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 to predict data trends through extrapolation or interpolation. 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 models using known data mining techniques, and those models are then used to create new data points that follow existing trends. This is in contrast to previous approaches which mostly seemed to focus on extrapolating data for specific applications or using purely numerical models. By using this approach, any data set which is sparse or exhibits 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, extrapolation

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) Build random data points within the ranges.

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 data density will increase per leaf node, but it will remain the same *relative* to other leaf nodes. Also, the generated data is random, but the values must fall within the ranges specified by data instances already classified by each leaf node. This ensures that any generated data will be classified the same way under the original data model.

3.2 Interpolated Data Validation

All interpolated data is validated through this process:

1. Interpolated data is combined with the original data.
2. A new decision tree is built based on the combined data.
3. Both the new and original decision trees are compared for accuracy against the combined 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 becomes apparent.

3.3 Experiment Procedure

Details about experiment variables go here.

4. RESULTS

Another section.

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.