Decision Trees and Their Uses

- Companies everywhere are developing databases that contain valuable information
- This information is valuable to the companies for developing a better understanding of their customer needs
- This is a focus in data mining – this is the study of hidden patterns in large datasets
- Decision trees make a great fit for mining datasets
- Decision trees are used in most of our daily web sessions.

- Netflix has a custom decision tree that they pair with a neural network to help predict rating for suggested movie titles
- There are many techniques used in data mining, one of which are decision trees
- Decision trees can accurately classify data and make effective predictions.
- Decision trees are used in many different industries

Basic Terminology Used in Making Decision Trees

- Root Node – start of the tree
- Splitting – dividing a node into two or more sub-nodes
- Decision Node – When a sub-node is split into more sub-nodes
- Terminal Node – Nodes that do not split a.k.a. leaf nodes
- Pruning – Sub-trees or sub-nodes are removed
- Sub-tree or branch – entire subsection of the tree

Types of Decision Trees

- Types of decision trees are base on the type of target variables

Two types of trees

- Categorical Variable Decision Trees – categorical target variable
- Continuous Variable Decision Trees – has a continuous target variable

Types of Decision Trees Cont.

Decision trees are one of the most used classification and prediction machine learning techniques used today

- Hierarchical structure – easy to understand by people even if they do not have a background in C.S. or Data Mining

Advantages and Disadvantages of Decision Trees

Advantages

- Output easy to understand for people
- Use graphical representation – very intuitive
- Does not require statistical knowledge to interpret them
- One of the fastest ways to identify the most significant variables and relationship between them
- Require less data cleaning compare to some other data modeling techniques (outliers, noise, and missing data)

Disadvantages

- Biggest disadvantage is overfitting

Algorithms Used in Decision Trees

Most used are:

- ID3 – categorical
- CART – categorical
- CHAID – categorical
- Reduction in Variance – continuous

Algorithms Used in Decision Trees

Others used are:

- Sprint, SLIQ – multiple sequential scans of data are ok for millions of examples
- VFDT – at most one sequential scan ok for billions of examples

Real world applications

- Engineering
- Business
- Financial
- Medical

Real world applications - Engineering

- Diagnosis of faulty machinery
- Common in rotary machines that rely on ball bearings – most important part
- Measure vibration and acoustic emission signals to detect faulty bearing (old way). This has many variables that contain noise
- Decision tree works well because there are ways of removing irrelevant data
- C4.5 algorithm is used

Real world applications – Engineering Cont.

- Utility industry
- Decision tree preferred method for finding customer needs and consumption

Real world applications – Business

- Used to improve customer experience and services
- Feasible way of extracting useful data
- Often done in online shopping to see frequency of purchases and then classify that data.

Real world applications – Healthcare

- Used for making predictions
- Help with diagnosis, treatment options, patient evaluation, and can identify special medical conditions
- Studies done in developmentally-delayed children
- Decision trees can find hidden knowledge in medical history
- Decision tree can identify majority of illnesses that result in delays of cognitive-development, language development, and motor skills
- Accuracies average around 88%

Other Areas Experimenting with Decision Trees

- Bioinformatics – assigning protein function and predicting splice sites.
- Psychology – exploring possible decision tree uses