



AVT-NBL: An Algorithm for Learning Compact and Accurate Classifiers from Attribute Value Taxonomies and Data

Jun Zhang and Vasant Honavar
Artificial Intelligence Research Laboratory
Iowa State University, USA

Presented by Dae-Ki Kang

This research is sponsored in part by grants from the National Science Foundation (IIS 0219699) and National Institutes of Health (GM066387)



Overview

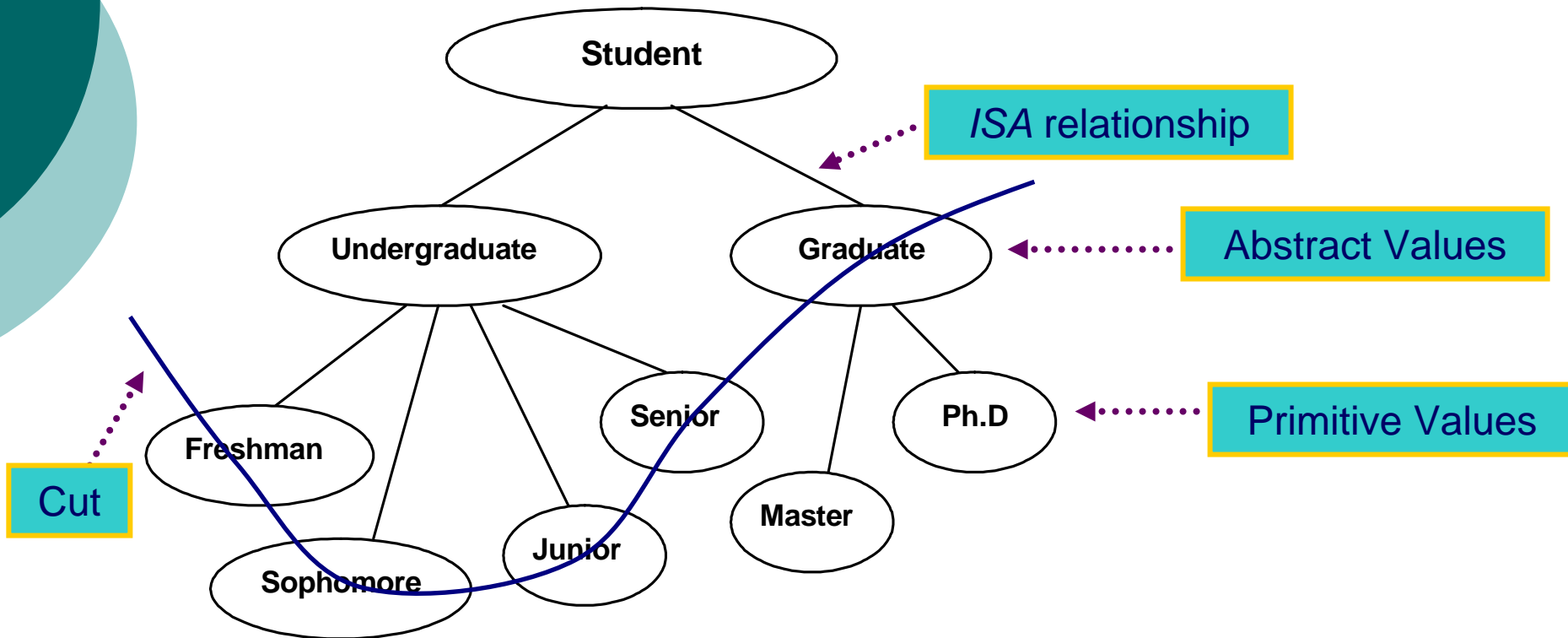
- Background and Motivation
- AVT-NBL Algorithm
- Experimental Results
- Summary

Paper Highlights

- Presents AVT-NBL, a natural generalization of the Naïve Bayes learner (NBL) that can effectively exploit **attribute value taxonomies**
- AVT-NBL is able to learn compact, accurate, and comprehensible classifiers from data, including **partially specified data**
- Experimental evaluation of AVT-NBL on a broad range of benchmark data sets, and synthetic partially specified data

Attribute Value Taxonomies (AVT)

-- ISA hierarchies



Attribute Value Taxonomy (AVT) for *student status*

Why explore AVT?

