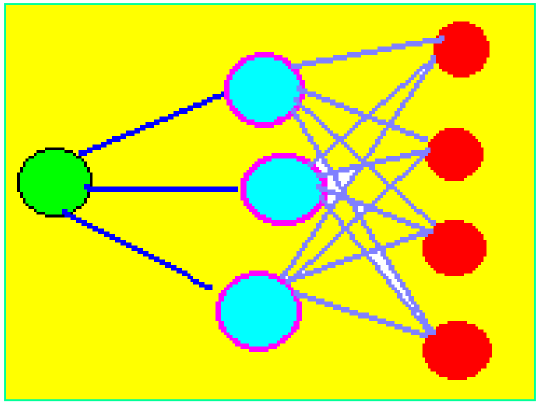
Lecture No. 6 – Info-Fuzzy Network

- IFN Overview
- Network Construction Procedure
- Prediction and Rule Extraction
- Main Characteristics
- Comparative Evaluation
- Software



About 15,500 results on Google (April 2019)

Info-Fuzzy Network (IFN)

Full Description:

1) O. Maimon and M. Last, *Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology*, Kluwer Academic Publishers, Boston, December 2000. **119 citations (April 2019)**

2) M. Last and O. Maimon, A Compact and Accurate Model for Classification, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 2, pp. 203-215, February 2004. **80 citations (April 2019)**

The Fragmentation Problem of Decision Trees

- In the top-down decision-tree construction procedure, the number of training instances at a node decreases with every split
- Conclusion
  - Recursive partitioning leads to statistically insignificant samples at each branch
- IFN (Info-Fuzzy Network) Solution to the Fragmentation Problem
  - Repetitive partitioning of *all* training instances in every layer
  - Statistical significance testing at every node

Lecture No. 6 – Info-Fuzzy Network

- IFN Overview
- Network Construction Procedure
- Prediction and Rule Extraction
- Main Characteristics
- Comparative Evaluation
- Software

IFN Example - Credit Approval Dataset

Source: http://archive.ics.uci.edu/ml/datasets/Credit+Approval

Candidate Input Features

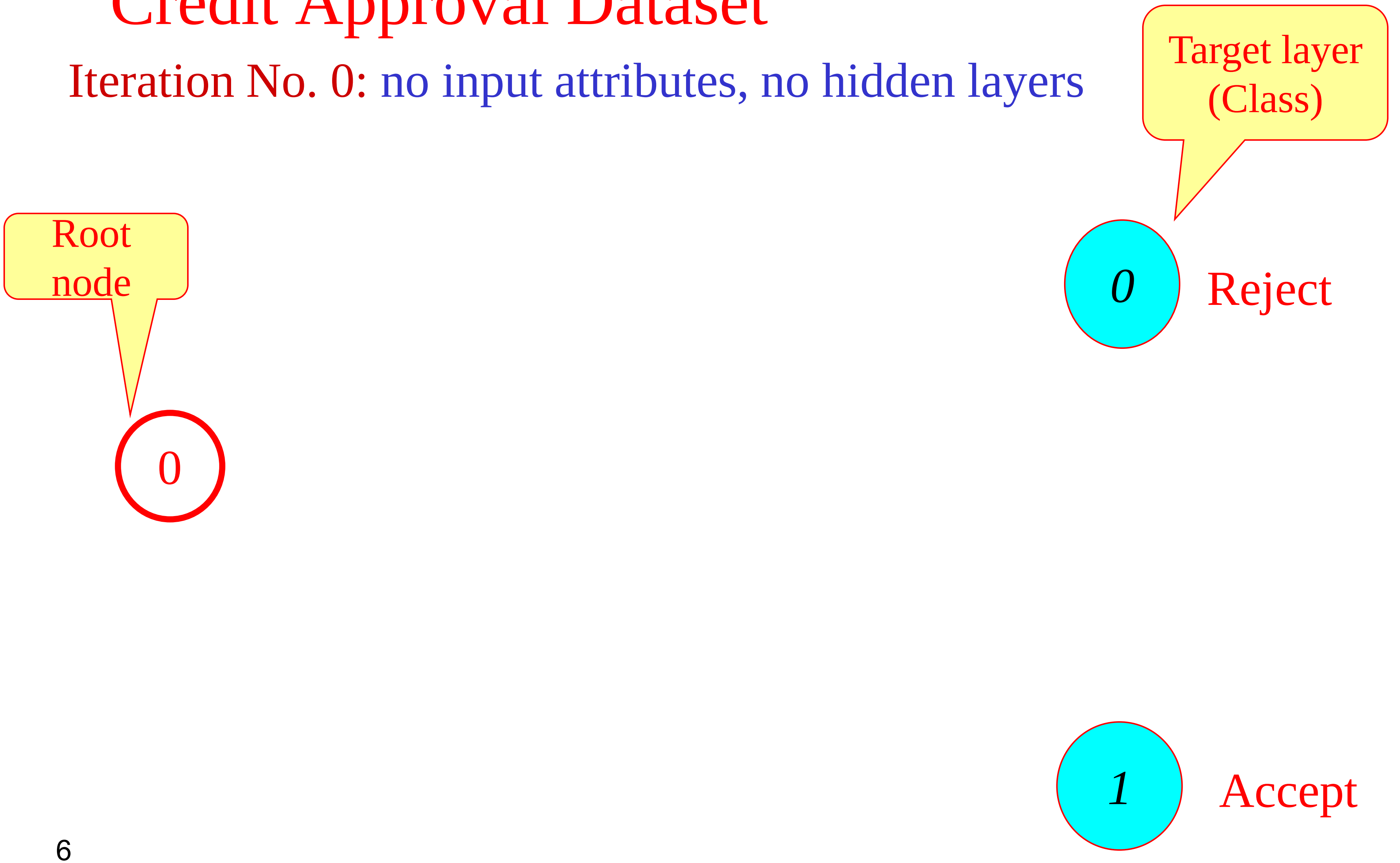
Target

Attribute	Domain	Type	Use in Network
A1 (Sex)	0, 1	Nominal	Candidate input
A2 (Age)	13.75 - 80.25	Continuous	Candidate input
A3 (Mean time at addresses)	0 - 28	Continuous	Candidate input
A4 (Home status)	1, 2, 3	Nominal	Candidate input
A5 (Current occupation)	1 - 14	Nominal	Candidate input
A6 (Current job status)	1 - 9	Nominal	Candidate input
A7 (Mean time with employers)	0 - 28.5	Continuous	Candidate input
A8 (Other investments)	0, 1	Nominal	Candidate input
A9 (Bank account)	0, 1	Nominal	Candidate input
A10 (Time with bank)	0 - 67	Continuous	Candidate input
A11 (Liability reference)	0, 1	Nominal	Candidate input
A12 (Account reference)	1, 2, 3	Nominal	Candidate input
A13 (Monthly housing expense)	0 - 2000	Continuous	Candidate input
A14 (Savings account balance)	1 - 100001	Continuous	Candidate input
Class (Accept / Reject)	0, 1	Nominal	Target

IFN Construction Procedure (0)

