KTH ROYAL INSTITUTE
OF TECHNOLOGY

Lecture 13
Machine Learning. Decision trees

On the Shoulder of Giants

Much of the material in this slide set is based upon:

”Automated Learning techniques in Power Systems” by L. Wehenkel, Université Liege

”Probability based learning” Josephine Sullivan, KTH

”Entropy and Information Gain” by F.Aiolli, University of Padova

Contents

Repeating from last time

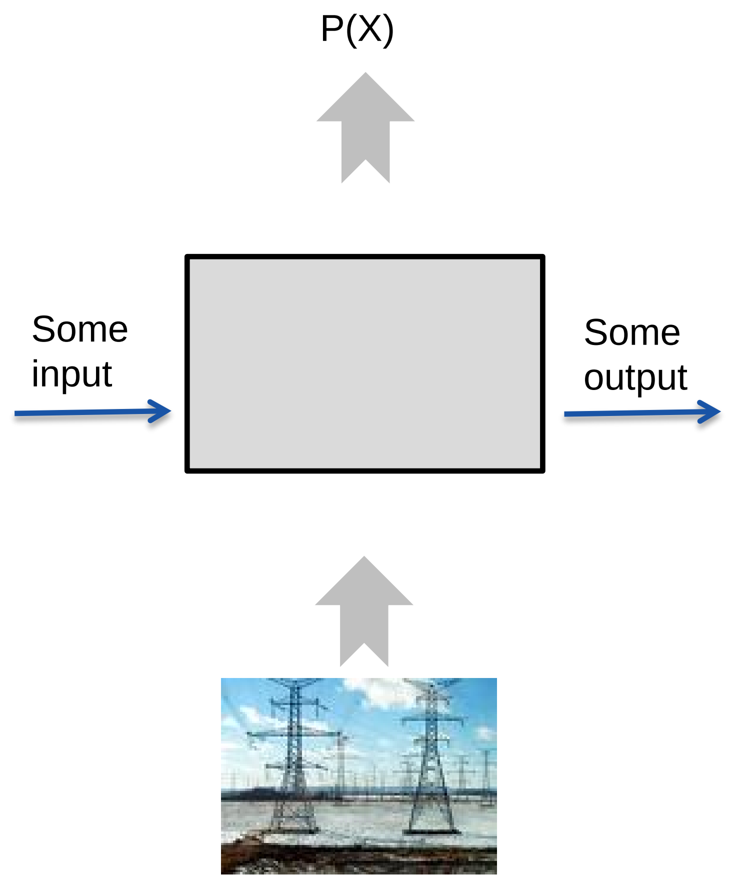
Decision Trees

Hands-On

Power Systems Analysis – An automated learning approach

Understanding states in the power system is established through observation of inputs and outputs without regard to the physical electrotechnical relations between the states.

Adding knowledge about the electrotechnical rules means adding heuristics to the learning.



Given a set of examples (the learning set (LS)) of associated input/output pairs, derive a general rule representing the underlying input/output relationship, which may be used to explain the observed pairs and/or predict output values for any new unseen input.

Classes of methods for learning

In **Supervised** learning a set of input data and output data is provided, and with the help of these two datasets the model of the system is created.

For this introductory course, our focus is here With a short look at unsupervised learning

In **Unsupervised** learning, no ideal model is anticipated, but instead the analysis of the states is done in order to identify possible correlations bewteen datapoints.

In **Reinforced** learning, the model in the system can be gradually refined through means of a utility function, that tells the system that a certain ouput is more suitable than another.

Classification vs Regression

Two forms of Supervised learning

Classification: The input data is number of switch operations a circuitbreaker has performed and tthe output is a notification whether the switch needs maintenance or not. ”Boolean”

Regression: Given the wind speed in a incoming weather front, the output is the anticipated production in a set of wind turbines. ”Floating point”

Supervised learning - a preview

In the scope of this course, we will be studying three forms of supervised learning.

- Decision Trees

  Overview and practical work on exercise session.

- Artificial Neural Networks

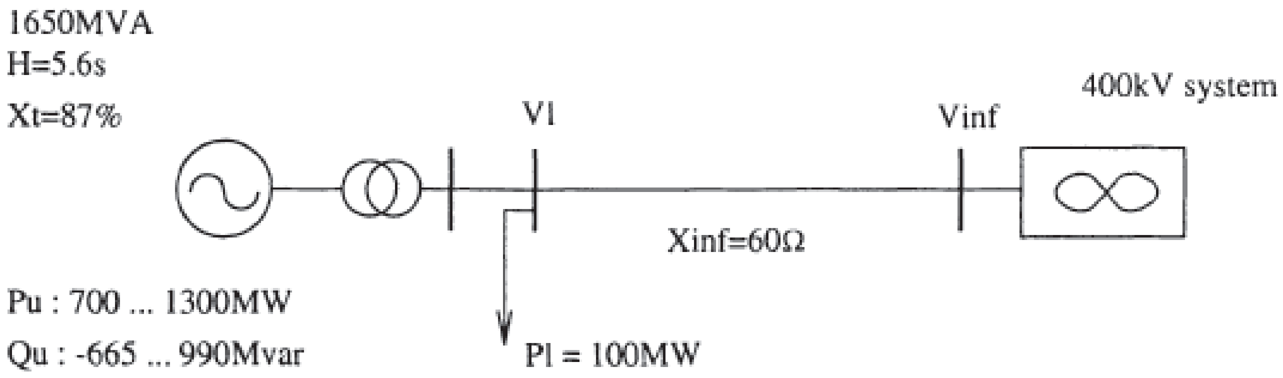
  Overview only, no practical work.

