INDUCTIVEBIASINDECISIONTREELEARNING

Inductive biasistheset of assumptionsthat,together withthe training data,deductivelyjustifytheclassificationsassignedbythelearnertofutureinstances

Givenacollectionoftrainingexamples,therearetypicallymanydecisiontreesconsistent with theseexamples.Which ofthese decisiontreesdoes ID3choose?

ID3searchstrategy

(a) selectsinfavourofshortertreesoverlongerones

(b) selectstreesthatplacetheattributeswithhighestinformationgainclosesttotheroot.

ApproximateinductivebiasofID3: Shortertreesarepreferredoverlargertrees

- Consideranalgorithmthatbeginswiththeemptytreeandsearches***breadthfirst*** throughprogressivelymorecomplextrees.
- Firstconsideringalltreesofdepth1,thenalltreesofdepth2,etc.
- Onceitfindsadecisiontreeconsistentwiththetrainingdata,itreturns thesmallestconsistenttreeatthatsearchdepth(e.g.,thetreewiththefewestnodes).
- LetuscallthisbreadthfirstsearchalgorithmBFS-ID3.
- BFS-ID3findsashortestdecisiontreeandthusexhibitsthebias"shortertrees arepreferredover longer trees.

- ID3 can be viewed as an efficient approximation to BFS-ID3, using a greedyheuristic search to attempt to find the shortest tree without conducting the entirebreadth-firstsearchthrough thehypothesis space.
- BecauseID3usestheinformationgainheuristicandahillclimbingstrategy, itexhibitsamorecomplexbiasthanBFS-ID3.
- Inparticular,itdoesnotalwaysfindtheshortestconsistenttree,anditisbiased tofavourtreesthatplaceattributeswithhighinformationgainclosesttotheroot.

RestrictionBiasesandPreferenceBiases

DifferencebetweenthetypesofinductivebiasexhibitedbyID3andbytheCANDIDATE-ELIMINATIONAlgorithm.

ID3

- ID3searchesacompletehypothesisspace
- Itsearchesincompletelythroughthisspace,fromsimpletocomplexhypotheses,untilitstermination conditionismet
- Itsinductivebiasissolelyaconsequenceoftheorderingofhypothesesbyitssearchstrategy. Itshypothesisspaceintroducesnoadditionalbias

CANDIDATE-ELIMINATIONAlgorithm

- TheversionspaceCANDIDATE-ELIMINATIONAlgorithmsearchesanincompletehypothesisspace
- Itsearchesthisspacecompletely,findingeveryhypothesisconsistentwiththetrainingdata.
- Its inductive bias is solely a consequence of the expressive power of its hypothesisrepresentation.Itssearchstrategyintroducesnoadditionalbias

RestrictionBiasesandPreferenceBiases

- The inductive bias of ID3 is a*preference* for certain hypotheses over others (e.g.,preference for shorter hypotheses over larger hypotheses), with no hard restrictiononthehypothesesthatcanbeeventuallyenumerated.Thisformofbiasis called a***preferencebias***or ***a search bias***.

- The bias of the CANDIDATE ELIMINATION algorithm is in the form of a*categorical* restriction on the set of hypotheses considered. This form of bias istypicallycalleda***restriction bias***or***alanguagebias***.

Whichtypeofinductivebiasispreferredinordertogeneralizebeyondthetrainingdata,apreferenc e bias or a restriction bias?

- Apreference bias is more desirable than arestriction bias, because it allows thelearner to work within a complete hypothesis space that is assured to contain theunknown target function.
- In contrast, a restriction bias that strictly limits the set of potential hypotheses isgenerally less desirable, because it introduces thepossibilityofexcluding theunknown target function altogether.

Occam'srazor

Occam's razor: is the problem-solving principle that the simplest solution tends to bethe right one. When presented with competing hypotheses to solve a problem, oneshouldselect the solutionwith the fewest assumptions.

Occam'srazor:"***Preferthesimplesthypothesisthatfitsthedata***".

WhyPreferShortHypotheses?

Argumentinfavour:

Fewershorthypothesesthanlongones:

- Shorthypothesesfitsthetrainingdatawhichare*lesslikely*tobe*coincident*
- Longerhypothesesfitsthetrainingdatamightbe*coincident*.

Manycomplexhypothesesthatfitthecurrenttrainingdatabutfailtogeneralizecorrectly tosubsequent data.

Argumentopposed:

- There are few small trees, and our priori chance of finding one consistent with anarbitrary set of data is therefore small. The difficulty here is that there are verymanysmallsetsofhypothesesthatonecandefine*butunderstoodbyfewerlearner*.
- The size of a hypothesisis determined by the representation used ***internally*** by thelearner.Occam'srazorwillproduce*twodifferenthypothesesfromthesametrainingexampleswhenitisappliedbytwolearners*,bothjustifyingtheircontradictory conclusions byOccam'srazor.OnthisbasiswemightbetemptedtorejectOccam'srazor altogether.