Credit Approval Dataset

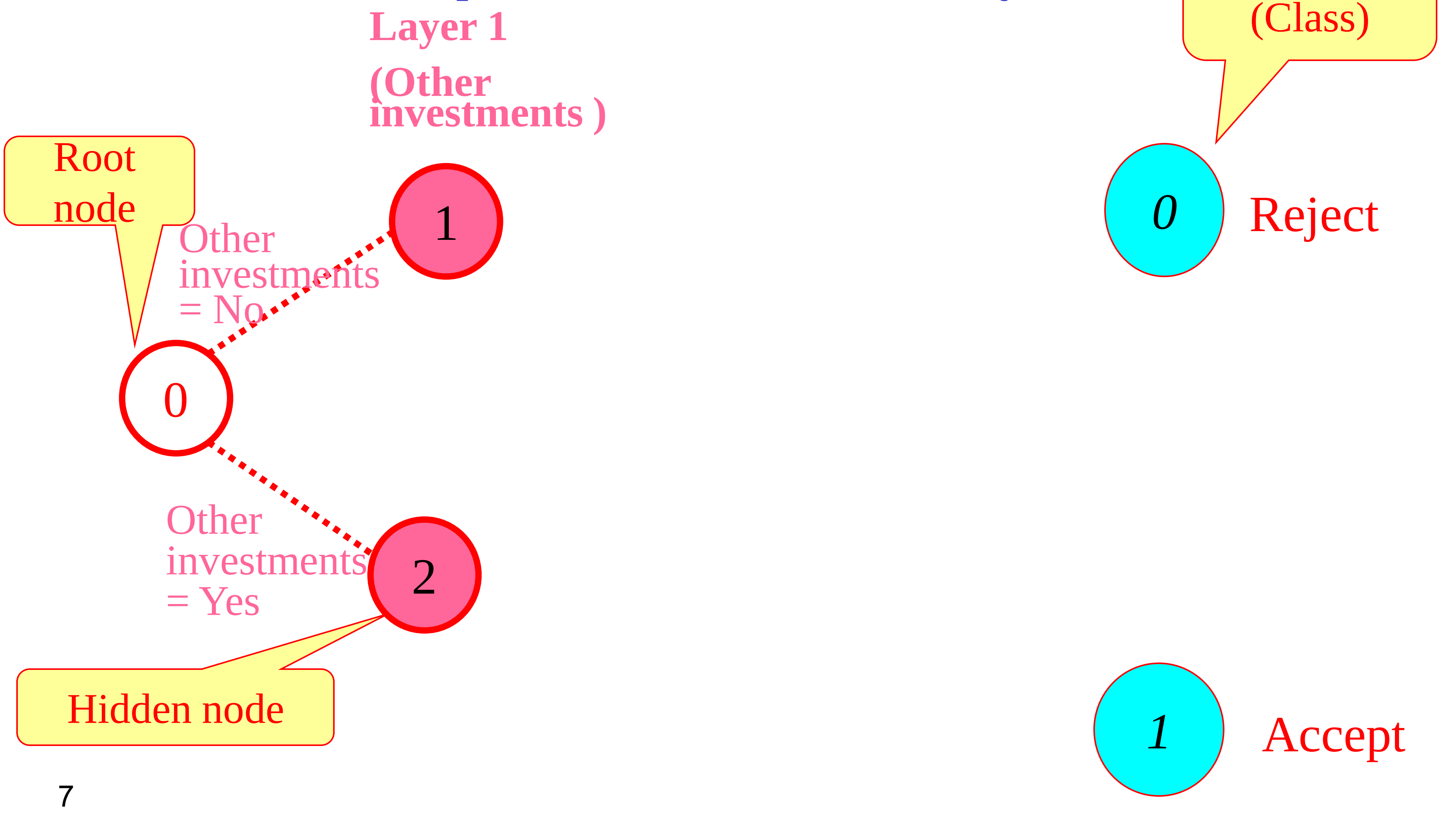
Iteration No. 0: no input attributes, no hidden layers



IFN Construction Procedure (1)

Credit Approval Dataset

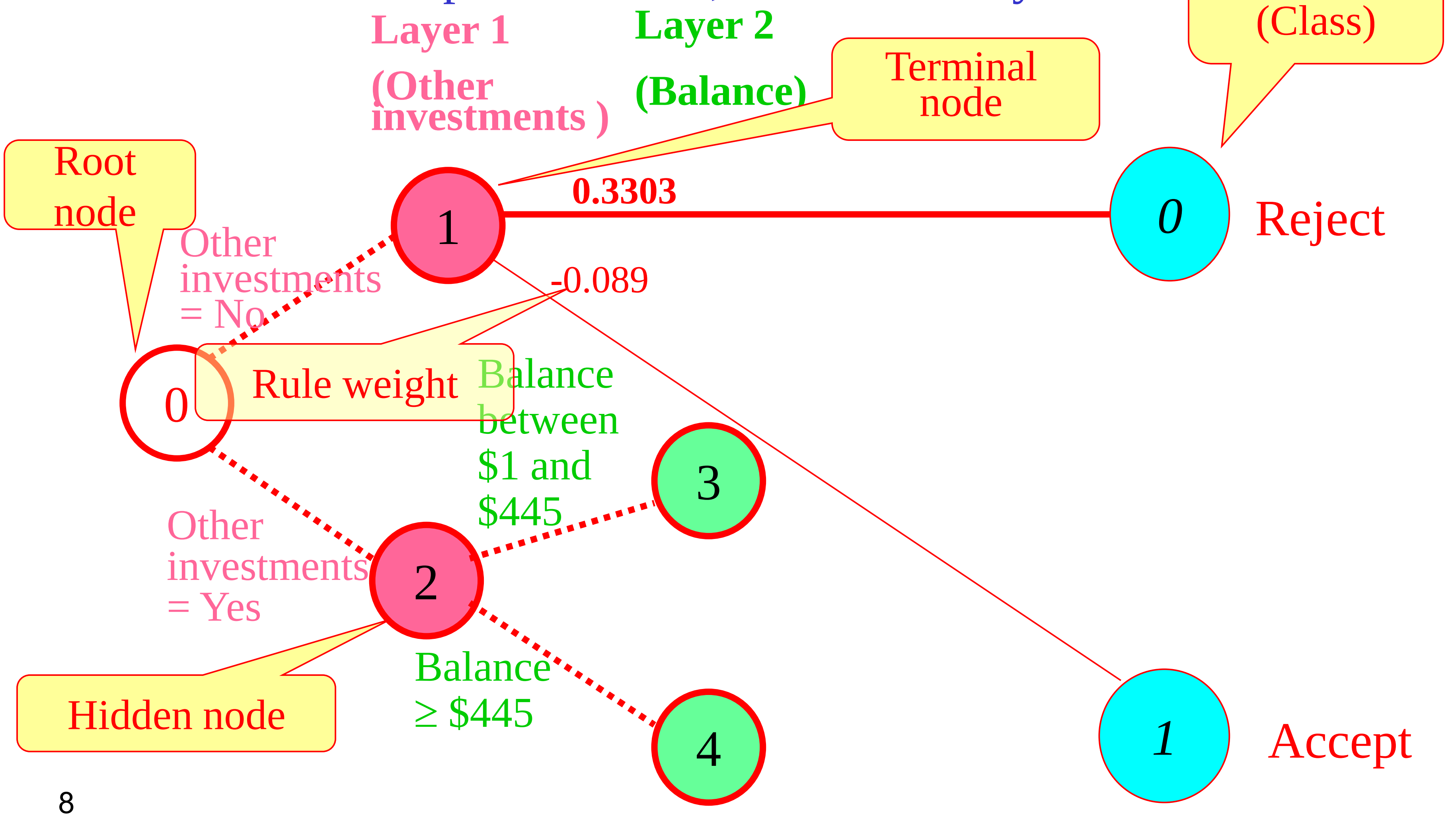
Iteration No. 1: one input attribute, one hidden layer



IFN Construction Procedure (2)

