Classification Trees with NPI

Rebecca Baker

Department of Mathematical Sciences,
University of Durham

1 The multinomial NPI model
- Motivation
- The probability wheel representation

2 Classification
- Classification trees
- Weka software

3 Finding the maximum entropy distribution for NPI
- The approximate algorithm
- The exact algorithm
- Comparison of NPI-M and A-NPI-M

4 Future work

The multinomial NPI model

Model for learning from multinomial data

- inferences about a future observation
- in form of a probability interval
- based entirely on past observations

Have observed $Y_1, ..., Y_n$**, want to find out about** Y_{n+1}

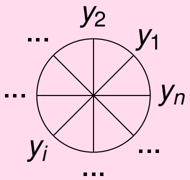
K **categories in total:** $c_1, ..., c_K$

Event of interest is ($Y_{n+1} \in E$**) where** E **is a subset of the** K **categories**

The probability wheel representation

Represent data on a **probability wheel**

- Y_{n+1} has probability $\frac{1}{n}$ of being in each slice

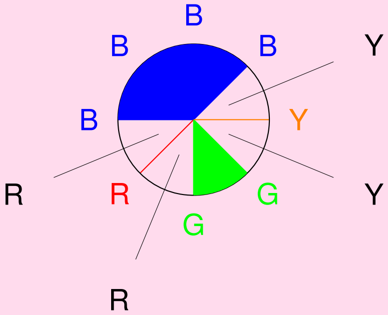


- Slice bordered by two observations in the same category is assigned to this category
- Slice bordered by two observations in different categories may be assigned to any available category

Note: Each category may only be represented by a single segment of the wheel.

Deriving lower probabilities

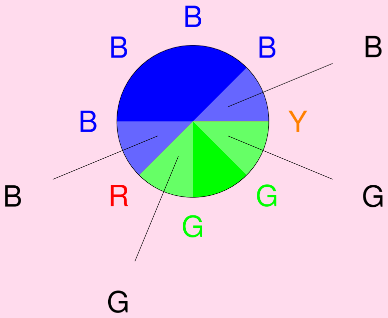
- Possible categories are blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O)
- Event $E = \{B, G, P\}$



- $\underline{P}(Y_{n+1} \in E) = \frac{4}{8}$

Deriving upper probabilities

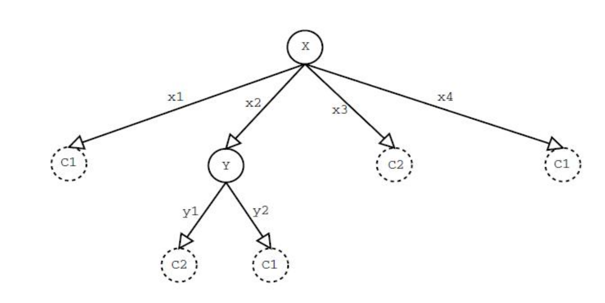
- Possible categories are blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O)
- Event $E = \{B, G, P\}$



- $\overline{P}(Y_{n+1} \in E) = 1$

Classification trees

- Hierarchical structure which defines classification rules



- Attributes at the nodes
- Category labels at the leaves

Building trees using imprecise probabilities

At each node:

- We need to select an attribute for splitting
- The generalised Shannon entropy measure S is employed, using the maximum entropy distribution p_{maxE}:

$$S = -\sum_{j=1}^{K} p_{maxE}(c_j) \log p_{maxE}(c_j)$$

- The information gain is measured for each attribute
- The most informative attribute is selected for splitting

Weka software

Weka software can be used to build classification trees

- One classifier can be analysed in detail
- Multiple classifiers can be compared in a number of ways

