Spatio-Temporal Relational Probability Trees: A Study of Brittleness

EM Project Checkpoint 5

Nathaniel P. Troutman
University of Oklahoma
ntroutman@acm.org

ABSTRACT

Spatio-temporal relational probability trees (SRPTs) are a recent technique for classifying relational data that is located in space and varying in time. Some of its applications include tornado-genesis prediction and prediction of drought propagation. The SRPT algorithm employs stochastic techniques to search the space of possible trees. A side effect of stochastic searches is that the solutions are brittle, with independent runs generating SRPTs with similar classification power, but with very different structures. Many of the domain experts in fields in which SRPTs could be used are wary of brittle solutions, regardless of their similar ability to classify. This paper studies the brittleness of SRPTs under a variety of conditions with the goal of being able to stabilize the solutions produced by the algorithm.

1. INTRODUCTION

Classification is one of the core problems of machine learning. Many approaches have been taken to solve the problem including: *support vector machines*, *Bayesian networks*, *hidden Markov models* and *decision trees*. One benefit of using decision trees for classification is that they are human-readable, which is important when applying machine learning techniques to domains with their own experts. The spatio-temporal relational probability tree (SRPT) is a recent extension of the classical decision tree which allows for the classification of relational data that is located in space and varying in time. Relational models have greater representational power than non-relational models and are able to easily capture specific attributes of data sets. The addition of space and time makes these models dynamic.

Due to the size of the search space of potential trees, the SRPT algorithm is stochastic. This stochastic element introduces brittleness into the solutions generated. Domain experts in such fields as meteorology are wary of the fact that independent runs over identical data and parameter sets can produce different solutions. However, brittleness and the wariness of domain experts are a concern not only in SRPTs but also in other machine learning techniques such as feature selection [Kalousis et al. 2007].

This paper studies the brittleness of solutions created by the SRPT algorithm. To form an idea of the effects of the learning algorithm parameters, we performed a factorial experiment over a range of parameter values and on three different datasets. The goal of understanding the brittleness inherent in the solutions produced using the SRPT algorithm is to find ways to reduce this brittleness and to provide evidence of the stability and robustness of the SRPT algorithm in a variety of domains.

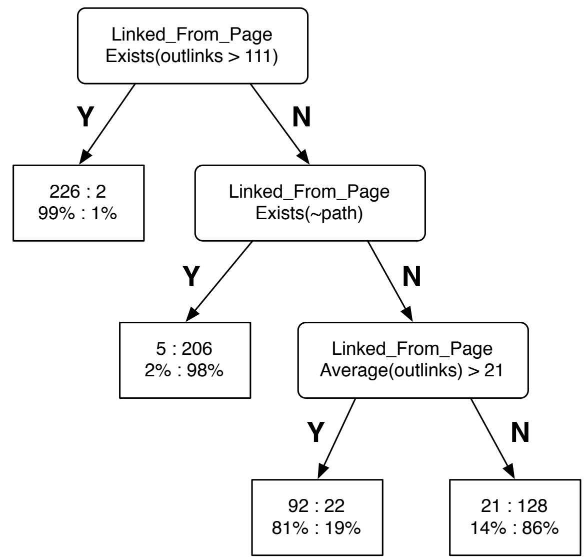

Figure 1: Example Probability Tree [Neville et al. 2003]

2. RELATIONAL PROBABILITY TREES

Relational probability trees are used for classification tasks. Like decision trees, each node that branches "asks" a question. In probability trees, these "questions" are called distinctions. A sample is a graph with objects connected to each other via relations. Both the objects and relations have attributes. The distinctions can be between the attributes of objects or between the relations joining them. Classification of a sample starts at the root node and proceeds down the tree, taking the branches that correspond to the answers to the questions asked by the distinctions. Once the search reaches the leaf nodes, the tree gives the probability that the data is or is not in a given class.

Figure 1 gives an example of a probability tree. The problem is to determine whether a webpage is or is not a student's homepage. The various web pages from the dataset connect to each other via hyperlinks which form relations. The root node contains the distinction "Exists(outlinks $>$ 111)" which asks the question if the number of outlink-relations is greater

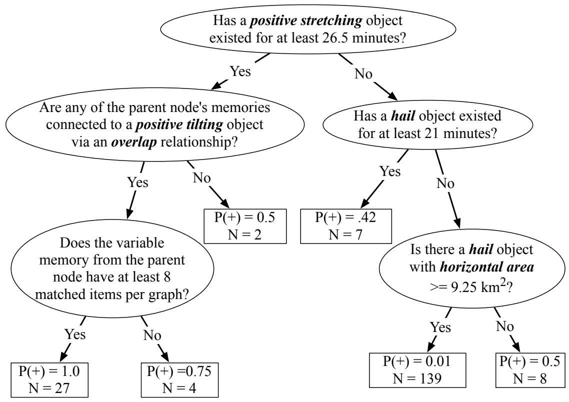

Figure 2: Example Spatio-Temporal Probability Tree [McGovern et al. 2008]

than 111. If so, there is 99% chance that the page is a student's homepage.

