SELECTING AMONG RULES INDUCED FROM A HURRICANE DATABASE

John A. Major and John J. Mangano
The Travelers Insurance Companies, Hartford, Connecticut 06183

Rule induction can achieve orders of magnitude reduction in the volume of data descriptions. For example, we applied a commercial tool (IXL[tm]) to a 1,819 record tropical storm database, yielding 161 rules. However, the human comprehension goals of Knowledge Discovery in Databases may require still more orders of magnitude. We present a rule refinement strategy, partly implemented in a Prolog program, that operationalizes "interestingness" into performance, simplicity, novelty, and significance. Applying the strategy to the induced rulebase yielded 10 "genuinely interesting" rules.

I. PURPOSE OF THE STUDY

At The Travelers Insurance Company, we are involved in applying statistics and artificial intelligence techniques to the solution of business problems. This work is part of an investigation into applications for Natural Hazards Research Services.

The purpose of this study is *not* to develop a hurricane model or predictor. It is, rather, to assess the utility of rule induction technology and our particular rule refinement strategy. The *object task* of the study is to develop rules that predict, from simple position and wind speed observations, whether a cyclone will make landfall on the U.S. coast.

II. BACKGROUND

The analysis of hurricanes is an important specialty of meteorology that feeds its results into actuarial science. Casualties and damage from hurricanes frequently cost insurers over $1 billion per year in claims.[1,2] Hurricane Andrew (1992) alone caused over $15 billion in property damage claims.[3]

A *cyclone* is "an atmospheric system in which the barometric pressure diminishes progressively to a minimum value at the center and toward which the winds blow spirally inward from all sides, resulting in a lifting of the air and eventually in clouds and precipitation.... The name does not signify any degree of intensity." A *hurricane* is a cyclone originating in the tropics with wind speeds of 64 knots or higher.[4]

IXL[tm] is a commercial knowledge discovery software tool marketed by IntelligenceWare, Inc. of Los Angeles, California. It accepts as input a relational database in any of several formats, plus control parameters and goals specified by the user. Its outputs include a text file of conjunctive rules with performance statistics.[5,6]

As Gaines[7] points out, a knowledge discovery tool can quickly generate many rules that will take a human days to understand. This motivated our development of a refinement strategy and its partial implementation in the REFINERY program.

III. THE DATA

The National Hurricane Center maintains a machine-readable file on all North Atlantic tropical cyclones since 1886.[8] Among other data items, the file contains positions and maximum sustained wind speeds of each cyclone at six-hour intervals. Due to the relative unreliability of observations before 1945, we limited our database to cyclones from 1945 to 1979. We reserve 1980 to 1992 data for validation.

Cyclones, being structured in space and time, are complex entities to represent in a database. While work has been done on induction on structured objects,[9] IXL is limited to unstructured entities. For this reason, we transformed cyclone tracks to point observations along the track, dropping the identity, and hence the continuity, of the cyclone. This introduces complications we will address in IX.B.

The attributes used in the study are presented in Figure 1.

DATE:	Real number representing the month+day as 1.01 through 12.31.
STORM.TYPE:	1="Tropical Storm or Hurricane," 3="Tropical Disturbance," and 5="Extratropical Storm." Over 90% of the records were 1.
WIND.SPEED:	Ranged from 15 to 155 knots.
LATIT:	Latitude ranged from 8.2 to 60.4 degrees (north).
LONGIT:	Longitude ranged from 8.5 to 101 degrees (west).
TRACK.SPD:	The forward speed of the storm system, 0 to 70.5 knots.
TRACK.ANG:	Direction in which the storm system is moving. Units were degrees clockwise from a direction going due west, -179 to 180.
DIST.COAST:	The distance to the nearest point on the U.S. coast. Ranged from 90 to 3378 nautical miles.
COAST.ANG:	Bearing from storm center to nearest point on the U.S. coast.
TRKCST.ANG:	The algebraic difference between the track angle and the coast angle, wrapped to range from -180 to 179. Positive means the storm is heading to the right of the nearest coastal point.
INW.SPEED:	The magnitude of the track speed vector projected onto the coast bearing vector. Ranged from -68.75 to 35.5 knots.
PAR.SPEED:	The magnitude of the track speed vector projected onto a vector 90 degrees clockwise to the coast bearing vector. Ranged from -19.75 to 53.25 knots. Positive means the storm is heading to the right of the nearest coastal point.
U.S.LAND:	Binary attribute used as goal. Equals one if the instance's storm did cross the U.S. coast.

Figure 1. Attributes of the hurricane database.

After selection and location edits (e.g., no observation closer than 90 miles from the coast) there were 334 cyclones, of which 100 crossed the U.S. coast. Those cyclones generated 1,819 records, of which 388 were from landfall cyclones.

IV. APPLYING IXL

IXL was given the goal of finding rules that conclude any value for U.S.LAND. Parameters specified a confidence factor of at least 30, coverage of at least 10, and a maximum rule length of 5 terms. (Confidence factor and coverage are discussed in

section V.) After 43 hours on an IBM PS/2 model 70 (Intel 80386 20Mhz), the run was interrupted. From the 529 rules induced up to that point, we extracted 161 that concluded U.S.LAND=1. We refer to these 161 as the landfall rules.

