Augmenting Medical Databases with Domain Knowledge

John M. Aronis, Bruce G. Buchanan, and Seok Won Lee
Intelligent Systems Laboratory
University of Pittsburgh, Pittsburgh, PA 15260

In this paper paper we discuss a method of linking databases with domain knowledge to provide an extended semantics for use with statistical, machine learning, and automated discovery programs. We focus on the use of data in conjunction with domain knowledge for automated discovery in medical databases, and show how an induction program can find new knowledge in a database by reasoning about classes and relationships that are implicit in the original data, but explicit in the representation of domain knowledge.

1 Introduction.

In this paper paper we discuss a method of linking databases with domain knowledge to provide an extended semantics for use with statistical, machine learning, and automated discovery programs. We focus on the use of data in conjunction with domain knowledge for automated discovery in medical databases, and show how an induction program can find new knowledge in a database by reasoning about classes and relationships that are implicit in the original data, but explicit in the representation of domain knowledge.

Programs for automated, inductive discovery have been shown to be effective in discovering patterns from data. Some discoveries have been made that are important enough to be published in the literature of the scientific subject domain. Although induction programs by themselves can make interesting discoveries, we focus here on removing the severe restriction that a learning program always works within a small, fixed, semantic bias. We illustrate these points in the domain of plant exposures, with the RL program extended and applied to a large, multi-year database of toxic and non-toxic plant exposures.

The present work is far from complete; however, it shows how knowledge bases and databases codified for other purposes can introduce an open-endedness to the bias within which an induction program operates. Our long-term view is to maintain access, perhaps over the internet, to large stores of background knowledge relevant to a given domain of inquiry. Our goal is that this background knowledge can be linked to the data for different induction problems to extend the semantic bias of the discovery program.

2 The Domain.

We are working to discover patterns in a large set of data recorded from calls to poison centers. The data we are focusing on, drawn from the American Association of Toxic

In AAAI Spring Symposium on Artificial Intelligence in Medicine, Stanford University, March 25–27, 1996.

Control Centers Toxic Exposure Surveillance System (AAPCC TESS), describe incidents of potentially toxic exposures of people to plants—most frequently, incidents of children eating parts of plants. The database contains about one million such records collected by poison control centers across the U.S. over the last 10 years. Because it is the policy of poison centers to record follow-up information we have a record of symptoms, recommended actions, actual actions, and outcome, as well as demographic information. Most of the time, poison centers are able to gather sufficient information about the plants to identify with high confidence the genus and species.

Of primary importance are patterns that indicate when a victim should be sent to a hospital emergency room and when it is safe to recommend waiting. However, as we point out, there are other interesting discoveries to me made with these data, especially when linked to botanical, geographical, and climatic knowledge bases.

3 The Need for Background Knowledge.

The representation most commonly used by statistical, inductive learning, and discovery systems to describe problems is the simple feature vector. However, scientific domain knowledge takes on a richer, more structured form. Prominent in any scientist's store of useful background knowledge are various taxonomies, categories and relationships between concepts. To illustrate this, consider the following two scenarios.

A sequence of potentially poisonous plant exposure cases may have related substances: one person ate Toxicodendron radicans, another Toxicodendron diversilobum, and another Toxicodendron vernix. Knowing the genus-species relationship in botanical naming, one would naturally say that these are all Toxicodendron exposures. This is important since the symptoms and treatment for exposure to the various species in the Toxicodendron genus are similar. More importantly, a scientist (or program) that knows the botanical taxonomy would see the commonality even if other, less obvious, names were used. Now suppose one is presented with exposures from various species in the Araceae family, and also from Rheum rhabarbarum (common rhubarb). Given that the irritant calcium oxalate is present throughout the Araceae family, and in the leaves of the R. rhabarbarum species, one might reason about common treatments for exposure to these plants jointly based on their common toxin.