Credit Approval Dataset

Iteration No. 2: two input attributes, two hidden layers

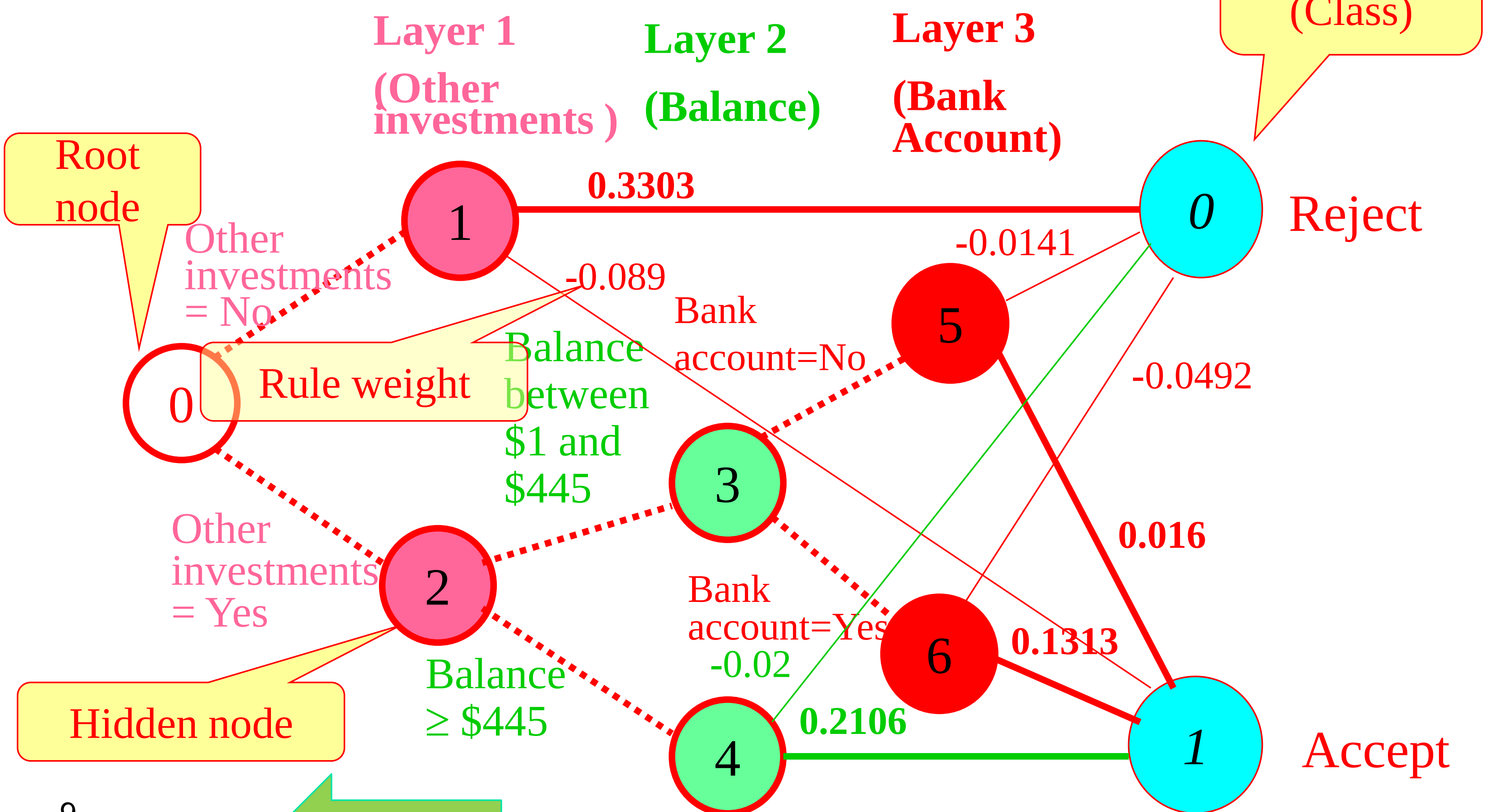


8

IFN Construction Procedure (3)

Credit Approval Dataset

Iteration No. 3: three input attributes, three hidden layers



9

Network Induction Algorithm

Input

- Extended relation schema (partition of attribute set)
- Set of training records
- Minimum significance level (default = 0.1%)

Output

- Set of selected input attributes
- Information-theoretic network

Step 1 - Initialize the information-theoretic network.

Step 2 - While the maximum number of hidden layers is not exceeded:

Step 2.1 - Find a candidate input attribute maximizing the statistically significant conditional mutual information (“the best candidate attribute”).

Step 2.2 - If the maximum conditional mutual information is greater than zero, make the best candidate attribute an **input attribute** and define a new layer of hidden nodes; else stop.

Step 3 – Return the set of selected attributes and the network structure

IFN: Conditional mutual information (MI) at a node z

$$MI\,(A_{i'}\,;\,A_i\,/\,z) = \sum_{j=0}^{M_i-1}\sum_{j'=0}^{M_{i'}-1} P(V_{ij};V_{i'j'};z)\bullet\log\frac{P(V_{i'j'}^{ij}/z)}{P(V_{i'j'}/z)\bullet P(V_{ij}/z)}$$

A_i - target attribute No. i

$A_{i'}$ - candidate input attribute No. i'

V_{ij} - value No. j of attribute A_i

z - network node (representing a conjunction of input attribute values)

$P\,(V_{ij}/\,z)$ - an estimated conditional (a posteriori) probability of V_{ij} given the node z.

$P\,(V_{i'j'}^{ij}/\,z)$ - an estimated conditional (a posteriori) probability of $V_{i'j'}$ and V_{ij} given the node z.

$P\,(V_{ij};\,V_{i'j'};\,z)$ - an estimated joint probability of $V_{i'j'}V_{ij}$, and the node z

11

Conditional MI Example

Other Investments (Node 0)

$$\sum_{j=0}^{M_i-1}\sum_{j'=0}^{M_{i'}-1} P(V_{ij};V_{i'j'};z)\bullet\log\frac{P(V_{i'j'}^{ij}/z)}{P(V_{i'j'}/z)\bullet P(V_{ij}/z)}$$

j'/ j	0	Cond.	Joint	1	Cond.	Joint	Total	Cond.
0	306	=306/690	=306/690	23	=23/690	=23/690	329	=329/690
1	77	=77/690	=77/690	284	=284/690	=284/690	361	=361/690
Total	383	=383/690		307	=307/690		690	

j'/ j	0	Cond.	Joint	1	Cond.	Joint	Total	Cond.
0	306	0.4435	0.4435	23	0.0333	0.0333	329	0.4768
1	77	0.1116	0.1116	284	0.4116	0.4116	361	0.5232
Total	383	0.5551		307	0.4449		690	

- *MI* $(A_{i'}=0\,;A_i=0\,/\,z)$ = 0.443 * $\log_2$(0.443 / (0.477*0.555))= 0.3303
- *MI* $(A_{i'}=1\,;A_i=0\,/\,z)$ = 0.112 * $\log_2$(0.112 / (0.523*0.555))= -0.1540
- *MI* $(A_{i'}=0\,;A_i=1\,/\,z)$ = 0.033 * $\log_2$(0.033 / (0.477*0.445))= -0.089
- *MI* $(A_{i'}=1\,;A_i=1\,/\,z)$ = 0.412 * $\log_2$(0.412 / (0.523*0.445))= 0.3384

Conditional mutual information *MI* $(A_{i'}\,;A_i\,/\,z)$ = **0.426 bits**

12

IFN: Testing Statistical Significance of *MI*

Likelihood-Ratio Statistic (Attneave, 1959):

$$G^2(A_{i'};A_i/z) = 2 \bullet (\ln 2) \bullet E^* \bullet MI(A_{i'};A_i/z)$$

- A_i - target attribute No. i
- $A_{i'}$ -candidate input attribute No. i'
- z - network node (representing a conjunction of input attribute values)
- E^* – total number of training cases
- $MI(A_{i'}; A_i/z)$ – conditional mutual information

$$H_0 : MI(A_{i'};A_i/z) = 0$$

$$G^2|_{H_0} \sim \chi^2((NI_{i'}(z)-1)\cdot(NT_i(z)-1))$$

A node is split if H_0 is rejected at the 0.1% significance level

$NI_{i'}(z)$ - number of values of a candidate input attribute i' at node z

$NT_i(z)$ - number of values of a target attribute i at node z

Likelihood-Ratio Example

Other Investments (Node 0)

j'/ j	0	Cond.	Joint	1	Cond.	Joint	Total	Cond.
0	306	0.4435	0.4435	23	0.0333	0.0333	329	0.4768
1	77	0.1116	0.1116	284	0.4116	0.4116	361	0.5232
Total	383	0.5551		307	0.4449		690	

- Conditional mutual information $MI\ (A_{i'}\, ; A_i\ /\ z)$ = 0.426 bits
- Likelihood-Ratio Statistic $G^2\ (A_{i'}\, ; A_i\ /\ z) = 2*ln2*690*0.426 = 407$
- Degrees of Freedom = (2-1)*(2-1) = 1
- Significance level >> 0.1%
- Conclusion: reject H_0 (consider *Other Investments* as the next input attribute)