V. INTERPRETING THE UNREFINED RESULTS

The first of the landfall rules, Rule 29, is presented in Figure 2.

```
% Rule 29
CF =  33.61
    "U.S.LAND" = "1"
IF
    "100" <= "WIND.SPEED" <= "115"
;
% Margin of Error: 8.8 %
% Applicable percentage of sample: 6.71 %
% Applicable number of records: 122
```

Figure 2. First of 161 induced rules, IXL format.

The concept embodied in the rule is the conjunction of the one or more terms following the "IF." In this case, the concept selects all instances in the database where the wind speed is between 100 and 115 knots.

The "Applicable number of records," or coverage, is the number of instances satisfying the concept. We say that the rule covers those instances. We sometimes use the letter G to represent coverage. In this case, G=122.

The CF (certainty factor) is the positive diagnostic power of the rule,[10] the percentage of covered instances with U.S.LAND equal to one (that is, true). We sometimes use the letter C to represent CF.

Even among the 51 rules presented in the Appendix, it is a nontrivial task for the human (domain expert or analyst) to grasp what has been revealed. Consider the following questions. Which rule has the highest coverage? CF? What rules outperform R366? Which rules mentioning coast.ang are disjoint; which are nested? Does R488 offer a significant improvement over its generalizations?

VI. CRITERIA FOR INTERESTINGNESS

A. Performance: <G,C> space and the performance frontier

Piatetsky-Shapiro[11] provides three performance axioms that rule interest measures $f(G,C)$ should satisfy. Let D be the overall occurrence rate in the database.

1: $f(G,D) = 0$.

2: $f(G_0,C)$ monotonically increases in C for fixed G_0.

3: $f(G,T_0/G)$ monotonically decreases in G for fixed T_0.

We would like to add a fourth, independent, axiom:

4: $f(G,C_0)$ monotonically increases in G for fixed $C_0>D$.

From axioms two and four, we derive a dominance relation between rules as points in <coverage,CF> space. If R1 is at <G,C> and R2 is at <F,B>, and G>F and C>B, then R1 *dominates* R2. R1 is certainly more interesting than R2, but we will withhold judgement for "off-diagonal" pairs. All rules that are not dominated by any other rules are said to be on the *performance frontier*. Such high-performance rules are, by that fact, potentially interesting.

B. Simplicity: concept lattice, cones, and MGR families

Possibly the most cited rule preference criterion, after performance, is simplicity. Simplicity can be explored by use of the rule lattice. This is a partial order on the set of induced rules where R1 subsumes (is more general than) R2 if the instances in R2's concept must lie in R1's concept.*

We refer to the set of specializations and generalizations of a rule as its *cone*. Given a high-performance rule (in the sense of section A), then the rules in its cone are also potentially interesting.

There will be at least one rule with no generalizations (other than the implicit rule **TOP**, whose concept covers everything, and which we add to the rule set). Such rules are Most General Rules (MGRs). The cone of an MGR is called its *family*. A family that contains many high-performance rules is probably worth examining in its entirety.

C. Novelty: hot spots, quasi-stars, and other redundancies

Another criterion is novelty,[12] which needs a knowledge context to operationalize it. Redundancy is a converse of novelty. A rule that adds little insight or performance to an existing set of rules has no novelty with respect to that set.

Say rules R1 and R2 have overlapping concepts, and a domain expert judges the formulation of R1 to make "more sense" with respect to known scientific theory than R2.

---

* Subsumption can be defined in two basic ways: intensional or extensional.

Intensional subsumption is a grammatical property of the concept formulations. Syntactic versions treat variables as independent; in the case of IXL concepts, it is a matter of matching variables and checking ranges. Logical versions must refer to an axiomatic theory of variable relations.

Extensional subsumption is a property of the instances covered by the concepts. Sample-based versions check whether all the instances *in a database* covered by a concept are among those covered by another concept. Population-based versions address the question with regard to *all possible* databases. When an axiomatic domain theory is true and complete, logical subsumption coincides with population-based subsumption.

Because REFINERY can only handle syntactic subsumption, we introduce some sample-based and logical analysis to move our understanding closer to the ideal.

If, further, the set of instances covered by R2 but not R1 have poor performance (in the sense of coverage and CF), then we consider R2 redundant and uninteresting.

In other databases we have found clusters of points in the instance space that have a higher than average occurrence rate of the criterion. We term such a region a hot spot. An induced rule set will tend to have many rules that cover overlapping subsets of the hot spot. We term such a set of rules a quasi-star (named after Michalski's stars[13]).

D. Significance: contrast statistics in a projected context

Another criterion is statistical significance.[14] Search techniques can always find extreme cases, but sound conclusions require an assessment of uncertainty.

Discussions of statistical tests use the terms *p-level, significance level* or *type I error rate* to refer to the probability, p, that any one test will incorrectly yield a positive result due to the effects of random sampling. If an induction tool generates 10,000 random rules and tests them on a database of pure noise, it will report about 100 of them as achieving a 1% significance.

This is a well-known problem,[15] and in the statistical literature, it is known as *multiple comparisons*. The oldest and simplest solution[16] is to use a significance level small enough to keep the expected number of false positives down to an acceptable level. Say, use $p=10^{-5}$ for 10,000 trials.

Our problem is in not knowing how many rules the induction tool tested on its way to finding the ones that are reported. Our solution is the *projected context*, an estimate of the size of the population of similar rules that had been examined.