- AVTs are domain **Ontologies** with explicit description of:
 - **concepts** (attribute values)
 - **interrelationships between concepts** (ISA relations)
 - **constraints of concepts** (a cut defines a partition of concepts)
- AVTs reflect prior knowledge and working assumptions in a specific application domain



Motivations for learning from AVT and data

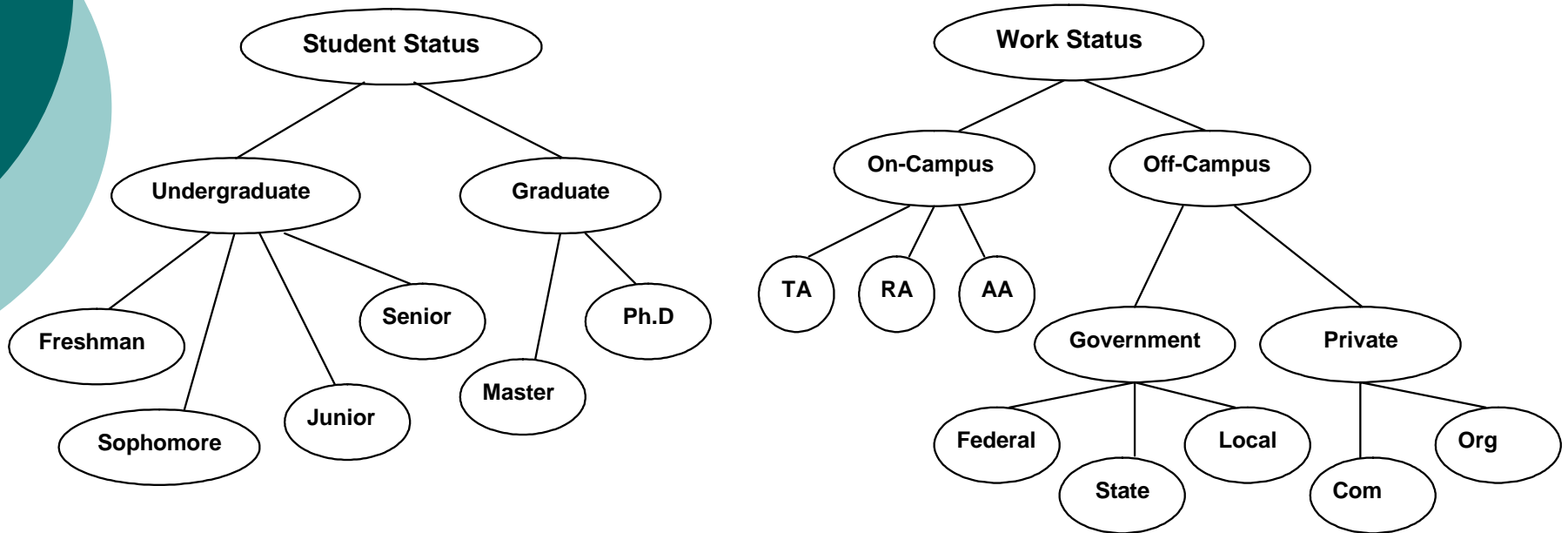
- Preference for simple, comprehensible, yet accurate and robust classifiers
- Classifiers that use *abstract* attribute values often provide simpler descriptions of the data
- When data are limited, statistics estimated from abstract values are often more reliable than statistics estimated from primitive values
- Use of AVT offers a simple way to perform to minimize over-fitting

Partially Specified Data

Partially specified instance: one or more of the attribute values are partially specified – correspond to *abstract* attribute values in AVT

- Common when data gathered by multiple, distributed and autonomous entities
- Unavoidable in information integration from semantically heterogeneous data sources with different ontologies

