanas. One could hypothesize that the four countries in the second group form a cohesive subgroup with a common bond worth exploring.

The previous examples were based on data sets extracted from the World Bank database. An experiment using the World Factbook's PEOPLE database demonstrated how INLEN can detect interesting facts within the subgroups it creates. While the subgroups in a demographic domain may indicate that member countries or regions that have something in common, notable exceptions may be exposed when one of the members of these constructed subsets shows a marked dissimilarity to the rest of the group. These exceptions in turn may prove be a springboard for further discovery.

INLEN discovered several rules from the PEOPLE database characterizing the countries with low (less than 1% per year) population growth rates by invoking the rule learning operator in characteristic mode.

One of the rules had three conditions that together were sufficient to distinguish 19 low growth countries from those with higher population growth rates:

Conditions characterizing Countries with Population Growth Rates below 1%:		Pos	Neg	Supp	
1	Birth Rate = 10 to 20 or Birth Rate • 50		46	20	69
2	Predominant Religion is not defined as Muslim or Mixed or Buddhist or Christian or Tibetan		40	68	37
3	Net Migration Rate • +20		32	104	23

The first (and therefore strongest) condition states that the country must have a low (under 20 per 1000 population) or very high (over 50) birth rate. The presence of a very high birth rate is extremely counterintuitive; examination of the 19 countries involved pointed out that 18 had birth rates below 20, while only one, Malawi, had this high birth rate.

INLEN had thereby identified an exception to normal patterns. When further learning was focused on Malawi, a massive outward net migration rate was discovered, by far the most extreme migration rate in the world. Further application of the knowledge discovery operators could explore the conditions unique to Malawi and hypothesize where else they might take place in the future.

4. The development status of INLEN

The examples shown above use just a few of the knowledge generation operators being implemented in INLEN. INLEN itself is growing steadily as operators are enhanced and added to its environment; while still in prototype form, it already has many functional operators.

Among the areas intended for future research and development are the addition of tools for creating new attributes for improved performance (constructive induction), based on the AQ17-MCI methodology (Bloedorn et al, 1993). This methodology combines data-driven construction of attributes based on the detection of relationships between attributes, hypothesis-driven construction of attributes based on patterns detected in preliminary rulesets, and statistically-based operators for quantization of continuous numerical attributes and summarization of notable groups of examples (e.g., a region's average per capita income).

Another enhancement will be the development and incorporation of a high-level language for knowledge discovery. With such a tool, a user will be able to program in sequences of operators, along with instructions detailing what conditions shall cause them to be invoked. The system will be able to follow such instructions as "If a new month's data shows less than a 95% consistency with the knowledge base, update the knowledge to incorporate the new data and then use characterization operators to seek out possible explanations why the behavior has changed in the new month" or "If a high-urgency network fault is detected, access the knowledge base to predict the location of the problem and the nature of a likely solution and report it immediately."

INLEN currently examines only one relational table at a time. An ongoing research effort is to develop methods for discovering knowledge that is not stored in one location, but rather distributed among several databases (Ribeiro et al, 1995). One approach avoids combining the data from the different sources, instead developing separate knowledge bases and deriving rule-based attributes in order to search for commonalities between the discovered knowledge and the other databases.

A fourth research effort applicable to INLEN is the development of a methodology to provide machine learning programs with enriched domain knowledge in the form of intricate conceptual hierarchies. The knowledge generation operators can then use this knowledge to help select the output knowledge's level of abstraction that will be most useful for the data analyst.

5. Conclusions

The INLEN methodology is based on the application of a large arsenal of machine learning and inference programs for the purpose of discovering knowledge from databases and providing concise conceptual explanations of their findings. These diverse machine learning and inference tools can work in conjunction with traditional statistical tools. Among the major advantages of INLEN is its emphasis on providing conceptually understandable results of data analysis due to the logic-style descriptions it generates. Another advantage is its modularity that makes it easy to add new knowledge generation operators.

Among the limitations of the current implementation of INLEN are the lack of integration with statistical data analysis methods and too much reliance on the data analyst for guidance in the selection of operators and the setting of parameters. Another limitation is portability; while future versions of INLEN will run on other platforms, it is currently limited to PC-based systems.