IFN: Conditional Mutual Information in a Layer *i'*

$$MI\,(A_{i'}\,;\,A_i) = \boxed{\sum_{\substack{z \in Layer_{i'} \\ Split(z)=true}} MI(A_{i'};A_i\,/\,z)}$$

A_i - target attribute No. i

$A_{i'}$ - candidate input attribute No. i'

z - network node

$MI\,(Ai'\,;\,Ai\,/\,z)$ – conditional mutual information between A_i and $A_{i'}$ given node z

Example: Layer $_{i'}$ = 0; $A_{i'}$ = *Other Investments*

MI (Other Investments; Class) = *MI (Other Investments; Class / z = 0)* = 0.426 bits

15

Global Discretization of Continuous Attributes

- *Input:* The first and the last distinct values in the interval S of a continuous attribute $A_{i'}$
- *Step 1* – For every distinct value T included in the interval S (except for the first distinct value) Do:
  - *Step 1.1* –For every node z of the final hidden layer Do:
  - *Step 1.1.1* - Calculate the likelihood-ratio test for the partition of the interval S at the threshold T and the target attribute A_i given the node z

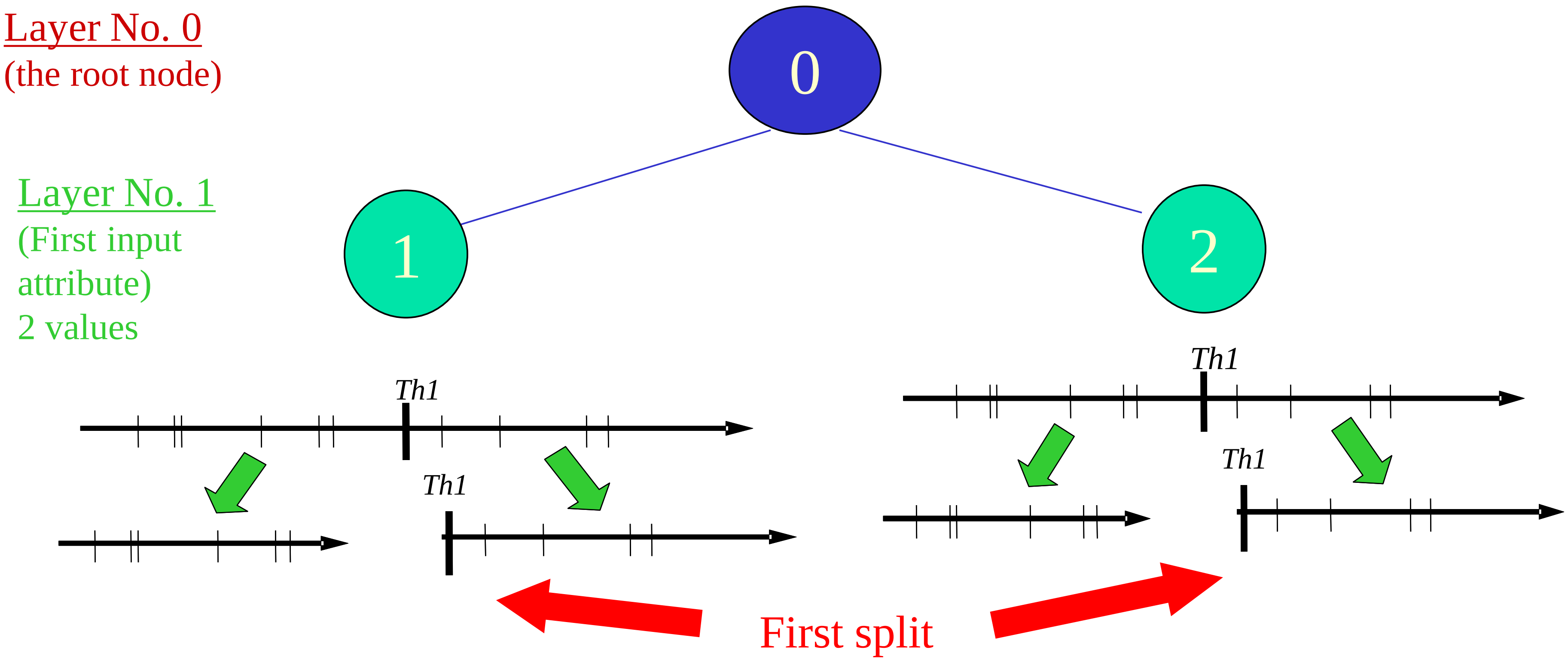
    First interval: $A_{i'} < T$; Second interval: $A_{i'} \geq T$
  - *Step 1.1.2* - If the likelihood-ratio statistic is significant, mark the node as "split" by the threshold T
  - *Step 1.1.3* - End Do
  - *Step 1.2* – End Do
- *Step 2* – Find the threshold T_{max} maximizing the sum of conditional mutual information over all nodes

Global Discretization Procedure (2)

- *Step 3* – If the maximum estimated conditional mutual information is greater than zero, then Do:
  - *Step 3.1* - For every node *z* of the final hidden layer Do:
  - *Step 3.1.1* – If the node *z* is split by the threshold T_{max}, mark the node as split by the candidate input attribute $A_{i'}$
  - *Step 3.2* - Partition each sub-interval of *S*
  - *Step 3.3* - End Do
- *Step 4* - Else return the list of threshold values for $A_{i'}$

Dynamic Discretization Procedure (cont.)

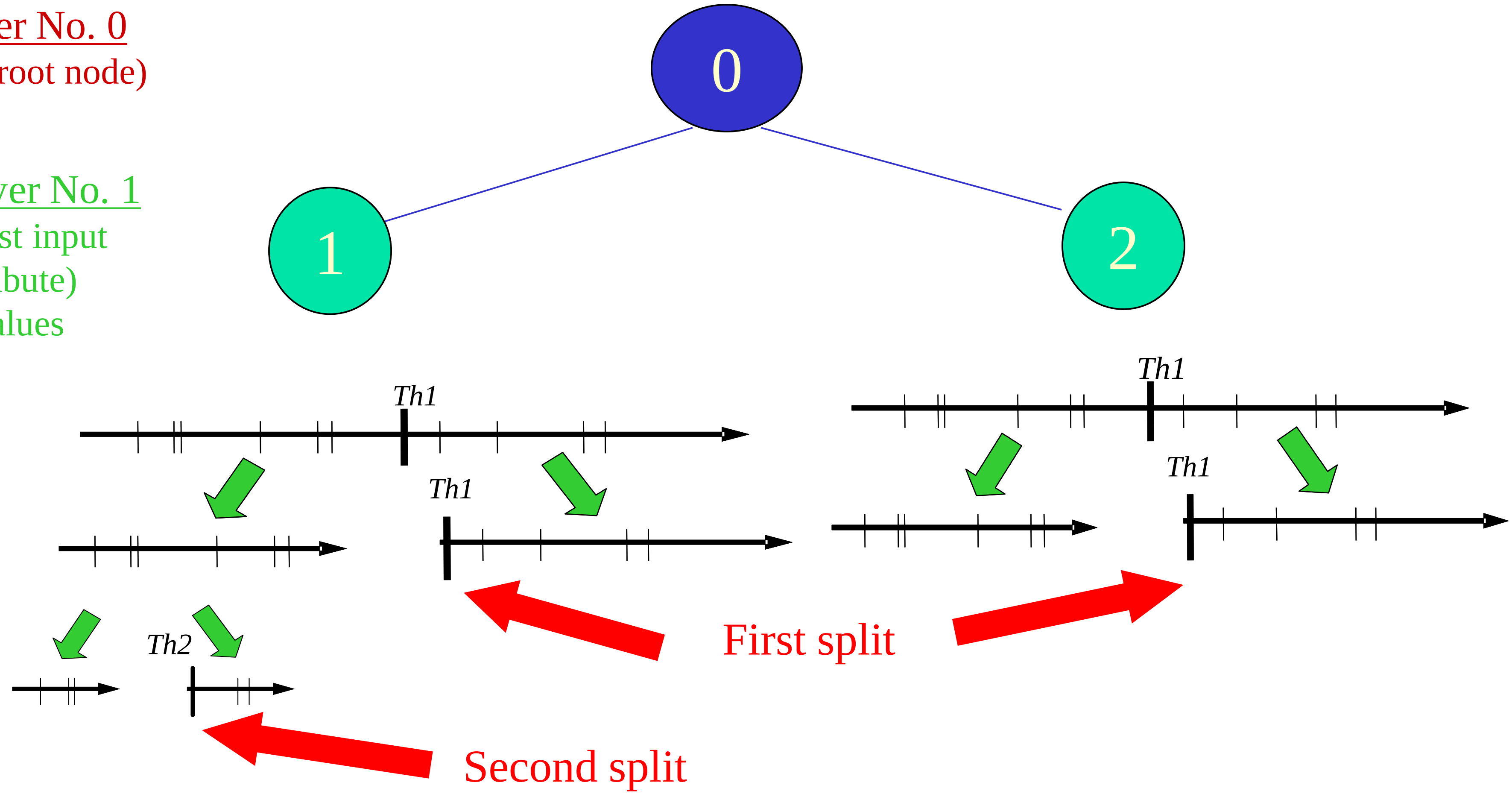
Example: discretization of the *second* input attribute in the network