- The software includes tools for pre-processing data

Weka software

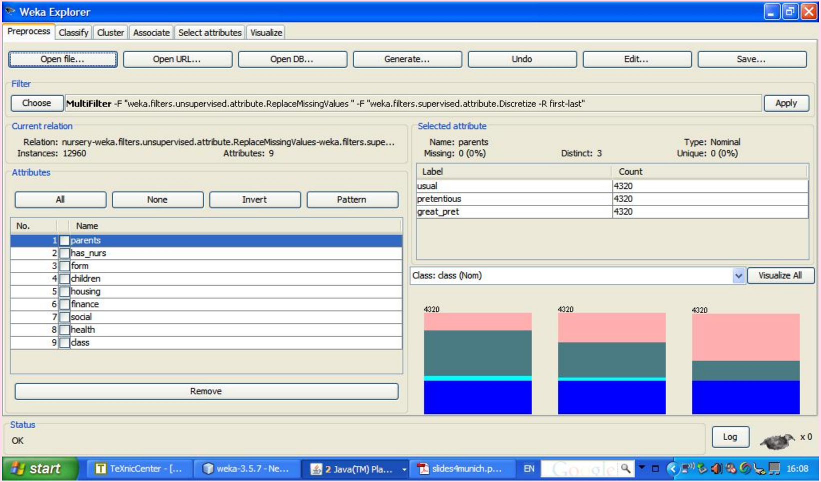


Figure: Weka Explorer: Pre-process tab

Weka software

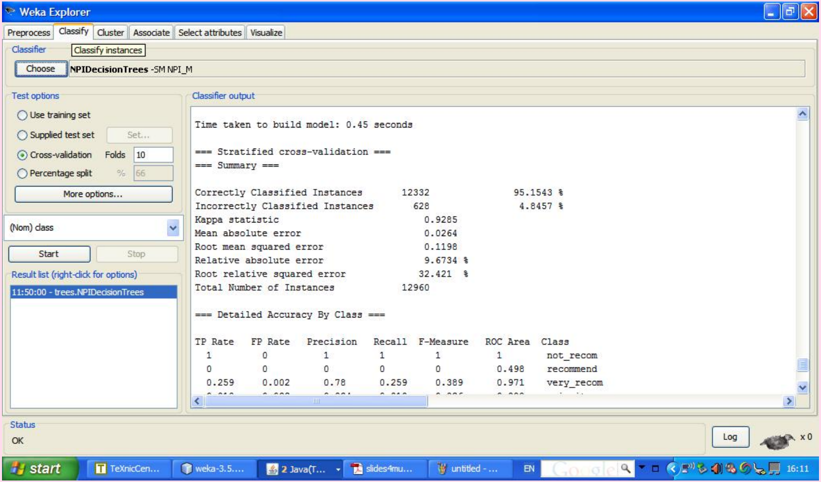


Figure: Weka Explorer: Classify tab

Weka software

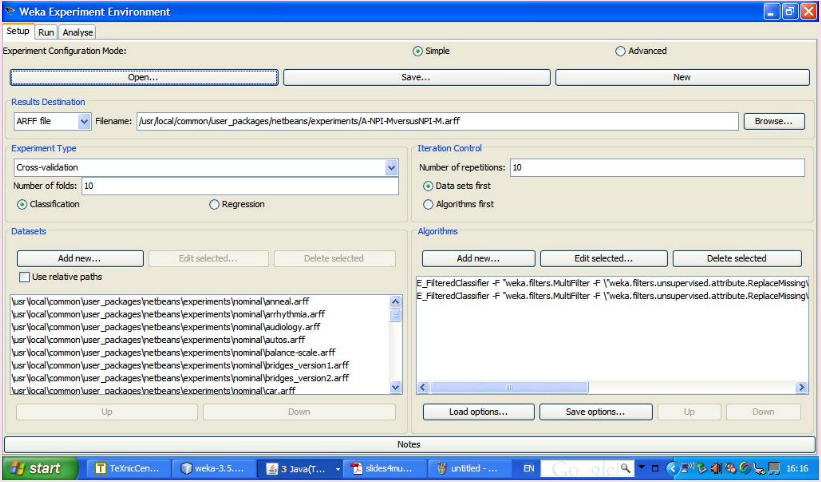


Figure: Weka Experimenter: Setup tab

Weka software

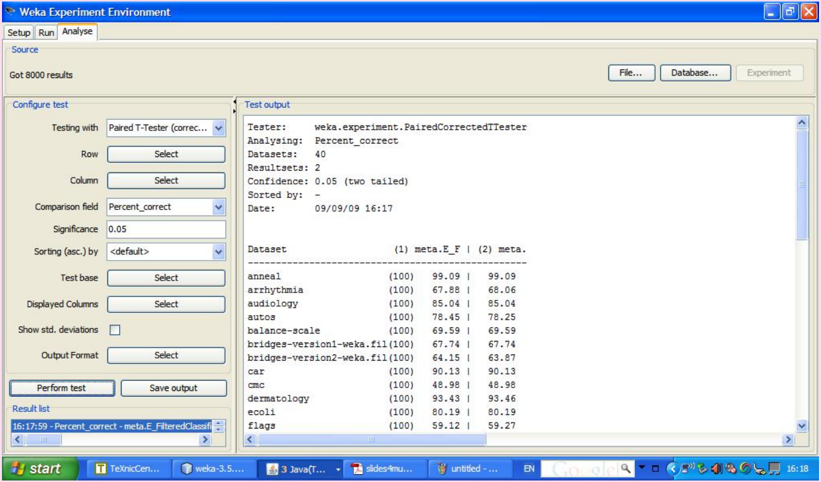


Figure: Weka Experimenter: Analyse tab

The approximate algorithm, A-NPI-M

Based on an algorithm by Abellan and Moral for finding the maximum entropy distribution within a credal set

- NPI gives set of probability intervals
  $Ł = [l_j, u_j] = [\underline{P}(c_j), \overline{P}(c_j)]$

  - These are F-probabilities
  - The probability of any event can be defined in terms of these single-category probabilities

- The credal set associated with the NPI lower and upper probabilities can be expressed by the set $Ł$

The approximate algorithm, A-NPI-M

The algorithm A-NPI-M is applied to the credal set Ł

- For each category, $p(c_j)$ is initially set to l_j
- The remaining probability mass is shared evenly between categories, beginning with those observed least often
- At each step, probabilities are increased by $\frac{1}{n}$ until they reach the value u_j or until all probability mass has been distributed

The resulting distribution is used to build classification trees

Comparison to other methods

Classification trees using A-NPI-M were compared to 4 other methods:

1. Trees using IDM
2. Trees with precise probabilities and IG split criterion
3. Trees with precise probabilities and IGR split criterion
4. More complex procedure involving pruning (J48)

- Experiment was carried out on 40 data sets
