Multistrategy Data Exploration Using the INLEN System: Recent Advances

Ryszard S. Michalski* and Kenneth A. Kaufman
George Mason University, Fairfax, VA

* Also with the Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

Recent advances in the development of the INLEN system for multistrategy data exploration are briefly reviewed. These advances include the development of a meta-level language for data mining and knowledge discovery, called *knowledge generation language* (KGL), and the employment of a new type of attributes, called *structured attributes*. These features are illustrated by an example concerned with determining economic and demographic patterns in a database containing facts about the countries of the world. The results demonstrate a high utility of INLEN for data mining and knowledge discovery.

Introduction

The availability of very large volumes of data in the electronic form has created a problem of deriving from them useful, task-oriented knowledge. Traditional data analysis techniques, which include statistical and numerical methods, are oriented primarily toward the extraction of quantitative data characteristics, and as such have inherent limitations. For example, statistical techniques cannot produce conceptual descriptions of dependencies among data items or explain reasons why these dependencies exist. Nor can they justify found relationships in the form of higher-level logic-style descriptions, or draw an analogy between the discovered regularity and a regularity in another domain.

To address such tasks as above, a data exploration system has to be equipped with a substantial amount of background knowledge, and be able to perform symbolic reasoning involving that knowledge and input data. To this end, researchers have turned to ideas and methods developed in machine learning that can acquire new knowledge using facts and background knowledge. These and related efforts have led to the emergence of a new research area called data mining and knowledge discovery (e.g., Michalski, Baskin and Spackman, 1982; Zhuravlev and Gurevitch, 1989; Michalski, 1991; Michalski et al, 1992; Van Mechelen et al, 1993; Fayyad et al, 1996; Evangelos and Han, 1996).

This paper briefly reviews a methodology for *conceptual data exploration*, by which we mean the derivation of high-level concepts and descriptions from data. The methodology, implemented in the INLEN system, integrates machine learning, data base and knowledge bases technologies in order to assist data analysts in determining task-oriented knowledge from data. The term *task-oriented* emphasizes the fact that an exploration of the same data may produce different knowledge; therefore, the methodology tries to connect the task at hand with the way data are explored. Such task-orientation naturally requires a multistrategy approach, because different tasks may require different data exploration and knowledge generation operators.

The aim of the methodology is to produce knowledge in a form that is close to that an expert might produce analyzing the same data. Such a form may include combinations of descriptions of different types, with a constraint that the end result should be easy to understand and interpret by an expert in the given domain, i.e., the produced descriptions should satisfy the "principle of comprehensibility" (Michalski, 1994). Our first efforts in developing such a methodology have been implemented in the system INLEN-1 (Michalski et al, 1992). The most recent version of the system, INLEN-3, combines a range of machine learning methods and tools with some traditional data analysis operators, in order to provide a user with powerful capabilities for data explorations and derivation of diverse kinds of knowledge from a database. Among the machine learning capabilities implemented in the current INLEN system are the ability to learn different types of rules from examples, conceptual clustering and hierarchy generation, automatic selection of most relevant attributes, rule editing by an expert, and automatic application of the learned or acquired rules to new cases (Michalski and Kaufman, 1997).

Important aspects of the INLEN approach that distinguish it from the most of existing data mining systems are that it employs a wide range of *knowledge generation operators* and is capable of *knowledge-intensive* discovery. It allows a user to incorporate and utlilize various aspects of domain knowledge, and is constructed in such a way that knowledge generated by one operator can serve as background knowledge for subsequent operators.

The recent version of INLEN includes several novel ideas and tools, specifically, an initial implementation of meta-level knowledge generation language, KGL, and methods for inductive reasoning with *structured* attributes (whose domains are hierarchically ordered, as opposed to conventional linearly ordered or unordered). By using KGL, a data analyst may plan complex experiments, in which sequences of very high level data mining and knowledge discovery operators are automatically executed. Individual operations can be pre-conditioned upon the results from previous ones. The following sections briefly describe the INLEN methodology, its recent advances, and an example of INLEN's application to discovery of economic and demographic patterns in a large database.

Integration of Multiple Operators in INLEN

To make data exploration operators applicable in sequences in which the output from one operation is an input to another one, programs performing these operators need to be integrated into one system. This idea underlies the INLEN system that integrates a variety of machine learning programs, statistical data analysis tools, a database, a knowledge base, inference procedures, and various supporting programs under a unified architecture and simple graphical interface (the name INLEN is derived from **in**ference and **le**ar**n**ing). The knowledge base is used for storing, updating and applying rules and other forms of knowledge that may be employed for assisting data exploration, and for reporting results from it.