These examples show the importance of using domain knowledge to describe and explain sets of items. In the first example, a naturally occuring *class* was used. In the second, we considered *a common feature of items in a set.* These forms of reasoning are particularly important in epidemiological problems where occurrences and distributions are analyzed in terms of commonalities introduced by demographic, geographic, and other relational factors.

4 Augmenting Databases with Inheritance Networks.

To automate the forms of reasoning illustrated above, we need to represent individual data items as well as their relationships to general concepts. We can do this in a uniform way using *inheritance networks*, which provide an efficient way to navigate and explore the space of relationships among data and concepts.

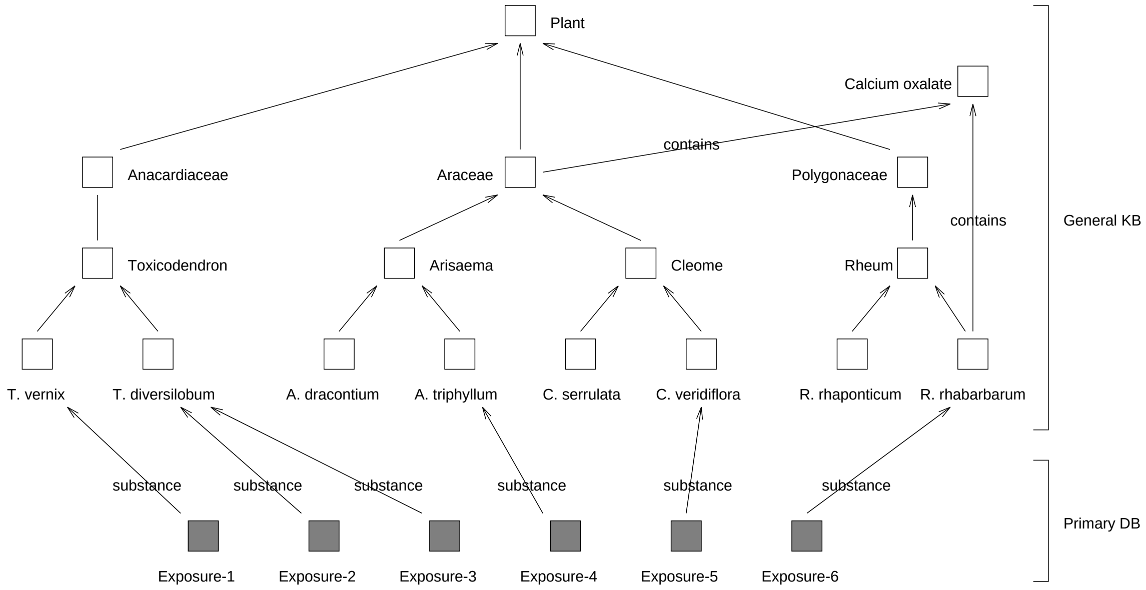

Figure 1: Linking Data to Botanical Knowledge.

Figure 1 illustrates how several exposure cases can be linked to part of a knowledge base of plants. Only a few records and their links to a small part of the botanical knowledge base are shown. Unlabeled arrows are *ISA links*, which can be interpreted as set inclusion. Thus, the link T. radicans → Toxicodendron means that every plant in the species T. radicans is also in the genus Toxicodendron. The *role link* Araceae $\xrightarrow{\text{contains}}$ 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.

For both automated discovery systems and database systems, the ability to scale up is very important. At this point, two aspects of the representation that facilitate scaling are apparent. First, inheritance networks are a compact representation. Background information is not duplicated across similar data items. Second, these predicates can be evaluated very quickly using marker-propagation algorithms. This is important since an inductive learning algorithm will evaluate many candidate predicates in its search. To find which exposures involved calcium oxalate, the program marks the node Calcium oxalate, then propagates markers down sequences of ISA links, and the roles contains and substance. The final markers will be attached to the extension of the predicate—in this case the nodes Exposure-4, Exposure-5, and Exposure-6. Special care must be taken if the network allows *nonmonotonic* links corresponding to exceptional cases, but the basic idea is the same.