- Classifiers were compared pairwise
- Numbers of correct classifications were compared

Results

Number of Wins, Ties and Losses (W/T/L) for each classifier:

	IDM	*NPI*	*IG*	*IGR*	*J48*
IDM	-	(19/2/19)	(18/2/20)	(15/2/23)	(17/1/22)
NPI	(19/2/19)	-	(18/2/20)	(15/2/23)	(17/1/22)
IG	(20/2/18)	(20/2/18)	-	(18/3/19)	(19/1/20)
IGR	(23/2/15)	(23/2/15)	(19/3/18)	-	(21/1/18)
J48	(22/1/17)	(22/1/17)	(20/1/19)	(18/1/21)	-
W-L	15	15	-4	-20	-8

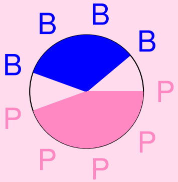
- The performance of A-NPI-M is similar to that of the IDM
- A-NPI-M performs better than the other classifiers here

Problem

- The A-NPI-M algorithm finds the maximum entropy distribution in the credal set Ł
- Some distributions in this set are not compatible with the probability wheel model

Example

- Possible categories $\{B, P, R, Y, O\}$ with observation counts $\{4, 5, 0, 0, 0\}$



- The credal set Ł is $\{[\frac{3}{9}, \frac{5}{9}]; [\frac{4}{9}, \frac{6}{9}]; [0, \frac{1}{9}]; [0, \frac{1}{9}]; [0, \frac{1}{9}]\}$
- A-NPI-M gives the distribution $\{\frac{3}{9}, \frac{4}{9}, \frac{2}{27}, \frac{2}{27}, \frac{2}{27}\}$
- There is no valid configuration of the wheel that corresponds to this distribution

The exact algorithm, NPI-M

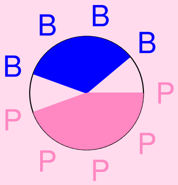
The exact algorithm finds the maximum entropy distribution consistent with the probability wheel model

- For each category, $p(c_j)$ is initially set to l_j
- The remaining probability mass is shared as evenly as possible between categories, beginning with those observed least often
- At each step, probabilities are increased by $\frac{1}{n}$ until they reach the value u_j or until all probability mass has been distributed