3. SPATIO-TEMPORAL RELATIONAL PROBABILITY TREES

Spatio-temporal relational probability trees (SRPTs) are an extension of relational probability trees. SRPTs add new distinctions that split based on the spatial, temporal, and spatio-temporal characteristics of the data. Of course, SRPTs require datasets that support such questions by including spatial and temporal data. Example distinctions include: "Does Object A wrap around Object B?" which is spatial; "Did Object A exist before Object B?" which is temporal; and "Is Object A moving closer to Object B" which is spatio-temporal. Figure 2 gives an example SRPT that is used for classifying a thunderstorm as tornado-producing or non-tornado-producing.

4. CONSTRUCTING PROBABILITY TREES

The construction of an SRPT follows a simple algorithm like the one outlined in Table 2. The tree grows recursively until adding a new node, or distinction, fails to significantly change the classification distribution. Growth of the trees is controlled by the parameters listed in Table 1. Growing an SRPT requires training using a set of pre-classified graphs. As the tree is built, the set of graphs is recursively split as it falls down the tree through each distinction. Choosing these distinctions is the main work of the learning algorithm.

4.1 Over-Fitting

All the parameters for the SRPT learning algorithm are capable of allowing the algorithm to *over-fit*. When building an SRPT, the data is broken into separate groups called "folds." A subset of the folds is used for training and the remaining folds for testing. The testing set is a stand-in for the universe and over-fitting occurs when the performance of the tree on the training set continues to improve while the performance on the testing set (the universe) becomes worse.

P – p value threshold
K – Number of samples
D – Max tree depth
S – Distinction set

Table 1: Algorithm Parameters

```
GrowTree(data, pValueThreshold, numSamples):
node = FindBestDistinction(data,
    pValueThreshold, numSamples)
yesSubset, noSubset = PartitionData(distinction, data)
node.addYesNode(
    GrowTree(yesSubset, pValueThreshold, numSamples))
node.addNoNode(
    GrowTree(noSubset, pValueThreshold, numSamples))
return node

FindBestDistinction(data, threshold, numSamples):
Loop numSamples:
    distinction = randomDistinction()
    chiSquared, pValue = X 2(distinction, data)
best = distinction yielding max chiSquared
    and p < threshold
if best is None return Leaf, else return best
```

Table 2: Pseudocode for Constructing an SRPT

4.2 P-Value Threshold

During the selection of the best distinction at each node in the tree, the distinctions are sampled stochastically and are used to split the subset of data at the current point in the tree. A contingency table is then built from the split. Using the pre-classification labels and the current classification of the growing tree, a p-value can be calculated to determine the significance of the split. As long as the p-value is below the threshold, growth continues along that branch of the tree. The p-value threshold is used to cut off the growth of a particular branch of the tree by determining what qualifies as a significant change in the distribution of the classes.

4.3 Sample Count

Since the space of possible distinctions is extremely large, stochastic sampling is used to find the best distinction. The sample count is chosen according to a heuristic that relates the number of samples to a desired confidence of ρ in sampling the top α percent to consider. Srinivasan's results say the number of samples chosen using this method is independent of the size of the search space [Srinivasan 1999].

$$K = \frac{\ln(1-\rho)}{\ln(1-\alpha)}$$

4.4 Max-Depth

As the tree develops it could potentially grow unbounded. This could result in over-fitting and unnecessary complexity. The maximum size of the tree is controlled by the *max-depth* parameter, which limits tree's depth by forcing it to terminate early. The depth can range from zero, having only a root distinction, to a depth limited only by machine memory. At times, the p-value threshold will cut off growth along a branch before the maximum depth is reached; on the other hand, max-depth can cut off the growth even if there still remain distinctions that would meet the p-value requirement.