Dynamic Discretization Procedure (cont.)

Layer No. 0
(the root node)

Layer No. 1
(First input attribute)
2 values



Conditional Mutual Information (Global Discretization Procedure)

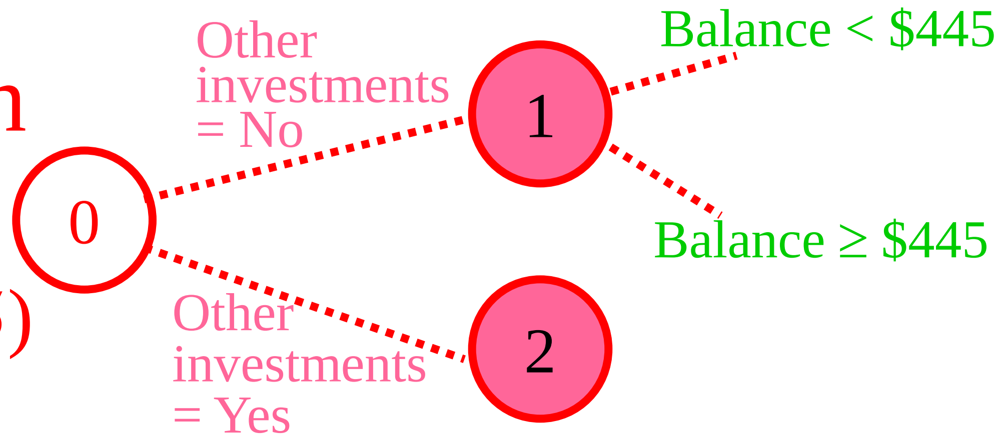
$$MI(Th;T/S,z)=$$

$$\sum_{t=0}^{M_i-1}\sum_{y=1}^{2} P(S_y;C_t;z)\bullet\log\frac{P(S_y;C_t\ /S,z)}{P(S_y/S,z)\bullet P(C_t/S,z)}$$

- $P\ (S_y/S,z)$ - an estimated conditional (a posteriori) probability of a sub-interval S_y given the interval S and the node z
- $P\ (C_t/S,z)$ - an estimated conditional (a posteriori) probability of a value C_t of the target attribute T given the interval S and the node z
- $P\ (S_y;C_t/S,z)$ - an estimated joint probability of a value of the target attribute T and a sub-interval S_y given the interval S and the node z
- $P\ (S_y;C_t;z)$ - an estimated joint probability of a value of the target attribute T, a sub-interval S_y, and the node z

Global Discretization Example (1)

Balance (Node 1, T = 445)

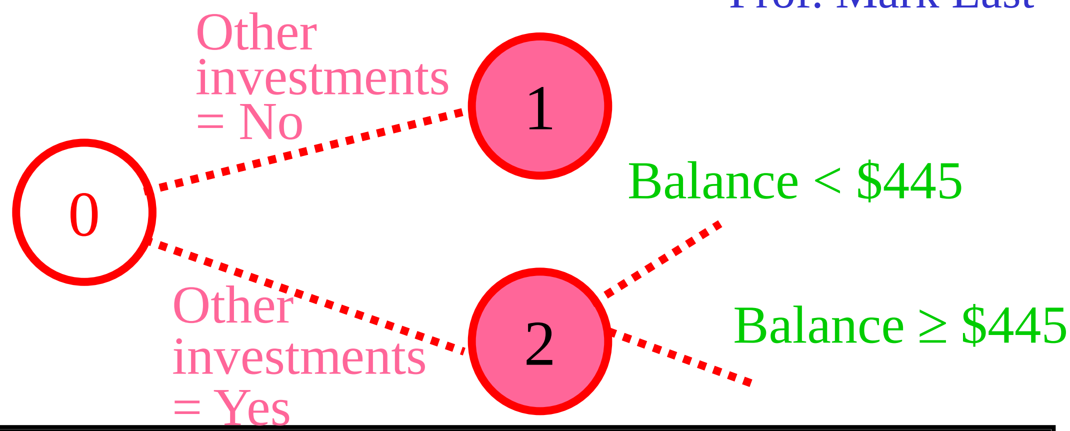


j' / j	0	Cond.	Joint	1	Cond.	Joint	Total	Cond.
<445	271	0.824	0.393	20	0.061	0.029	291	0.8845
>=445	35	0.106	0.051	3	0.009	0.004	38	0.1155
Total	306	0.930		23	0.070		329	

- Conditional mutual information $MI\ (A_{i'} ; A_i / z)$ = 0.0000546 bits
- Likelihood-Ratio Statistic $G^2\ (A_{i'} ; A_i / z) = 0.0522$
- Degrees of Freedom = (2-1)*(2-1) = 1
- Significance level < 18%
- Conclusion: do not reject H_0 (**do not split** the *Balance* attribute at this node for the specified threshold)

Global Discretization Example (2)

Balance (Node 2, T = 445)



j' / j	0	Cond.	Joint	1	Cond.	Joint	Total	Cond.
< 445	74	0.205	0.107	156	0.432	0.226	230	0.637
>= 445	3	0.008	0.004	128	0.355	0.186	131	0.363
Total	77	0.213		284	0.787		361	

- Conditional mutual information $MI\ (A_{i'}\ ; A_i\ /\ z)$ = 0.0592 bits
- Likelihood-Ratio Statistic $G^2\ (A_{i'}\ ; A_i\ /\ z)$ = 56.6564
- Degrees of Freedom = (2-1)*(2-1) = 1
- Significance level >> 0.1%
- Conclusion: reject H_0 (**split** the *Balance* attribute at this node for the specified threshold)

Credit Dataset - Layer 0

Attribute	**Significant Conditional Mutual Information**
A1 (Sex)	0
A2 (Age)	0.023
A3 (Mean time at addresses)	0.041
A4 (Home status)	0.03
A5 (Current occupation)	0.109
A6 (Current job status)	0.05
A7 (Mean time with employers)	0.123
A8 (Other investments)	0.426
A9 (Bank account)	0.156
A10 (Time with bank)	0.214
A11 (Liability reference)	0
A12 (Account reference)	0
A13 (Monthly housing expense)	0.051
A14 (Savings account balance)	0.123

Credit Dataset - Layer 1

Attribute	**Significant Conditional Mutual Information**
A1 (Sex)	0
A2 (Age)	0
A3 (Mean time at addresses)	0.041
A4 (Home status)	0
A5 (Current occupation)	0
A6 (Current job status)	0
A7 (Mean time with employers)	0.018
A9 (Bank account)	0.055
A10 (Time with bank)	0.055
A11 (Liability reference)	0
A12 (Account reference)	0.022
A13 (Monthly housing expense)	0.027
A14 (Savings account balance)	0.059