The general architecture of INLEN is presented in Figure 1. The system consists of a database (DB) connected to a knowledge base (KB), and a set of operators. The operators are divided into three classes:

- *DMOs:* Data Management Operators, which operate on the database. These are conventional data management operators that are used for creating, modifying and displaying relational tables.
- *KMOs:* Knowledge Management Operators, which operate on the knowledge base. These operators play a similar role to the DMOs, but apply to the rules and other structures in the knowledge base.
- *KGOs*: Knowledge Generation Operators, which operate on both the data and knowledge bases. These operators perform symbolic and numerical data exploration tasks. They are based on various machine learning and inference programs, on conventional data exploration techniques, and on visualization operators for displaying graphically the results of exploration.

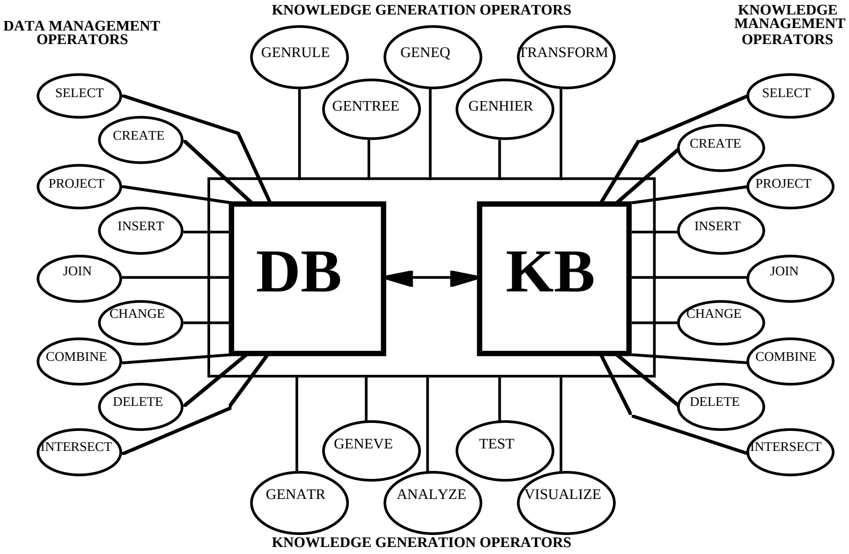


Figure 1. Architecture of the INLEN system for multistrategy data exploration.

The KGOs are the heart of the INLEN system. To facilitate their use, the concept of a *knowledge segment* was introduced (Kaufman, Michalski and Kerschberg, 1991). A knowledge segment is a structure that links one or more relational tables from the database with one or more structures from the knowledge base. KGOs can be viewed as modules that perform some form of inference or transformation on knowledge segments and, as a result, create new knowledge segments. Knowledge segments may be both inputs to and outputs from the

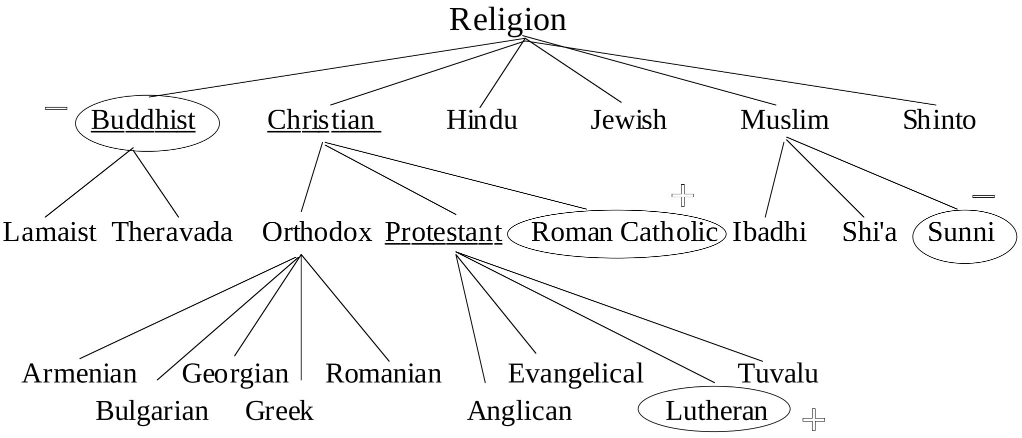
KGOs. Thus, they facilitate the passage of data and knowledge from one knowledge generation operator to another.