Partially Specified Data - example



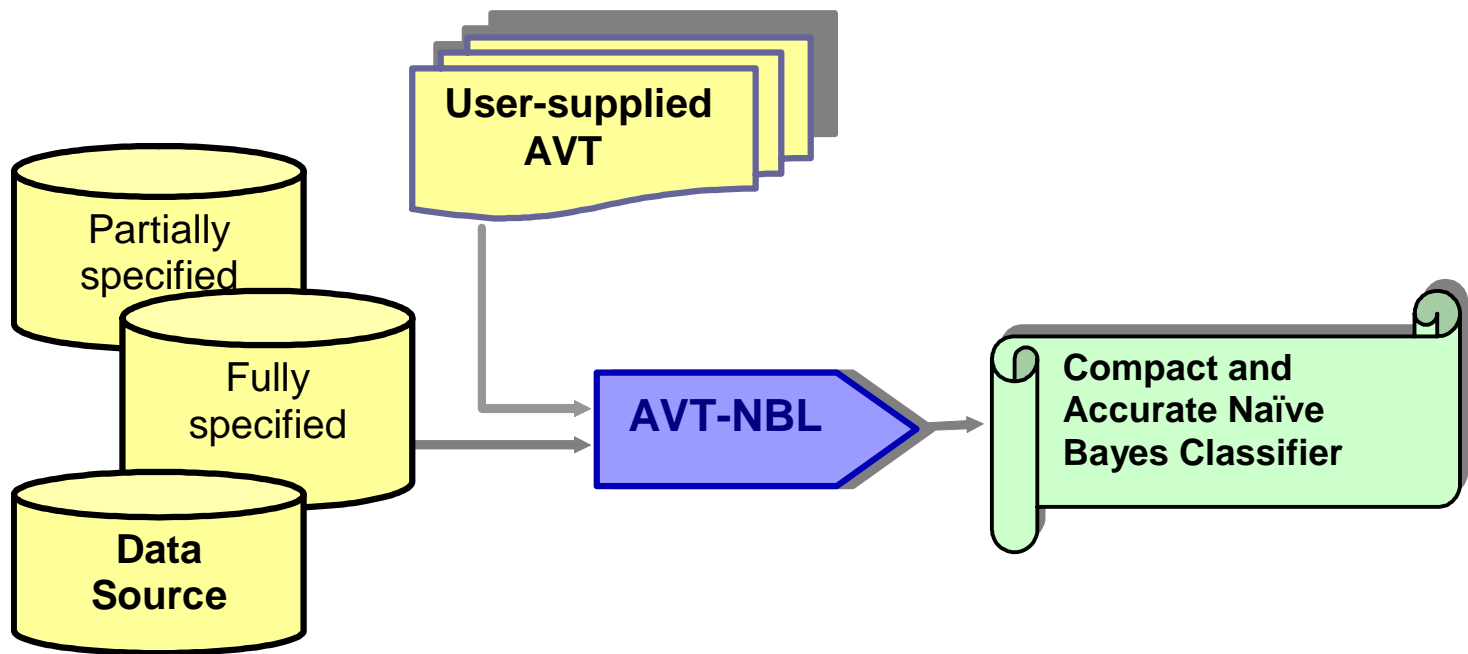
Examples of partially specified instances:

(Undergraduate, RA)

(Freshman, Government)

(Graduate, Off-Campus)

Our Learning Scenario



Related Work

Using attribute value taxonomies/Class Taxonomy in learning

Núñez, M. (1991); Dhar & Tuzhilin (1993); Han & Fu (1996); Taylor, Stoffel, & Hendler (1997); desJardins, Getoor, & Koller, (2000); Blockeel, et al (2002)...

Learning attribute value taxonomies

Pereira, Tishby & Lee (1993), Yamazaki, Pazzani, Merz, (1995) ...

Gathering statistics from data with missing attribute values

Quinlan (1992), McClean, Scotney & Shapcott (2001)

Using set valued features

Quinlan (1992), Cohen (1996)



