C4.5
Assessing the Improvements of ID3

Lauren Flanakin

Math 3210

Outline

- Assessed Improvements
- Datasets
- C4.5 Structure
- Decision Tree
- The Focus
- The Hypothesis
- Quinlan's Issues
- Is C4.5 an Improvement?

Assessed Improvements

- Processes data with missing attribute values.
- Processes noisy data.

www.wikipedia.com

Datasets used to Assess C4.5

- Diabetes
  - Noisy and has missing values
- Wine
  - Noisy

C4.5 Structure

- Modifying datasets into comma delimited form.
- Using the Entropy formula:

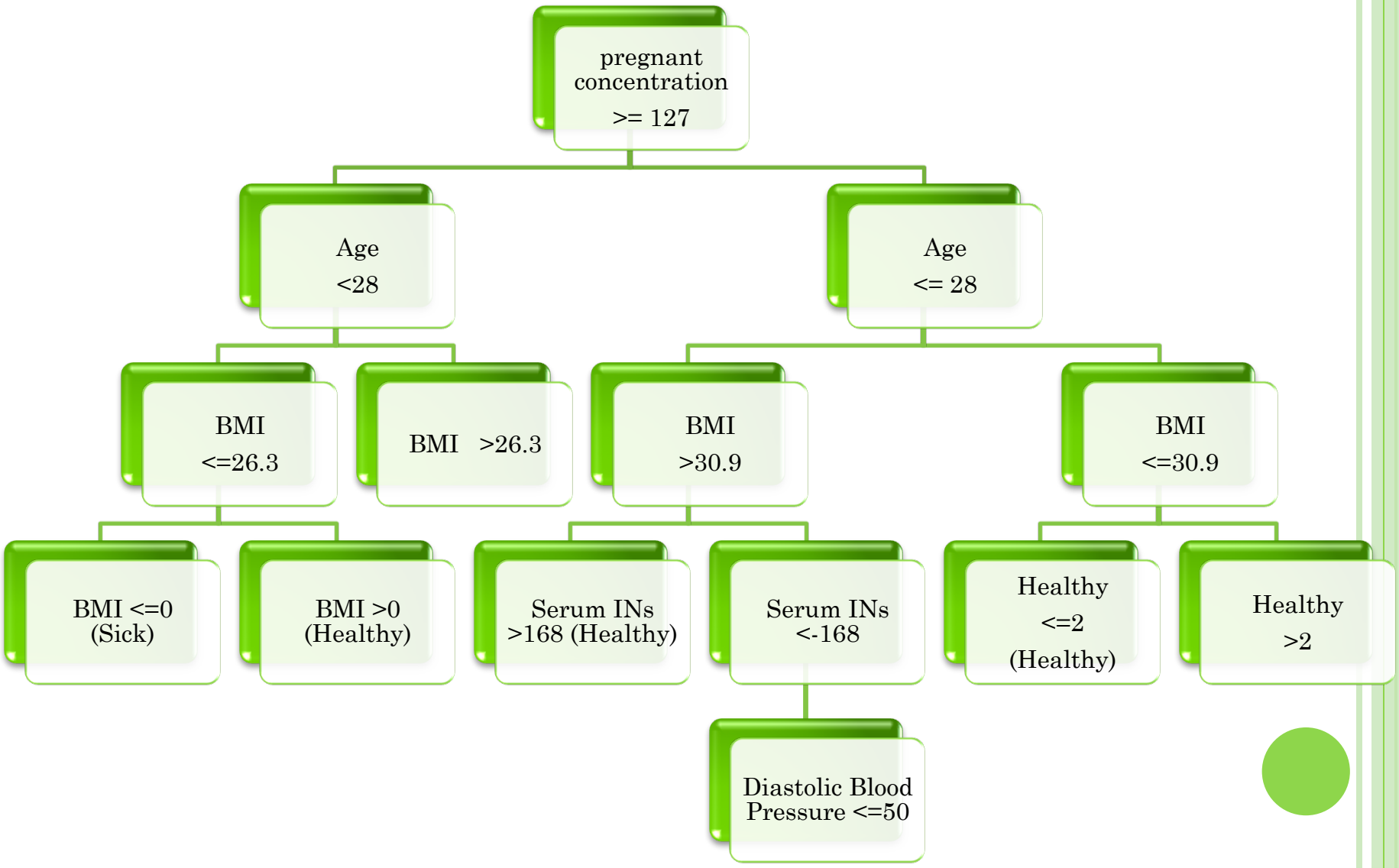
$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

- Entropy finds the attribute with the highest probability in class X.

- Create Decision tree.

www.wikipedia.com

Diabetes Decision tree



The Focus

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes test	1/100%	0.0%	0.0%
Diabetes test	2/150%	10.0%	-1%
Diabetes test	1/150%	0.0%	-15.9%

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes 2	1/100%	0.0%	0.0%
Diabetes 2	2/500%	3.6%	-1%
Diabetes 2	1/150%	0.0%	-8.7%

The Hypothesis

Quinlan's Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes test	3/15%	15%	48%
Diabetes test	4/10%	15%	52.9%

Quinlan's Default

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes test	2/25%	20%	39.7%

My Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes test	2/100%	10%	12.5%

The Hypothesis

Quinlan's Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes 2	4/10%	12.6%	24.2%
Diabetes 2	3/15%	12.3%	22.5%

Quinlan's Default

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes 2	2/25%	7.5%	18.1%

My Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes 2	1/25%	4.4%	17.3%

The Hypothesis

Quinlan's Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Wine	3/10%	2%	10.1%
Wine	3/15%	2%	8.7%

Quinlan's Default

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Wine	2/25%	0.7%	5.8%

My Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Wine	2/100%	1%	7.6%

Quinlan's Issues

- "When the total amount of data is moderate (several hundred cases), different divisions of the data into training and test sets can produce surprisingly large variations into error rates on unseen cases."

J. Ross Quinlan

Quinlan's Issues

- Use of Default
  - Default is used when a dataset does not have any selected rules.

J. Ross Quinlan

Is C4.5 an Improvement?

Quinlan's Default

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes 2	2/25%	7.5%	18.1%

My Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Diabetes 2	1/25%	4.4%	17.3%

Is C4.5 an Improvement?

Quinlan's Default

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Wine	2/25%	0.7%	5.8%

My Suggestion

Dataset	Weight/Pruning	Percentage of Error	Estimated Percentage of Error
Wine	2/100%	1%	7.6%

Review

- Assessed Improvements
- Datasets
- C4.5 Structure
- Decision Tree
- The Focus
- The Hypothesis
- Quinlan's Issues
- Is C4.5 an Improvement?

REFERENCES

- Dunham, M. (2006). *Data Mining: Introductory and advanced topics*. India: Delhi, Dorling Kindersley Pvt. Ltd.
- Seidler, T. (2004). *The C4.5 Project: An overview of the algorithm with results of experimentation.* www.mercury.webster.edu/aleshunas Retrieved 12/7/09
- Quinlan, Ross J. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufman Publishers, Inc.
- C4.5. www. Mercury.webster.edu/aleshunas.
- www.wikipedia.com Retrieved 11/27/09