Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases, Seattle, WA, pp. 229-236, August, 1994.

COMPARING INTERNATIONAL DEVELOPMENT PATTERNS USING MULTI-OPERATOR LEARNING AND DISCOVERY TOOLS

Kenneth A. Kaufman

Center for Artificial Intelligence
George Mason University
4400 University Drive
Fairfax VA 22030-4444

kaufman@aic.gmu.edu

Abstract

The multistrategy knowledge discovery tool, INLEN, is applied to databases consisting of economic and demographic facts and statistics about the countries of the world. Preliminary experiments focus 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, but they are only a beginning in exploring such data. By discovering patterns and exceptions such as the ones presented, domain experts may have new insights into national development patterns, predict future developments in certain countries, or use these discoveries to influence national policies. Users who are not experts in the domain may also make interesting discoveries with INLEN. The results of these initial experiments are presented and future paths of research in this domain are proposed.

1 Introduction

The domain of international economic and demographic statistics has a strong potential to be a fruitful application area for knowledge discovery techniques. Specifically, a database of such statistics is fertile ground for mining. One may find relationships among seemingly independent features in the data, and use that knowledge to better understand the conditions in various countries. Individual countries may be compared and contrasted, and also analyzed for clues as to their potential future development. Two countries in similar situations may wish to follow similar policy paths. Conversely, two dissimilar countries may require vastly different approaches.

Since many of the facts in the database are updated periodically, one can also inspect the development of a country over several decades and learn from the discovered patterns. A country may find itself in a state similar to that of another country some time ago. Decision-makers can then trace the progress of the latter nation and determine whether or not they want their land to follow a comparable course.

This paper discusses initial experiments using such a database and the results they have produced. The discoveries were made using INLEN-1, a program for making discoveries in databases using an array of machine learning tools. Section 2 describes the INLEN-1 program and the knowledge discovery operators used in these experiments. In Section 3 the contents of the database are discussed in more detail. Section 4 describes the preliminary experiments and results, and Section 5 outlines future work in this domain.

It should be emphasized that the INLEN discovery system is still under development, and that its use in exploration of the data described in this paper has just begun. Hence, the results shown here are preliminary ones, and are more indicative of the knowledge discovery capacity of this methodology than they are substantial contributions to understanding the domain. Through further interaction with the interested domain experts, the discovery process will be better focused on their specific areas of interest.

2 The INLEN Knowledge Discovery System

INLEN is designed to serve as a platform for knowledge discovery in databases that overcomes some of the limitations of traditional knowledge discovery systems. Many of the tools in the INLEN package utilize symbolic machine learning methods which, unlike purely statistical methods, are able to conceptually explain their findings and the conditions under which they hold. Thus, one does not have to be a domain expert to get a deep understanding of what a particular set of results means.

Furthermore, traditional methods are not able to alter the feature space in the database to improve prospects for discovery. If a database contains information on the lengths, widths and heights of objects but not their volumes, such methods will have difficulty in making discoveries that depend closely on the volumes themselves. The INLEN architecture includes several operators for modifying the representation space through constructive induction, including operators for creating new attributes based upon either the data itself or intermediate discovered knowledge, selecting subsets of a larger feature space that are better-suited for discovery than the entire initial feature set, and combining portions of separate, related databases into promising views of the data.

One weakness of a symbolic machine learning-oriented approach is that many such learning programs have the limitation of being task-inflexible, i.e., they can only perform one specific type of learning or discovery task. A clustering program will not be able to discover simple equations that characterize a table of numeric data while a program that can learn such equations will fail to make discoveries in non-numeric databases. INLEN is designed to provide a multistrategy environment for discovery, with different operators available for use depending on the particular task at hand. INLEN also features a modular architecture, in which separate programs are integrated into the system as individual operators or groups of operators. By introducing new programs into the system, one can add to its knowledge discovery capabilities. Similarly, a program in INLEN can be replaced by another that performs a similar task better according to some measure. To compare the performance for a given task of one program to another is beyond the scope of this paper; the software-oriented emphasis instead lies upon the integration of diverse components into a multistrategy knowledge discovery system.

