Knowledge Discovery Using Concept-Class Taxonomies

Venkateswarlu Kolluri[1], Foster Provost[2], Bruce Buchanan[3], and Douglas Metzler[3]

[1] Chitika, Inc., MA 01545
vkolluri@chitika.com
[2] New York University, NY 13576
fprovost@stern.nyu.edu
[3] University of Pittsburgh, PA 15260
buchanan@cs.pitt.edu, metzler@sis.pitt.edu

Abstract. This paper describes the use of taxonomic hierarchies of concept-classes (dependent class values) for knowledge discovery. The approach allows evidence to accumulate for rules at different levels of generality and avoids the need for domain experts to predetermine which levels of concepts should be learned. In particular, higher-level rules can be learned automatically when the data doesn't support more specific learning, and higher level rules can be used to predict a particular case when the data is not detailed enough for a more specific rule. The process introduces difficulties concerning how to heuristically select rules during the learning process, since accuracy alone is not adequate. This paper explains the algorithm for using concept-class taxonomies, as well as techniques for incorporating granularity (together with accuracy) in the heuristic selection process. Empirical results on three data sets are summarized to highlight the tradeoff between predictive accuracy and predictive granularity.

1 Introduction

The importance of guiding the discovery process with domain knowledge has long been recognized [12], but most existing data mining systems do not exploit explicitly represented background knowledge. Recently, taxonomic background knowledge of attributes and attribute value has received attention ([1], [2], [8]). However, while it has been recognized ([9], [11], [12]) that there is often sufficient domain knowledge to generate hierarchies over the (dependent variable) concept-class values as well, in most classification learning research the concept-class variable is assumed to comprise a simple set of discrete values determined by a domain expert. Concept-class, attribute and attribute-value taxonomies are structurally similar, but are distinguished by the role that they play in a particular learning situation as dependent or independent variables.

The practice of leaving to domain experts or data analysts the task of selecting the appropriate levels of analysis [e.g., 8, 17] in situations involving large sets of concept-class values that are inherently hierarchically structured is problematic. Human choosing of appropriate levels of concepts to learn is an inherently labor intensive task that is compounded by the fact that there is not, in general, one ideal

G.I. Webb and Xinghuo Yu (Eds.): AI 2004, LNAI 3339, pp. 450–461, 2004.

generalization level for a given problem. The most useful level is a function not only of the desired conceptual outputs, but also of the data available to drive the learning process. The effects of data availability and the utility of concept class taxonomies are most evident when the data set is small and the concept class value set is relatively large. For instance, in the "bridges" domain, [16], if each type of bridge has only a few examples in a given data set it might be difficult to find sufficient evidence for all specific types of bridges, while there might be evidence for more general types (Figure 1). Thus hierarchical classification learning is a two-fold process. In addition to the search for a set of patterns that best describe a given concept, hierarchical classification involves a search over the concept-class taxonomy to find the concepts that represent the best tradeoffs concerning usefulness and degree of support in the given data set. The approach described here provides for the simultaneous search for rules to predict concepts at all levels of a taxonomic hierarchy, while allowing the user to bias the system to varying degrees of specificity vs. accuracy.

In this paper we describe HRL (Hierarchical Rule Learner), an extension to an existing rule-learning system, RL [14] that demonstrates the feasibility of learning within concept-class taxonomies. We describe why existing methods for evaluating classification models are not sufficient to evaluate hierarchical classification models, introduce the concept of *prediction granularity* of a model that needs to be considered along with predictive accuracy, and show how *granularity* can be incorporated in the learning process. Empirical results on three data sets are summarized to highlight the tradeoffs between predictive accuracy and predictive granularity.