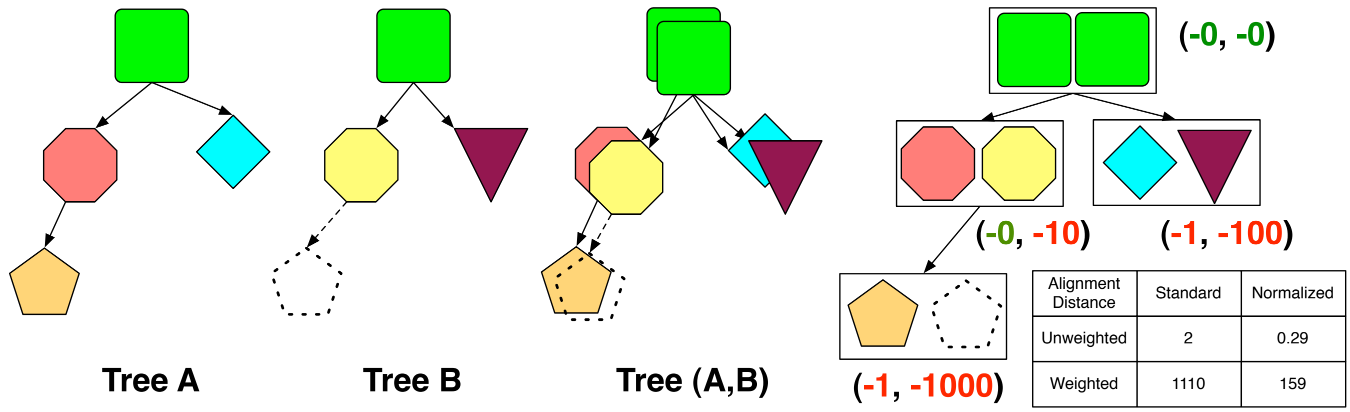

Figure 3: Calculating Alignment Distance

4.5 Distinction Set

At each node of the tree the best distinction is chosen by stochastic sampling. The possible distinctions that could be sampled are controlled by the distinction set. In order to study the SRPT algorithm under various constraints, the algorithm can be limited to using only a subset of the possible distinctions. The currently-implemented distinctions are: *Attribute*, *Count Conjugate*, *Exists*, *Partial Derivative*, *Spatial Gradient*, *Spatial Relation Change*, *Structural Conjugate*, *Temporal Exists* and *Temporal Ordering*. Any subset of the distinctions would be a valid constraint, but the three subsets are commonly used *base*, *non-temporal* and *all*. This work will only use the *all* subset.

Base	exists, temporal-exists, attribute, partial derivative, spatial-relation-change
Non-Temporal	exists, attribute, spatial-relation-change count-conjugate, structural-conjugate

5. BRITTLENESS OF TREES

The primary focus of this project is to study brittleness in the context of the SRPT algorithm. A solution is considered to be "brittle" if a small change, or even no change at all, to the parameters used by the algorithm produces radically different solutions. The differences between solutions can be both in their *structure* and *classification power*. The structure of an SRPT solution is the shape and choice of nodes in the tree produced. Since structure determines classification power, structural stability will also result in stable classification power. So in this study, we focus on structural brittleness and how the parameters of the SRPT algorithm (see Table 1) affect this.

For this study, the classification power of a SRPT is measured using the true skill statistic (TSS) which may vary from -1 to 1 [Sharan and Neville 2008]. A TSS of -1 indicates "intentionally" classifying incorrectly; a value of 0 is equivalent to the performance of a random classifier; and a value of 1 indicates perfect classification. TSS was chosen because it was designed for measuring performance on multi-class classification problems, and some of the data sets used in this study contain more than two classes.

5.1 Measuring Brittleness

Studying brittleness under parameter and experimental changes requires the choice of a metric of similarity between solutions. Since our solutions are trees, the metric must measure similarity between trees. One possibility is the *alignment-distance* metric, which measures the "distance" between two trees based upon the number of changes, deletions, and insertions required to change one tree into the other [Bille 2003]. Each of these operations is assigned an associated cost, and the total distance between trees is the total weighted sum of the operations needed to change one tree into the other. For the SRPT the change operation will be divided into three weighted operations: adding/deleting a missing/extra distinction; changing the distinction type; and changing the parameters of the distinction.

An example of calculating the alignment distance between trees is illustrated in Figure 3. In this example the shapes represent the different types of distinctions and the color represents the parameters for the distinctions. Tree's A and B are aligned on top of each other to form the new tree A, B whose nodes are pairs of distinctions. The alignment is a simple operation as the tree has an inherent ordering of the nodes, with yes and no always being one the same side. The distance between the trees is calculated by taking the sum of a distance function applied to each pair of nodes.

5.2 Picking the Distance Function

First, calculating the alignment-distance requires modifying the trees by adding to them a number of null-nodes such that the two trees become isomorphic when ignoring the actual distinctions made by the real nodes. Next, the two trees are overlapped, which produces a pair of distinctions at each node. The distance function $\gamma(\delta_1, \delta_2)$ gives the distance between the distinctions δ_1 and δ_2. The alignment-distance of the trees is the sum of the distances between all the pairs.