A detailed description of INLEN's architecture and the operators from which it is built is given by Michalski et al. (1992b). 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). QUIN was designed both as a stand-alone system, and as a subsystem of ADVISE, a large-scale inference system for designing expert systems (Michalski and Baskin, 1983; Michalski et al, 1987; Baskin and Michalski, 1989). The INLEN architecture expands on the architecture used by QUIN by incorporating a number of new learning and inference operators. Additionally, it maintains a more complex knowledge base than QUIN's, which was designed primarily with expert system rule bases in mind.

In the last few years, new tools have been developed – in particular, more advanced inductive learning systems, e.g., AQ15 (Michalski et al, 1986), AQ17 (Bloedorn, Wnek and Michalski, 1993) and ABACUS-2 (Greene, 1988), and expert database systems (Kerschberg, 1986, 1987, 1988). The above systems have influenced the development of INLEN and its components. INLEN also draws upon the experiences of working with AGASSISTANT, a shell for developing agricultural expert systems (Katz, Fermanian and Michalski, 1987) and AURORA, a general-purpose PC-based expert system shell with learning and discovery capabilities, designed by Michalski and Katz (INIS, 1988). The kernel of INLEN-1 (the prototype version of the system upon which we have conducted these experiments) has grown directly out of these.

INLEN-1 includes a knowledge base of simple decision rules, a prototype relational database, modules for easy management of the data, and an extensive user-oriented menu-based graphical interface. The knowledge generation operators (KGOs) used by INLEN-1 consist of a subset of the full set of operators that will be incorporated in later versions of the system. The particular operators that were used in the experiments presented in this paper, CLUSTER, CHARSET, IMPROVE and TEST, are described below:

CLUSTER: Conceptual Clustering

CLUSTER is an operator used to discover different ways of classifying an input set of examples. It does so by creating logical divisions of the data into two or more groups, forming conceptual descriptions of the groups, and evaluating the quality of those descriptions. The descriptions can be used in selecting particular groupings of the data. In the example presented in Section 4, a clustering is chosen that divides the examples based on one attribute and then partitions each subgroup according to another attribute in a manner that suggests a large set of representative examples and a small set of exceptions in each subgroup. The program to perform these clustering tasks is based on CLUSTER/2 (Michalski, Stepp and Diday, 1981; Michalski and Stepp, 1983; Stepp, 1983, 1984).

CHARSET: Characterize and Differentiate Sets of Events

CHARSET learns decision rules from sets of examples of different classes. Parameters can be set to learn characteristic rules (very specific rules that fully characterize the examples of an input class), discriminant rules (maximally general rules that provide sufficient information to differentiate between the classes of examples), or rules of intermediate generality that best satisfy other user criteria. The method used in these experiments employs the AQ15 learning program (Michalski et al, 1986) to learn characteristic rules. AQ15 discovers classification rules by examining different generalizations of a "seed" example according to user-defined selection criteria and creating a rule based on an optimal such generalization.

IMPROVE: Improve Knowledge Base

The IMPROVE operator can be used in several ways, each with the goal of improving existing knowledge. One form of this operator is incremental learning, in which newly available data is matched up with the existing knowledge, and inconsistencies are resolved through fine-tuning of the knowledge base. Another form is rule optimization, demonstrated in Section 4, in which decision rules, rather than examples are the input to the AQ15 learning module. Here, the portion of the event space covered by a rule serves as the seed, and it is maximally generalized so as not to extend into areas covered by rules for other classes. In the example shown here, discriminant rules are generated from existing characteristic ones. By doing so, we can often learn more succinct knowledge than if we had learned discriminant rules directly from examples (Cuneo, 1975).

TEST: Test Knowledge

A knowledge-testing operator implemented using the ATEST program (Reinke, 1984), TEST, was also used in these experiments. A set of rules and some testing examples are the input to this operator. ATEST examines each example and determines whether or not it is correctly classified by the ruleset (based on exact match with its proper rule). If it is not, the example is set against every rule in the ruleset to determine the degrees of match between the rules and the example. The system judges the rule with the highest degree of match as being the best match for the example. ATEST can summarize its findings in three ways: by showing the percentages of examples that match their target class both exactly and through best fit, by creating a table showing the percentage of examples of each class classified by the rules into each class, and by showing the degrees of match for each class for the individual examples.