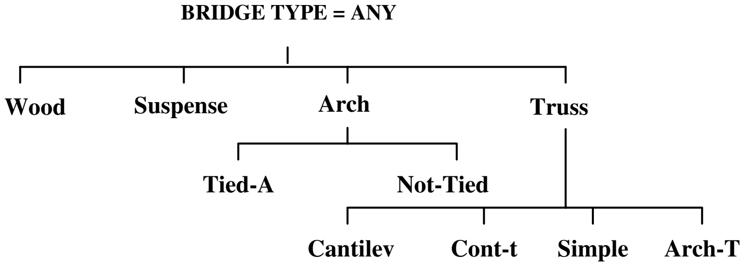

Fig. 1. Bridge concept class taxonomy

2 Hierarchical Rule Learning Using Concept Class Taxonomies

The Hierarchical Rule Learning (HRL) algorithm discovers appropriate concepts within taxonomically structured (dependent) concept-class taxonomies. HRL is an extension of the BFMP learning technique [3], which in turn is a marker-passing based extension of the RL induction system, a descendent of the MetaDENDRAL system [6]. RL is a generate-and-test rule-learning algorithm that performs heuristic search in a space of simple IF-THEN rules containing conjunctions of features (attribute-value pairs).

Figure 2(a) contains a simple database. The location and occupation fields (attributes) contain independent attribute values. The car field is the dependent concept to be learned. In this case the concept values are binary (US made vs. imported). The task consists of learning a set of rules to reliably predict the dependent concept values, e.g., the set of people who own imported cars (i.e., Sam and Mary). Typical top-down inductive learners such as MetaDENDRAL-style learners [6], or decision-tree learners such as C4.5 [15] and CART [5], start with a null rule (which covers everything) and generate additional conjuncts to specialize it, such as *Location = Pittsburgh*, *Occupation = Research* etc. Each conjunct is matched against the data, and statistics are gathered. The statistics are fed to an evaluation function that decides which conjuncts should be further specialized on the next iteration.

Name	**Location**	**Occupation**	**Car**
Sam	*Pittsburgh*	*Research*	*Imported*
John	*Harrisburg*	*Business*	*US made*
Bob	*San Francisco*	*Research*	*US made*
Tim	*Pittsburgh*	*Business*	*US made*
Mary	*Pittsburgh*	*Research*	*Imported*

Fig. 2(a). Cars database

Car
Toyota
Dodge
Dodge
Ford
Honda

Fig. 2(b). Cars database class values

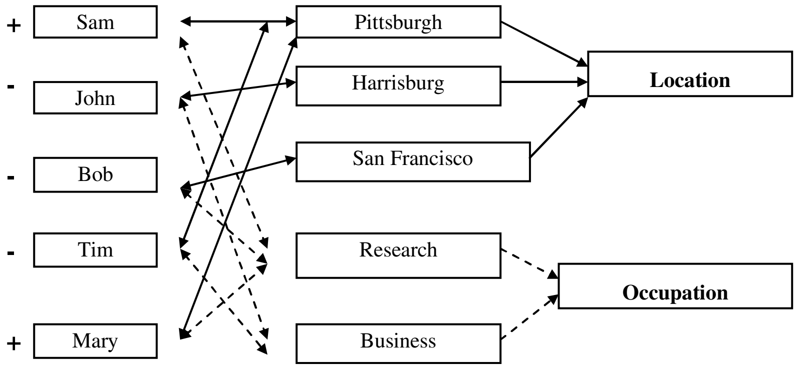

Fig. 2(c). Network representation of the Cars database

BFMP, [3] replaces the matching portion of RL and related algorithms by breadth-first marker propagation, and the collection and counting of markers. In Figure 2(c), attribute values are represented by pointers into the space of values (with a different type of pointer for each type of attribute). BFMP places a class marker on each data item (e.g. Sam) and propagates these markers along attribute links to the value nodes. BFMP then checks the coverage of predicates by counting how many positive and negative markers accumulate on the corresponding value nodes, thereby replacing the matching step in the rule learners. In binary cases such as this, positive markers represent support for the utility of adding that node to specialize the rule under consideration and negative markers represent negative evidence. E.g., the rule

(Location = Pittsburgh ➔ US made) receives two positive markers and one negative. Aronis and Provost [3] showed that rule learning using such marker propagation techniques increases performance over standard pattern matching approaches and is especially effective when dealing with datasets containing large attribute-value sets (with and without attribute-value taxonomies).

