Addressing Knowledge Discovery Problems in a Multistrategy Framework

Kenneth A. Kaufman

Machine Learning and Inference Laboratory,
George Mason University,
Fairfax, Virginia, 22030, USA
kaufman@aic.gmu.edu

Abstract

This paper discusses a methodology for multistrategy data analysis based on the application of diverse learning and discovery programs and tools and how it approaches some of the difficulties posed by the knowledge discovery task. Research in the area of integrated learning systems has led to the development of INLEN, an intelligent assistant for discovering knowledge in large databases. The architecture of INLEN is based on the interaction of a number of *knowledge generation operators* – manifestations of diverse learning tools within a uniform environment. Examples of the system's application to databases consisting of world economic and demographic facts demonstrate its operation. During its development, INLEN has encountered problems inherent in the application of symbolic learning programs to database analysis that do not appear in the laboratory environment; such problems are described, and the responses to these problems that have been built into INLEN are discussed.

Introduction

As the amount of electronically available information has grown, it has become both critically important and increasingly difficult to analyze the data to derive desired knowledge from them. Traditionally, tools for data analysis have employed 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 – 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.

Because of this, the machine learning community has taken an interest in the problem of knowledge extraction from databases. Machine learning approaches can overcome some of the limitations inherent in traditional data analysis methods. For instance, constructive induction and deduction methods can improve the data description space based on the nature of the data itself, the knowledge learned by the discovery system, and/or the background knowledge provided by the domain expert. Symbolic learning methods have the advantage of representing their knowledge in such a way that it is very easy for users to understand and explain the meaning of the discovered knowledge.

Machine learning researchers have developed an assortment of domain-independent programs that can each perform a narrow set of symbolic learning tasks. The weakness of these programs is the fact that they are so task-specific. No conceptual clustering program, for example, can generate equations governing quantitative data, create rules distinguishing between classes of objects, select representative examples from a larger database, or improve a ruleset based on new data.

In general, depending on the situation and the data itself, an analyst may be seeking:

- The most important factors influencing the observed behavior of a system or a process and their relationship to this behavior.
- The functional or logical dependencies that exist among concepts and attributes in the database.
- A means of determining whether a certain condition is present.
- The most illustrative examples of a given behavior.
- A listing of the elements of a system that are not behaving according to the assumed model.
- The best actions to be taken given the observed behavior.
- An understanding of how a system is changing over time, and what types of future behavior can be anticipated.
- A concise description of the data that will highlight the important trends or exceptions.

- An organization of the data into a useful hierarchy of categories.
- A meaningful consolidation of facts from different sources that can be of use in pattern discovery.
- A collection of elements extracted from a large pool of data that the analyst is likely to find interesting or useful.

Recent research in multistrategy learning (e.g., Michalski & Tecuci 1991; 1993) has attacked the problem of integrating diverse learning tools into composite systems whose wholes may exceed the sums of their parts, in order that single systems may provide many of these answers on request.

One such approach to this problem, called INLEN, applies an integrated system to the specific problem of knowledge discovery in databases (Kaufman, Michalski & Kerschberg 1991). It was designed to overcome some of the limitations of statistical data analysis by applying advanced methods of machine learning. Its architecture integrates database, knowledge base and machine learning technologies into a single package for data analysis and knowledge discovery. In doing so, it offers a data analyst a powerful tool for discovering patterns of non-statistical nature, determining logic-style data descriptions, and producing justifications or explanations of the discovered patterns (Michalski et al. 1992).

Symbolic learning programs, such as those that comprise many of INLEN's available tools, 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, this methodology is well-suited for many domains in which databases contain a large number of records and attributes, and in which the results of analysis must be understood by both experts and non-experts.

The conceptual architecture – employing a large set of specialized operators as tools that may be called upon when needed – has been employed in domains other than knowledge discovery in databases. For instance, CONDOR (Strat 1992) follows a similar philosophy in the domain of computer vision. Here, different operators (e.g., edge detection) are invoked by a heuristic-based control engine in an attempt to recognize different viewed objects.