The execution of a KGO usually requires background knowledge, and is guided by control parameters (if some parameters are not specified, default values are used). The background knowledge may contain general knowledge, as well as knowledge specifically relevant to a given application domain, such as a specification of the value sets and types of attributes, the constraints and relationships among attributes, initial rules hypothesized by an expert, etc. KGOs currently implemented or under implementation in INLEN include those for rule generation, decision structure creation, equation discovery, creation of concept hierarchies, transformation of knowledge, modification of the representation space or example set, generating a numerical analysis of data, testing knowledge, and visualizing data or knowledge.

Summarizing, INLEN integrates a large set of operators for performing various types of tasks on the data base, on the knowledge base, or on the data and knowledge combined.

Learning Rules with Structured Attributes

In addition to conventional symbolic and numerical attributes, INLEN supports a new kind of attributes, called *structured*. Such attributes have value sets ordered into hierarchies (Michalski, 1980). In order to utilize structured attributes during inductive learning, new inductive generalization rules have been developed.



Circled nodes marked by + and – are values occurring in positive and negative examples, respectively. Underlined nodes are anchor nodes .

Figure 2. The domain of a structured attribute "Religion".

To illustrate the problem, consider the structured attribute "Religion" shown in Figure 2. Each non-leaf node denotes a concept that is more general than its children nodes. These relationships need to be taken into consideration when generalizing facts involving different religions. Suppose that the concept to be learned is exemplified by the statements, “There is a Lutheran member in a group,” “There is a Roman Catholic member in the group,” “The group does not include any Buddhists” and “The group does not include any Sunni Muslims.” There are many consistent generalizations of these facts, for example, “The group consists of Lutherans and Roman Catholics”, “The group consists of Protestants and Roman Catholics”, “The group consists of Christians”, “The group consists of people who are not Buddhist or Muslim”, or “The group consists of people who are not Buddhist or Sunni Muslim.”

The first statement above represents the *maximally specific description*, the last statement represents the *maximally general description*, and the remaining ones represent intermediate levels of generalization. A problem arises in determining the generalization that is most appropriate for the given situation. We approach this problem by drawing insights from human reasoning, namely that people prefer generalizations to certain intermediate levels in a generalization hierarchy (e.g., Rosch et al, 1976), and that the choice of the generalization level is based on the concept typicality and the context in which the concept is being used.

To provide a mechanism for capturing such preferences, INLEN allows a user to define *anchor nodes* in a generalization hierarchy. Such nodes reflect the interests of a given application (Kaufman and Michalski, 1996). To illustrate this idea, consider Figure 2 again. In this hierarchy, Lutheranism is a denomination of Protestantism, which is a type of Christianity. In everyday usage, we will describe a Lutheran by one of these three terms depending on the context. We can designate the most appropriate term for the given problem domain as an anchor node. Using information about anchor nodes, task-oriented criteria for hypothesis selection can be specified. The above reasoning applies to structuted attributes used as independent (input) variables.

In many applications, it is desirable to use structured attributes also as dependent (output) variables. For example, when deciding which personal computer to buy, one may first decide the general type of the computer—whether it is to be IBM PC-compatible or Macintosh-compatible. After deciding the type, one can focus on a specific model of the chosen type. The above two-level decision process is easier to execute than a one-level process in which one has to directly decide which computer to select from a large set.

For this reason, INLEN supports also the use of structured attributes as dependent (output) variables. Structured dependent attributes represent hierarchies of decisions or classifications that can be made about an entity. Through the use of structured output attributes, INLEN’s learning module can determine rules at different levels of generality. When a dependent variable is structured, the learning operator focuses first on the top-level values (nodes), and creates rules for them. Subsequently, it creates rules for the descendant nodes in the context of their ancestors. This procedure produces decision rules that are simpler and easier to interpret than rules learned with a flat (nominal) organization of the decision attribute.

The KGL Language for Knowledge Discovery

The previous versions of INLEN made it easy for the user to apply various operators, but their application required explicit involvement of the user at every step. Specifically, the analyst had to inspect the results of each step of the process in order to determine which operator to apply next. Such a process can be laborious and time-consuming, and the analyst will be prone to errors.

If one could define general rules for controling the application of operators, then these rules could be embedded within a discovery system, and automatically applied to knowledge discovery tasks. Some control rules may be highly domain- or task-dependent. Therefore, a user should be able to articulate her/his needs and interests to the system, so that the system can automatically perform desirable sequences of operators.