The definition proceeds as follows: Consider a concept Q and a more specific concept R. Let I be the coverage of Q (or the number of instances in the database if Q=TOP). Let V be the number of variables available to formulate specializations of Q (the number available in the database less any that are restricted to single values in Q). Let R cover G instances and its defining concept consist of T conjuncts. Then we consider R one of $J = (I/G) \cdot V!/(T!(V-T)!)$ specializations of Q examined by the induction mechanism.

Our significance measure is $S(R|Q) = -\log_{10}(A \cdot J)$, where A is a numerical approximation to the one-tail significance level of the Chi-square test that R's covered instances are drawn randomly from Q's. Among *reported* rules with significance s, we can expect one in 10^s to be spurious.

As an example, consider rules R56 and R334 (refer to the Appendix). R334, "R," is a specialization of R56, "Q."

R56 covers I=513 instances; R334 covers G=228 instances. (I/G) = 2.25 is the number of subsets in any attempt to partition R56's instances into mutually exclusive subsets of size G. R334 uses T=2 out of V=12 available variables. V!/(T!(V-T)!) = 12!/(2!·10!) = 66 is another index of freedom in generating "comparable" specializations: the number of ways of selecting T out of V variables for use in writing

the terms of a rule.* The product of these two indices, 148.5, is J.

Statistical significance can be assessed by a contingency table as follows:

	U.S.LAND=0	U.S.LAND=1	TOTAL
R56-R334	209	76	285
R334	103	125	228
R56	312	201	513

A test of independence yields a Chi-square value of 42.145 on one degree of freedom. The one-tail significance level of this statistic is $4.237 \cdot 10^{-11} = A$ (Abramowitz & Stegun, Table 26.2,[17] four-point interpolation).

Our measure, S(R334|R56) = -log(AJ) = 8.2012, is well beyond the 2.0 minimum we set for significance. (REFINERY uses an approximation for A and reports S as 7.9552.)

VII. FORMALIZING THE STRATEGY

Following the above considerations, our analysis strategy proceeds in three phases.

A. Phase 1: Identify potentially interesting rules

Potentially interesting (PI) rules are those that satisfy the performance criterion or are closely related to rules that do. Specifically:

A rule R is PI if
R is on the performance frontier, or
Q is on the frontier and R is in the cone of Q, or
Q is a Most General Rule, and there are at least 3 frontier rules in Q's family, and R is also in Q's family.

B. Phase 2: Identify technically interesting rules

Technically interesting (TI) rules are selected among PI rules according to a recursively defined principle based on the simplicity and statistical significance criteria. If a potentially interesting rule is a spurious specialization of a technically interesting rule, it is therefore uninteresting. Specifically, **TOP** is considered TI, to start the recursion, and:

A rule R is TI if
R is PI, and
for all Q such that Q is TI and R specializes Q: S(R|Q) > 2.

Relative to the TI rule R, we refer to those rules Q (in the definition) as R's *Most Specific TI Generalizations* (MSTIGs). PI rules that are significant in contrast to all their MSTIGs are TI.

The first two phases are algorithmic. The next phase requires some judgement.

---

* It might be argued that if rule Q has U variables a more appropriate index is $2^U \cdot (V-U)!/((T-U)!(V-T)!)$. The above formula is more conservative.

C. Phase 3: Remove rules that are not genuinely interesting

The TI rules are each examined for redundancy. This consists of two aspects.

First, remove most rules from a quasi-star. Keep the simplest and/or most general rule that adequately covers the hot spot, and possibly some other extremely high-performance rules in the quasi-star. Discard all others.

Second, remove a rule R that is similar to another TI rule Q if Q makes much more sense to an expert and the difference R-Q does not perform well.

VIII. EXECUTING THE STRATEGY ON THE LANDFALL RULES

Parts of the above strategy were implemented in a Prolog program, REFINERY. This program parsed the IXL rule file, identified subsumptions, and computed statistical contrasts between all related pairs. Other parts of the strategy were carried out with the help of a geographical information system (GIS) operating on a 50% sample.

A. Potentially interesting rules

There were 17 rules on the performance frontier (Figure 3). One MGR family (R56) had six frontier rules, another (R119) had four. R119 itself was on the frontier. These two families were expanded into 19 and 10 PI rules, respectively. Expanding cones generated 22 additional PI rules. A total of 51 potentially interesting rules were identified. These are presented in the Appendix in detail.

B. Technically interesting rules

In family R56, R56 was significant in contrast to **TOP.** An immediate specialization which was on the frontier, R59, and two other rules had significant contrasts to R56. This reduced the 19 PI rules down to four TI rules. Under family R119, R119 was significant. Two of its immediate specializations had significant contrasts. This reduced its 10 PI rules to three TI rules. Of the remaining 22 additional PI rules, 14 were TI, bringing the total to 21 technically interesting rules. (Note: Figure 3 does not mark R114 or R351 as TI.)

The 21 TI rules were used as queries in the GIS as an aid to understanding. We observed that R114's term, "coast.ang in [84,176]," implied a localization of the instances near or in the Gulf of Mexico. R114 was in fact a specialization of R56. The contrast between R114 and R56 was not significant, therefore R114 was not TI. Further search revealed that R351 specialized R334 and was not significant.

There were, finally, 19 technically interesting rules. Nine were on the frontier.

C. Genuinely interesting rules

The localization to the Gulf of Mexico was seen in seven of the TI rules. The simplest, most representative among them was R59. Among the other rules in the Gulf quasi-star, R352 had the highest CF and a highly significant contrast to its MSTIG, R119. The other five Gulf rules, R107, R350, R357, R484, and R481, were dropped.