The choice of the distance function is an important one. In this study two different functions are used: γ is an unweighted function and γ_w is a weighted one. The distinctions have a type, τ, such as *Exists*, *Structural-Change*, etc., and a set of parameters, ρ. A sample distinction would be "Does there *exist* an object of type *acyclic-downdraft*. In this example τ would be *exist* and *rho* would be *acyclic-downdraft*. For the null-nodes, $\tau = \rho = \emptyset$ The distance functions are defined as follows:

$$\gamma(\delta_{\rho_1}^{\tau_1}, \delta_{\rho_2}^{\tau_2}) = \begin{cases} 1 \text{ if } \tau_1 \neq \tau_2 \\ 0 \text{ otherwise} \end{cases}$$

and

$$\gamma_w(\delta_{\rho_1}^{\tau_1}, \delta_{\rho_2}^{\tau_2}) = \begin{cases} 1000 \text{ if } \tau_1 = \emptyset \text{ or } \tau_2 = \emptyset \\ 100 \text{ if } \tau_1 \neq \tau_2 \\ 10 * |(\rho_1 \cup \rho_2) - (\rho_1 \cap \rho_2)| \text{ if } \rho_1 \neq \rho_2 \\ 0 \text{ otherwise} \end{cases}$$

5.3 Removing Bias

The alignment-distance is biased towards the depth of the tree. Since deeper trees have more nodes, there are more possibilities for discrepancies between trees. Thus, shallower trees have a smaller maximum-alignment distance than deeper trees. To correct for this bias, a normalization term of $1/\sum_{i=0}^{D} 2^i = 2^{D+1} - 1$ is applied to the alignment-distance. This results in a *normalized brittleness metric* that is a per-node distance, which makes possible the comparison of trees of different maximum depths.

6. DATA SETS

In this paper, brittleness is studied over three datasets: one is synthetic; the other two represent real-word problems. The *Shapes-World* data set is a very simple synthetic data set. The simulated worlds from which the data are derived contain several three-dimensional shapes that exist in time and have temporally-varying attributes. The shapes are either a ball, a pyramid, or a cube, and are colored either red, green or blue. The shapes' geometric form and color is static throughout the extent of the world, but they have a temporally-varying volume. Each shape is in at least one relation with another shape in the world. The relations between shapes are either Nearby or OnTopOf, and these have an associated distance which varies temporally.

The *Reality* dataset is a collection of data about the cell phone usage of approximately 100 people at MIT. It was collected by MIT's Reality Mining group.[1] Data were collected by recording which cell phones were in the vicinity of each other, when the phones were in use, their proximity to a tower and the proximity of any blue-tooth devices. This dataset has temporal and spatial aspects with a very small number of relation and attribute types. The cell-phones are classified as belonging to a student, a faculty/staff member, a Sloan associate, or unknown. Following the work of Sharan and Neville, the data are aggregated into 14-day periods to reduce the resolution [Sharan and Neville 2008].

The *Tornado* dataset is a very large one used for predicting which storms might be tornado-producing. The dataset was generated by the Advanced Regional Prediction System (ARPS) which used a non-hydrostatic model [Xue et al. 2003] with a selection of appropriate environmental variables [Rosendahl 2008] to produce 163 simulated supercell thunderstorms. The model was run for three hours, with the state of the simulation saved every 30 seconds of simulated time. Multiple storms which occurred in a single simulation were tracked separately, so as to allow the prediction of tornado formation in the individual storms. The storms contain sixteen different types of objects and four possible relations between the objects. The objects, their attributes, and the relations vary temporally.

[1] Reality Mining: http://reality.media.mit.edu

7. SEARCH SPACE

The size of the search space of distinctions at a given node depends upon two factors: the set of distinctions being searched, and the data set. The SRPT algorithm makes a total of ten distinctions, divided into two classes, *base* and *conjugate*, as shown below:

Base	exists, temporal-exists, attribute, partial derivative, spatial-relation-change
Conjugate	count-conjugate, structural-conjugate, temporal-ordering

The following calculations yield the upper bound on the size of the search space for each dataset. $N_{objTypes}$ refers to the number of object types and $N_{relTypes}$ refers to the number of relation types. $\max(graphExtent)$ is the maximum extent of all the graphs. $N_{attributeTypes}$ is the number of attribute types and $N_{temporalAttributeTypes}$ is the number of attribute types which are temporal. $N_{attributeValues}$ is the number of unique attribute values of all types across all timesteps. $N_{baseDist}$ is the total possible number of unique distinctions of the types in the base subset.