3 Data Description

These experiments were performed upon a database provided by the World Bank that contains information on 171 countries. For each country, 95 attributes were measured (when available) for each year between 1965 and 1990 inclusive. Many more attributes could be constructed by measuring changes in values from one year to another or by combining related features. For example, although no information was explicitly available on the average education attained by women in a country's labor force, this could be generated from the male and overall education statistics as well as the male/female labor force ratio.

In general, the features in the database deal with the quality of life and the economic conditions in the various countries. Of these attributes, the ones used in one or both of the experiments described below (since the focus of the first experiment shown here was more demographic, and the focus of the second was more economic, the feature sets used were not identical) included the following:

- Annual Population Growth Rate
- Population Density
- Urban Population Percentage
- Percentage of Population Between Ages 16 and 64
- Change in the Percentage of the Population in the Labor Force during the 1980s
- Percentage of the Labor Force in Industry
- Percentage of the Labor Force in Agriculture
- Percentage of Land Devoted to Agriculture
- Life Expectancy
- Fertility Rate
- Death Rate
- Infant Mortality Rate
- Per Capita GNP

- Percentage of GDP devoted to Agriculture
- Percentage of GDP devoted to Medicine
- Percentage of GDP devoted to Education
- Percentage of Children Completing Primary School

Each of these features is represented in the database by numerical values, typically with one decimal place where appropriate. To facilitate the learning, the values were quantized into intervals, with the intervals being selected based on either the distribution of the values or groupings that had extra semantic meaning for a human user (for example, by tens), as best suited the data. In the results shown below, the values are given in a simplified but equivalent form for ease of understanding.

4 Initial Experiments on the World Bank Data

The motivation and focus of these experiments were based on discussions with the domain experts with whom we have been collaborating. It is through these discussions that we chose to concentrate on certain parts of the world and on certain indicators tending to have direct relevance to the quality of life in those areas. The selection of the learning operators was user-driven based on both the presence of specific goals at the current stage of the knowledge discovery process and the output from previous steps.

Our first experiment focused on discovering characterizations for countries in different parts of the world by learning and comparing patterns for different regions and searching for interesting behavior in those countries that did not follow the local pattern. We chose a set of quality-of-life indicators from the database and all of the countries from Eastern Europe and the Far East for which most of that information was available.

The goal of this experiment was to generate descriptions contrasting the "Eastern European" model from the "Far Eastern" one. Secondarily, we were interested in seeing if any of the European countries appeared to have conditions similar to those in the Asian pattern of development. If so, they might be well-advised to examine the policies of the Asian countries and try to follow the paths of the more successful ones while avoiding the pitfalls of the struggling nations. Conversely, the European nations that showed far different conditions from those in the Far East would likely be best served by the implementation of a different set of policies, given that they would have to operate under a different set of assumptions.

Another goal was to identify some of those conditions that would tend to determine whether or not a country is "Far-East-like" in order to ascertain easily whether a country is converging with this model or moving away from it. It may be the case, for example, that Bulgaria and Malaysia have little in common right now in terms of these key indicators, but that they may become more similar several years in the future. In this way, using some concise decision rules, similarities and dissimilarities may be tracked.

The first step in this experiment was to apply conceptual clustering to the set of countries. Most of the sets of groups created by the clustering program did not highlight the differences between the regions, but one four-group division did exactly that. Two of the groups consisted of Eastern European countries while the other two held only Far Eastern ones. Each pair of groups from the same region was divided into two individual groups by (among other things) the change in each country's labor force participation percentage. That is, one of the Eastern European groups consisted of countries in which that figure was below a certain threshold, while the other held ones with higher values. Similarly, the Asian countries were divided into two by this feature, albeit with a different threshold marking the division point.

It should be emphasized that differences in other features also distinguished the individual groups. However, the region and the change in labor force participation are focused upon here due to the clarity with which they defined group membership.