Turning to rule comparisons, we found R334 dominated R339 in <G,C> space and they shared a term. Cross-classifying instances in the GIS, we saw R339 covered about

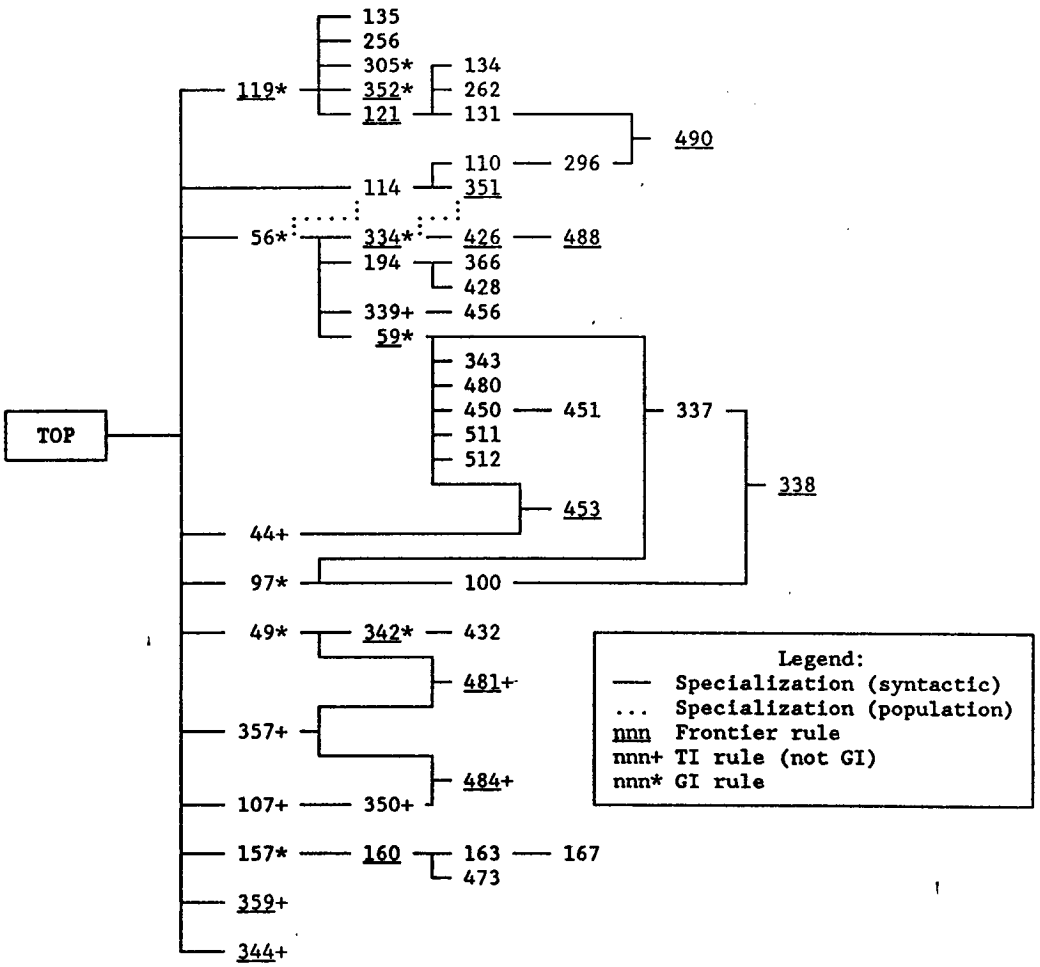

Figure 3. Specialization lattice of Potentially Interesting rules.

half the instances that R334 did, and added a few new ones. The instances covered by R334 but not R339 had about the same proportion of U.S.LAND (CF) as all of R334's instances. However, the other difference set, R339-R334, had a lower CF than R339. R339 seemed sufficiently redundant and inferior that we dropped it. (The logic behind R334's success is discussed in section D.)

Similarly, R342 and R359 shared over 400 covered instances, with R342 contributing a superior difference set. R359 was therefore dropped.

R344 had no generalizations nor specializations among the 161 original rules, yet was related to R342 and R359. R344 selects slowly moving storms, but the other two select storms that are moving *towards the coast.* R344's performance was not superior, so it was not considered interesting.

Another TI rule, R44, "latit in [22.3,28.4]," had a CF of 30.4%. The upper limit

made climatological sense (see discussion of recurve in section D), but landfall rates are 30% or higher at latitudes well below 22.3. Although there was no similar rule to favor, R44 certainly did not achieve a reasonable combination of performance and sensibility. Therefore, we dropped it.

This took the 19 TI rules down to 10 genuinely interesting rules, of which five were on the frontier. The GI rules are identified by an asterisk in Figure 3 and by italics in the Appendix.

D. Observations

Longitude (R49, R56, R59, etc.) and distance to coast (R97, R359) are important. This is obvious. An inward speed term (R359) excludes instances where the storm is moving away from the coast. This is also obvious.

Two sets of terms begin to bring knowledge in the sense of Frawley.[12] A typical Atlantic hurricane starts in the low latitudes and moves west under the influence of the trade winds, with some tendency to drift north due to the latitudinal gradient of the Coriolis force. If it hasn't dissipated by the time it reaches 30 degrees latitude, it will probably recurve, moving east under the influence of the prevailing westerlies.[18] The track-coast angle terms (R119, R305, and R352), and parallel speed term (R157) select storms whose current track is inward and to the left of the nearest coastal point. Such storms are more likely to reach land before recurving than storms which veer to the right.