5 Automatic Discovery Using Domain Knowledge.

We now have the machinery to describe the Knowledge-Based Rule Learner (KBRL) with an example. Consider the network in Figure 2. Six examples of Datura exposures are shown, connected to a database of geographical knowledge with location links. The locations in the geographic knowledge base are, in turn, connected to a database of temperature zones with zone links in the diagram. (Poisonings are also connected to several smaller knowledge bases, including one for times and dates, showing the relationships of dates, months, and seasons.) Datura exposures normally occur in August-October; here we are interested in characterizing an anomalous subset of toxic exposures that occur in May. Thus, we direct the system to search for a predicate that is true of exactly exposures 2-4.

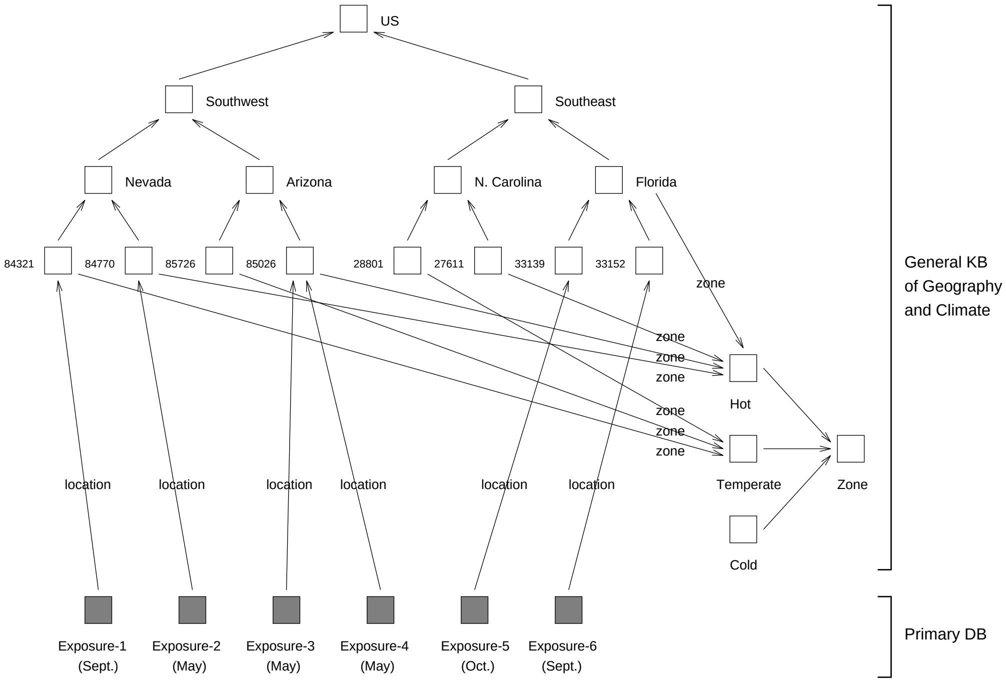

Figure 2: Characterizing May Datura Exposures.

The RL algorithm is a top-down (general-to-specific) inductive learning algorithm, so the system starts with general predicates and attempts to specialize them. The user of the system defines criteria with which the system will judge a discovery to be interesting. For this example, let us use the simple criteria: an interesting pattern is one that covers all of the May exposures, and none of the other exposures. (Of course, discovering a pattern characterizing a concept is seldom this easy. Predicates have to be evaluated statistically, and the concept will usually be covered only partially or covered by a disjunction of predicates.) The system starts with the general predicate:

$$\mathsf{location(x) = US}$$

This is specialized to:

$$\mathsf{location(x) = Southwest}$$

The new rule still does not capture the concept, so an additional conjunct is added:

$$\mathsf{location(x) = Southwest\ \&\ zone(location(x)) = Any\text{-}Zone}$$

This is further specialized to become:

$$\mathsf{location(x) = Southwest\ \&\ zone(location(x)) = Hot}$$

This rule is accepted since it covers the data items in the concept. In practice, rules are evaluated statistically and the system may have to learn several rules to adequately cover a concept.