Interestingly, most of the European countries (there were only two exceptions - Albania and Romania) had labor force percentage changes below the threshold at which the region’s countries were divided. Similarly, all but two of the Asian countries (Cambodia and Laos) had changes above their threshold. One hypothesis to consider was that the two European countries with above-threshold labor force changes might be following closer to the general Asian model in other respects, and the two Asian countries with below-threshold changes could be more European-like in outlook. Hence our next step was to regroup the countries based on whether they were above or below their region’s threshold and then use AQ15 to learn rules describing the two groups. These are the four rules that AQ15 discovered:

Class is Class1 (Asian-Like) if:

A.1.Change in Labor Force Participation • slight_gain,
2.Percentage of Labor Force in Industry • 40%,
3.Population Growth Rate • 0,
4.Life Expectancy • 60, *(total 9, unique 9)*
or
B.1.Region is Far East,
2.Percentage of Labor Force in Industry • 30%,
3.Life Expectancy is in 60s,
4.Population Growth Rate > 2%,
5.Working Age Population < 64%,
6.Agricultural Land Area is not 30% to 50%,
7.Population Density is sparse,
8.Change in Labor Force Participation is near 0. *(total 2, unique 2)*

Class is Class2 (Asian-Unlike) if:

A.1.Change in Labor Force Participation is near 0 or decreasing,
2.Population Density • moderate,
3.Population Growth Rate is not 1% to 2%,
4.Percentage of Labor Force in Industry is not 20% to 30%,
5.Life Expectancy is not in 60s,
6.Agricultural Land Area is not 30% to 50% *(total 7, unique 7)*
or
B.1.Percentage of Labor Force in Industry > 40,
2.Region is Eastern Europe,
3.Agricultural Land Area > 50%,
4.Change in Labor Force Participation is near 0,
5.Population Density is moderate,
6.Population Growth Rate < 1%,
7.Working Age Population • 64%,
8.Life Expectancy is in 60s. *(total 1, unique 1)*

The numbered conditions in each rule are ordered on the basis of expected informativeness, based on the number of examples of the class and the number of overall examples covered by that condition. Thus in Rule 2A, Change in Labor Force Participation being near 0 or decreasing is much stronger evidence that a country is “Asian-Unlike” than if its Agricultural Land Area is not between 30% and 50%.

The “total” and “unique” weights for each rule show respectively how many of the countries were described by the entire rule, and how many were described by that rule and none of the other ones learned. In each of the classes, Rule A is a “heavyweight” rule, i.e., it covers the majority of the examples of the class. Rule B, then, accounts for the countries not described by Rule A.

Rule 1A, in particular, gives a concise characterization of what Albania and Romania have in common with most of the Far Eastern nations. In addition to the defining condition of change in labor force participation, they also share a certain commonality in terms of industrial labor force, population growth rate and life expectancy.

Rule 2A describes most of the countries that do not follow this Asian model. They can be characterized by low population densities, decreases in labor force participation, and a lack of moderate values in many of the other attributes.

To try to better encapsulate this knowledge, we then took the learning a step further by applying “rule optimization” - using AQ to learn from the rules it created earlier. In doing so, AQ simplified the rules and diminished the effects of tautological conditions. The optimization of the above characteristic rules into discriminant ones generated the following:

Class is Class1 (Asian-Like) if:
A.1.Change in Labor Force Participation • slight_gain,
or
B.1.Working Age Population • 64%,
2.Life Expectancy is in 60s.

Class is Class2 (Asian-Unlike) if:
A.1.Life Expectancy is not in 60s,
2.Change in Labor Force Participation is near 0 or decreasing,
or
B.1.Percentage of Labor Force in Industry • 40.

In this case, the key elements to look for were found to consist of the change in labor force participation, life expectancy, working age population, and industrial labor force percentage. These characteristics were found to most directly link conditions in Albania and Romania to those in the nations of the Far East, and to distinguish those in the rest of Eastern Europe from them.

The second experiment involved a larger set of attributes from the database and a larger set of countries, based upon the suggestion of a domain expert working with the World Bank. For this experiment, we defined seven country groups (North America (USA and Canada), Western Europe (including Great Britain and Scandinavia), Southern Europe (the Iberian and Hellenic peninsulas), Northeastern Europe (Eastern Europe north of the Balkans), Southeastern Europe, Latin America and the Far East). In each of these regions, one or more “core countries” were selected as being typical of the region.