In R305, the wind speed term excludes many low-intensity storms that die out before reaching the coast. There are very few instances with wind speeds over 140, so the upper limit is not much of a constraint.

The track angle terms (R334, R352) appeared anomalous, but made sense in the GIS. Track angles between 90 and 175 (most of R352's range) signify a storm that has already recurved. However, the track-coast angle term places these instances in the Gulf of Mexico, where recurving tracks are *more* likely to strike the U.S. In R334, the track angle specifies (mostly) a non-recurving situation. Why does this add significance to R56? Instances in those longitudes with track angles less than 27 are mostly non-recurving tracks in the Gulf. The southern Gulf tracks have a good chance of passing south of Florida and avoiding U. S. landfall. Instances with track angles greater than 116 are mostly recurving southern Gulf or Atlantic tracks. The Atlantic tracks are therefore moving away from the coast.

IX. ASSESSMENT

The application of the three-phase rule analysis to the 161 rules took less than two days. It succeeded in refining them to a modest set of high-performing, significant, simple, and nonredundant rules. We expect the process would have taken only a few hours with a more fully developed version of REFINERY.

However, two dependency problems introduced difficulties in our analysis. These are likely to appear, to some degree, in any real database.

A. Dependency among variables

In VIII.B we saw that "coast.ang in [84,176]" was a specialization of "longit in

[73.1,97.3]." There are more such implications derivable from variable dependencies. Distance to coast and coast angle are each functions of latitude and longitude. Track-coast angle is a linear function of track angle and coast angle. Inward speed and parallel speed are functions of track speed and track angle.

Since the algorithm in REFINERY was uninformed of such relations, it did not find all the links that exist in the (population-based) concept lattice.

B. Dependency among instances

The derivation of storm points from tracks violates the independence assumptions underlying our rule significance measure. Adjacent points in a track share the same value for U.S.LAND and have correlated values for all other variables.

To assess the impact of this, we recomputed significance levels assuming all track instances were multiple occurrences of the same row of values. On this basis, a one in 100 misreporting rate required $S > 13.5$, rather than $S > 2$. Rules R49, R56, R119, R157, and R359 were still significant. R342 and R352 didn't quite make the cut. R59, R97, R305, and R334 fell short and had to be removed. No new TI rules were introduced due to the removal of MSTIGs.

This is the other extreme of the dependence spectrum. The truth lies in between.

X. RELATED WORK

KDD applications for insurance include Major[19] and Piatetsky-Shapiro.[20] A case-based reasoning approach to hurricane modeling based on the NOAA data was taken by Hope.[21] More recently, AI work in meteorology was discussed by Moninger.[22]

Gaines[7] addresses the problem of post-processing induced rules, but in the context of noise-free data. With noisy data,[23] he uses the significance of a test of a binomial sampling model. This is asymptotic to a Chi-square test when coverage is large yet much smaller than the database.

Gebhardt[24] develops a refinement mechanism based on measures of performance and rule similarity. A rule R will be suppressed by a rule Q if the ratio of their performance measures is less than their similarity measure. While offering a number of alternatives, performance is always a one-dimensional measure, and choice of similarity measure is based on pragmatic considerations. There does not appear to be an adjustment for search context, although the general mechanism permits it.

Most other work is predicated on processes within, rather than downstream from, the induction mechanism. Weiss[10] presents preference heuristics that are used to prune a set of candidate rules; two are similar to our performance and redundancy criteria. Quinlan[25] addresses decision trees. Piatetsky-Shapiro[11] infers rule accuracy (CF) in the entire database from samples where CF=1. A radical approach to the multiple comparisons problem is outlined by Jensen.[26]

A set of induced rules is not a knowledge base. KBs have rules concluding a variety of attributes, and a relatively low level of redundancy. Shen[27] addresses regularities in a large KB. Ginsberg[28] presents a metalanguage of knowledge base refinement concepts based on rule-chaining relations and error rates.

Interestingness is still wide open. Klosgen[29] presents a mechanism by which users specify patterns for what they consider interesting statements. The landfall rules we analyze are of the type: "share of units of a <target group> is significantly larger in a <subgroup> than in a <total population>." Lenat[28] provides considerable food for thought on the operationalization of interestingness.

XI. FURTHER WORK

One potential shortcoming of this approach is the shadowing of some possibly interesting rules by unrelated rules that outperform it. For example, storm.type=1 could dominate the performance frontier, leaving us ignorant of some important predictors involving storm.type=3. In other work with REFINERY, we have addressed this issue by applying the induction mechanism iteratively, removing instances covered by the best rules at each step. Another approach would be to use rule disjointness in the criteria for interestingness.

The variable dependency discussed in IX.A might be helped with additional term subsumption machinery.[7,31] For example, say Z=f(X,Y) monotonically increases in X and Y. Then from a rule that includes "X in [A,B]," we should be able to derive "Z in [$f(A,Y_{min})$, $f(B,Y_{max})$]" to use in search of generalizations.

Observation dependency is thornier. The use of block assumptions to compute an upper bound on required significance scores does not address the real issue here. Storms are evolving entities, structured in time. An approach combining induction on structured concepts[9,32] with Bayesian methods[33] might be fruitful.