HRL extends BFMP to learn appropriate concepts within taxonomically structured (dependent) concept-class value sets. Figure 2(b) shows an alternate dependent concept class (car) for the data set in Figure 2(a) containing individual values, which can be hierarchically ordered as in Figure 4. In such a non-binary learning task, individual markers, (e.g., *T*, *D*, *F*, and *H* respectively in this case) replace the binary +/- markers, and the negative evidence for a given class at a particular node must be computed as the sum of the markers belonging to the compliment set of class markers.

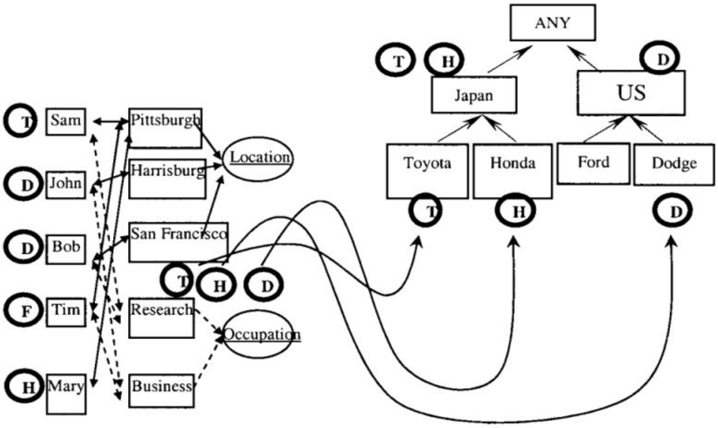

Fig. 3. Extended Car data set network with concept class taxonomy

The left side of Figure 3 shows a network like that shown in Figure 2(c). The individual markers (T, D, F and H) are shown instead of the binary (+/-) markers of Figure 2(c), and the accumulation of those markers are shown for the occupation=research node. These markers are then passed to the class taxonomy shown on the right hand side of Figure 3., and passed up that hierarchy. Although the evidence so far is weak for any rule predicting a particular manufacturer based on occupation=research, it would appear that evidence is gathering for a rule that predicts country=Japan based on occupation=research. The concept class node "Japan" now represents the rule: (Occupation = Research) ➔ (Car = Japan), and it has a "positive" coverage of 2. To calculate the negative coverage for this rule the concept class markers are propagated to the top-level root node "ANY". The difference between the total of all markers gathered at "ANY" and those gathered at the class node "Japan" represents the "negative" coverage of the rule (Occupation = Research) ➔ (Car = Japan). Hence the final coverage statistics for this rule are as follows: *Total* Coverage = 3; Confidence = 2/3= 0.66 Positive coverage = 2, Negative coverage = 1.

HRL can produce multi-level rule sets (models) with rules predicting concept class categories at various levels in the taxonomies. Such hierarchical classification models can make high-level, coarse-grained predictions when insufficient information is available to make precise low-level predictions. The individual rules themselves are useful since they capture structural relationships between the independent features and the dependent concept values at various levels in the concept class taxonomy. Unlike top down learning methods (e.g., [8]), HRL is inherently bottom up, discovering all rules for which there is sufficient support.

3 Working with (Dependent) Concept Class Taxonomies

Heuristics that guide the selection of "*interesting*" rules over hierarchical concept class value sets are fundamentally different from the heuristics needed to select rules with hierarchical attribute value sets. In general, more general attribute values (e.g., location = Pennsylvania rather than location = Pittsburgh) are preferred in the case of attribute value hierarchies, (assuming they are predicting the same concept and have similar statistical evidence), since the coverage of the more general rule will be greater. However, in the case of concept class hierarchies (i.e., a hierarchy used as a set of predicted or dependent variables), more specific values (e.g., Pittsburgh rather than Pennsylvania) will be preferred since the rules will be more informative[1].

