MACHINE INTELLIGENCE

Issues in Decision Tree Learning and solutions to it

K.S.Srinivas
Department of Computer Science and Engineering

MACHINE INTELLIGENCE

Issues in Decision Tree Learning and solutions to it

Srinivas K S.
Associate Professor, Department of Computer Science

MACHINE INTELLIGENCE

Issues in Decision Tree Learning

- The most important issues in decision tree comes at two place
- over fitting of data
- handling continuous attributes
- we will discuss each of the problem and try to find an appropriate solutions for it

MACHINE INTELLIGENCE

Over fitting of data

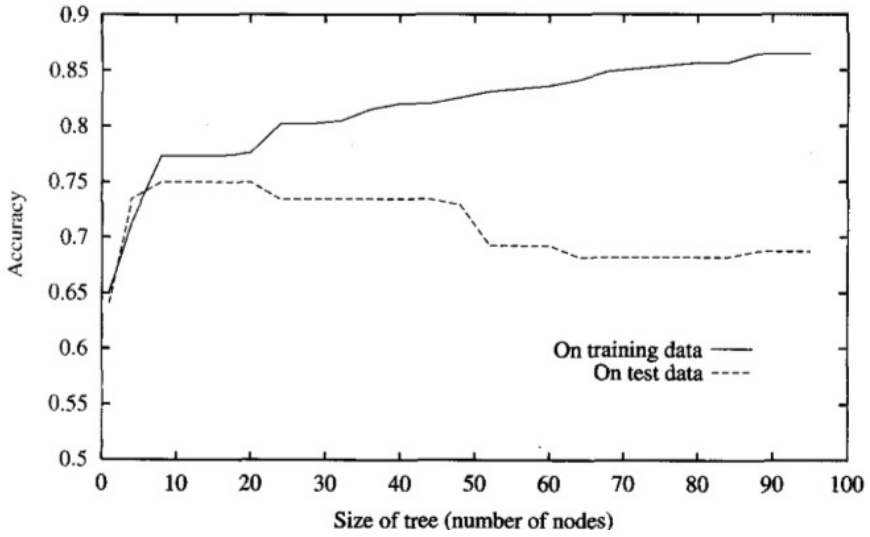
- The ID3 algorithm grows each branch of the tree just deeply enough to perfectly classify the training examples
- It can lead to difficulties when there is noise in the data, or when the number of training examples is too small to produce a representative sample of the true target function. In either of these cases, this simple algorithm can produce trees that over fits the training examples.
- We will say that a hypothesis over fits the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances

- **Definition:**Given a hypothesis space H, a hypothesis h € H is said to over fit the training data if there exists some alternative hypothesis h' € H, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances.

MACHINE INTELLIGENCE

Over fitting of data

- The graph illustrates the impact of over fitting in a typical application of decision tree learning.
- In this case, the ID3 algorithm is applied to the task of learning which medical patients have a form of diabetes
- The horizontal axis of this plot indicates the total number of nodes in the decision tree, as the tree is being constructed. The vertical axis indicates the accuracy of predictions made by the tree.
- The solid line shows the accuracy of the decision tree over the training examples, whereas the broken line shows accuracy measured over an independent set of test examples (not included in the training set).
- Predictably, the accuracy of the tree over the training examples increases monotonically as the tree is grown. How- ever, the accuracy measured over the independent test examples first increases, then decreases. As can be seen, once the tree size exceeds approximately 25 nodes
- further elaboration of the tree decreases its accuracy over the test examples despite increasing its accuracy on the training examples.



MACHINE INTELLIGENCE

Avoiding Over fitting of data

- One way this can occur is when the training examples contain random errors or noise.
- There are several approaches to avoiding over fitting in decision tree learning. These can be grouped into two classes:
    1. approaches that stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data,
    2. approaches that allow the tree to over fit the data, and then post-prune the tree.
- Although the first of these approaches might seem.more direct, the second approach of post-pruning over fit trees has been found to be more successful in practice.
- This is due to the difficulty in the first approach of estimating precisely when to stop growing the tree.

MACHINE INTELLIGENCE

Avoiding Over fitting of data

- Regardless of whether the correct tree size is found by stopping early or by post-pruning, a key question is what criterion is to be used to determine the correct final tree size.
- Approaches include:
    1. Use a separate set of examples, distinct from the training examples, to evaluate the utility of post-pruning nodes from the tree.
    2. Use all the available data for training, but apply a statistical test to estimate whether expanding (or pruning) a particular node is likely to produce an improvement beyond the training set. For example, Quinlan (1986) uses a chi-square test to estimate whether further expanding a node is likely to improve performance over the entire instance distribution, or only on the current sample of training data.
    3. Use an explicit measure of the complexity for encoding the training examples and the decision tree, halting growth of the tree when this encoding size is minimized.

MACHINE INTELLIGENCE

Dealing with continuous value

- Our initial definition of ID3 is restricted to attributes that take on a discrete set of values.
- First, the target attribute whose value is predicted by the learned tree must be discrete valued.
- Second, the attributes tested in the decision nodes of the tree must also be discrete valued
- This second restriction can easily be re- moved so that continuous-valued decision attributes can be incorporated into the learned tree.
- This can be accomplished by dynamically defining new discrete-valued attributes that partition the continuous attribute value into a discrete set of intervals.
- In particular, for an attribute A that is continuous-valued, the algorithm can dynamically create a new Boolean attribute A_c, that is true if A < c and false otherwise. The only question is how to select the best value for the threshold c.

MACHINE INTELLIGENCE

Dealing with continuous value

- As an example, suppose we wish to include the continuous-valued attribute Temperature in describing the training example days in the learning task in the given table

TEMP	40	48	60	72	80	90
Play Sport	no	no	yes	yes	yes	no

- What threshold-based Boolean attribute should be defined based on Temperature?
- Clearly, we would like to pick a threshold, c, that produces the greatest information gain.
- By sorting the examples according to the continuous attribute A, then identifying adjacent examples that differ in their target classification, we can generate a set of candidate thresholds midway between the corresponding values of A.
- It can be shown that the value of c that maximizes information gain must always lie at such a boundary
- These candidate thresholds can then be evaluated by computing the information gain associated with each.

MACHINE INTELLIGENCE

Dealing with continuous value

TEMP	40	48	60	72	80	90
Play Sport	no	no	yes	yes	yes	no

- In the current example, there are two candidate thresholds, corresponding to the values of Temperature at which the value of Play Tennis changes: (48 + 60)/2, and (80 + 90)/2.
- The information gain can then be computed for each of the candidate attributes
- $\text{Temperature}_{>54}$ and $\text{Temperature}_{>85}$, and the best can be selected $\text{Temperature}_{>54.}$
- This dynamically created Boolean attribute can then compete with the other discrete-valued candidate attributes available for growing the decision tree.
- Further extension to this can be splitting data into multiple intervals rather then just two intervals based on a single threshold.

THANK YOU

K.S.Srinivas
srinivasks@pes.edu
+91 80 2672 1983 Extn 701