The phase three criteria (VII.C) are not yet algorithmic. Quasi-star removal could be made so; the redundancy/performance analysis could be automated as far as identifying candidates for removal. Gebhardt's[24] mechanisms could be applied here as well. Also, a better coupling with graphical data analysis[34] would help.

This work was sponsored by The Travelers Insurance Companies, Hartford, Connecticut. The views and conclusions contained in this report are those of the authors and do not represent Travelers positions. The authors acknowledge a great debt to the work of Don Friedman. They also thank Dan Riedinger for his technical support, and the reviewers for their helpful comments.

VI. REFERENCES

1. D. G. Friedman, "Natural hazard risk assessment for an insurance program," *The Geneva Papers on Risk and Insurance*, 9, 57-128 (1984).

2. D. G. Friedman, "Is Hugo a forerunner of future great hurricanes?" *Research Review, Journal of the Society of Insurance Research*, July, 1990.

3. Property Claim Services, *Catastrophe Bulletin*, February 24, 1993, Rahway, NJ: American Insurance Services Group.

4. G. E. Dunn and B. I. Miller, *Atlantic Hurricanes*, Louisiana State University Press, Baton Rouge, 1964.

5. *IXL: The Machine Learning System User's Manual*, IntelligenceWare, 1990.

6. K. Parsaye et. al., *Intelligent Databases: Object-Oriented, Deductive, Hypermedia Technologies*, Wiley, New York (1989).

7. B. R. Gaines, "Refining Induction into Knowledge," In *Proceedings of the 1991 AAAI Workshop on Knowledge Discovery in Databases*, G. Piatetsky-Shapiro (Ed.), AAAI, Anaheim, CA, 1991, pp. 1-10.

8. B. R. Jarvinen, C. J. Neumann, and M. A. S. Davis, *A Tropical Cyclone Data Tape for the North Atlantic Basin*, Technical Memorandum NWS NHC 22, National Oceanic and Atmospheric Administration and National Weather Service, 1984.

9. R. S. Michalski, "A theory and methodology of inductive learning," In *Machine Learning: An Artificial Intelligence Approach*, R. Michalski, J. Carbonell, and T. Mitchell (Eds.), Tioga, Palo Alto, 1983, pp. 83-133.

10. S. M. Weiss, R. S. Galen and P. V. Tadepalli, "Maximizing the predictive value of production rules," *Artificial Intelligence* 45, 47-71 (1990).

11. G. Piatetsky-Shapiro, "Discovery, analysis and presentation of strong rules," In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley (Eds.), AAAI/MIT Press, 1991, pp. 229-248.

12. W. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, "Knowledge discovery in databases: an overview," In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley (Eds.), AAAI/MIT Press, 1991, pp. 1-27.

13. R. S. Michalski and R. E. Stepp, "Learning from observation: conceptual clustering," In *Machine Learning: An Artificial Intelligence Approach*, R. Michalski, J. Carbonell, and T. Mitchell (Eds.), Tioga, Palo Alto, 1983, pp 331-363.

14. D. Pregibon, "A statistician's view of knowledge discovery in data (KDD) - what are important long term directions?" In "Panel Positions on 'Hilbert' problems in KDD," addendum to *Proceedings of the 1991 AAAI Workshop on Knowledge Discovery in Databases*, G. Piatetsky-Shapiro (Ed.), AAAI, Anaheim, CA, 1991, pp. 9-10.

15. R. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

16. H. L. Harter, "Early history of multiple comparison tests," In *Handbook of Statistics*, P.R. Krishnaiah (Ed.), North-Holland, New York, 1980, pp. 617-622.

17. M. Abramowitz and I. A. Stegun (Eds.), *Handbook of Mathematical Functions*, U. S. Department of Commerce, National Bureau of Standards, 1972.

18. G. W. Cry, *Tropical Cyclones of the North Atlantic Ocean*, Technical Paper No. 55, U. S. Department of Commerce, Weather Bureau, 1965.

19. J. A. Major and D. R. Riedinger, "EFD: A hybrid knowledge/statistical-based system for the detection of fraud," *International Journal of Intelligent Systems*, 7, 687-703 (1992).

20. G. Piatetsky-Shapiro and C. J. Matheus, "Knowledge discovery workbench for exploring business databases," *International Journal of Intelligent Systems*, 7, 675-686 (1992).

21. J. Hope and C. J. Neumann, "An operational technique for relating the movement of existing tropical cyclones to past tracks," *Monthly Weather Review*, 98, 925-933 (1970).

22. W. R. Moninger et. al., "Shootout-89, a comparative evaluation of knowledge-based systems that forecast severe weather," *Bulletin of the American Meteorological Society*, 72, 1339-1354 (1991).

23. B. R. Gaines, "The trade-off between knowledge and data in knowledge acquisition," In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley (Eds.), AAAI/MIT Press, 1991, pp. 491-505.

24. F. Gebhardt, "Choosing among competing generalizations," *Knowledge Acquisition*, 3, 361-380 (1991).

25. J. R. Quinlan, "Generating production rules from decision trees," In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, 1987, pp. 304-307.

26. D. Jensen, "Knowledge discovery through induction with randomization testing," In *Proceedings of the 1991 AAAI Workshop on Knowledge Discovery in Databases*, G. Piatetsky-Shapiro (Ed.), AAAI, Anaheim, CA, 1991, pp. 148-159.