Among other efforts to apply multistrategy methods to data analysis are several systems incorporating symbolic learning. For example, Alexander, Bonissone and Rau (1993) have developed a system for discovering knowledge that can be used to improve marketing strategies. It combines the C4.5 decision tree learning program (Quinlan 1990) with a statistical analysis system in order to extract information from a sales database. RECON (Simoudis et al. 1994) combines inductive and deductive reasoning into a general-purpose knowledge discovery environment. In comparison with these systems, INLEN focuses more on symbolic learning operators, and is designed for the incorporation of a large number of tools, rather than just one or two for inductive reasoning. By necessity, this leads to an increase in the complexity of the system's infrastructure and its knowledge base.

This paper describes the INLEN architecture, presents examples of INLEN's discovery process through its application to the analysis of several databases consisting of world economic and demographic facts, and then presents some of the problems inherent in applying multistrategy symbolic learning to knowledge discovery and the way in which they have affected the development of INLEN. We conclude by summarizing the development status of INLEN, the advantages and current limitations of this methodology, and outline the plans for future research.

The INLEN Methodology

The INLEN system integrates a database, knowledge base and machine learning technologies into a unified knowledge discovery environment (Figure 1). The system's components consist of a relational database, a knowledge base, and three sets of operators (Kaufman, Michalski & Kerschberg 1991). The database is used for storing data, and can be modified by the user through *data management operators.* The 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 is made up of *knowledge segments,* which store declarative, procedural, and exemplary knowledge about the application. This knowledge may include the domains of the attributes, background knowledge relevant to the problem, 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 discovery engine of the system consists of 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. These operators 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 through feature selection or constructive induction, testing a set of knowledge for consistency with respect to new data, and predicting values for missing data.

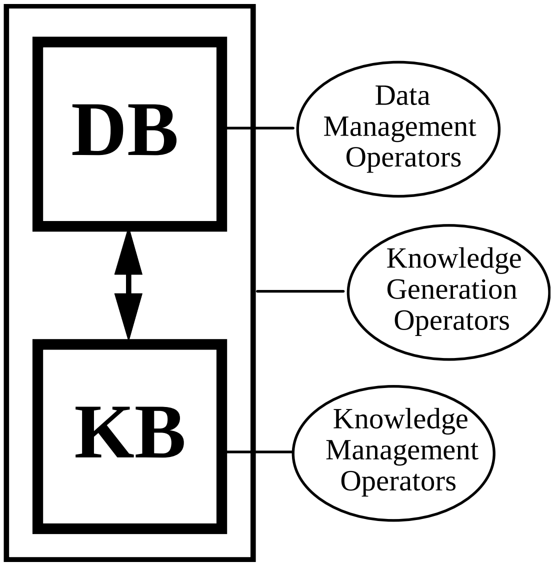
DB
KB
Data Management Operators
Knowledge Generation Operators
Knowledge Management Operators

Figure 1. High-Level Architecture of INLEN

The application of these operators is controlled by the user, while the system extracts the necessary information from the data and knowledge bases to complete the input to the learning operator. In the knowledge base more information is maintained than the operators might return in stand-alone mode; facts and links that might be useful to another operator in the future are stored in the output knowledge segments. For instance, if a rule learning operator is used to find conditions distinguishing between two sets of data records, INLEN will retain not only the discovered rules, but also the learning mode and parameters used, an estimate of the informational value of the rules and their component conditions, and links to the records that satisfy the rule. This information may be later accessed by a user or another operator, such as a prediction engine or a program for selecting representative examples from the dataset.

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 & Spackman 1982; Michalski & Baskin 1983; Spackman 1983).