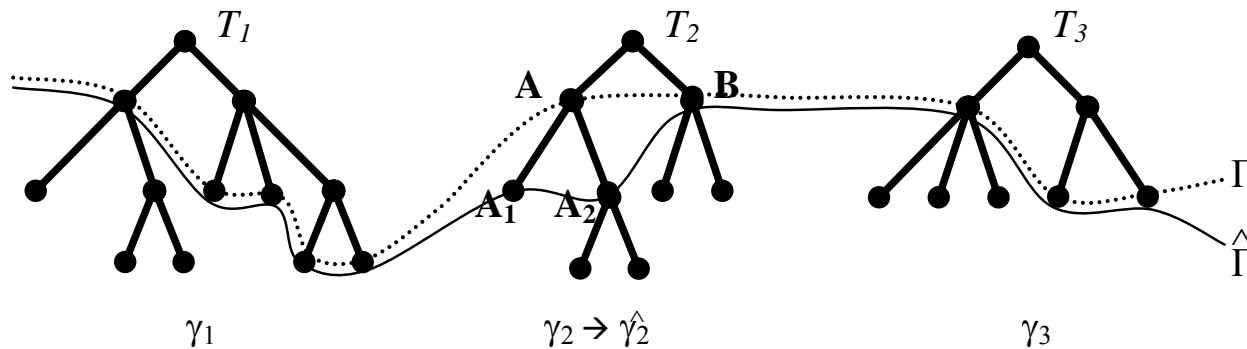
Overview

- Background and Motivation
- AVT-NBL Algorithm
- Experimental Results
- Summary

AVT-DTL algorithm

- Find the most accurate Naïve Bayes classifier with the least specification parameters
 - Same assumption as NBL that each attribute is independent of the other attributes given the class
 - Starting with the NBL that is based on the most abstract value of each attribute and successively refining the classifier (hypothesis)
 - Using a tradeoff criterion between the accuracy and complexity of the resulting classifier
 - Handling partially specified data

Cut Refinement (Def.)

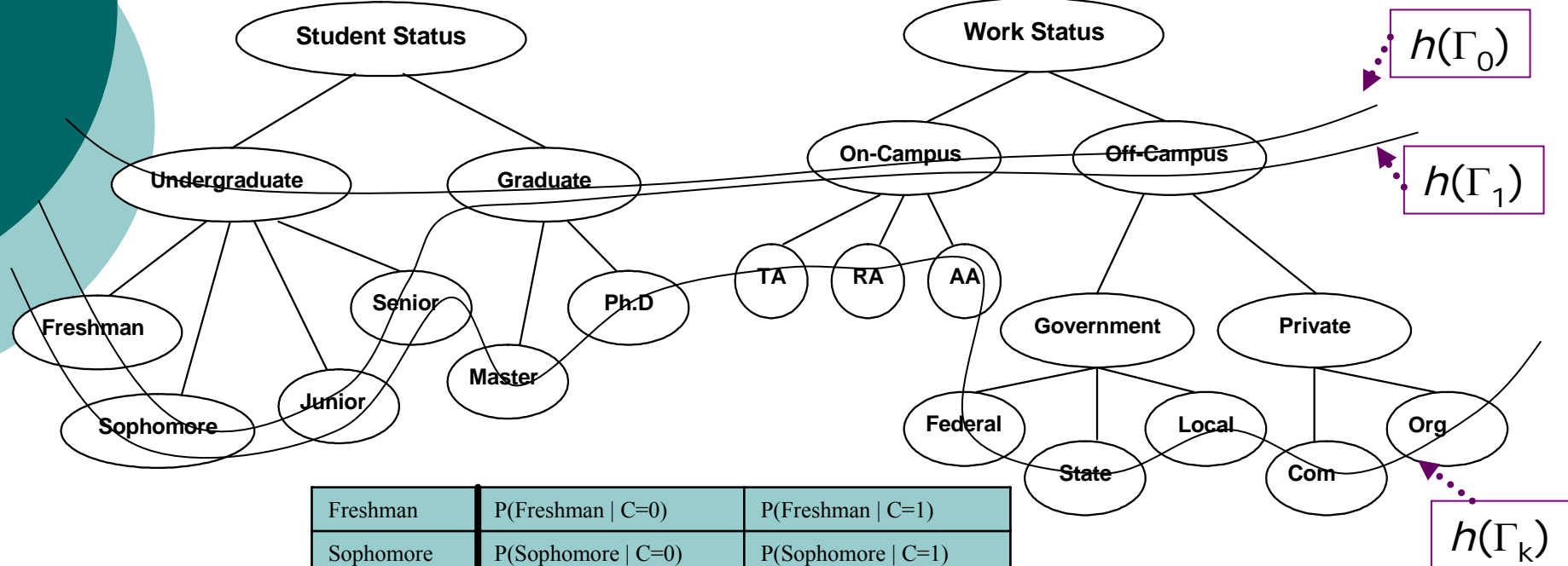


- A cut defines a partition (a set of mutual exclusive attribute values) over the primitive values of an attribute
- One cut is a refinement of another cut if at least one previous attribute value has been replaced with its descendants
- Each cut specifies the entries of class conditional probability table

AVT-Induced Instance Space (Def.)