The goals of this experiment were to discover some characteristics of the various regions in terms of quality of life, organizational level, way of life, etc.; to detect interesting differences between the regions; and to examine the behavior of countries outside the core groups and analyze how they compared to the more representative nations of their region.

First we applied AQ15 to the individual regions, both through rule learning and rule optimization, to learn what characterizes them and what differentiates them from other parts of the world. Some strong trends were discovered. South American countries (all of the Latin American core countries were South American) could be characterized by very high population growth and fertility rates, high infant mortality and low per-capita GNP and life expectancy. Western European core countries had high per-capita GNP, low infant mortality 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. Southern Europe showed low allocation of resources to education, moderate per capita GNP and low fertility. The core countries in the Far East were characterized by a very low death rate.

We then applied the ATEST operator to these characterizations by applying them to some of the other countries in the region and examining how well the descriptions matched them. The "core" countries had tended to be moderate-to-large ones in their regions – representative while not dominating. The additional countries were smaller or larger, or were regarded as not being typical of their region (for example, Italy was considered to be closer in nature to its Western European neighbors than to the Mediterranean countries with whom it was geographically lumped. Similarly, Mexico was thought to have more in common with South American countries than with the other large nations of North America.)

We learned that the South American pattern was very strong throughout the continent and up to Mexico. Paraguay deviated farthest from it, but did not come close to fitting any other patterns with the possible exception of the Southeastern European pattern. Developing East Asian nations also tended toward the South American model. As expected, cultural similarity outweighed geographical similarity with respect to Italy and Mexico and the People's Republic of China, the latter more closely resembling the formerly Communist countries in Southeastern Europe than its Far Eastern neighbors. Interestingly, Canada more closely resembled the developed nations of the Far East than it did the United States. Perhaps due to being isolated by water, Iceland and Ireland did not resemble Western and Northern Europe as much as other test countries from the region did, including non-aligned Switzerland. (The United Kingdom, as part of the core training set, had already had its characteristics incorporated into the knowledge base.)

5 Discussion of Results and Future Research

The results from these initial experiments have been encouraging. A few simple operations have generated results that were interesting and sometimes unexpected. The discovered knowledge showed evidence that would confirm some hypotheses about the countries of the world, while providing evidence that contradicted others.

Clearly, this has not been anything near an exhaustive search through the data, but rather an initial foray into a rich area of knowledge. We intend to progress to experiments with new and larger sets of data from this database, and also to work with similar data from other sources.

Countries from other parts of the world such as the Middle East, Sub-Saharan Africa, South Asia, Oceania and the Caribbean can be incorporated into future work.

We also plan to begin constructing and working with derived attributes that may be better suited toward pointing out patterns in the data, and start to incorporate temporal factors by including data taken from different years and using statistical and machine learning-based programs to determine which years and differentials thereof have the greatest potential for leading us to interesting discoveries.

INLEN is also growing, and as such will contain more tools for different forms of knowledge discovery. From the current version of the system, which can be applied to larger data sets than before, further expansion is taking place. Operators are being added for attribute selection and construction, for example selection, for discovering relationships within the knowledge base, for knowledge representation and use by decision trees, and for statistical processing. The entire system is being also converted into a platform-independent tool that will allow for direct interaction with more databases.

The results described here provide further support to the hypothesis (Michalski et al, 1992a) that a tool such as INLEN can be used for making discoveries in data by non-experts in the data domain. By putting a robust knowledge discovery tool whose operation can easily be learned into the hands of domain experts, the experts will be able to conduct wide or narrow searches through databases in order to make discoveries that they will find interesting and important.

Acknowledgments

The author thanks the World Bank for providing the data used in this research and Tibor Vamos for both his insights in approaching this domain and acting as a liaison between the INLEN development team and the World Bank. The author also thanks Eric Bloedorn for his comments and criticism.

This research was conducted in the Center for Artificial Intelligence at George Mason University. The Center's research is supported in part by the Advanced Research Projects Agency under the grant No. N00014-91-J-1854, administered by the Office of Naval Research, and under the grant No. F49620-92-J-0549, administered by the Air Force Office of Scientific Research, in part by the Office of Naval Research under the grant No. N00014-91-J-1351, and in part by the National Science Foundation under the grant No. IRI-9020266.