This leads to a distribution which is as uniform as possible but still corresponds to a valid configuration of the wheel

Example

- Possible categories $\{B, P, R, Y, O\}$ with observation counts $\{4, 5, 0, 0, 0\}$



- NPI-M gives the distribution $\left\{\frac{3}{9}, \frac{4}{9}, \frac{1}{9}, \frac{1}{18}, \frac{1}{18}\right\}$
- This is as close to uniform as possible while still corresponding to a valid configuration of the wheel

Comparison of NPI-M and A-NPI-M

We implemented NPI-M for building classification trees in Weka

- Comparison of NPI-M and A-NPI-M was carried out on 40 data sets
- Numbers of correct classifications were compared

Results

Percentage of correct classifications for each method:

Dataset	(1)	(2)	
anneal	99.09	99.09	
arrhythmia	67.88	68.06	
audiology	85.04	85.04	
autos	78.45	78.25	
balance-scale	69.59	69.59	
bridges-version1	67.74	67.74	
bridges-version2	64.15	63.87	
car	90.13	90.13	
cmc	48.98	48.98	
dermatology	93.43	93.46	
ecoli	80.19	80.19	
flags	59.12	59.27	
hypothyroid	99.33	99.33	
iris	93.40	93.40	
letter	78.77	78.77	
lung-cancer	41.33	41.33	
lymphography	73.68	73.68	
mfeat-factors	81.71	81.68	
mfeat-fourier	68.90	68.92	
mfeat-karhunen	73.14	73.15	
mfeat-morphological	69.78	69.78	
mfeat-pixel	79.99	79.92	
mfeat-zernike	64.19	64.24	
nursery	95.15	94.99	●
optdigits	78.95	78.98	
page-blocks	96.08	96.10	
pendigits	89.37	89.37	
postoperative-patient-data	71.11	71.11	
primary-tumor	39.21	39.48	
segment	94.18	94.20	
soybean	93.29	93.35	
spectrometer	43.32	43.33	
splice	93.25	93.25	
sponge	94.48	94.48	
tae	46.78	46.78	
vehicle	69.39	69.39	
vowel	75.92	75.95	
waveform	73.99	73.99	
wine	92.02	92.02	
zoo	95.53	95.53	

○, ● statistically significant improvement or degradation

- Peformance is not significantly different on most data sets
- NPI-M performs significantly better on ‘nursery’ data set

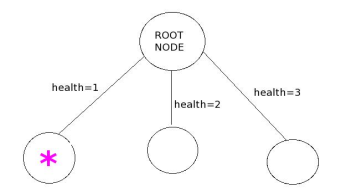
Nursery data set

Data set taken from applications for places at a private nursery school

- Applicants classified in terms of how likely they are to be accepted
- 5 categories: c_1, c_2, c_3, c_4, c_5
- 8 attribute variables

Nursery data set

Most informative attribute is 'health'



At '*', counts in $\{c_1, c_2, c_3, c_4, c_5\}$ are $\{0, 0, 0, 1854, 2466\}$

- A-NPI-M and NPI-M both give $p_{maxE}(c_4) = \frac{1853}{4320}$ and $p_{maxE}(c_5) = \frac{2465}{4320}$
- A-NPI-M gives equal probability $\frac{1}{6480}$ to c_1, c_2 and c_3
- NPI-M gives probabilities $\{\frac{1}{4320}, \frac{1}{8640}, \frac{1}{8640}\}$ to $\{c_1, c_2, c_3\}$

In the branch of the tree beginning at '*', A-NPI-M and NPI-M will always give different distributions

Future work

- Investigation into the use of the maximum entropy algorithm for NPI with subcategories
- Study of classifiers which use NPI with various different uncertainty measures

References

- Abellán, J. and Moral, S. (2003) Maximum entropy for credal sets *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **11(5)**, 587-597.
- Augustin, T. and Coolen, F.P.A. (2004) Nonparametric predictive inference and interval probability *Journal of Statistical Planning and Inference*, **124**, 251-272.
- Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model *ISIPTA '05*, 125-134.
- Coolen, F.P.A. (2006) On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, **15**, 21-47.
- Coolen, F.P.A. and Augustin, T. (2009) A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories *International Journal of Approximate Reasoning*, **50**, 217-230.