Among the knowledge generation operators employed by INLEN are ones for rule generation (from sets or sequences of examples), decision tree generation (from examples or rules), equation generation (from quantitative and qualitative data), conceptual cluster and taxonomy formation, knowledge transformation (e.g., through abstraction, generalization, or incremental learning), representation space modification (through feature selection, constructive induction, and attribute quantization), event set generation (through example selection, simulation or prediction), relational analysis (through statistical and non-statistical metrics), knowledge testing for consistency and completeness, and concept visualization. The programs that form the foundations for these operators are cited and described individually in (Michalski et al. 1992).

These operators also represent various knowledge transmutations as are catalogued in Michalski’s Inferential Theory of Learning (Michalski 1993). For instance, learning classification rules from examples involves an inductive generalization transmutation; conceptual clustering is an act of inductive agglomeration; representation space modification can consist of generation, selection or abstraction transmutations; and knowledge-based prediction of missing values incorporates a similization transmutation.

An important feature of the 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 and pinpoint which type of knowledge is likely to be most useful. Two examples below illustrate the choices of parameters when using a knowledge generation operator.

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 either highly specific characterizations, short rules for discrimination, a set of maximally simple rules, or rules that tend to incorporate or avoid specific attributes if at all possible. The user can also create and specify a different set of preference criteria.

Another knowledge generation operator is based on an extension of 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 learn from the subset of the data consisting of only the attributes likely to be more relevant to the learning task and, in doing so, improve the efficiency of the learning process and decrease the likelihood of spurious knowledge being discovered. This operator can be applied 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. As such, it is useful in the building of decision trees or other procedural structures. 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 utility in classifying the examples.

Exemplary Application: Discovery in Economic and Demographic Domains

INLEN and the programs that have been adapted for use as its operators have been applied to learning and discovery problems in such domains as engineering design, disease diagnosis, intelligence gathering and economic analysis (e.g., Arciszewski et al. 1992; Michalski et al. 1992). 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 to an analyst the opposite changes in two related categories? 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 raw data.

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. Two examples of such experiments and results follow. The first illustrates the passage of knowledge from operator to operator as a concise description of the goal concept is formulated. The second illustrates how the background information in the knowledge base may expose anomalous data.

Distinguishing Between Two Regions' Development Patterns

An experiment that serves as an example of the linkage of different learning and discovery methods focused on distinguishing between development patterns in Eastern Europe and East Asia (Kaufman 1994). A sequence of operators consisting of attribute selection, conceptual clustering, rule generation based on set characterization, and rule optimization combined to generate the output conclusions. After the feature set was pared down to those economic indicators deemed more likely to differentiate the two regions, a conceptual clustering operator based on CLUSTER/2 (Michalski & Stepp 1983) 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 (Wnek et al. 1995), was then called upon twice – first in characteristic mode to categorize the commonalities among the Asian-like countries (those above their regional thresholds) and among the European-like countries (those below their regional thresholds), and then in discriminant rule-optimizing mode to condense the lengthy characterizations from the previous set (4-8 conditions per rule) 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. *(2 countries)*

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.

Identification of an Unusual Example

Because INLEN's knowledge base maintains records of the training data that supports the discovered rules, this information can be used to group the records within a particular class. For example, when the AQ15c rule generation operator was called upon to characterize the 13 countries of South America, it came up with two distinct characterizations – one describing the majority of the countries, and the second describing the other four: Ecuador, French Guyana, Peru and Venezuela. By suggesting that these two subgroups of countries may have significant commononalities among themselves, the program has proposed a classification scheme for further investigation.

Another experiment clearly indicates 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 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 additional discovery.

INLEN discovered several rules from the World Factbook PEOPLE database characterizing the countries with low (less than 1% per year) population growth (Kaufman & Michalski 1996a). One of the rules had three conditions that together were sufficient to distinguish 19 low growth countries from all of the countries with higher population growth rates. The rule is shown here with three weights attached to each condition: *Pos* represents the number of positive examples (countries with population growth rates below 1%) satisfying the condition, *Neg* represents the number of negative examples (countries with population growth rates above 1%) satisfying the condition, and *Supp*, defined as Pos / (Pos + Neg) in percent, represents an approximate measure of the degree of support that the condition alone provides for the conclusion that a country might have a population growth rate below 1%.