References

Baskin, A.B. and Michalski, R.S., “An Integrated Approach to the Construction of Knowledge-Based Systems: Experiences with ADVISE and Related Programs,” in *Topics in Expert System Design*, Guida, G. and Tasso, C. (eds.), Elsevier Science Publishers B. V., Amsterdam, 1989, pp. 111-143.

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

Cuneo, R.P, “Selected Problems of Minimalization of Variable-Valued Logic Formulas,” Master’s Thesis, Department of Computer Science, University of Illinois, Urbana, 1975.

Greene, G., “Quantitative Discovery: Using Dependencies to Discover Non-Linear Terms,” Master’s Thesis, Department of Computer Science, University of Illinois, Urbana, IL, 1988.

International Intelligent Systems, Inc., “User’s Guide to AURORA 2.0: A Discovery System,” Fairfax, VA, International Intelligent Systems, Inc., 1988.

Katz, B., Fermanian, T.W. and Michalski, R.S., “AgAssistant: An Experimental Expert System Builder for Agricultural Applications,” Report No. UIUCDCS-F-87-978, Department of Computer Science, University of Illinois, Urbana, IL, Oct. 1987.

Kaufman, K., Michalski, R.S. and Kerschberg, L., “Mining For Knowledge in Data: Goals and General Description of the INLEN System,” in Piatetsky-Shapiro, G. and Frawley, W.J. (Eds.), *Knowledge Discovery in Databases*, AAAI Press, Menlo Park, CA, 1991, pp. 449-462.

Kerschberg, L. (ed.), Expert Database Systems: Proceedings from the First International Workshop, Benjamin/Cummings Publishing Company, Menlo Park, CA, 1986.

Kerschberg, L. (ed.), Expert Database Systems: Proceedings from the First International Conference, Benjamin/Cummings Publishing Company, Menlo Park, CA, 1987.

Kerschberg, L. (ed.), Expert Database Systems: Proceedings from the Second International Conference, George Mason University, Fairfax, VA, 1988. (to appear in book form, Benjamin/Cummings Publishing Company, Menlo Park, CA, 1988.

Michalski, R.S. and Baskin, A.B., “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, 1983, pp. 256-258.

Michalski, R.S., Baskin, A.B. and Spackman, K.A., “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, 1982, pp. 792-796.

Michalski, R.S., Baskin, A.B., Uhrik, C. and Channic, T., “The ADVISE.1 Meta-Expert System: The General Design and a Technical Description”, Report No. UIUCDCS-F-87-962, Department of Computer Science, University of Illinois, Urbana, IL, 1987.

Michalski, R.S., Kerschberg, L., Kaufman, K. and Ribeiro, J., “Searching for Knowledge in Large Databases,” *Proceedings of the First International Conference on Expert Systems and Development*, Cairo, Egypt, April 1992.

Michalski, R.S., Kerschberg, L., Kaufman, K. and Ribeiro, J., “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, August 1992, pp. 85-113.

Michalski, R.S., Mozetic, I., Hong, J. and Lavrac, N., “The AQ15 Inductive Learning System: An Overview and Experiments,” Report No. UIUCDCS-R-86-1260, Department of Computer Science, University of Illinois, Urbana, IL, 1986.

Michalski R.S., and Stepp, R.E., “Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983.

Michalski, R.S., Stepp, R.E. and Diday, E., “A Recent Advance in Data Analysis: Clustering Objects into Classes Characterized by Conjunctive Concepts,” in *Progress in Pattern Recognition*, Vol. 1, L. N. Kanall and A. Rosenfeld (Eds.), New York: North-Holland, 1981, pp. 33-56.

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

Spackman, K.A., “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, 1983.

Stepp, R.E., “A Description and User’s Guide for CLUSTER/2, a Program for Conceptual Clustering,” Department of Computer Science, University of Illinois, Urbana, IL, 1983.

Stepp, R.E., “Conjunctive Conceptual Clustering: A Methodology and Experimentation,” Ph.D. Thesis, Department of Computer Science, University of Illinois, Urbana, IL, 1984.