ISSUESINDECISIONTREELEARNING

1. AvoidingOverfittingtheData

   Reduced error

   pruningRulepost-pruning

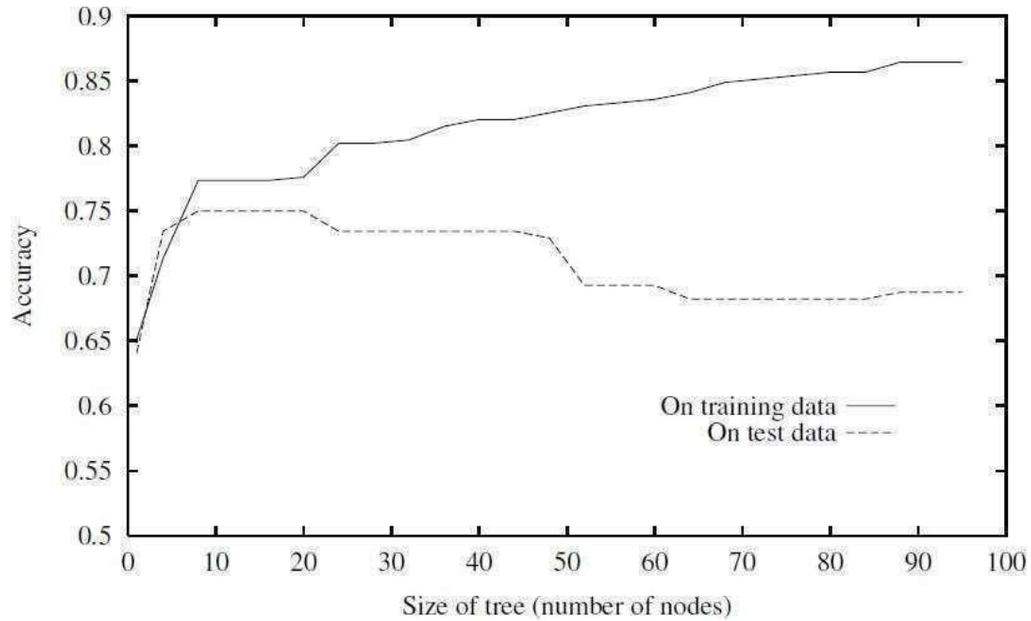
2. IncorporatingContinuous-ValuedAttributes
3. AlternativeMeasuresforSelectingAttributes
4. HandlingTrainingExampleswithMissingAttributeValues
5. HandlingAttributeswithDifferingCosts

1. AvoidingOverfittingtheData

- The ID3 algorithm grows each branch of the tree just deeply enough to perfectlyclassify the training examples but it can lead to difficulties when there is noise inthe data, or when the number of training examples is too small to produce arepresentative sample of the true target function. This algorithm can produce treesthat***overfit***the training examples.

- ***Definition-Overfit:***GivenahypothesisspaceH,ahypothesis$h \in$Hissaidtooverfit the training data if there exists some alternative hypothesis $h' \in$H, suchthat*h*hassmallererrorthan*h'*overthetrainingexamples,but*h'*hasasmallererrorthan *h*over the entire distribution of instances.

- Thebelowfigureillustratestheimpactofoverfittinginatypicalapplicationofdecisiontreelearning.



- The**horizontalaxis**ofthis plotindicatesthetotalnumberofnodesinthedecisiontree,asthetreeisbeingconstructed.The**verticalaxis**indicates theaccuracyofpredictionsmadebythetree.
- The**solidline**showstheaccuracyofthedecisiontreeoverthetrainingexamples.The**brokenline**showsaccuracymeasuredoveranindependentsetoftestexample
- The**accuracy**ofthetreeoverthe**trainingexamplesincreases**monotonicallyasthetreeisgrown.The **accuracy**measuredovertheindependenttestexamples**firstincreases**,**thendecreases**.

Howcanitbepossiblefortreehtofitthetrainingexamplesbetterthanh',butforittoperformmorepoorlyoversubsequentexamples?

1. Overfittingcanoccurwhenthetrainingexamplescontainrandomerrorsornoise
2. Whensmallnumbersofexamplesareassociatedwithleafnodes.

NoisyTrainingExample

Example15:<*Sunny,Hot,Normal,Strong,*->

- Exampleisnoisybecausethecorrectlabelis+
- Previouslyconstructedtreemisclassifiesit