Credit Dataset - Layer 2

Attribute	**Significant Conditional Mutual Information**
A1 (Sex)	0
A2 (Age)	0
A3 (Mean time at addresses)	0.027
A4 (Home status)	0
A5 (Current occupation)	0
A6 (Current job status)	0
A7 (Mean time with employers)	0
A9 (Bank account)	0.031
A10 (Time with bank)	0.031
A11 (Liability reference)	0
A12 (Account reference)	0
A13 (Monthly housing expense)	0.022

Credit Dataset - Layer 3

Attribute	**Significant Conditional Mutual Information**
A1 (Sex)	0
A2 (Age)	0
A3 (Mean time at addresses)	0
A4 (Home status)	0
A5 (Current occupation)	0
A6 (Current job status)	0
A7 (Mean time with employers)	0
A10 (Time with bank)	0
A11 (Liability reference)	0
A12 (Account reference)	0
A13 (Monthly housing expense)	0

Lecture No. 6 – Info-Fuzzy Network

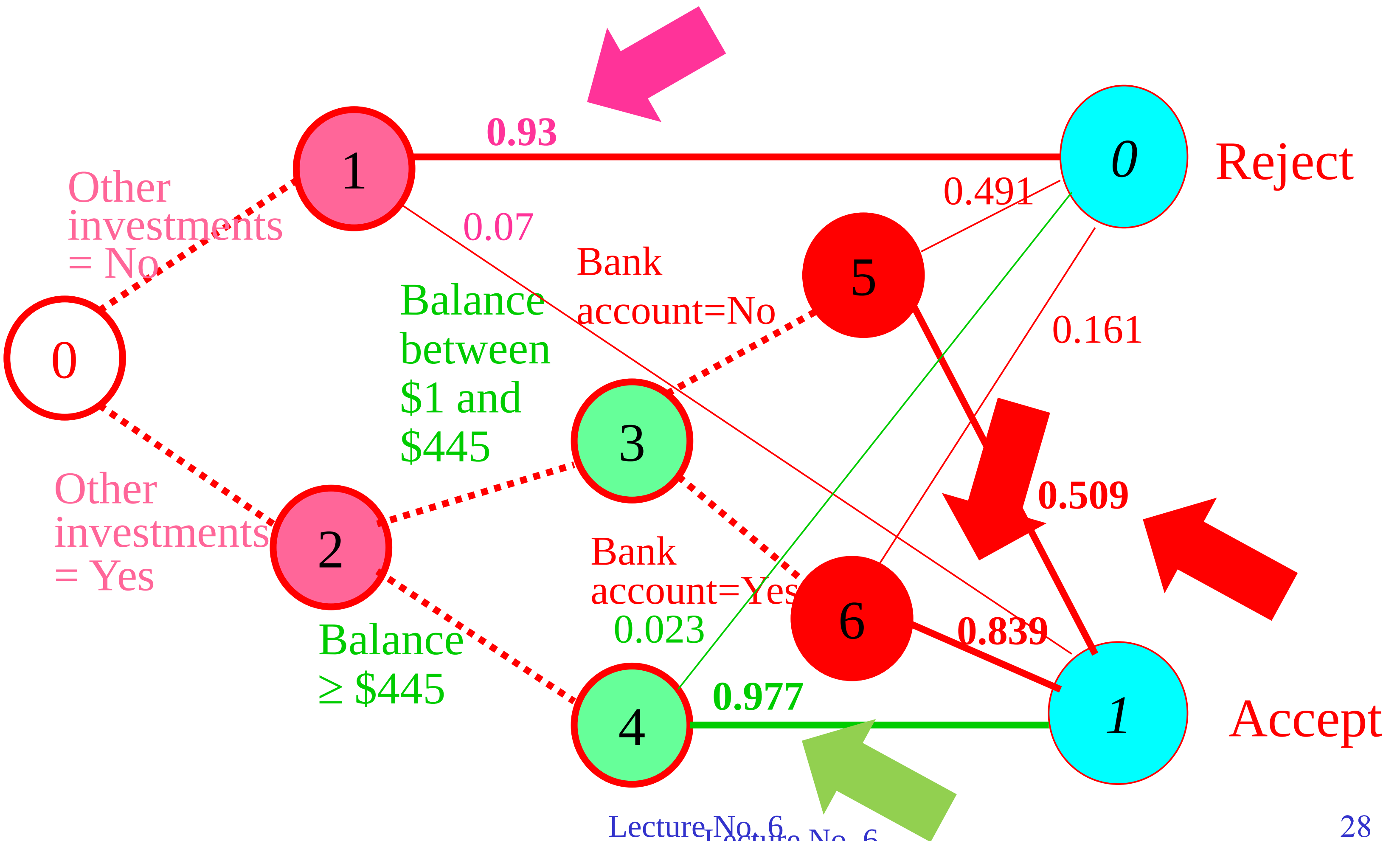
- IFN Overview
- Network Construction Procedure
- Prediction and Rule Extraction
- Main Characteristics
- Comparative Evaluation
- Software

Prediction

Predicted Value (maximum *a posteriori*):

(of the target attribute A_i at the node z)

$$j^* = \arg\max_j P(V_{ij} / z)$$



Rule Extraction and Scoring

Connection Weight:

Use ?

$$w_z^{ij} = P(V_{ij}; z) \bullet \log \frac{P(V_{ij} / z)}{P(V_{ij})}$$

V_{ij} - value No. j of target attribute A_i

$P(V_{ij}; z)$ - an estimated joint probability of V_{ij} and the node z

$P(V_{ij})$ - an estimated unconditional (a priori) probability of V_{ij}

$P(V_{ij}/ z)$ - an estimated conditional (a posteriori) probability of V_{ij} given the node z

Interpretation: mutual information between the node z and the value j of the target attribute A_i

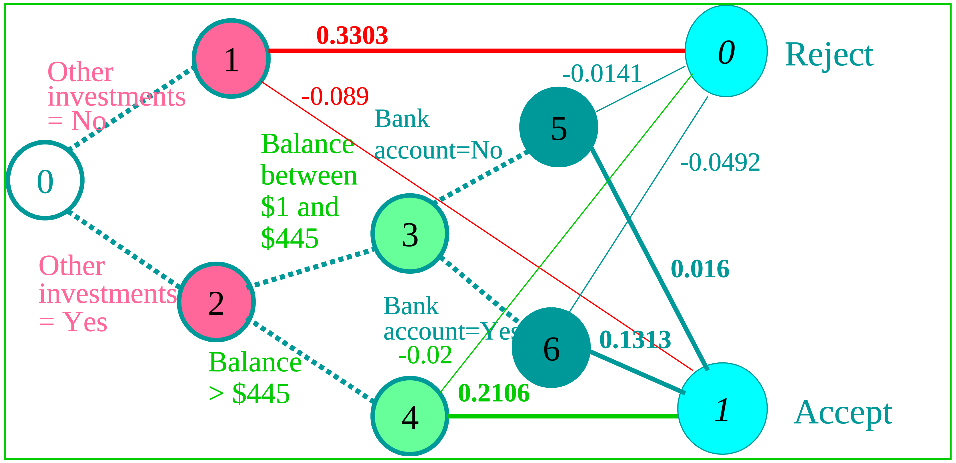
$w_z^{ij} > 0$: The probability of V_{ij} at z is higher than average

$w_z^{ij} < 0$: The probability of V_{ij} at z is lower than average

Rule Extraction Example

Credit Dataset

$$w_z^{ij} = P(V_{ij};z) \bullet \log \frac{P(V_{ij}/z)}{P(V_{ij})}$$



Example – Rule No. 1 (Connection 1 → 0)

Other investments = No: 329 records

Other investments = No and Class = Reject: 306 records

Class = Reject: 383 records

Total records: 690

P (V_{ij}) = 383 / 690 = 0.5551 (unconditional probability)

P (V_{ij}/ *z*) = 306 / 329 = 0.9301 (conditional probability)

P (V_{ij}; *z*) = 306 / 690 = 0.4435 (joint probability)

Rule Weight: 0.4435*log (0.9301 / 0.5551) = 0.3303 units?