To this end, we have initiated the development of KGL (**K**nowledge **G**eneration **L**anguage), a meta-level language for specifying knowledge discovery experiments using INLEN operators (Kaufman and Michalski, 1997). Specifically, the language allows the user to create plans of experiments and specify instructions for automatically guiding the system through sequences of steps and contingencies. The language is designed to support writing simple KGL programs that could perform very complex data mining and knowledge discovery tasks. These programs may be executed once, periodically, or on the occurrence of some conditions or events, such as the perception of some pattern in the data or some property in the acquired knowledge, or an infusion of new information into a database. The requirements of such a language include:

- Invoking different types of programs for learning and knowledge discovery as single operators with user- or program-specified parameters.
- Looping and branching abilities similar to those found in conventional programming languages.
- Discrimination among the different properties and types of attributes in the database. The user should be able to classify the attributes into groups based on the importance, the type, the size of the attribute domain, etc.
- Discrimination among the different rules, rulesets, decision structures, etc. that make up a domain’s knowledge base. The user should be able to select rules based on their complexity, support by the data, typicality, etc.
- Data-driven control strategies, triggered by changes to a database beyond a given threshold level. Among the patterns that must be detectable are missing values and conflicts with the existing knowledge base.
- Knowledge-driven control strategies, triggered, for example, by the discovery of especially strong patterns or exceptions. The program must be able to examine discovered knowledge and identify some of the attributes of the knowledge itself.

Earlier efforts to develop a language to assist in knowledge discovery tasks have been almost exclusively logic-based, using Prolog-style queries (e.g., LDL). One exception is an extension to SQL, called M-SQL (Imielinski, Virmani and Abdulghani, 1996), which allows a user to query for certain types of rules and invoke a rule-generating operator. KGL differs from M-SQL in that it is able to call upon many different types of knowledge generation operators, and also in that it is designed to be less tightly coupled with SQL (although an SQL interface for invoking data management operators is planned). KGL more closely resembles a programming language than a query language in the facilities it offers for flow control. Details about the current implementation of KGL have been described in (Kaufman and Michalski, 1997).

An Illustrative Application: Seeking Demographic and Economic Patterns in a Database of World Facts

To exemplify the INLEN methodology, we will consider its application to the problem of searching for demographic and economic patterns in a world database. In this application, we used the PEOPLE dataset from the 1993 World Factbook, published by the Central Intelligence Agency. In INLEN's implementation of this database, the predominant religion of a country was defined as a structured attribute (much of the structure is shown in Figure 2).

In one experiment, INLEN was asked to characterize the dependence of population growth rate on the other attributes. One of the generated rules characterized 19 of the 55 countries with low population growth rates (less than 1% in 1993) in the following way:

Countries with Population Growth Rate < 1% are characterized by:

1 *Birth Rate* < 20 or *Birth Rate* > 50 per 1000 people

2 *Predominant Religion* is Orthodox or Protestant or Hindu or Shinto

3 *Net Migration Rate* < +20 per 1000 people

Notice how structuring the predominant religion attribute has made the rule easier to understand in comparison to how it would look if all the representative branches of Orthodox Christianity and Protestantism were listed.

When this rule was generated, it was noticed that the first condition was unusual. A low birth rate occurring in conjunction with a low population growth rate is expected, but how can a very high birth rate be explained? Upon further examination of the data, it was found that 18 of the 19 countries characterized by the rule had the expected low birth rate, but one, Malawi, had a birth rate above 50. Focusing INLEN upon Malawi revealed an explanation: the country had by far the world's highest negative net migration rate.

This kind of discovery can be made automatically through an application of KGL. To do so, appropriate background knowledge needs to be provided to the KGL program, either by a user or as a result of previous steps. In this case, such knowledge could include a specification of known relationships among given attributes, such as the existence of a positive correlation between population growth rate and birth rate in most of the countries.

Figure 3 presents an example of a KGL program that invokes INLEN operators that lead to the discovery of Malawi's unusual characteristics, and report them to the user. Terms in bold are calls to INLEN's operators; underlined terms are names of attributes provided in the database, or constructed by the KGL program.

```
open PEOPLE
CHAR(PopGrRate)
anomalous = 0
GENATR(name=unusual, type=boolean)
if #rules(PopGrRate, contradicts_BK(PopGrRate, BirthRate)) >0
    begin
    forall examples
        if contradicts_BK(PopGrRate, BirthRate)
            begin
            anomalous = anomalous + 1
            print CountryName
            unusual = true
            end
        else
            unusual = false
    print “Total: “, anomalous, “ unusual countries“
    CHAR(unusual)
    end
```

Figure 3. A KGL program that is able to discover Malawi’s unusual characteristics regarding population growth rate and birth rate.

