

# 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

There are times when the only class of data available for analysis is measured or simulated data. Collecting such data can be time and labor intensive, and the data points may be sparse at best. As such, it can be worthwhile to use techniques that increase the utility of data points which have already been collected. Interpolation in particular can help to clarify trends within the data by generating new points according to the patterns established by existing data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

Prior methods of interpolation all work by utilizing different mathematical functions to fit patterns of values. This process can be ineffective if different dimensions or classifications of the data do not fit to a single formula, such as linear, logarithmic, polynomial, et cetera, as then certain formulas would only perform accurately for specific segments of the data.

Our proposed solution first requires modelling the data using decision trees. By doing this, it is possible to group data points by their classification and similar characteristics at various leaf nodes. We then interpolate new data instances based upon the distribution of instances within each of the leaf nodes. This produces interpolated data with a similar distribution and attribute values when compared to the original data set. Also, the new instances retain roughly the same proportion as the original data regarding the number of instances per leaf node.

By using this method against multiple well known test data sets, we demonstrate a proof of concept for its applicability across multiple problem domains.

## 2. RELATED WORK

Early works on the interpolation of scattered data evaluate a variety of different computational methods that focus on obtaining a smooth function  $F(x, y)$  to follow the data set. They utilize numerous different mathematical methods including the following: inverse distance weighted method, rectangle based blending method, triangle based blending method, finite element based method, Foley's method, global basis function type method, and modified Maude method [3]. All of these techniques focus only on developing a function in order to interpolate scattered data sets.

Later works build off of that approach, by taking classical radial basis functions, such as Duchon's thin plate splines and Hardy's multiquadratics, and compressing them in order to shorten the excessive computation times that result from applying these functions to large data sets, while trying to maintain a smooth data fitting [2].

There are reapplications of some of these interpolation functions to generate continuous surfaces from irregularly distributed data, in attempts to analyze which function best for spatial analysis. The methods include: inverse square distance method, Kriging method, tension finite difference method, and Hardy's multiquadric method [1].

Several cases can be found in which these interpolation functions are modified to more accurately apply to specific data sets. One example is the use of a combination of the thin plate smoothing spline and Kriging method in spa-

tial interpolation in order to create a more comprehensive archive of Australian climate data [4]. Another uses spatial interpolation in improving the MODIS global data sets for terrestrial gross and net primary production [5].

### 3. METHODOLOGY

Decision Tree Interpolation follows this process:

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.

All interpolated data is validated through this process:

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.

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.

Result data is shown in Table 1. The first two columns list the data source and the maximum depth for generated decision trees. Note that two decision trees are generated per row, one for the original data and one for the new (interpolated) data. The last four columns list the accuracy of both the original and new decision trees against the original

### 4.1 adult sample

## 4.2 car

### 4.3 iris

#### 4.4 lung-cancer

## 4.5 tic tac toe

## 4.6 voting

## 5. DISCUSSION

## 6. CONCLUSION AND FUTURE WORK

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

text text text text. Text text text text text text. Text text  
text text text text text. Text text text text text. Text text  
text text text text text text.

## 7. REFERENCES

- [1] C. Caruso and F. Quarta. Interpolation methods comparison. *Computers & Mathematics with Applications*, 35(12):109–126, 1998.
- [2] M. S. Floater and A. Iske. Multistep scattered data interpolation using compactly supported radial basis functions. *Journal of Computational and Applied Mathematics*, 73(1):65–78, 1996.
- [3] R. Franke. Scattered data interpolation: Tests of some methods. *Mathematics of computation*, 38(157):181–200, 1982.
- [4] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick. Using spatial interpolation to construct a comprehensive archive of australian climate data. *Environmental Modelling & Software*, 16(4):309–330, 2001.
- [5] M. Zhao, F. A. Heinsch, R. R. Nemani, and S. W. Running. Improvements of the modis terrestrial gross and net primary production global data set. *Remote sensing of Environment*, 95(2):164–176, 2005.