The search heuristic of classification learning systems usually attempts to optimize to cover as many positive examples of a class while producing as few false positives as possible [7]. If a classifier needs to select nodes from a class hierarchy however, simple classification accuracy alone is not sufficient as can be seen by noting that the top level rule, (null ➔ ANY, where ANY is the top-level concept in the taxonomy), has 100% coverage and accuracy, but no informative value. On the other hand, fine-grained predictions at the lowest level in a taxonomy are informative but may not be easy to learn if there is insufficient data at that level. In general, more specific rules are more informative but less accurate. Hence the search for "interesting" rules essentially involves a tradeoff between maximizing accuracy and vs. seeking rules of maximum *granularity* (taxonomic specificity).

3.1 Granularity Metrics

In order to capture the intuitive preference for fine-grained (low-level) rules, a measure is needed of taxonomic depth or specificity (which we refer to as *granularity).* A granularity metric should be 1) proportional to the depth of the rule's concept class in the concept class taxonomy, and 2) independent of the given rule's coverage and accuracy (i.e. data-dependant coverage statistics).

[1] Other factors may mitigate against these general rules however. Turney [18] pointed out that attribute values may have costs associated with them and a rule that predicts a more general concept may be preferred over a more specific one if the costs of obtaining the predictive attribute values are lower. Model complexity might also influence a user to prefer general rules, since a bias towards lower levels specific rules might generate too many rules.

Simple Granularity Metric: A *simple* approach would be to define the granularity score as the ratio of *i,* the number of links between the node and the root, and *d,* the maximum depth of the path along which the node exists. This provides an intuitive scoring scheme that satisfies the two conditions, but leads to inconsistencies in the general case of trees of varying depth.

$$\text{SimpleGranularity} = \text{i/d}$$

Absolute Granularity Metric: The *"absolute* granularity" of a node *n* is defined as:

$$\text{AbsoluteGranularity(n)} = \frac{\text{N} - \text{S(n)}}{\text{N}}$$

where N is the total number of non-root nodes in the taxonomy and S(n) is the number of nodes subsumed under the node *n.* This assigns a value of 0 to the root node and a value of 1 for all leaf nodes. This is generally intuitive, however if different parts of a taxonomy vary in how well developed they are, one might not want to consider all leaves as equivalently specific. Moreover, this metric is susceptible to changes in the taxonomy, and such changes, if they occur dynamically during a learning process, will invalidate the previously determined information based on the now-changed granularity score.

Relative Granularity Metric:

$$\text{RelativeGranularity(n)} = (1 - \frac{1}{\text{d}+1})$$

This measure is sensitive only to *d,* the depth of n from the root. The root is still 0 while other nodes approach 1 as depth increases. This measure is also susceptible to changes in the taxonomy but only to changes above a node, since it does not involve the number of nodes subsumed under a particular node. On the assumption that taxonomies are more likely to grow dynamically from the leaves or lower conceptual nodes, this measure will less frequently be affected mid-process than would be the absolute granularity metric.

3.2 Rule Quality Measure

Symbolic rule learning systems typically use accuracy (also known as confidence) of individual rules as a selection criterion to pick "interesting" rules. But as explained above, accuracy by itself is not sufficient when dealing with taxonomically structured concept class values. In that case one must utilize a measure of rule granularity (R_g) as well as a measure of rule accuracy (R_{Acc}). Although one could simply sum these, a *weighted linear* combination is more desirable, since it allows users to select an appropriate weighting scheme that suits their preferences:

$$\text{LQ}(\text{R},w) = w(\text{R}_{\text{Acc}}) + (1\text{-}w)(R_g)$$

This allows users to explore the models (rule sets) produced by the system under varying emphases on predictive granularity vs. predictive accuracy.