Lecture No. 6 – Info-Fuzzy Network

- IFN Overview
- Network Construction Procedure
- Prediction and Rule Extraction
- Main Characteristics
- Comparative Evaluation
- Software

IFN Characteristic 1

The overall decrease in conditional entropy of the target attribute is equal to the sum of drops in conditional entropy across the network hidden layers (based on the Chain Rule)

$$\boxed{\begin{aligned} &MI(A_i ; I_i) = \\ &\sum_{s=1}^{m} MI(A_{i'}(s); A_i / A_{i'}(1), \ldots, A_{i'}(s-1)) \end{aligned}}$$

A_i – target attribute i

I_i - set of input attributes in the network of the target attribute i

m - total number of layers (input attributes)

$A_{i'}(s)$ –input attribute i' associated with the layer s

Implication: IFN can be constructed <u>incrementally</u>

IFN Characteristic 1 - Example

Credit Dataset - Summary Table

Iteration	Attribute Name	Mutual Information	Conditional MI	Conditional Entropy	Split Nodes	MI to Attributes
0	Other investments (A8)	0.426	0.426	0.566	1	0.426
1	Balance (A14)	0.485	0.059	0.506	1	0.243
2	Bank account (A9)	0.516	0.031	0.475	1	0.172

$$MI(A_i; I_i) = \sum_{s=1}^{m} MI(A_{i'}(s); A_i / A_{i'}(1),\ldots, A_{i'}(s-1)) = 0.426 + 0.059 + 0.031 = 0.516$$

IFN Characteristic 2

The sum of connection weights at all terminal nodes is equal to the estimated mutual information between the set of input attributes I_i and the target attribute A_i (based on the definition of mutual information)

$$\boxed{MI(A_i; I_i) = \sum_{z \in F} \sum_{j=0}^{M_i - 1} w_z^{ij}}$$

F – set of terminal nodes z

M_i – number of distinct values (classes) of the target attribute A_i

Implication: the rule weights represent the contribution of each terminal node to the overall mutual information

IFN Characteristic 2 – Example

Credit Dataset - Extracted Rules

No.	**Rule**	**Weight**
1	If Other investments is 0 then Class is 0	0.330
2	If Other investments is 0 then Class is not 1	-0.089
3	If Other investments is 1 and Balance is more than 445.00000 then Class is not 0	-0.020
4	If Other investments is 1 and Balance is more than 445.00000 then Class is 1	0.211
5	If Other investments is 1 and Balance is between 1.00000 and 445.00000 and Bank account is 0 then Class is not 0	-0.014
6	If Other investments is 1 and Balance is between 1.00000 and 445.00000 and Bank account is 0 then Class is 1	0.016
7	If Other investments is 1 and Balance is between 1.00000 and 445.00000 and Bank account is 1 then Class is not 0	-0.049
8	If Other investments is 1 and Balance is between 1.00000 and 445.00000 and Bank account is 1 then Class is 1	0.131
	Total	0.516

$$\boxed{MI(A_i;I_i)=\sum_{z\in F}\sum_{j=0}^{M_i-1} w_z^{ij}}$$

41

IFN Characteristic 3

Minimum Prediction Error *Pe* of a given info-fuzzy network can be estimated based on Fano's inequality:

$$\boxed{H(A_i / I_i) \le H(P_e) + P_e \log_2(M_i - 1)}$$

A_i – target attribute i

I_i - set of input attributes in the network of the target attribute i

M_i – number of distinct values (classes) of the target attribute A_i

Implication: no testing set is needed to estimate the maximum achievable accuracy of a given network

IFN Characteristic 3 – Example Credit Dataset

Conditional entropy: $H\,(A_i\,/\,I_i\,) = 0.475$

Number of classes $M_i = 2$

$$\boxed{\begin{aligned}&H(P_e)+P_e\,\log_2(M_i-1)-H(A_i/I_i)=\\&-P_e\,*\log_2(P_e)-(1-P_e)*\log_2(1-P_e)+P_e\,*\log_2(2-1)-0.475=0\end{aligned}}$$

Min $P_e = 0.102$

$$\boxed{-0.102*\log_2(0.102)-0.898*\log_2(0.898)+0.102*\log_2(2-1)-0.475=0}$$

Mean P_e (10-fold cross-validation) = 0.159 > 0.102

Lecture No. 6 – Info-Fuzzy Network

- IFN Overview
- Network Construction Procedure
- Prediction and Rule Extraction
- Main Characteristics
- Comparative Evaluation
- Software

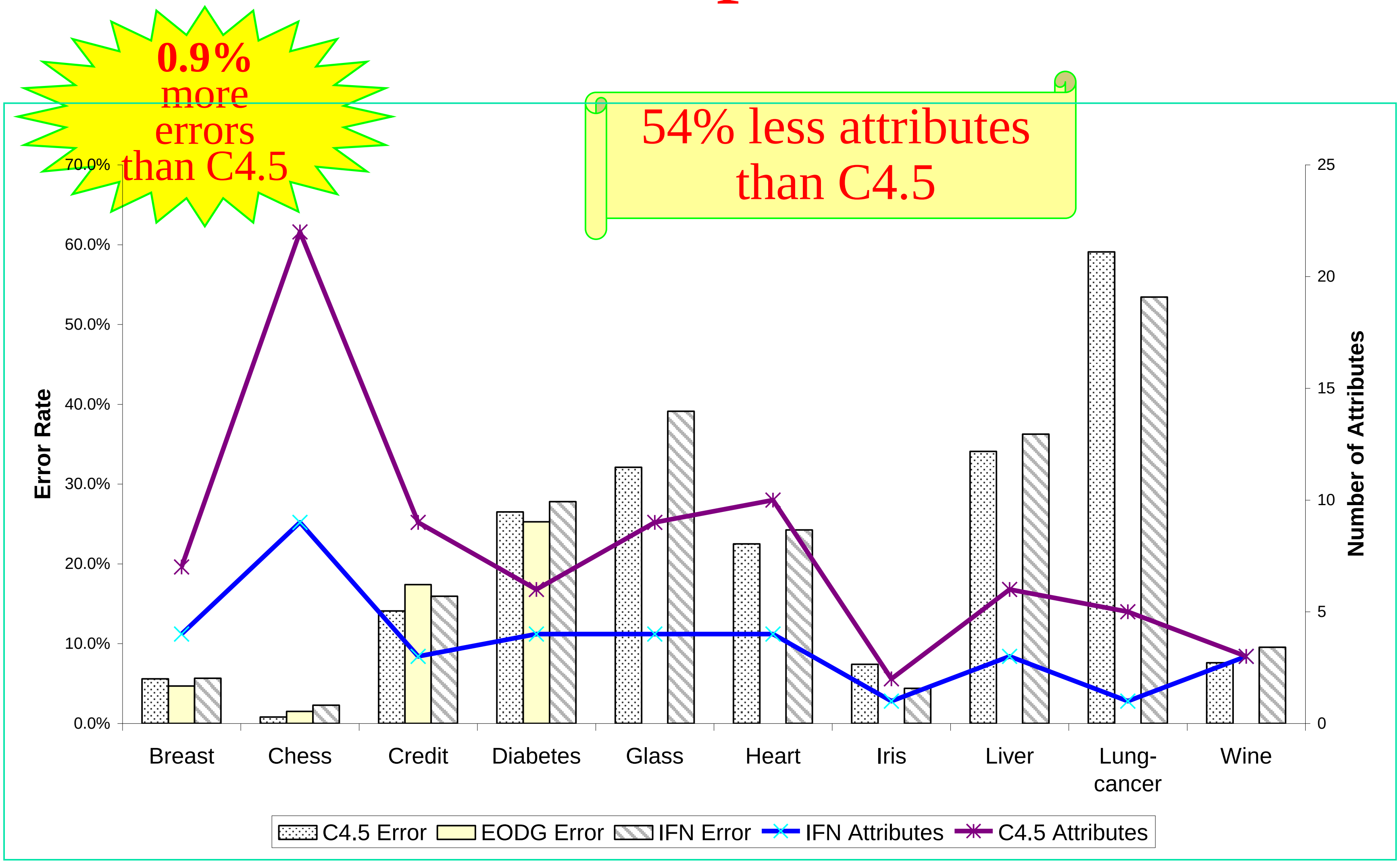
Comparison to Other Methods

From M. Last, O. Maimon, E. Minkov, "Improving Stability of Decision Trees", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, No. 2, pp. 145-159, 2002

