# On Building A Data Fitting System Using Ad Hoc Models

Amy Briggs
abbr5@mst.edu

Andrew Fallgren
ajffk6@mst.edu

George Rush
gdr34b@mst.edu

## ABSTRACT

One class of data is measured or simulated data with error estimation. This data can consist of many continuous dimensions for which values are available only at discrete points. Increasing the number of discrete points at which the data is available can be expensive or even impossible to obtain, but it can still be useful for predicting data trends. Unfortunately, this is difficult when the various dimensions do not follow the same type of fit (linear, logarithmic, polynomial, etc.). Our approach focuses on building decision trees and using them to interpolate new data points that follow existing trends. This is in contrast to previous methods which focused on extrapolating data for specific applications or using purely numerical regression models. By using this approach, sparse data sets or those that exhibit unusual patterns can be analyzed effectively.

### Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*

### General Terms

Algorithms

### Keywords

data mining, sparse data, interpolation

## 1. INTRODUCTION

Outline goes here.

- The first item
- The second item
- The third etc ...

### 1.1 Stuff

This is a subsection.



#### *1.1.1 More Stuff*

This is a subsubsection.

## 2. RELATED WORK

Another section. I am citing something random [1].

## 3. METHODOLOGY

### 3.1 Decision Tree Interpolation

Decision Tree Interpolation follows this process:

1. Build a decision tree from the original data.
2. For each leaf node in the tree:
   (a) Obtain all attribute values for associated data instances.
   (b) Define ranges for attribute values.
      - For numeric attributes, define the range using minimum and maximum values.
      - For discrete attributes, define the range as all distinct values.
   (c) Calculate the distribution of all associated data instances.
   (d) Create new data points within the ranges that match the statistical distribution.

Note that the number of data points created per leaf node is proportional to the number of data points already classified by that leaf node. This ensures that any interpolated data will follow the overall data distribution, at least relative to the data density per leaf node.

### 3.2 Interpolated Data Validation

All interpolated data is validated through this process:

1. A new decision tree is built based on the interpolated data. Note that the original data is *not* included here.
2. Both the new and original decision trees are compared for accuracy against the new and original data sets.

Note that any decision tree with an arbitrarily large maximum depth can classify data with perfect accuracy. Defining a low maximum depth means that classification is imperfect, and it is under these conditions that differences in the quality of different decision trees become apparent.

Table 1: Experiment Result Summary

| Data Set | Max Tree Depth | OT -> OD | OT -> ND | NT -> OD | NT -> ND |
|---|---|---|---|---|---|
| adult_sample | 1 | 0.805527123849 | 0.780737704918 | 0.804503582395 | 0.782786885246 |
| adult_sample | 2 | 0.808597748209 | 0.809426229508 | 0.249744114637 | 0.813524590164 |
| adult_sample | 3 | 0.816786079836 | 0.850102669405 | 0.801432958035 | 0.852156057495 |
| adult_sample | 4 | 0.822927328557 | 0.794661190965 | 0.787103377687 | 0.784394250513 |
| adult_sample | 5 | 0.822927328557 | 0.84052532833 | 0.792221084954 | 0.84052532833 |
| car | 1 | 0.700231481481 | 0.710648148148 | 0.700231481481 | 0.710648148148 |
| car | 2 | 0.777777777778 | 0.783564814815 | 0.774305555556 | 0.789351851852 |
| car | 3 | 0.824074074074 | 0.809027777778 | 0.824074074074 | 0.815972222222 |
| car | 4 | 0.894097222222 | 0.903935185185 | 0.889467592593 | 0.915509259259 |
| car | 5 | 0.96412037037 | 0.966981132075 | 0.938078703704 | 0.982311320755 |
| iris | 1 | 0.666666666667 | 0.693333333333 | 0.666666666667 | 0.693333333333 |
| iris | 2 | 0.96 | 0.973333333333 | 0.946666666667 | 0.986666666667 |
| iris | 3 | 0.973333333333 | 0.959459459459 | 0.953333333333 | 0.972972972973 |
| iris | 4 | 0.98 | 0.959459459459 | 0.946666666667 | 1.0 |
| iris | 5 | 1.0 | 1.0 | 0.966666666667 | 1.0 |
| lung-cancer | 1 | 0.59375 | 0.6 | 0.375 | 0.666666666667 |
| lung-cancer | 2 | 0.625 | 0.571428571429 | 0.4375 | 0.714285714286 |
| lung-cancer | 3 | 0.625 | 0.642857142857 | 0.5625 | 1.0 |
| lung-cancer | 4 | 0.6875 | 0.428571428571 | 0.53125 | 1.0 |
| lung-cancer | 5 | 0.78125 | 0.538461538462 | 0.53125 | 1.0 |
| tic_tac_toe | 1 | 0.699373695198 | 0.68267223382 | 0.699373695198 | 0.68267223382 |
| tic_tac_toe | 2 | 0.705636743215 | 0.690376569038 | 0.703549060543 | 0.696652719665 |
| tic_tac_toe | 3 | 0.769311064718 | 0.779874213836 | 0.755741127349 | 0.758909853249 |
| tic_tac_toe | 4 | 0.831941544885 | 0.82264957265 | 0.745302713987 | 0.856837606838 |
| tic_tac_toe | 5 | 0.918580375783 | 0.907284768212 | 0.83611691023 | 0.933774834437 |
| voting | 1 | 0.95632183908 | 0.923766816143 | 0.95632183908 | 0.923766816143 |
| voting | 2 | 0.95632183908 | 0.956896551724 | 0.95632183908 | 0.956896551724 |
| voting | 3 | 0.963218390805 | 0.913357400722 | 0.95632183908 | 0.927797833935 |
| voting | 4 | 0.963218390805 | 0.892156862745 | 0.928735632184 | 0.90522875817 |
| voting | 5 | 0.972413793103 | 0.866071428571 | 0.937931034483 | 0.895833333333 |

### 3.3 Experiment Parameters

We completed 30 experiments based on two variables: data source and maximum depth of the decision tree. The maximum depth ranged from one to five, and there were six data sources pulled from Orange's documentation data sets.

## 4. RESULTS

Describe data in table 1.

## 5. DISCUSSION

Another section.

## 6. CONCLUSION AND FUTURE WORK

Last section.

## 7. REFERENCES


[1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.