3.3 Model Evaluation

Just as classification accuracy alone is insufficient to guide the rule selection process during learning within taxonomic classes, it is also insufficient for evaluation of the performance of a generated model (rule set). A lower-level, more specific rule is more informative at a given level of accuracy than is a more general rule. Instead of giving equal weight to all correct predictions, a weighted value proportional to the depth of the predicted class node along the path between the root node and the actual class value node can be used to formulate a quality score for evaluating the classification performance:

$$HierarchicalWeighted\ Accuracy = 100\ x\Sigma c(i/d)/N$$

where for each prediction: c = correctness, *i* = the level of the predicted class in the class taxonomy, and *d* = the total depth of the tree path on which the predicted class is, and *N* = total number of predictions made by the classifier. For example consider a test set in the cars database, (Figure 4) with 10 instances, all of which having the same concept class value, Ford.

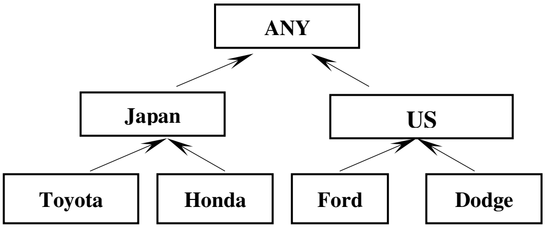

Fig. 4. Cars database taxonomy

If the model predicted all 10 cases as FORD, then its quality value would be: 100 x (10 x 1) / 10 = 100%. If all 10 were predicted to be TOYOTA the quality value would be 0. But if 5 of the 10 predictions were US and 5 were Ford, then the accuracy of the model is: (100 x ((5 x 0.5) + (5 x 1))) / 10 = 75%. In this example, each prediction of type Ford received a value of 1 and each prediction of type US received a value of 0.5, since the node US is halfway between the node FORD and the root node in the concept class taxonomy. Such differential weighting schemes proportional to the depth of the class node capture the notion that there is some information loss when a “correct” higher-level coarse-grained prediction is made. However this approach fails to distinguish between misclassification errors among sibling class nodes vs. error between more distant nodes or the differential costs of different misclassifications, a subject for future work.

4 Empirical Study

We conducted an empirical study to demonstrate the tradeoffs between predictive accuracy and predictive granularity. Three real world data sets, the Soybean data set, the Pittsburgh Bridges data set, and the Imports data set obtained from the UCI-Irvine

ML data repository [4], were used. The data sets were chosen based on the following criteria: large number of concept-classes, and available taxonomic grouping over the set of concept-classes.

In the Soybean domain data set there are 683 instances. The task is to diagnose soybean diseases. There are 19 classes and 35 discrete features describing leaf properties and various abnormalities. The set of 19 classes have been grouped into a two-level concept-class taxonomy as shown in Figure 5.

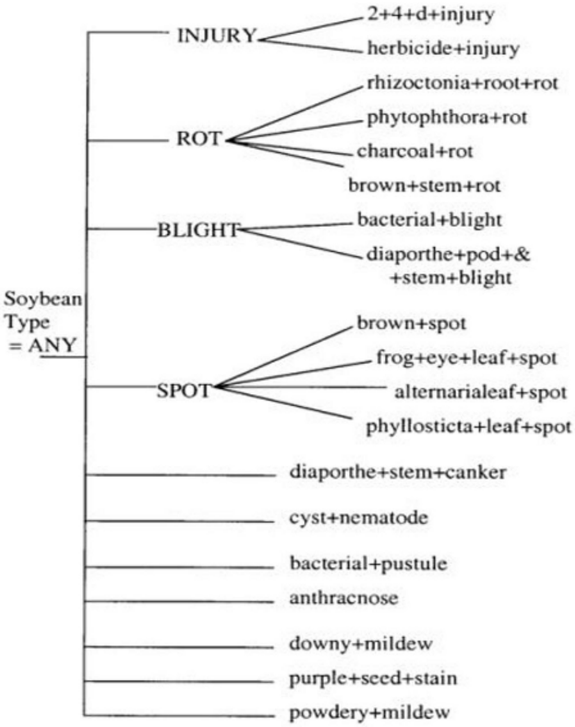

Fig. 5. Soybean concept-class taxonomy

In the Pittsburgh Bridges data set [16], each bridge is described using seven specification properties (e.g., the *river* and *location* of the bridge), and five design properties (e.g., the *material* used). There are 108 instances in this data set. In this study the learning task was to predict the *type* of bridge, given the seven specification properties. There are eight types of bridges: *wood, suspense, tied-a, not-tied, cantilev, cont-t, simple* and *arch*. The Bridges data set's eight concept-class values can be grouped into a two level taxonomy as shown in Figure 1.