Exists splits on the existence of a graph containing a relation or object of the given type:
$N_{exists} = N_{objTypes} + N_{relTypes}$

Temporal-exists splits on the existence of a object or relation with a given type having lasted for a given time:
$N_{TExists} = \max(graphExtent) * (N_{objTypes} + N_{relTypes})$

Attribute splits on an object's having an attribute with a statistical value *(MEDIAN, MODE, MAX, MIN, MEAN, GREATERTHAN)* with a relation to a given splitValue:
$N_{attrib} = N_{attributeTypes} * N_{attributeValues} * N_{splitStatistics}$

Partial-derivative splits on the temporal change of an attribute value being greater than a given threshold:
$N_{PD} = N_{uniqueValueDifferences} * \max(graphExtent)$

Spatial-relation-change splits on a graph's having a relation of type A between an object of type X and another of type Y, that changes to a relation of type B at some point during the extent of their containing graph:
$N_{SRChange} = (N_{objTypes})^2 * (N_{relTypes})^2$

Count-conjugate splits on the number of items returned by a base distinction being greater than a given threshold value:
$N_{countConj} = N_{baseDist} * \max(\max(N_{objects}, N_{relations}) \ \forall \text{ graphs})$

Structural-conjugate splits on an object of a given type from a base distinction being related via a relation of a given type:
$N_{structuralConj} = N_{baseDist} * N_{objTypes} * N_{relTypes}$

Temporal-Ordering checks to see if any objects or relations returned by two base distinctions have a particular temporal order *(BEFORE, MEET, OVERLAP, EQUAL, START, FINISH, DURING)*:
$N_{TOrdering} = (N_{baseDist})^2 * N_{temporalOrderings}$

7.1 Shapes

Dataset Schema	Values
# Object Types	3
# Relation Types	2
# Attribute Types	8
# Temporal AttributeTypes	5
max($graphExtent$)	10

Distinction Sizes	
$N_{exists} =$	5
$N_{TExists} =$	45
$N_{attrib} =$	$1,344$
$N_{PD} =$	$34,020$
$N_{SRChange} =$	13
$N_{baseDist} =$	5
$N_{countConj} =$	$318,843$
$N_{structuralConj} =$	$212,562$
$N_{TOrdering} \approx$	8.786×10^9
$total \approx$	8.786×10^9

7.2 Reality

Dataset Schema	Values
# Object Types	4
# Relation Types	4
# Attribute Types	4
# Temporal Attribute Types	3
max($graphExtent$)	16

Distinction Sizes	
$N_{exists} =$	8
$N_{TExists} =$	128
$N_{attrib} \approx$	$5,376$
$N_{PD} \approx$	$2,397,696$
$N_{SRChange} =$	32
$N_{baseDist} \approx$	$2,403,240$
$N_{countConj} \approx$	5.566×10^9
$N_{structuralConj} \approx$	38451840
$N_{TOrdering} \approx$	4.043×10^{13}
$total \approx$	4.043×10^{13}

7.3 Tornado

Dataset Schema	Values
# Object Types	14
# Relation Types	5
# Attribute Types	299
# Temporal Attribute Types	291
max($graphExtent$)	340

Distinction Sizes	
$N_{exists} =$	18
$N_{TExists} =$	$6,102$
$N_{attrib} \approx$	1.577×10^9
$N_{PD} \approx$	7.627×10^{16}
$N_{SRChange} =$	212
$N_{baseDist} \approx$	7.627×10^{16}
$N_{countConj} \approx$	8.160×10^{18}
$N_{structuralConj} \approx$	4.271×10^{18}
$N_{TOrdering} \approx$	4.071×10^{31}
$total \approx$	4.071×10^{31}

8. EFFECTS OF SAMPLING ON TREE GROWTH

SRPTs are grown using stochastic sampling to choose distinctions. For two identical trees to be grown, assuming the same dataset, the choices for each distinction must be the same. It is not necessary that the distinctions be sampled in the same order at corresponding points in the different trees, nor is it necessary that exactly the same set of distinctions be sampled; it is only necessary that the best distinction be sampled in both cases.

The probability of the same distinction being sampled at corresponding points in the tree is directly related to the size of the distinction space and the number of samples taken. Further, when extended to producing identical trees, ordering is required such that the distinctions ultimately chosen for each node in the tree will be sampled in the same order. In addition, the probability of generating identical trees is dependent not only on the size of the distinction space and number of samples, but also on the size of the tree as well.