- I : Original Instance Space – defined by fully specified instances using all primitive attribute values
- I_{Γ} : Abstraction Instance Space – instances using abstract attribute values
- $I_{\mathbf{A}}$: AVT-Induced Instance Space – the union of instance spaces induces by all of the cuts through the set of AVTs
- AVT-NBL search for classifier h_{Γ} : $I_{\mathbf{A}} \neq C$
- Resulting a structured hypothesis space H_{Γ} needs to be searched efficiently

Example



Freshman	$P(\text{Freshman} \mid C=0)$	$P(\text{Freshman} \mid C=1)$
Sophomore	$P(\text{Sophomore} \mid C=0)$	$P(\text{Sophomore} \mid C=1)$
...
Master	$P(\text{Master} \mid C=0)$	$P(\text{Master} \mid C=1)$
Ph.D	$P(\text{Ph.D} \mid C=0)$	$P(\text{Ph.D} \mid C=1)$
TA	$P(\text{TA} \mid C=0)$	$P(\text{TA} \mid C=1)$
RA	$P(\text{RA} \mid C=0)$	$P(\text{RA} \mid C=1)$
...
Com	$P(\text{Com} \mid C=0)$	$P(\text{Com} \mid C=1)$
Org	$P(\text{Org} \mid C=0)$	$P(\text{Org} \mid C=1)$

Two Major Steps in AVT-NBL

- Calculate class conditional frequency counts on AVTs with the presence of partially specified data – *Sufficient Statistics* for AVT-NBL
- Search for compact Naïve Bayes classifier (Hypothesis refinement) based on *Conditional Minimum Description Length* (CMDL) score

Class Conditional Frequency Counts on AVTs

- Fully specified attribute values
 - Simply aggregating the frequency counts of lower levels to higher level
- Partially missing attribute values
 - Aggregate the (non-zero) counts upward from each such node to its ancestors
 - Propagate and Update the fractional counts downward for partially specified attribute values

Cascaded Naïve Bayes classifiers

- Any global cut $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_N\}$ completely specifies a Naïve Bayes classifier $h(\Gamma)$
- Each cut γ_i has an associate CPT table $\text{CPT}(\gamma_i)$ by doing Laplace estimates
- $h(\Gamma) = \{\text{CPT}(\gamma_1), \dots, \text{CPT}(\gamma_N)\}$
- Each $h(\Gamma)$ has its own parameters and accuracy

Search for Compactness and Accuracy

- Using scoring functions to rank alternative models (classifiers), and making tradeoff between complexity and accuracy
 - A complex model is likely to result in poor parameter estimations with *high variance (overfitting)*
 - A simplified model with the right structure is likely to yield more reliable estimates and better tradeoff.

Conditional Minimum Description Length Score

- Complexity part: capture the number of parameters w (corresponding to entries in CPT, and relate to the levels of abstraction)
- Accuracy part: capture how accurate the classifier $h(\Gamma)$ is (based on data dependent likelihood calculation)

CMDL Score Calculations

$$\begin{array}{c}
 \text{complexity} \quad \text{accuracy} \\
 \left. \vphantom{\frac{\log|D|}{2}} \right\} \quad \left. \vphantom{size(h)} \right\} \\
 CMDL(h | D) = \left(\frac{\log|D|}{2} \right) size(h) - CLL(h | D)
 \end{array}$$

where

$$CLL(h | D) = |D| \sum_{p=1}^{|D|} \log P_h(c_p | a_{1p}, \dots, a_{Np})$$

$$CLL(h | D) = |D| \sum_{p=1}^{|D|} \log P_h(c_p | a_{1p}, \dots, a_{Np}) = |D| \sum_{p=1}^{|D|} \log \left(\frac{P(c_p) \prod_i P_h(a_{ip} | c_p)}{\sum_{j=1}^{|C|} P(c_j) \prod_i P_h(a_{ip} | c_j)} \right)$$

decompose because of
independence assumption

Search/Refinement Procedure

- Efficiently optimizing the criterion independently for each attribute
- It terminates when none of the candidate refinements of the classifier yield statistically significant improvement in the CMDL score
- Output final Naïve Bayes classifier $h(\Gamma^*)$

Alternative approaches to learning from AVTs and partially specified data

- Treat each partially specified attribute values as if it were totally missing
- AVT-based propositionalization method: construct a set of Boolean attributes for each value in AVT; strong dependencies exist among the Boolean attributes derived from AVT



Overview

- Background and Motivation
- AVT-NBL Algorithm
- Experimental Results
- Summary

Experimental Settings

- Performance comparisons between:
 - AVT-NBL
 - NBL (Naïve Bayes Learner)
 - PROP-NBL (NBL applied to propositionalized data)
- Data sets:
 - Benchmark datasets from UCI Machine Learning Repository
 - Synthetic datasets with different pre-specified percentages of totally/partially missing attribute values
- Use 10-fold cross validation and calculate 90% confidence interval on error rates