The Imports data set [4] consists of 15 continuous, 1 integer and 10 nominal-valued attributes. For this study, the following attributes were used to predict the "make" of the car: *symboling, fuel-type, aspiration, num-of-doors, body-style, drive-wheels, engine-location, engine-type, num-of-cylinders,* and *fuel-system*. The data set contains 114 unique instances. The 22 possible concept-class values (i.e., the make of the car) were grouped into a two-level hierarchy (Figure 6).

To highlight the tradeoffs between predictive accuracy and predictive granularity of the resulting hierarchical classification models, a series of experiments were conducted using the *weighted linear quality metric* (Section 3.2) to guide the HRL system. The weight *w*, ranging from 0 to 1, can be used to vary emphasis on either the rule confidence, R_{Acc}, or the rule granularity R_g. (Higher w-scores bias the system toward higher predictive accuracy; lower w-scores bias the system towards higher predictive granularity.)

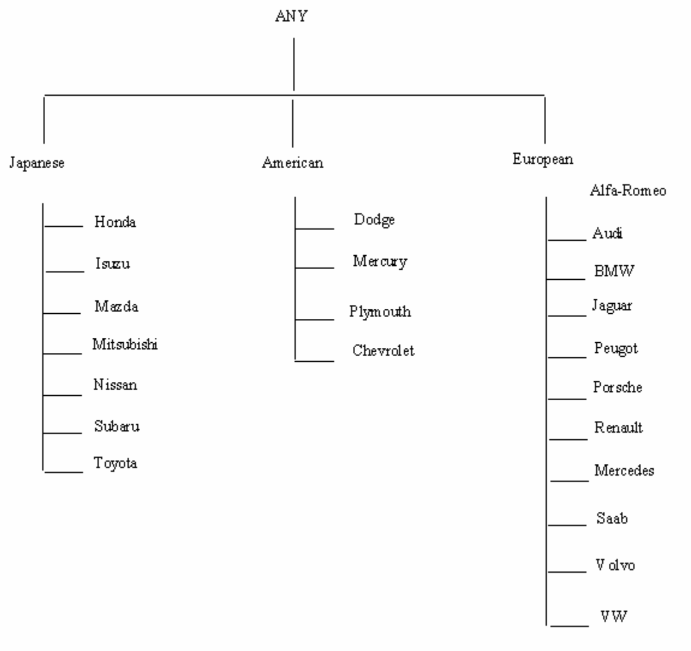

Fig. 6. Imports data set taxonomy

To explore the utility of the various proposed granularity metrics, sets of experiments were conducted using each of the three granularity metrics (Section 3.1). The results obtained with the three granularity metrics at different settings of *w* in the three domains are summarized in Table 1. The table shows the predictive accuracy and the predictive granularity of the models (rule sets) generated (Section 3.3). The model granularity is the sum of the granularity of all rules in the model. For comparison, the simple-granularity metric was used to compute the predictive granularity of all final models. The results for soybean data using the simple granularity metric are plotted (Figure 7) to highlight the tradeoffs between granularity and accuracy scores.

Results obtained from the HRL experiment with w-score = 1 can be considered as experiments using the “flattened” concept-class value set, using all concept classes in the concept-class taxonomy, but ignoring the structural relationships. As expected, the experiment with w-score of 1 resulted in models with highest accuracy but lowest granularity values, because the learning system was biased to ignore the semantic information implicit in the structural relationships among class nodes in the concept-class taxonomy. But when the w-score was decreased, forcing the learning system to consider the prediction granularity along with the prediction accuracy, a significant increase in the predictive granularity was observed (for all three sets of experiments using different granularity metrics) with a corresponding loss in accuracy.

