Exploiting Background Knowledge in Automated Discovery

John M. Aronis
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
aronis@cs.pitt.edu

Foster J. Provost
NYNEX Science and Technology
400 Westchester Avenue
White Plains, NY 10604
foster@nynexst.com

Bruce G. Buchanan
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
buchanan@cs.pitt.edu

Abstract

Prior work in automated scientific discovery has been successful in finding patterns in data, given that a reasonably small set of mostly relevant features is specified. The work described in this paper places data in the context of large bodies of background knowledge. Specifically, data items are connected to multiple databases of background knowledge represented as inheritance networks. The system has made a practical impact on botanical toxicology research, which required linking examples of cases of plant exposures to databases of botanical, geographical, and climate background knowledge.

Introduction

Discoveries made by computer programs have been characterized as human/computer discoveries because the discovery process is far from being completely automated (Valdes-Perez, 1995). One area where the human component has been vital is in guiding the discovery system based on background knowledge. In this paper we augment a standard inductive learning program by connecting data items to background knowledge represented as inheritance networks with role links and a limited form of non-monotonic inheritance, extending the ability of the program to make discoveries by using the semantics of the features describing the data items.

Representing Background Knowledge

Scientific domain knowledge takes on a rich, structured form. Prominent in any scientist's store of useful background knowledge are various taxonomies, categories, and relationships between concepts. To automate discovery using these forms of domain knowledge we must represent and reason about classes and relationships, and be able to bring the knowledge to bear on the discovery process. *Inheritance networks* are an efficient way to implement this kind of reasoning, because they can represent class structure and complex relational knowledge, yet can be navigated efficiently (Fahlman, 1979).

Figure 1 illustrates how some knowledge about plant families and their properties can be represented using standard inheritance network notation. A few records from a database of potentially toxic plant exposures and a small part of a botanical knowledge base are shown. Unlabeled arrows are *ISA links*, which can be interpreted as set inclusion. Thus, the link T. radicans $\rightarrow$ Toxicodendron means that every plant in the species T. radicans is also in the genus Toxicodendron. The link Toxicodendron $\rightarrow$ Anacardiaceae means that the genus Toxicodendron is a subset of the family Anacardiaceae. The *role link* Araceae $\overset{\text{contains}}{\rightarrow}$ Calcium-oxalate means that plants in the Araceae family contain calcium oxalate. Since calcium oxalate is present throughout the Araceae family we put the link at the family level, and let lower nodes *inherit* it. Calcium oxalate is specific to R. rhabarbarum (within its family), so the contains link is put directly on that species' node. These data are not in the primary database, but can be found in other databases.

Nodes and links can be used to form predicates. For instance, Toxicodendron(x) is true of everything in the genus Toxicodendron. Roles represent relations and can be multivalued; an exposure can have more than one substance link. We can use predicates to characterize sets of data items in terms of the knowledge base. For instance, Toxicodendron(substance(x)) characterizes the exposures 1-3. The more complicated predicate Calcium-oxalate(contains(substance(x))) characterizes exposures 4-5.

We note several advantages of this representation. First, inheritance networks provide a natural way to represent domain knowledge. For instance, our system allows a limited form of nonmonotonic inheritance to represent and reason about default and incomplete information. Second, since the representation does not duplicate domain knowledge for each database record there is a huge gain in both time and space efficiency. Third, inheritance networks are sufficient to represent multi-table relational databases, with role composition representing joins between tables. Finally, using inheritance networks for inductive learning connects automated discovery to work in knowledge representation.