First set of experiments

- Compare performance on original fully specified data (8 datasets from UCI)
- AVTs on *Mushroom*, *Soybean* and *Nursery* were supplied by domain experts
- For the rest 5 datasets, AVTs were generated by AVT-Learner (Dae-Ki Kang et al, 2004)

Table 1. Comparison of error rate and size of classifiers generated by NBL, PROP-NBL and AVT-NBL on benchmark data

% Error rates using 10-fold cross validation with 90% confidence interval; The size of the classifiers for each data set is constant for NBL and Prop-NBL, and for AVT-NBL, the size shown represents the average across the 10-cross validation experiments.

DATA SET	NBL		Prop-NBL		AVT-NBL	
	error	size	error	size	error	size
Audiology	26.55 (± 5.31)	3696	27.87 (± 5.39)	8184	23.01 (± 5.06)	3600
Breast-Cancer	28.32 (± 4.82)	84	27.27 (± 4.76)	338	27.62 (± 4.78)	62
Car	14.47 (± 1.53)	88	15.45 (± 1.57)	244	13.83 (± 1.50)	80
Dermatology	2.18 (± 1.38)	876	1.91 (± 1.29)	2790	2.18 (± 1.38)	576
Mushroom	4.43 (± 1.30)	252	4.45 (± 1.30)	682	0.14 (± 0.14)	202
Nursery	9.67 (± 1.48)	135	10.59 (± 1.54)	355	9.67 (± 1.48)	125
Soybean	7.03 (± 1.60)	1900	8.19 (± 1.72)	4959	5.71 (± 1.45)	1729
Zoo	6.93 (± 4.57)	259	5.94 (± 4.25)	567	3.96 (± 3.51)	245

Results Shown from Table 1

- AVT-NBL yields lower error rates than NBL and PROP-NBL on the original fully specified data
- PROP-NBL generally produces classifiers with higher error rates than NBL
- AVT-NBL yields classifiers that are substantially more compact than those generated by PROP-NBL and NBL



Second set of experiments

- Explore the performance on datasets with different percentage (10%, 30%, 50%) of partially missing and totally missing attribute values
- Assume the missing values are uniformly distributed on the nominal attributes

Table 2. Comparison of error rates on data with partially or totally missing values

% Error rates using 10-fold cross validation with 90% confidence interval							
DATA		Partially Missing			Totally Missing		
Methods		NBL	Prop-NBL	AVT-NBL	NBL	Prop-NBL	AVT-NBL
Mushroom	10%	4.65(±1.33)	4.69(±1.34)	0.30(±0.30)	4.65(±1.33)	4.76(±1.35)	1.29(±0.71)
	30%	5.28 (±1.41)	4.84(±1.36)	0.64(±0.50)	5.28 (±1.41)	5.37(±1.43)	2.78(±1.04)
	50%	6.63(±1.57)	5.82(±1.48)	1.24(±0.70)	6.63(±1.57)	6.98(±1.61)	4.61(±1.33)
Nursery	10%	15.27(±1.81)	15.50(±1.82)	12.85(±1.67)	15.27(±1.81)	16.53(±1.86)	13.24(±1.70)
	30%	26.84(±2.23)	26.25(±2.21)	21.19(±2.05)	26.84(±2.23)	27.65(±2.24)	22.48(±2.09)
	50%	36.96(±2.43)	35.88(±2.41)	29.34(±2.29)	36.96(±2.43)	38.66(±2.45)	32.51(±2.35)
Soybean	10%	8.76(±1.76)	9.08(±1.79)	6.75(±1.57)	8.76(±1.76)	9.09(±1.79)	6.88(±1.58)
	30%	12.45(±2.07)	11.54(±2.00)	10.32(±1.90)	12.45(±2.07)	12.31(±2.05)	10.41(±1.91)
	50%	19.39(±2.47)	16.91(±2.34)	16.93(±2.34)	19.39 (±2.47)	19.59(±2.48)	17.97(±2.40)