9. EXPECTATIONS

9.1 Max-Depth

Max-depth exerts a potentially strong influence on the structure of solutions. In instances where significant changes to the class distribution can still be made, a low max-depth will cut off the growth of the tree before it reaches its full classifying potential. As max-depth increases, the size of the tree increases. As mentioned above in the discussion on sampling and tree growth, it is expected that as the max-depth increases, structural brittleness will also increases. As max-depth increases, the number of nodes in the tree grows exponentially. So, if the other parameters are held constant, it follows that the relation between max-depth and brittleness is non-linear.

9.2 Sample count

Sample count affects the coverage of the distinction space. Given a distinction space of fixed size, the higher the sample count, the better the coverage attained over the space. As coverage of the distinction space increases, the probability of picking the same distinction twice in independent runs increases as well. With the increased probability of selecting the same distinction comes a decrease in structural brittleness. Consequently, if the sample count decreases, the structural brittleness increases. It is expected, then, that the relation between sample count and brittleness will be sub-linear and asymptotic as brittleness approaches zero.

9.3 P-value Threshold

P-value threshold determines the point at which adding new nodes to the tree is no longer considered significant. Since this threshold affects the depth of the tree, it is expected that changing the p-value threshold will yield results similar to those produced by changing max-depth. As the threshold increases, deeper trees will be produced, making it less likely that identical trees will be produced on independent runs. However, in cases where max-depth is forcing the cut-off of tree growth, a decrease in the p-value threshold will not result in larger trees and it will not affect the structural brittleness of the results. Since the best distinction is always chosen, lowering the threshold will simply al-

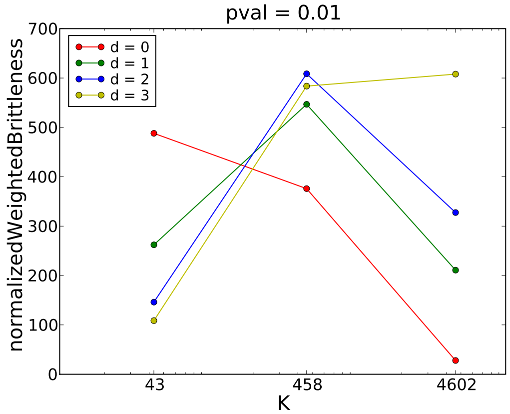

Figure 4: Shapes Results

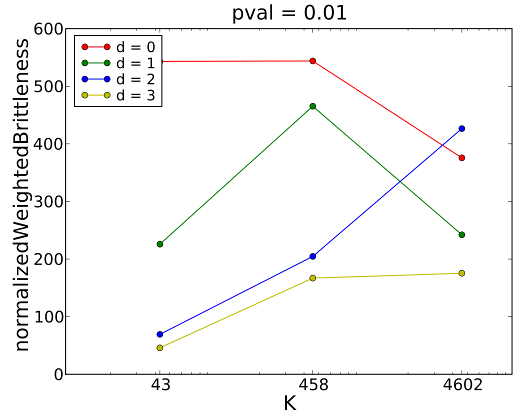

Figure 5: Reality Results

low lower threshold distinctions to be considered. However the best distinction will still be chosen, regardless of the existence of lower p-value choices. Moreover, since the algorithm counts samples regardless of their meeting the p-value threshold, the distinctions sampled will not be affected. So, the expected relation between p-value-threshold and brittleness will be non-linear.

9.4 Distinction Set

The *distinction set* directly influences the size of the distinction space being sampled. As the size of the distinction space increases, a higher number of samples is required to get the same coverage when sampling distinctions. As the distinction set grows larger, the brittleness will increase, unless the sample count is increased along with the distinction space. Since the distinction space is sampled at every node in the tree, assuming that the sampling count remains the same, we can expect to see a power relation between increased distinction set size and increased brittleness. Adding new distinctions to the set may yield drastic improvements in accuracy if the new distinctions provide useful splits when partitioning the data.

9.5 Interactions

Both the max-depth and the p-value threshold can cut off the growth of the tree. Since brittleness is affected by the structure and depth of the tree, the factors that control tree depth are likely to influence brittleness as well. Thus we expect to see some interaction between max-depth and p-value threshold and their respective effects on brittleness. We also expect to see an interaction effect between the number of samples and max-depth. The deeper the tree grows, the more samples will be needed to pick the same distinctions at each position in the tree in order to form a duplicate tree.