	Conditions characterizing Countries with Population Growth Rates below 1%:	Pos	Neg	Supp
1	Birth Rate = 10 to 20 or over 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 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; using the links in the knowledge base, one may examine the 19 countries involved. Such an inspection points out that 18 have birth rates below 20, while only one, Malawi, has the 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 can then explore the conditions unique to Malawi and hypothesize where else they might take place in the future.

Database Analysis Based on Multistrategy Learning

The previous sections discussed and provided examples of the architecture and major components of INLEN. Through the sequential application of KGOs, a user can link different learning and discovery programs into a stream of tasks. Many of the operators are based on machine learning programs that were not designed for the analysis of large databases, but were instead written based on the assumption that they would be operating in a supervised learning environment. As a result, they often have the following characteristics:

(1) Their inference methods are tailored to simple nominal or linear feature domains without rich domain knowledge.

(2) It is assumed that most of an example's attribute values will be provided.

(3) All of the information relevant to a problem will be available in one location.

Unfortunately, these conditions do not always hold when exploring a real-world database. The attributes in the data may be based on complex hierarchies, lattices and gradations of concepts. Many of the fields in the data may be missing due to incomplete information. And the goal knowledge may be only attainable through the extraction and combination of information from multiple sources.

These problems have been addressed by the knowledge generation operators in INLEN. INLEN supports learning with complex structured data types. Not only can a user define a data attribute to be hierarchically structured, one can also designate nodes within the hierarchy as *anchor nodes* – especially significant foci for learning, generalization and specialization (Kaufman & Michalski 1996b). The justification behind the selection of such nodes is that we tend to weight the significance of nodes in a classification hierarchy unevenly. For instance, a red delicious is an apple, which is a kind of fruit, which is a type of food. In everyday usage, we will not think of a given red delicious at each of those different levels of abstraction with equal frequency.

Cognitive scientists speak of *basic* level nodes within a generalization hierarchy whose children share many sensorially recognizable commonalities (Rosch et al. 1976). Other factors that help to characterize a node's utility compared to those at higher or lower levels of abstraction are concept typicality (how common are the features of this concept among its sibling concepts), and the context in which the concept is being used (Klimesch 1988; Kubat, Bratko & Michalski 1996). Each of these factors affects the selection of a particular level of abstraction in making descriptions.

By encoding the relative utility of the nodes into the knowledge representation of a discovery system, the system can present discovered knowledge that focuses on the more useful levels of abstraction when possible. We will typically prefer to see the classification rules “An object belongs to Class 1 if it is an apple” or “An object belongs to Class 1 if it is a fruit” instead of “An object belongs to Class 1 if it is a red delicious” or “An object belongs to Class 1 if it is food.”

Another important feature in INLEN’s structured data representation is its ability to work with multiple views of the data (Kaufman & Michalski, 1996b). Consider an application in which a marketing specialist is trying to target the customers who are most likely to be interested in a new product. A customer database may have extensive information including the model of the automobile driven by the customer. Automobiles may in turn be organized according to manufacturer, type (e.g., sedan, sports car, station wagon), price, year, etc. One may not know prior to a learning task which classification view will provide the most concise and useful knowledge. Also, more than one view may generate the best rules for determining whether a customer is likely to buy the product. A decision rule, for example, may include the condition that likely buyer will often drive European station wagons, while the specification of the type of vehicle or manufacturer alone may not give an accurate representation of the customer’s propensities. INLEN’s generalization engine automatically selects from the possible ways of expressing a set of attribute values (in this case the automobile models that are European station wagons) a concise representation of the knowledge.