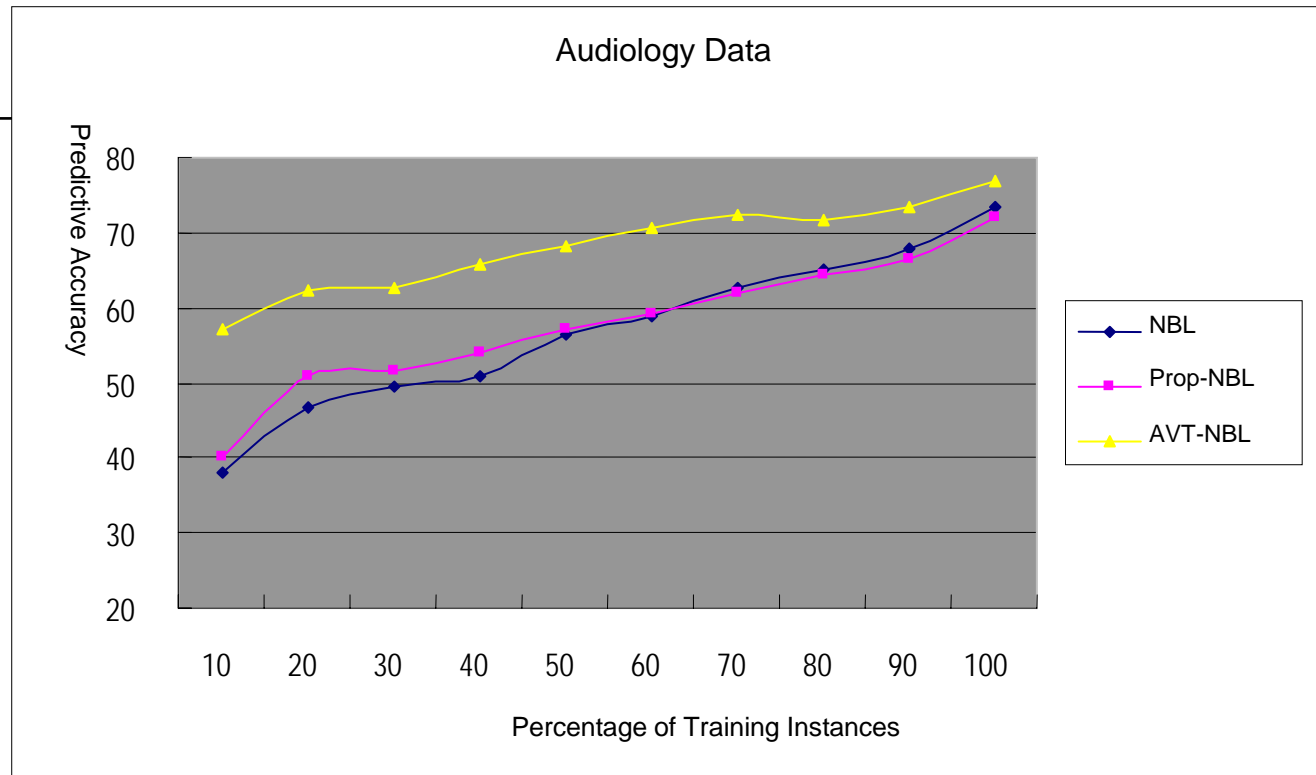
Results Shown from Table 2

- AVT-NBL yields significantly lower error rates than NBL and PROP-NBL on partially specified data and data with totally missing values.
- The differences are more pronounced at higher percentages of partially or totally missing attribute values.

Third set of experiments

- Investigate the performance of classifiers as a function of the training set size
- Training Pool + Test Pool
 - Sampled training sets of different sizes: 10%, 20%, ..., 100% of Training Pool
 - Resulting classifiers were evaluated on Test Pool

Classifier accuracy as a function of training set size



- AVT-NBL produces more accurate classifiers than NBL and Prop-NBL for a given training set size
- AVT-NBL produces classifiers that outperform those produced by NBL using substantially fewer training examples.

Summary

- AVT-NBL offers an effective approach to learning compact (hence more comprehensible) accurate classifiers from data – including data that are *partially specified*
- AVT-NBL is more efficient in its use of training data, it produces classifiers using substantially fewer training examples

Future Work

- Development AVT-based variants of machine learning algorithms for learning classifiers from distributed, semantically heterogeneous and partially specified data sources
- Extension of the algorithms like AVT-DTL and AVT-NBL to handle taxonomies defined over ordered and numeric attribute values
- Further experimental evaluation of AVT-based learning algorithms on a broad range of data sets in scientific knowledge discovery applications



Thank You!

Questions?