The above program opens the PEOPLE dataset (which contains in this case demographic information about 190 countries), and then applies the CHAR operator to characterize the relationship between the population growth rates of different countries and other attributes in the database. Subsequently, it creates a new binary attribute, called “unusual”, that is appended to records in the PEOPLE database. This attribute is set to true if a given country contradicts the typical relationship between the birth rate and the population growth rate (i.e., the program’s background knowledge). The “unusual” countries are counted, and their names are listed. In the final phase, INLEN invokes the CHAR operator again in order to characterize the found unusual countries (in this case, just Malawi) in terms of other known attributes. The characterization serves as an explanation for the anomalous behavior (in this case, the world’s highest net outward migation rate).

Summary

The INLEN system integrates a wide range of data mining and knowledge discovery operators. Recent advances in the development of INLEN include two novel features, one concerned with the use of structured attributes, and the second with the implementation of the meta-language KGL.

Structured attributes provide a useful method for providing learning programs with background information about the domain of application. The structuring of attributes and the introduction of anchor nodes facilitate the process of detecting and expressing patterns at an appropriate level of generality.

The knowledge generation language, KGL, supports the planning of data mining and knowledge discovery experiments involving a variety of learning and knowledge processing operators. The initial, partial implementation of the language has shown that even simple programs in that language can execute quite complex data mining and knowledge discovery tasks.

The INLEN system provides a powerful environment for experiments in searching for solutions to practical learning and discovery tasks. The results from experiments done so far are very promising, and indicate that the system can be of high practical utility.

Acknowledgments

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

References

Evangelos S. and Han, J. (eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR, 1996.

Fayyad, U.M. Piatetsky-Shapiro, G. Smyth, P. and Uhturusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*, San Mateo, CA, AAAI Press, 1996.

Imielinski, T., Virmani, A. and Abdulghani, A., "DataMine: Application Programming Interface and Query Language for Database Mining," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR, 1996, pp. 256-261.

Kaufman, K. and Michalski, R.S., "A Method for Reasoning with Structured and Continuous Attributes in the INLEN-2 Knowledge Discovery System," *2nd International Conference on Knowledge Discovery and Data Mining*, Portland, OR, 1996b, pp. 232-237.

Kaufman, K. and Michalski, R.S., "KGL: A Language for Learning," *Reports of the Machine Learning and Inference Laboratory*, MLI 97-2, George Mason University, Fairfax, VA, 1997

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.

Michalski, R.S., "Inductive Learning as Rule-Guided Generalization and Conceptual Simplification of Symbolic Descriptions: Unifying Principles and a Methodology," *Workshop on Current Developments in Machine Learning*, Carnegie Mellon University, Pittsburgh, PA, 1980.

Michalski, R.S., "Searching for Knowledge in a World Flooded with Facts," in *Applied Stochastic Models and Data Analysis,* Vol. 7, pp. 153-l63, January, 1991.

Michalski, R.S., “Inferential Theory of Learning: Developing Foundations for Multistrategy Learning,” In Michalski, R.S. and Tecuci, G. (eds.), *Machine Learning: A Multistrategy Approach,* San Francisco: Morgan Kaufmann, 1994, pp. 3-61.

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. and Kaufman, K., “Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach,” Chapter in Michalski, R.S., Bratko, I. and Kubat, M. (eds.), *Machine Learning and Data Mining: Methods and Applications*, London, John Wiley & Sons, 1997 (to appear).

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*, 1, August 1992, pp. 85-113.

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

Van Mechelen, I., Hampton, J., Michalski, R.S. and Theuns, P. (eds.), *Categories and Concepts: Theoretical Views and Inductive Data Analysis*, London, Academic Press, 1993.

Zhuravlev, Y.I. and Gurevitch, I.B., “Pattern Recognition and Image Recognition,” Chapter in Zhuravlev, Y.I. (ed.), *Pattern Recognition, Classification, Forecasting: Mathematical Techniques and their Application*. *Issue 2*, Nauka, Moscow, pp. 5-72 (in Russian), 1989.