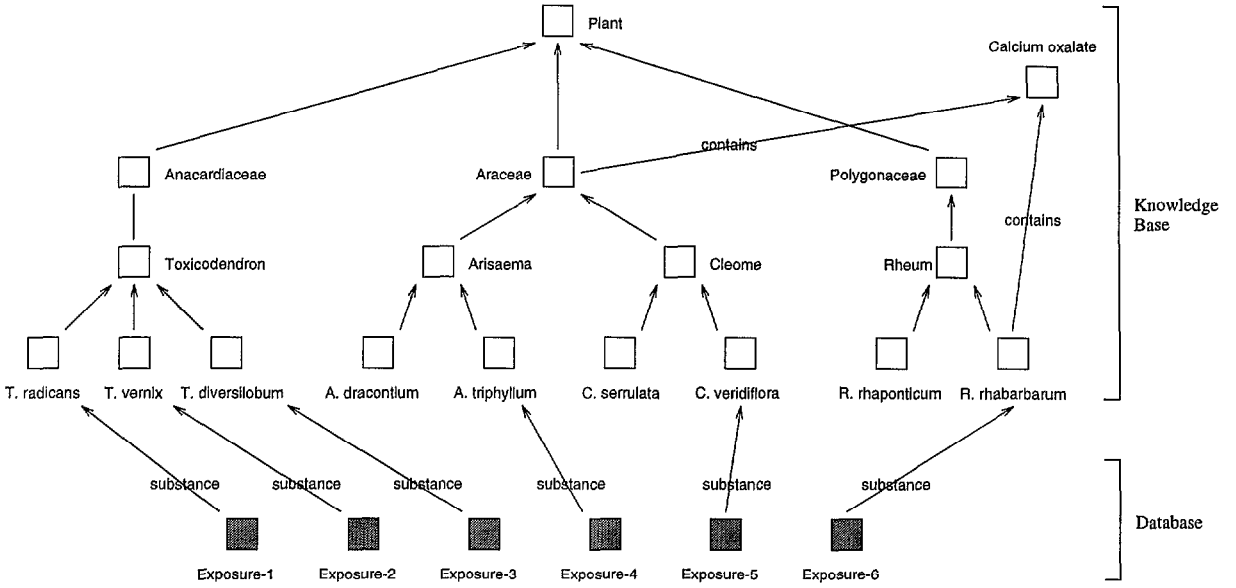

Figure 1: Linking Data to Botanical Knowledge.

An Illustrative Example

Consider the network in Figure 2. Six examples of Datura exposures are shown, connected to a database of geographical and climate knowledge. Datura exposures normally occur in August-October; here we are interested in characterizing an anomalous subset of toxic exposures that occur in May. The Knowledge-Based Rule Learner (KBRL) starts with general predicates and attempts to specialize them. The user defines criteria with which the system will judge a discovery to be interesting. For this example, we use the simple criteria: an interesting pattern is one that covers all of the May exposures, and none of the others.[1]

The search starts with the general predicate US(location(x)). Since testing reveals that this is an overly general characterization, its *specializations* are formed from relationships in the network, and are immediately tested:

Southeast(location(x))

Southwest(location(x))

The first predicate fails to cover any members of the concept class in the database, so the system rejects it. The second correctly excludes some of the complement of the concept class, while still covering the incidents we are interested in categorizing, so the system retains it. However, since this predicate still covers part of the complement, the system tests each of its specializations:

Nevada(location(x))

Arizona(location(x))

Neither of these have adequate coverage—they reject items in the concept class—so the system rejects them and looks for other ways to specialize the current hypothesis. The system cannot use the hierarchy of locations to refine its hypothesis any further, so it tries the zone link. Retaining the predicate already found, it forms the rule:

Southwest(location(x)) & AnyZone(zone(location(x)))

Again, the additional predicate is vacuous, so it is specialized to create the three hypotheses:

Southwest(location(x)) & Hot(zone(location(x)))

Southwest(location(x)) & Mild(zone(location(x)))

Southwest(location(x)) & Cold(zone(location(x)))

Checking each of these verifies that the first characterizes the May incidents perfectly, so it is retained as a characterization that satisfies the system's criteria for an interesting discovery.

Some Details of the Algorithm

KBRL, based on the RL learning program (Clearwater & Provost, 1990), performs a general-to-specific heuristic search for a set of conjunctive rules that satisfy user-defined *rule evaluation criteria.* At each stage of the search, KBRL *specializes* the currently most promising rules by either restricting their predicates, or adding new ones to the conjunction on the left-hand side of the rule. KBRL starts with the rule $T(x) \rightarrow C(x)$, where T(x) is the most general concept in the knowledge base (true of everything), asserting that everything is a member of the concept C. KBRL performs

[1] Of course, discovering a pattern characterizing a concept is seldom this simple. Predicates have to be evaluated statistically, and the concept will usually be covered only partially or covered by a disjunction of predicates.