27. W. M. Shen, "Discovering regularities from knowledge bases," *International Journal of Intelligent Systems*, 7, 623-635 (1992).

28. A. Ginsberg, S. M. Weiss, P. Politakis, "Automatic knowledge base refinement for classification systems," *Artificial Intelligence*, 35, 197-226 (1988).

29. W. Klosgen, "Problems for knowledge discovery in databases and their treatment in the statistics explorer Explora," *International Journal of Intelligent Systems*, 7, 649-673 (1992).

30. D. B. Lenat, "The role of heuristics in learning by discovery: three case studies," In *Machine Learning: An Artificial Intelligence Approach*, R. Michalski, J. Carbonell, and T. Mitchell (Eds.), Tioga, Palo Alto, 1983, pp. 243-306.

31. W. Buntine, "Generalized subsumption and its applications to induction and redundancy," *Artificial Intelligence*, 36, 149-176 (1988).

32. P. H. Winston, "Learning structural descriptions from examples," In *The Psychology of Computer Vision*, P. H. Winston (Ed.), McGraw-Hill, New York, 1975, ch. 5.

33. J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

34. G. Grinstein, et. al., "Visualization for knowledge discovery," *International Journal of Intelligent Systems*, 7, 637-648 (1992).

APPENDIX

Potentially Interesting rules and selection commentary.

(Note: Genuinely Interesting rules are italicized.)

RULE	DEFINITION	CVG	CF	PI, TI, GI COMMENTARY
R44 latit	in [22.3,28.4]	405	30.4%	PI: Gen:{R453}. Note: MGR TI? Yes, S|TOP=4.9 GI? No, low performance+sense.
R49 longit	*in [62.9,95.7]*	*922*	*30.0%*	*PI: Gen:{R342,R481}. Note: MGR* *TI? Yes, S|TOP=17.0* *GI? Yes*
R56 longit	*in [73.1,97.3]*	*513*	*39.2%*	*PI: Gen:{R59,R334,R338,R426,R453, R488}. Note: MGR* *TI? Yes, S|TOP=26.5* *GI? Yes*
R59 longit	*in [83.5,97.3]*	*252*	*47.6%*	*PI: Frontier* *TI? Yes, S|R56=2.9* *GI? Yes. Chosen as Gulf hot spot representative.*
R97 dist.coast	*in [90,450]*	*538*	*30.5%*	*PI: Gen:{R338}. Note: MGR* *TI? Yes, S|TOP=7.7* *GI? Yes.*
R100 dist.coast	in [90,360]	415	32.8%	PI: Gen:{R338} TI? No, S|R97<2
R107 coast.ang	in [44,166]	424	37.5%	PI: Gen:{R484. Note: MGR TI? Yes, S|TOP=17.1 GI? No, Gulf hot spot.
R110 coast.ang	in [112,176]	117	54.7%	PI: Gen:{R490} TI? No, S|R114<2

R114 coast.ang	in [84,176]	237	46.0%	PI: Gen:{R351, R490}. Note: MGR TI? No. S|TOP=18.6, but found to specialize R56 with S|R56<2.
R119 trkcst.ang	*in [-180,15]*	*942*	*30.6%*	*PI: Frontier and gen:{R121,R352, R490}. Note: MGR* *TI? Yes, S|TOP=20.1* *GI? Yes.*
R121 trkcst.ang	in [-180,1]	818	30.8%	PI: Frontier TI? No, S|R119<2
R131 trkcst.ang	in [-38,1]	495	32.1%	PI: In family R119 TI? No, S|R119<2
R134 trkcst.ang	in [-60,-23]	296	33.1%	PI: In family R119 TI? No, S|R119<2
R135 trkcst.ang	in [-22,3]	325	32.9%	PI: In family R119 TI? No, S|R119<2
R157 par.speed	*in [-19.75,1]*	*884*	*30.2%*	*PI: Gen:{R160}. Note: MGR* *TI? Yes, S|TOP=16.4* *GI? Yes*
R160 par.speed	in [-19.75,-1]	717	31.0%	PI: Frontier TI? No, S|R157<2
R163 par.speed	in [-19.75,-3.25]	525	31.1%	PI: Spec:{R160} TI? No, S|R157<2
R167 par.speed	in [-7.5,-3.25]	336	31.3%	PI: Spec:{R160} TI? No, S|R157<2
R194 date &longit	in [1.01,8.22] in [73.1,97.3]	142	45.8%	PI: Spec:{R56} TI? No, S|R56}<2
R262 date &trkcst.ang	in [8.14,9.16] in [-180,-24]	192	38.5%	PI: Spec:{R119} TI? No, S|R119<2
R256 date &trkcst.ang	in [1.01,9.07] in [-180,15]	448	35.5%	PI: Spec:{R119} TI? No, S|R119<2
R296 storm.type &coast.ang	= 1 in [112,176]	114	55.3%	PI: Gen:{R490} TI? No, S|R114<2
R305 trkcst.ang *&wind.speed*	*in [-180,15]* *in [80,140]*	*187*	*42.8%*	*PI: Spec:{R119}* *TI? Yes, S|R119=2.1* *GI? Yes*
R334 longit *&track.ang*	*in [73.1,97.3]* *in [27,116]*	*228*	*54.8%*	*PI: Frontier* *TI? Yes, S|R56=8.0* *GI? Yes*