Kohavi (1995)

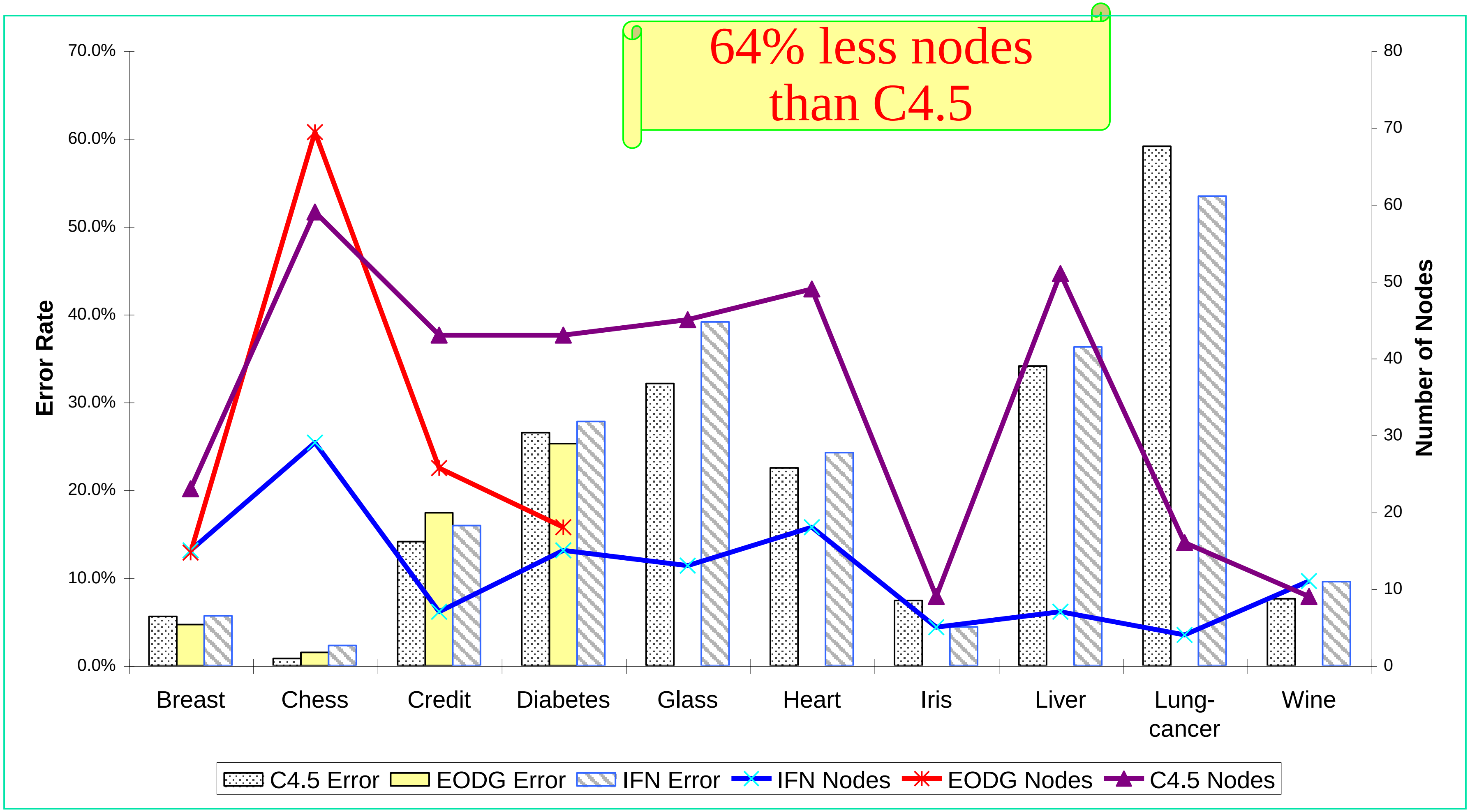
Property	**CART / C4.5**	**EODG**	**IFN**
Tree construction strategy	Recursive partitioning of a subset of training instances at each node	Repetitive partitioning of all training instances in every level	Repetitive partitioning of all training instances in every layer (except for instances at unsplit nodes)
Feature selection	The best feature is selected for every node	All nodes in a given level are split on the same feature	All nodes in a given layer are split on the same feature
Splitting criteria	CART: Gini, Twoing, Entropy C4.5: Gain Ratio	Adjusted Mutual Information	Conditional Mutual Information
Splits on continuous features	Binary (threshold) splits only The same feature may be tested at different levels	Binary (threshold) splits only The same feature may be tested at different levels	Multi-way splits The same feature is not tested at more than one layer
Pre-pruning criteria	Minimum number of cases for each outcome at a node	The instances are split on all features	Likelihood-Ratio Test
Post-pruning criteria	CART: cost-complexity pruning C4.5: Reduced error bottom-up pruning	Bottom-up error-based pruning Top-down merging of nodes	No post-pruning
Target (Category) layer	No	Yes	Yes

Number of Input Attributes



46

Network Size (Number of Nodes)



47

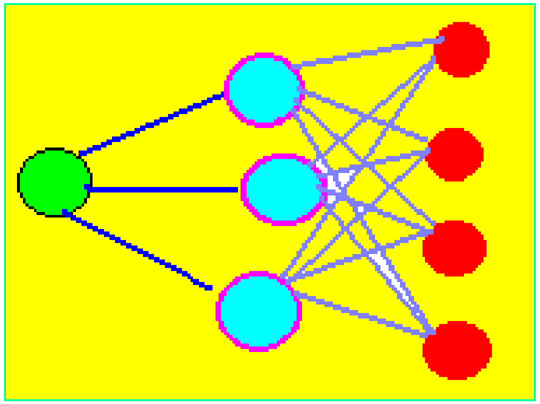
IFN – Selected References

- O. Maimon and M. Last, “Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology”, Kluwer Academic Publishers, Boston, December 2000.
- O. Maimon, A. Kandel, and M. Last, “Information-Theoretic Fuzzy Approach to Data Reliability and Data Mining”, Fuzzy Sets and Systems, Vol. 117, No. 2, pp. 183-194, Jan. 2001.
- M. Last, Y. Klein, A. Kandel, “Knowledge Discovery in Time Series Databases”, IEEE Transactions on Systems, Man, and Cybernetics, Volume 31: Part B, No. 1, pp. 160-169, Feb. 2001.
- M. Last, A. Kandel, O. Maimon, “Information-Theoretic Algorithm for Feature Selection”, Pattern Recognition Letters, 22 (6-7), pp. 799-811, 2001.
- M. Last, “Online Classification of Nonstationary Data Streams”, Intelligent Data Analysis, Vol. 6, No. 2, pp. 129-147, 2002.
- M. Last, O. Maimon, E. Minkov, “Improving Stability of Decision Trees”, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16, No. 2, pp. 145-159, 2002.
- M. Last and O. Maimon, “A Compact and Accurate Model for Classification”, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 2, pp. 203-215, February 2004.

Summary

- Info-Fuzzy Network is constructed by repetitive partitioning of all training instances in every layer
- Each network layer is uniquely related to a single input attribute
- No testing set is needed to estimate the maximum achievable accuracy of a given network
- The IFN algorithm produces much more compact models than C4.5
  - Recommended when interpretability overweighs accuracy!

IFN SOFTWARE



Location: Moodle