10. THE EXPERIMENT

A factorial experiment was carried out on each dataset with varying parameters for the SRPT algorithm. For max-depth the values used were {0, 1, 2, 3}. P-value thresholds applied were one of {1%, 10%, 50%}. The distinction set used contained all the types of distinctions. The sample counts were {46, 460, 4605}. This corresponds to sampling

Shapes

Factor	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
d	**0.000**	**0.000**	**0.000**	**0.000**
p	0.877	0.176	0.829	0.345
k	0.789	0.348	0.764	0.276
d×p	0.582	0.179	0.533	0.173
d×k	0.971	0.730	0.954	0.564
p×k	0.153	0.054	0.143	**0.034**

Reality

Factor	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
d	**0.000**	**0.003**	**0.000**	**0.003**
p	**0.039**	0.131	**0.017**	0.238
k	**0.022**	0.083	**0.012**	0.110
d×p	**0.012**	**0.003**	**0.008**	**0.003**
d×k	**0.000**	**0.002**	**0.000**	**0.002**
p×k	**0.026**	**0.047**	**0.031**	0.051

Tornado

Factor	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
d	**0.000**	**0.000**	**0.000**	**0.000**
p	**0.011**	**0.002**	**0.015**	**0.002**
k	**0.000**	**0.000**	**0.000**	**0.000**
d×p	0.134	0.161	0.186	0.219
d×k	0.971	0.730	0.954	0.564
p×k	0.807	0.709	0.759	0.601

All

Factor	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
d	**0.000**	**0.000**	**0.000**	**0.000**
p	0.96	0.346	0.918	0.551
k	0.481	**0.038**	0.479	**0.033**
d×p	0.985	0.172	0.978	0.200
d×k	0.846	0.707	0.853	0.674
p×k	0.732	0.164	0.767	0.175

Table 3: ANOVA P-Values: ***P-Values < .05 highlighted in bold***

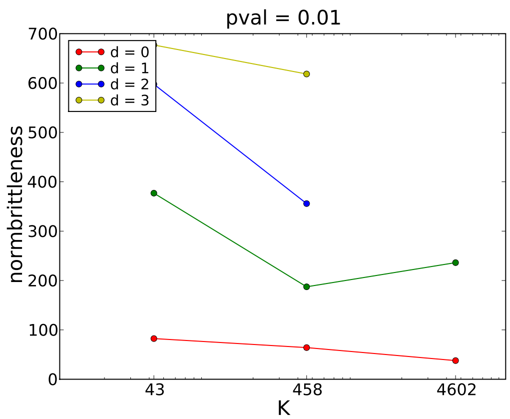

Figure 6: Tornado Results

with 99% confidence of the top {10%, 1%, 0.1%}. The parameter values for the experiments were chosen to cover a wide range of reasonable values yield a total of 108 parameter sets. Since the SRPT algorithm is stochastic and the predictions are on independent runs with varying and constant parameter sets, each set was run 30 times.

11. RESULTS

Some results from the Shapes, Reality and Tornado data sets are shown in Figure 4, 5 and 6, respectively. Each figure has brittleness shown along the vertical-axis and K along the horizontal-axis. The color of the data points indicates the max-depth. Looking at the figures, it can be seen that *max-depth* appears to confirm the predictions of its effects on brittleness. There are fairly clear layers of brittleness that increase at what appears to be an exponential rate with respect to *max-depth*. Looking at figure Figure 7, we see a lack of structure when brittleness is plotted against K with varying values of the p-value threshold.

12. ANALYSIS

12.1 Determining Significant Factors

A visual analysis provides a "feel" for the experimental results, but to check whether the trends that appear to exists are real, we used a 3-way ANOVA. The results of the 3-way ANOVA for the Shapes, Reality and Tornado data sets are shown in Table 9.5. The ANOVA reveals some very interesting information. For Shapes, of the three single factors, only depth matters at the 5% confidence level. Interestingly, when the brittleness is normalized an interaction effect appears between the p-value threshold and the number of samples. This is an unexpected result.

On the Reality data set depth again shows a significant effect on brittleness. In addition both pval and the number samples affect brittleness, but only when it is not normalized. There are also interaction effects between all the factors save for an interaction between pval and number samples under the normalized weighted distance metric. The reality results confirm our hypothesis regarding all the factors and their interaction affects in the case where a non-normalized distance function is used.