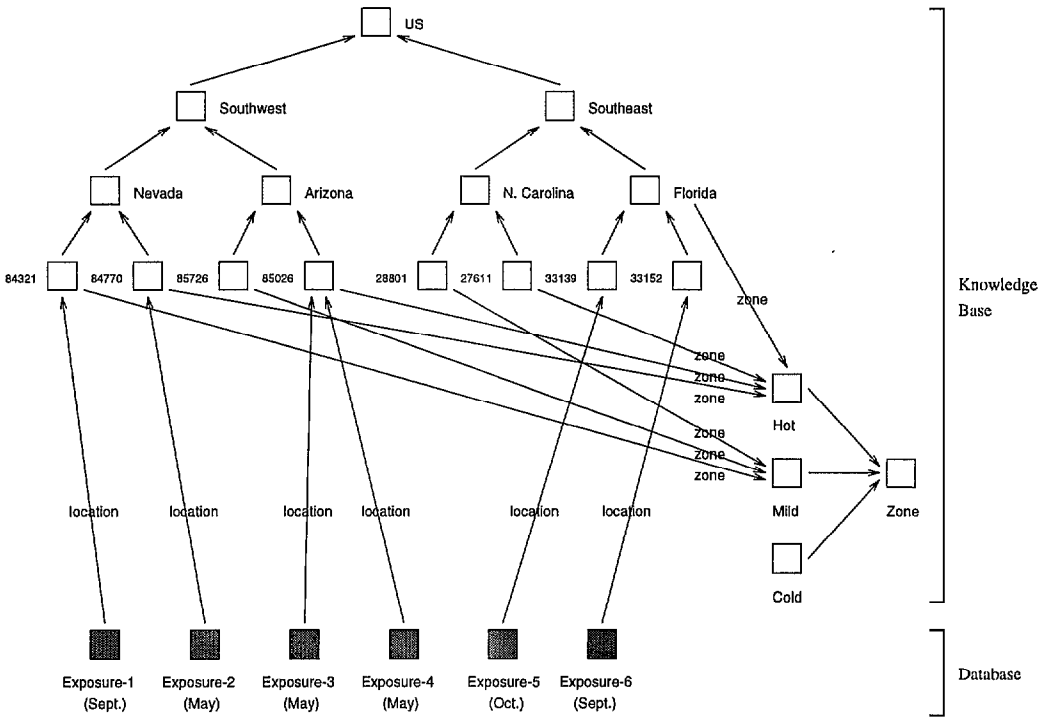

Figure 2: Characterizing May Datura Exposures.

an n-best search through the space of rules defined by the following *specialization* operators:

1. The rule $\ldots P(f_n \ldots f_1(x)) \ldots \rightarrow C(x)$ can be specialized to the rule $\ldots P(f_n \ldots f_1(x)) \ldots \; \& \; T(x) \rightarrow C(x)$.
2. Given a rule of the form $\ldots P(f_n \ldots f_1(x)) \ldots \rightarrow C(x)$ and ISA links $P_1 \rightarrow P, \ldots, P_n \rightarrow P$ in the network, the rules $\ldots P_1(f_n \ldots f_1(x)) \ldots \rightarrow C(x)$ through $\ldots P_n(f_n \ldots f_1(x)) \ldots \rightarrow C(x)$ are specializations.
3. If the node P has f role values which are restricted to P', the rule $\ldots P(f_n \ldots f_1(x)) \ldots \rightarrow C(x)$ specializes to $\ldots P(f_n \ldots f_1(x)) \; \& \; P'(f_{n+1} f_n \ldots f_1(x)) \ldots \rightarrow C(x)$.

The first operator—*Add a Predicate*—allows us to add additional predicates to a rule. This allows us to form rules with several conjuncts. The second operator—*Specialize a Predicate*—searches downward through a network identifying classes of the concept. It is important to note that in some cases there will be several different classifications of items. In botany, for example, there are different hierarchies based on different approaches to classification. The KBRL search algorithm explores all of these, specializing predicates according to each hierarchy and using heuristics to guide the search down paths that make meaningful distinctions in the current context. The third operator—*Restrict a Role*—selects a set of items based on their relationship to other parts of the knowledge base. Notice that the third operator is recursive, and we can restrict the predicate $P(x)$ to $P'(f(x))$, $P''(gf(x))$, etc. Thus, we can talk about concepts such as "the average annual rainfall of the location of the exposure."

Membership in interesting classes may be determined by exceptional information, so it is important to incorporate and use some form of nonmonotonic information. We currently use a simple form of default inheritance that allows role values to be overridden by more specific information. Consider the diagram in Figure 3. The items in the concept, marked by "+", are characterized by the predicate $Q_2(f(X))$. This includes every item in P_2, which all have f's that default to Q_2, as well as I_3, which has an exceptional f value.

An Application to Botanical Toxicology

We have been working with botanical toxocologists to analyze a database of potentially toxic plant exposures (Krenzelok, et al., 1995a,b,c). KBRL was applied to these data linked to a knowledge base of geographic areas and their climates constructed from several sources on the World Wide Web.