INLEN’s knowledge generation operators have also been enhanced to cope with the problem of incomplete data sets. Experiments with sparse data exposed some of the limitations of these operators, leading to modifications in which the logical implications of unknowns are more rigorously encoded into the learning algorithm. For example, in the AQ family of programs, attributes with unknown values are represented as having all of their values under consideration. A stipulation in earlier versions of the program required that generalizations of examples with unknown values for some attributes maintain consistency by permitting those attributes to take any value. While this will guarantee rigorous consistency with the data, trouble arises when many examples have only a few known attribute values. Generalizing just a few of them together creates a situation in which nothing can be assumed about any feature.

As an example of a domain in which this may be a real problem, intelligent agents are being developed to scan text and summarize it based on key words in several categories of interest to the user. Articles often will only contain key words in a few of these categories, leading to a very empty database. The relationships among entries in the various categories will often be tenuous ones given that for much of the data, one or more of these fields will be empty. In such a domain a discovery program must be able to sift through the information that is present without getting lost in that which is missing. At the expense of some additional computational complexity, the knowledge generation operators in INLEN have been modified in such a way that they can now generate knowledge consistent with what facts have been made available, without adhering to the assumption that unknowns must be generalized to take on all values. With the relaxation of this condition, a learning engine can detect more substantial relationships.

Another aspect of this research approaches the problem of knowledge extraction from distributed sources. The INSIGHT program (Ribeiro, Kaufman & Kerschberg. 1995) is being developed as an operator to perform a knowledge-driven search through multiple databases. The combinatorial cost of combining separate data sets is avoided through INSIGHT’s mechanism of finding relationships between a database and the knowledge generated from another database.

In order to facilitate the interface with INSIGHT and other operators, INLEN maintains information on the database records relevant to each rule in its knowledge base. As was shown above in the population growth example, a use for these links is to enable the identification of significant clusters or exceptions. Another use is to allow the measurement of the degree of match between a rule and a set of records in a second database matching a given set of conditions. A high degree of match may suggest a linkage between the two concepts. For example, a rule describing the climate of a country may be cross-referenced with a database of natural disasters. If a class of natural disaster occurred in a set of countries similar to the set of countries covered by the climate rule, it may suggest a relationship between the climate and that kind of disaster.

Conclusion

The INLEN methodology is based on the application of a wide variety 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 this methodology 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.

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

This report describes a work in progress in which new capabilities have been added over the course of its development. Among the limitations of the current implementation are the facts that it still awaits integration with statistical data analysis methods and that there is still too much reliance on the data analyst for guidance in the selection of operators and the setting of parameters. In particular, the methodology is strongly human-driven, and an area of ongoing research is seeking to develop a reliable means to automate the discovery process. 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 current research include the development of methods for automated data abstraction and the analysis of distributed databases. Future projects include 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 over a 10 year period).

Another enhancement will be the development and incorporation of a high-level language for knowledge discovery. Such a tool will attack one of the limitations of the system mentioned above, namely that the knowledge discovery process must be closely supervised at present. With the addition of a knowledge discovery language, a user will be able to program in sequences of operators, along with instructions or heuristics detailing what conditions should 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."

In the presented examples, INLEN's knowledge generation operators were applied to databases in various economic and demographic domains. The system's searches have unearthed some surprising facts. The experiments described here illustrate some of the potential capabilities of the application of integrated learning strategies to large databases and indicate that such an application has the potential for determining important but heretofore unknown findings within the data.

Whether this or any other multistrategy architecture is used for data analysis, the task of database exploration presents certain problems not often faced by the learning components. When learning tools are turned toward such knowledge discovery problems, they should be equipped to represent rich domain knowledge, handle very sparse data sets, and be prepared to integrate data from different sources. This paper has described approaches to each of these problems.

The goal of all of these features is to facilitate INLEN's compatibility with real-world databases. The tools that serve as knowledge generation operators can then perform at a high level on real-world problems as well as on carefully supervised data sets, while their integration results in a multistrategy system not limited to a narrow class of discovery tasks.