Table 1. Accuracy-Granularity scores for experiments using Int-HRL in Soybean, Bridges and Imports Domains, respectively

w-score	**SimpleGranularity**		**AbsoluteGranularity**		**RelativeGranularity**	
	Accuracy	***Granularity***	***Accuracy***	***Granularity***	***Accuracy***	***Granularity***
1	82.64	58.78	82.70	58.04	84.03	59.44
0.9	83.45	78.54	82.11	76.65	83.02	77.37
0.8	83.88	80.37	80.36	79.66	82.52	79.83
0.7	80.90	85.41	80.17	80.50	82.03	82.27
0.6	79.62	90.71	79.92	86.44	76.26	80.22
0.5	74.08	98.84	79.43	86.05	75.05	80.74
0.4	69.60	100.00	76.69	86.48	72.42	83.87
0.3	52.74	100.00	76.36	88.17	45.19	100.00
0.2	49.15	100.00	50.94	100.00	39.35	100.00
0.1	45.97	100.00	48.47	100.00	36.97	100.00

w-score	**SimpleGranularity**		**AbsoluteGranularity**		**RelativeGranularity**	
	Accuracy	***Granularity***	***Accuracy***	***Granularity***	***Accuracy***	***Granularity***
1	72.02	57.05	69.96	57.47	66.80	59.36
0.9	57.58	80.73	59.04	78.48	59.68	83.35
0.8	60.52	85.82	57.55	86.09	58.33	80.24
0.7	54.36	90.54	57.55	88.03	57.23	84.16
0.6	55.85	95.30	58.44	91.34	53.77	86.62
0.5	49.48	100.00	56.64	95.21	52.25	92.88
0.4	51.37	100.00	52.75	99.16	50.89	94.68
0.3	44.27	100.00	41.91	100.00	39.86	100.00
0.2	41.36	100.00	40.35	100.00	39.27	100.00
0.1	36.43	100.00	42.64	100.00	29.44	100.00

w-score	**SimpleGranularity**		**AbsoluteGranularity**		**RelativeGranularity**	
	Accuracy	***Granularity***	***Accuracy***	***Granularity***	***Accuracy***	***Granularity***
1	63.96	52.84	66.54	50.00	63.36	50.00
0.9	62.63	69.43	58.94	64.92	60.73	66.25
0.8	57.40	64.18	53.92	65.81	59.60	62.58
0.7	51.35	70.04	61.88	66.29	52.66	68.98
0.6	45.13	86.76	53.41	69.94	54.04	66.55
0.5	31.22	94.58	51.09	79.18	41.34	71.23
0.4	30.18	100.00	38.27	86.46	37.53	84.57
0.3	25.95	100.00	28.87	98.68	21.85	100.00
0.2	26.32	100.00	21.81	100.00	26.07	100.00
0.1	30.39	100.00	28.01	100.00	20.11	100.00

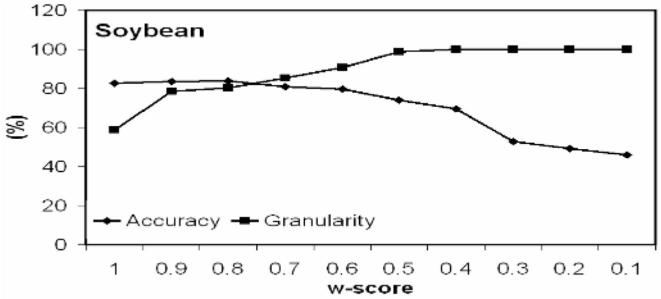

Fig. 7. Accuracy vs. Granularity scores for models generated using the Int-HRL system. Each data point is the average of 10 tests in a 10-fold cross validation experiment

5 Discussion

HRL demonstrates the ability to learn in a space of taxonomically structured (dependent) classes. It produces hierarchical classification models that can be used to classify new instances at differing levels of generality depending on the information available. The use of concept class hierarchies (as opposed to attribute and attribute-value hierarchies) introduces new research issues concerning the heuristics used to estimate the quality of rules. A tradeoff exists between rule accuracy and granularity (specificity) and measurement of the latter is somewhat ambiguous. We introduced three possible metrics for concept-class granularity, each with advantages and disadvantages but it is not yet clear if one is consistently superior to the others. Preliminary results highlight the tradeoffs between predictive granularity and predictive accuracy and indicate similar behavior for the three granularity metrics in each of the domains.