As described above, topics of further research include an integration of methods for the automated improvement of the representation space and discovery of new attributes more relevant to the task at hand, the development of a high-level conceptual language to assist the user in planning and executing complex data analysis tasks, development of methods for automated data abstraction, and analysis of distributed databases.

In the presented study, INLEN's knowledge generation operators have been applied to databases of economic and demographic indicators. Findings obtained by INLEN confirmed some known facts while unearthing some surprising ones. These experiments illustrate some of the INLEN's abilities and clearly indicate that this system has a potential for determining important but unknown economic findings in large database of facts.

While future experimental investigation of the INLEN methodology will continue to explore the international demographic datasets described in this paper, the application of this system to other economic problems, such as international trade data analysis and stock market trend detection and prediction is also planned.

Acknowledgments

The authors thank Eric Bloedorn for his comments and suggestions regarding earlier drafts of this paper.

This research was conducted in the Machine Learning and Inference (MLI) Laboratory at George Mason University. MLI research is supported in part by the National Science Foundation under Grants No. CDA-9309725, IRI-9020266, and DMI-9496192, in part by the Advanced Research Projects Agency under Grant No. N00014-91-J-1854, administered by the Office of Naval Research, and Grant No. F49620-92-J-0549, administered by the Air Force Office of Scientific Research, and in part by the Office of Naval Research under Grant No. N00014-91-J-1351.

References

Baim, P.W. (1982). The PROMISE Method for Selecting Most Relevant Attributes for Inductive Learning Systems. Report No. UIUCDCS-F-82-898, Department of Computer Science, University of Illinois, Urbana IL.

Bloedorn, E., Wnek, J., and Michalski, R.S. (1993). Multistrategy Constructive Induction: AQ17-MCI. *Proceedings of the Second International Workshop on Multistrategy Learning*, Harpers Ferry, WV, pp. 188-203.

Central Intelligence Agency (1993). 1993 World Factbook.

Kaufman, K. (1994). Comparing International Development Patterns Using Multi-Operator Learning and Discovery Tools. *Proceedings of AAAI-94 Workshop on Knowledge Discovery in Databases*, Seattle, WA, pp. 431-440.

Michalski, R.S. (1993). Inferential Theory of Learning as a Conceptual Basis for Machine Learning. *Machine Learning*, **11**, pp. 111-151.

Michalski, R.S. and Baskin, A.B. (1983). Integrating Multiple Knowledge Representations and Learning Capabilities in an Expert System: The ADVISE System. *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, pp. 256-258.

Michalski, R.S., Baskin, A.B. and Spackman, K.A. (1982). A Logic-based Approach to Conceptual Database Analysis. Sixth Annual Symposium on Computer Applications in Medical Care (SCAMC-6), George Washington University Medical Center, Washington, DC, pp. 792-796.

Michalski, R.S., Kerschberg, L., Kaufman, K. and Ribeiro, J. (1992). Mining for Knowledge in Databases: The INLEN Architecture, Initial Implementation and First Results. *Journal of Intelligent Information Systems: Integrating AI and Database Technologies*, Vol. 1, No. 1, pp. 85-113.

Quinlan, J.R. (1990). Probabilistic Decision Trees. In *Machine Learning: An Artificial Intelligence Approach, Volume III*, Y. Kodratoff and R.S. Michalski, (Eds.), Morgan Kaufmann, San Mateo, CA.

Reinke, R.E. (1984). Knowledge Acquisition and Refinement Tools for the ADVISE Meta-Expert System, Master's Thesis, Department of Computer Science, University of Illinois, Urbana, IL, 1984.

Ribeiro, J.S., Kaufman, K.A. and Kerschberg, L. (1995). Knowledge Discovery From Multiple Databases. *First International Conference on Knowledge Discovery and Data Mining*, Montreal PQ.

Spackman, K.A. (1983). QUIN: Integration of Inferential Operators within a Relational Database. ISG 83-13, UIUCDCS-F-83-917, M. S. Thesis, Department of Computer Science, University of Illinois, Urbana, IL.

Wnek, J., Kaufman, K., Bloedorn, E. and Michalski, R.S. (1995). Selective Induction Learning System AQ15c: The Method and User's Guide, *Reports of the Machine Learning and Inference Laboratory*, MLI 95-4, Machine Learning and Inference Laboratory, George Mason University.