Acknowledgments

The author thanks Ibrahim Imam and Eric Bloedorn for their constructive suggestions regarding earlier versions of this paper.

This research was conducted in the Machine Learning and Inference Laboratory at George Mason University. The Laboratory's research is supported in part by the Defense Advanced Research Projects Agency under Grant No. N00014-91-J-1854 administered by the Office of Naval Research, in part by the Defense Advanced Research Projects Agency under Grants No. F49620-92-J-0549 and F49620-95-1-0462 administered by the Air Force Office of Scientific Research, in part by the Office of Naval Research under Grant No. N00014-91-J-1351, and in part by the National Science Foundation under Grants No. DMI-9496192 and IRI-9020266.

References

Alexander, W.P., Bonissone, P.P. and Rau, L.F. 1993. Preliminary Investigations into Knowledge Discovery for Quick Market Intelligence. In Proceedings of AAAI-93 Workshop on Knowledge Discovery in Databases, 52-60. Washington, DC.

Arciszewski, T., Bloedorn, E., Michalski, R.S., Mustafa, M. and Wnek, J. 1992. Constructive Induction in Structural Design. *Reports of the Machine Learning and Inference Laboratory*, MLI-92-7, George Mason University, Fairfax, VA.

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. In Proceedings of the Second International Workshop on Multistrategy Learning, 188-203, Harpers Ferry, WV..

Central Intelligence Agency 1993. *1993 World Factbook*.

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

Kaufman, K. and Michalski, R.S. 1996a. A Multistrategy Conceptual Analysis of Economic Data. In Proceedings of the Fourth International Workshop on Artificial Intelligence in Economics and Management, Tel Aviv.

Kaufman, K. and Michalski, R.S. 1996b. A Method for Reasoning With Structured and Continuous Attributes in the INLEN-2 Knowledge Discovery System. *Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR (to appear).

Kaufman, K. Michalski, R.S., and Kerschberg, L. Mining for Knowledge in Databases: Goals and General Description of the INLEN System. Chapter in Piatetsky-Shapiro, G. and Frawley, W.J. (eds.) *Knowledge Discovery in Databases*, Cambridge, MA: AAAI Press, 449-462

Klimesch, W. 1988. *Struktur und Aktivierung des Gedaechtnisses. Das Vernetzungsmodell: Grundlagen und Elemente einer uebergreifenden Theorie*. Bern: Verlag Hans Huber.

Kubat, M., Bratko, I. and Michalski, R.S. 1996. A Review of Machine Learning Techniques. Chapter in *Methods and Applications of Machine Learning and Discovery*. Forthcoming

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

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

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

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* 1:85-113.

Michalski, R.S., and Stepp, R.E. 1983. Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5:396-410.

Michalski, R.S. and Tecuci, G. eds. 1991. *Proceedings of the First International Workshop on Multistrategy Learning*. Harpers Ferry, WV.

Michalski, R.S. and Tecuci, G. eds. 1993. *Proceedings of the Second International Workshop on Multistrategy Learning*. Harpers Ferry, WV.

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

Ribeiro, J.S., Kaufman, K.A. and Kerschberg, L. 1995. Knowledge Discovery From Multiple Databases. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 240-245, Montreal PQ.

Rosch, E., Mervis, C., Gray, W., Johnson, D. and Boyes-Braem, P. 1976. Basic Objects in Natural Categories. *Cognitive Psychology*, 8: 382-439.

Simoudis, E., Livezey, B., and Kerber, R. 1994. Integrating Inductive and Deductive Reasoning for Database Mining. In Proceedings of AAAI-94 Workshop on Knowledge Discovery in Databases, 37-48, Seattle, WA.

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.

Strat, T.M. 1992. *Natural Object Recognition*. New York: Springer Verlag.

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, George Mason University, Fairfax, VA.