The ANOVA for Tornado differs from that of Shapes and

Shapes

Max Depth	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
0-1	0.0000	0.0000	0.0000	0.0000
1-2	0.0000	0.0000	0.0000	0.0000
2-3	0.0000	0.0000	0.0000	0.0000

Reality

Max Depth	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
0-1	0.0000	0.0278	0.0000	0.4614
1-2	0.0000	0.0000	0.0000	0.0000
2-3	0.1776	0.0000	0.1931	0.0000

Tornado

Max Depth	Unweighted	Normalized Unweighted	Weighted	Normalized Weighted
0-1	0.0000	0.0000	0.0000	0.0000
1-2	0.0000	0.0000	0.0000	0.0000
2-3	0.0000	0.0000	0.0000	0.0000

Table 4: Two-Sample Student-T Test P-Values for Differences in Mean Brittleness Between Successive Max-Depths

Reality. All three of the factors show a significant effect on brittleness, regardless of the distance function or normalization. The interaction effect between the p-value threshold and the number of samples isn't seen in the Tornado data set; instead there is an interaction effect between the max-depth and the number of samples. The ANOVA confirms the interaction that had been predicted. Though all three factors show an effect, it is interesting to note that the p-value threshold's effect is the least confident of the three.

The final ANOVA that was run combines all the results together to look at the factors and interactions regardless of the dataset. Again we see a different table than for Shapes, Reality or Tornado. Depth still shows a significant effect on the brittleness of the solutions. The number of samples also shows an effect, but only the normalized distance metrics. This is contrary to what is seen in the Reality and Shapes results, but inline with the Tornado results.

Each set of experiments has produced very different ANOVA tables. The only statistically supported result for all three datasets is that the maximum depth affects brittleness. This confirms our initial hypothesis regarding depth. The other two factors and all three interaction effects do not show a consistent significant impact on brittleness.

12.2 Effects of Depth

ANOVA reveals that depth is the only significant factor across all datasets. Table 4 lists the results of running a two-sample Student's T-Test on brittleness between consecutive depths for each data set. Both Shapes and Tornado show a significant change in mean brittleness for each pair of consecutive max-depths and each distance function at a 5% confidence level. Reality only shows a significant change between each pair of max-depths for the normalized unweighted distance function. It appears that the normalization of the distance function does allow comparisons between different max-depths to be statistically significant.

13. FUTURE WORK

While this work is mainly concerned with brittleness, it is still preferable to study problems for which the SRPT algorithm achieves a modest TSS. An alternate data set from Oklahoma's Mesonet may be used in place of Reality. So far, this study has analyzed brittleness only under the SRPT algorithm. A baseline algorithm would provide a way of gauging whether the SRPT algorithm is particularly brittle, or if the brittleness results from the nature of the problem itself and the stochastic sampling needed for extremely large search spaces. A baseline could be constructed using stochastic feature selection and a naïve Bayes classifier. A method for comparing brittleness in the naive classifier and the SRPT algorithm would be needed in order to make direct comparisons. However, even without a comparison between approaches, a baseline would still be enlightening.

14. ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under Grant No. NSF/IIS/0746816, NSF/CISE/REU 0453545 (a seed grant from the University of Oklahoma's College of Engineering), the Oklahoma Space Grant, and NASA Grant number NNG05GN42H.

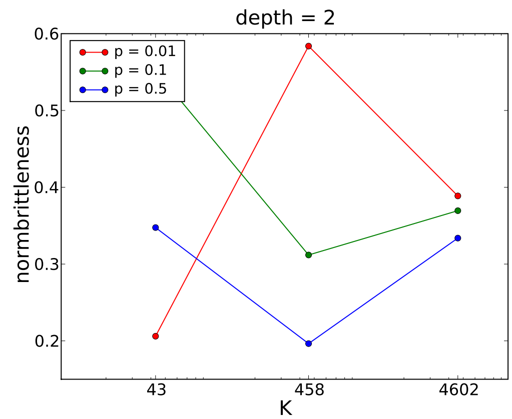

Figure 7: Lack of Structure

15. REFERENCES

Bille, P. (2003). Tree Edit Distance, Alignment Distance and Inclusion.

Kalousis, A. Prados, J. and Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge Information Systems* 12(1):95–116.

McGovern, A. Hiers, N. Collier, M. Gange II, D. J. and Brown, R. A. (2008). Spatiotemporal Relational Probability Trees. *to appear in the proceedings of ICDM 2008.*

Neville, J. Jensen, D. Friedland, L. and Hay, M. (2003). Learning relational probability trees. *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.*

Rosendahl, D. H. (2008). Identifying precursors to strong low-level rotation with numerically simulated supercell thunderstorms: A data mining approach. Master's thesis University of Oklahoma.

Sharan, U. and Neville, J. (2008). Temporal-relational classifiers for prediction in evolving domains. *to appear in the proceedings of ICDM 2008.*

Srinivasan, A. (1999). A study of two probabilistic methods for searching large spaces with ILP.

Xue, M. Wang, D. Gao, J. Brewster, K. and Droegemeier, K. K. (2003). The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteorology and Atmospheric Physics* (82):139–170.