The final learned rule is important in two respects. First, by appealing to classes mentioned in the knowledge base individual cases are covered at once. Thus, exposures 2-4 are all found to be in hot regions, although in the database they are listed as occurring in separate locations. Second, the rules are phrased in underlying, fundamental terms, rather than specific surface features.

6 Advantages of Our Representation of Background Knowledge.

Scientific discovery involves finding regularities and generalizations that are potentially interesting, often expressed in terms of known categories and relationships. We may not know *a priori* which categories or relationships are needed, so the knowledge base covering information that is possibly relevant to the domain may be quite large.

The use of background knowledge consisting of categories and relations has two complementary functions, adding to the expressive power and efficiency of any system that accesses them:

> *Categories and relations enable inductive generalization based on more complex reasoning than simple pattern matching of features.* If all or most of the items we have seen thus far in a category have a certain property, we can induce that the other items in that category also have that property.
>
> *Categories and relations focus exploration.* Linking data to domain knowledge provides a strong bias on the set of predicates. Furthermore, these predicates are hierarchically arranged, imposing a logical top-down search strategy.

Scientific knowledge is often based on taxonomic relationships, with inheritable properties linked to classes. Inheritance networks are a natural, intuitive method for representing these forms of knowledge. They are easily specified and easily understood.

In standard feature-vector learners the semantic bias only includes features specifically associated with the items. For instance, a poisoning incident may contain values for age, sex, location, substance, *etc.*, but there is no natural way to include information *about* the location, substance, its constituents, *etc.* For such learners, the only option is to flatten out the knowledge base by creating a new feature for each possible predicate that might be used to describe an example.

In addition to general categorical and relational information, scientific background knowledge often contains *exceptional cases*, requiring the ability to represend and reason with nonmonotonic information. The ability to reason nonmonotonically also facilitates the use of partial databases. For example, a particular database may have important properties specified at a very fine granularity (*e.g.*, at the individual zip-code level), but may cover only a portion of the data. The use of defaults allows areas not covered at such a fine granularity to inherit properties from coarser-grained entities (*e.g.*, from the state level).

Since our objective is to provide a context of domain knowledge for a pattern-discovery engine, it is essential that we can represent compactly a very large amount of background knowledge, and reason with it efficiently. This is necessary, because it is impossible to know exactly what will be relevant to the next discovery; assuming that one will be able to specify a small amount of background knowledge that will be sufficient begs the very question of discovery. A related advantage of using inheritance networks to represent background knowledge is that they are efficient to search. Relational predicates can be evaluated against a combination database and knowledge base using fast marker-propagation algorithms.

7 Applications in Botanical Toxicology.

The American Association of Toxic Control Centers Toxic Exposure Surveillance System (AAPCC TESS) database lists exposures in terms of species, specific location, *etc.* In order to extend the generality of the potential patterns discovered with this data, we connected the AAPCC TESS data to a knowledge base of geographic areas and their climates. We also used a knowledge base of botanical species, genera, and families.

The AAPCC TESS database has been used with these additional knowledge structures for our work with collaborators in toxicology and botany. Most of this work is standard statistical analysis that is facilited by the additional categories. We have also used both the RL and KBRL learning system with this data/knowledge base to characterize Datura poisonings in terms of basic environmental factors classes of poisonings.

8 Acknowledgements.

This research was supported by National Science Foundation grant IRI-9412549, National Institutes of Health grant 2-P41-RR06009-06 (administered through the Pittsburgh Supercomputing Center), and the W.M. Keck Foundation. We also thank our collaborators Edward P. Krenzelok, Director of the Pittsburgh Poison Center, and Terry D. Jacobsen of the Hunt Institute for Botanical Documentation.