References

1. Almuallim, H., Akiba, Y., and Kaneda, S. (1995). On handling tree-structure attributes in decision tree learning. In *Proc. of the 12th Intl. Conf. on Machine Learning*, Morgan Kaufmann
2. Aronis, J. M., Provost, F. J. and Buchanan, B. G. (1996). Exploiting background knowledge in automated discovery. In *Proc. of the 2nd Intl. Conf. on Knowledge Discovery and Data Mining*, pp: 355--358, Menlo Park, CA, AAAI Press.
3. Aronis, J. M. and Provost, F. J. (1997). Efficient data mining with or without hierarchical background knowledge. *In Proc. of the 3rd Intl. Conf. on Knowledge Discovery and Data Mining*, New Port Beach, CA.
4. Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
5. Breiman, L., Friedman, J. H., Olsen, R. A., and P. J. Stone (1984). *Classification and regression trees.* Wadsworth International Corp., CA.

6. Buchanan, B. G. and Mitchell, T. M. (1978). *Model-directed learning of production rules*. In D Waterman and F Hayes-Roth, editors, Pattern Directed Inference Systems. Academic Press., New York, NY.
7. Fürnkranz. J. (1999) Separate-and-Conquer Rule Learning. *Artificial Intelligence Review* 13(1) pp:3-54, 1999.
8. Kaufmann, K. A. and Michalski, R. S. (1996). A Method for Reasoning with Structured and Continuous Attributes in the INLEN-2 Multistrategy Knowledge Discovery System. In *Proc. of the 2nd Intl. Conf. on Knowledge Discovery and Data Mining*, pp: 232-238
9. Koller, D. and Sahami, M. 1997 Hierarchically Classifying Documents Using Very Few Words. In Proc. of the 14th Intl. Conf. on Machine Learning, pp. 170-178, San Francisco, CA: Morgan Kaufmann.
10. Krenzelok, E., Jacobsen T., and Aronis J. M. (1995) Jimsonweed (datura-stramonium) poisoning and abuse: an analysis of 1,458 cases. In *Proc. of North American Congress of Clinical Toxicology*, Rochester NY.
11. McCallum, A., Rosenfeld, R., Mitchell, T. and Nigam, K. (1998) Improving Text Classification by Shrinkage in Hierarchy of Classes. In *Proc. Of the 15th Intl. Conf. in Machine Learning*.
12. Michalski, R. S. (1980). Inductive Rule-Guided Generalization and Conceptual Simplification of Symbolic Descriptions: Unifying Principles and Methodology. Workshop on Current Developments in Machine Learning. Carnegie Mellon University, Pittsburgh, PA.
13. Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T. and Brunk, C. (1994). Reducing Misclassification Costs. In *Proc of the 11th Intl. Conf. of Machine Learning*, New Brunswick. Morgan Kaufmann
14. Provost, F. J. and Buchanan, B.G. (1995). Inductive Policy: The Pragmatics of Bias Selection. *Machine Learning (20).*
15. Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
16. Reich, R. & Fenves. R. (1989). Incremental Learning for Capturing Design Expertise. Technical Report: EDRC 12-34-89, Engineering Design Research Center, Carnegie Mellon University, Pittsburgh, PA.
17. Taylor, M. G., Stoffel K., and Hendler J. A. (1997) Ontology-based Induction of High Level Classification Rules. In *Proc. of the SIGMOD*
18. Turney, P. D., (1995). Cost-sensitive classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm, *Journal of Artificial Intelligence Research*, 2, March, 369-409.