As an example of the flexibility of learning with background knowledge represented as an inheritance hierarchy, consider the geographic knowledge base, which consists of approximately 1000 geographic regions. The smallest, most specific region is a "zip code area"—a geographically contiguous set of zip codes

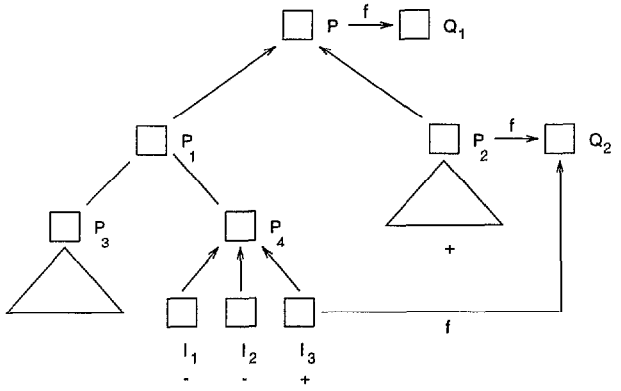

Figure 3: A Relation with an Exception.

that share the same first three digits. Zip code areas are arranged into a hierachy, with upper levels for states and geographic regions. Climate data, including rainfall, solarization, soil conditions, and temperatures, are linked to nodes in the geographic hierarchies by role links. In most cases, KBRL reasons about similarities in climate conditions by utilizing the exposure records' zip-code fields. However, some exposures are missing the zip code, but do have the telephone area code. Fortunately, KBRL can key into the geographic knowledge base at a less detailed level using the telephone area code.

The inheritance hierarchy allows the use of the most specific climatic information possible. The nodes at the lowest level do not all have complete information, so some information must be inherited from the state level. Although information at the state level is complete, it tends to be approximate. We also used a knowledge base of botanical species, genera, and families adapted from a U. S. Department of Agriculture database. Several small hierarchies of demographic factors, treatment patterns, etc., were also used.

One area of investigation in which KBRL took part was a study of exposures to Datura species. Many of the rules KBRL found refined the existing model of a seasonal spread of Datura exposures, but were not surprising to our botanical and toxicology collaborators. Rules showing that Datura exposures peak later in colder areas than in warm areas are a reflection of the fact that plants take longer to mature in colder climates. Other rules, such as a surprising degree of Datura abuse in some states, were unexpected but could have been found by other methods. A new rule was found that characterizes an unexpected set of May exposures in terms of basic enviromental conditions. This new rule was judged significant by our collaborators in botany and toxicology (Krenzelok, et al., 1995b).

Discussion

KBRL extends the notion of tree-structured attributes (Almuallim, Akiba & Kaneda, 1995) by allowing values to reference into multiple ISA hierarchies, complete with role relations and inheritance. However, the expressiveness of KBRL's language is currently limited to binary relational terms, and thus is not as expressive as some existing inductive logic programming systems (Muggleton, 1992). The design of KBRL purposely chose efficiency over expressiveness when it came to decisions about particularly expensive constructs, such as n-ary and recursive relational terms. On the other hand, because it was crucial for applications with incomplete data, KBRL incorporates default inheritance, which is difficult to deal with naturally in other relational systems. However, KBRL's form of nonmonotonic inheritance is limited, and will be difficult to extend since we want to allow multiple values and multiple ISA inheritance with negation.

Acknowledgements

This research was supported by National Science Foundation grants IRI-9412549, BES-9315428, and by the W. M. Keck Foundation.

References

Almuallim, H., Akiba, Yasuhiro, A., and Kaneda S. (1995). On Handling Tree-Structuree Attributes in Decision Tree Learning. In *Proceedings of the Twelfth International Conference on Machine Learning (ML-93)*. Morgan Kaufmann.

Clearwater, S. and Provost, F. (1990). RL4: A Tool for Knowledge-Based Induction. In *Proceedings of the Second International IEEE Conference on Tools for Artificial Intelligence*, 24–30. IEEE C.S. Press.

Fahlman, S. (1979). *NETL: A System for Representing and Using Real-World Knowledge*. MIT Press, Cambridge, MA.

Krenzelok, E.P., Jacobsen, T.D. and Aronis, J.M. (1995a). Mistletoe Exposures ...The Kiss of Death. To appear in *American Journal of Emergency Medicine*.

Krenzelok, E.P., Jacobsen, T.D. and Aronis, J.M. (1995b). Jimsonweed (Datura stramonium) Poisoning and Abuse ...An Analysis of 1,458 Cases. Submitted to *American Journal of Emergency Medicine*.

Krenzelok, E.P., F.J. Provost, Jacobsen, T.D., Aronis, J.M., and Buchanan, B.G. 1995c. Poinsettia (Euphorbia pulcherrima) Exposures Have Good Outcomes ...Just As We Thought. To appear in *American Journal of Emergency Medicine*.

Muggleton, S. (Editor) (1992). *Inductive Logic Programming*, Academic Press.

Valdes-Perez, R., (1995). Some Recent Human-Computer Discoveries in Science and What Accounts for Them. *AI Magazine*, 16(3), pp. 37–44.