R337 longit	in [83.5,97.3]	188	51.6%	PI: Gen:{R338}
&dist.coast	in [90,450]			TI? No, S|R59<2
R338 longit	in [83.7,97.3]	144	54.9%	PI: Frontier
&dist.coast	in [90,360]			TI? No, S|R59<2
R339 longit	in [73.1,97.3]	142	52.8%	PI: Spec:{R56}
&trkcst.ang	in [-10,84]			TI? Yes, S|R56=2.0
				GI? No, redundant against R334
R342 longit	*in [62.9,95.7]*	*518*	*39.8%*	*PI: Frontier*
&inw.speed	*in [2.5,22.5]*			*TI? Yes, S|R49=10.3*
				GI? Yes
R343 longit	in [83.7,97.3]	115	53.0%	PI: Spec:{R56}
&inw.speed	in [0,6.75]			TI? No, S|R59<2
R344 dist.coast	in [90,684]	560	33.4%	PI: Frontier. Note: MGR
&track.spd	in [0,12.75]			TI? Yes, S|TOP=13.4
				GI? No, redundant against R342
R350 track.ang	in [59,175]	176	48.3%	PI: Gen:{R484}
&coast.ang	in [44,166]			TI? Yes, S|R107=2.1
				GI? No, Gulf hot spot
R351 track.ang	in [27,116]	115	60.0%	PI: Frontier
&coast.ang	in [84,176]			TI? No, S|R114=2.8, but S|R334<2
R352 trkcst.ang	*in [-180,15]*	*93*	*68.8%*	*PI: Frontier*
&track.ang	*in [59,175]*			*TI? Yes, S|R119=13.1*
				GI? Yes, Gulf, but high-performance
R357 par.speed	in [-19.75,3.25]	154	50.7%	PI: Gen:{R481,R484}. Note: MGR
&track.ang	in [59,175]			TI? Yes, S|TOP=16.0
				GI? No, Gulf hot spot
R359 dist.coast	in [90,822]	466	39.9%	PI: Frontier. Note: MGR
&inw.speed	in [2.5,22.5]			TI? Yes, S|TOP=24.4
				GI? No, redundant against R342
R366 date	in [1.01,8.22]	135	46.7%	PI: Spec:{R56}
&storm.type	= 1			TI? No, S|R56<2
R426 date	in [8.23,10.04]	95	67.4%	PI: Frontier
&longit	in [73.1,97.3]			TI? No, S|R334<2
R428 date	in [1.01,8.22]	124	47.6%	PI: Spec:{R56}
&longit	in [73.1,97.30]			TI? No, S|R56<2
R432 date	in [1.01,9.07]	239	43.9%	PI: Spec:{R342}
&longit	in [62.9,95.7]			TI? No, S|R342<2
&inw.speed	in [2.5,22.5]			

```
R450 storm.type = 1              149  49.7% PI:  Spec:{R56}
    &wind.speed in [15,50]                 TI?  No, S|R59<2
    &longit     in [83.5,97.30]

R451 storm.type = 1              126  50.0% PI:  Spec:{R56}
    &wind.speed in [15,45]                 TI?  No, S|R59<2
    &longit     in [83.7,97.3]

R453 storm.type = 1              114  61.4% PI:  Frontier
    &latit      in [22.3,28.4]             TI?  No, S|R59<2
    &longit     in [83.7,97.3]

R456 storm.type = 1              139  54.0% PI:  Spec:{R339}
    &longit     in [73.1,97.30]            TI?  No, S|R339<2
    &trkcst.ang in [-10,84]

R473 wind.speed in [110,155]      10 100.0%PI:  Spec:{R160}
    &coast.ang  in [-26,29]                TI?  No, S|R160<2
    &par.speed  in [-19.75,-1]

R480 longit     in [84.4,97.3]    10 100.0%PI:  Spec:{R56}
    &trk.speed  in [10,13.25]              TI?  No, S|R59<2
    &coast.ang  in [84,111]

R481 longit     in [62.9,95.7]   131  58.8% PI:  Frontier
    &par.speed  in [-19.75,3.25]           TI?  Yes, S|R49=10.5, S|R357=3.7
    &track.ang  in [59,175]                GI?  No, Gulf hot spot

R484 coast.ang  in [44,166]      101  67.3% PI:  Frontier
    &track.ang  in [59,175]                TI?  Yes, S|R350=5.9, S|R357=5.6
    &par.speed  in [-19.75,3.25]           GI?  No, Gulf hot spot

R488 date       in [8.23,10.04]   93  68.8% PI:  Frontier
    &storm.type = 1                        TI?  No, S|R334<2
    &longit     in [73.1,97.3]
    &trk.angle  in [27,116]

R490 date       in [9.27,10.22]   12 100.0%PI:  Frontier
    &storm.type = 1                        TI?  No, S|R114<2
    &coast.ang  in [112,176]
    &trkcst.ang in [-38,1]

R511 storm.type = 1               10 100.0%PI:  Spec:{R56}
    &latit      in [26.3,35.5]             TI?  No, S|R59<2
    &longit     in [83.5,97.3]
    &track.spd  in [9.75,16.5]

R512 storm.type = 1               10 100.0%PI:  Spec:{R56}
    &latit      in [26.3,35.5]             TI?  No, S|R59<2
    &longit     in [83.5,97.3]
    &inw.speed  in [5,14.25]
```