- Statistical methods – k-Nearest Neighbour

  Overview and practical work on exercise session. Also included in Project Assignment

  kNN algorithm can also be used for <u>unsupervised</u> clustering.

Example from Automatic Learning techniques in Power Systems



One Machine Infinite Bus (OMIB) system

- Assuming a fault close to the Generator will be cleared within 155 ms by protection relays
- We need to identify situations in which this clearing time is sufficient and when it is not
- Under certain loading situations, 155 ms may be too slow.

Source: Automatic Learning techniques in Power Systems, L. Wehenkel

OMIB – further information

In the OMIB system the following parameters influence security

- Amount of active and reactive power of the generator (
- Amount of load nearby the generator (Pl)
- Voltage magnitudes at the load bus and at the infinite bus Short-circuit reactance Xinf, representing the effect of variable topology in the large system represented by the infinite bus.

In the example, Voltages at generator and Infinite bus are assumed similar and constant for simplicity

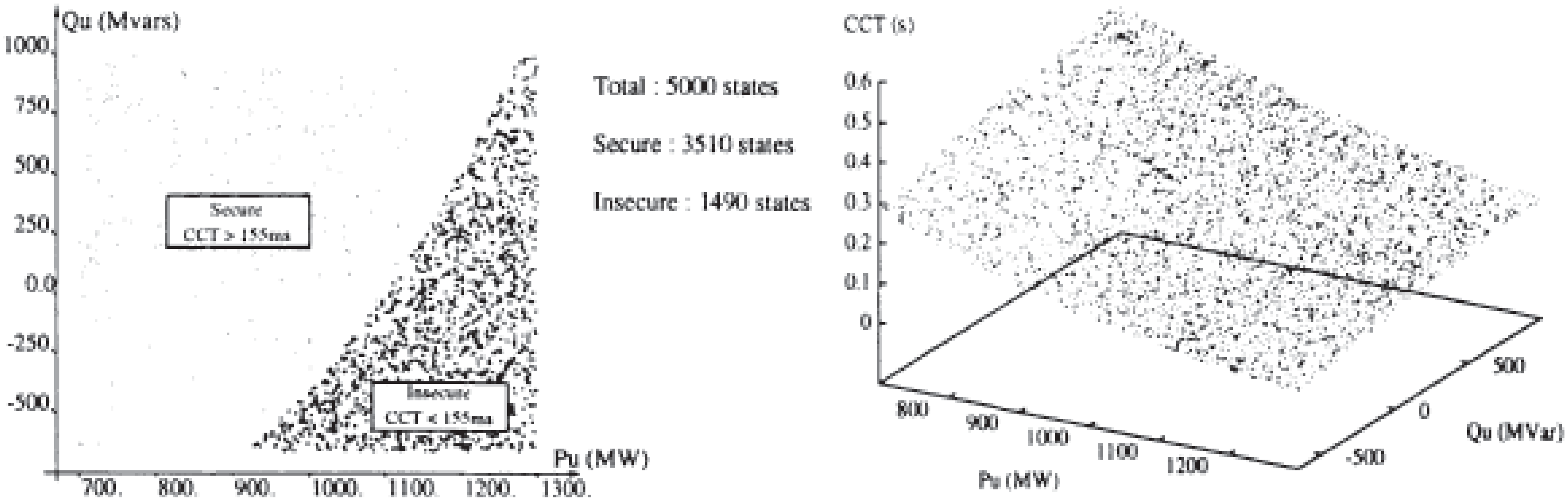
Our database of objects with attributes

In a simulator, we randomly sample values for P_u and Q_u creating a database with 5000 samples (objects) and for each object we have a set of attributes $(P_u, Q_u, V_1, P_1, V_{inf}, X_{inf}, CCT)$ as per below.

Table 1.1. Sample of OMIB operating states

State Nb	Pu (MW)	Qu (MVAr)	V1 (p.u.)	P1 (MW)	Vinf (p.u.)	Xinf (Ω)	CCT (s)
1	876.0	-193.7	1.05	-100	1.05	60	0.236
2	1110.9	-423.2	1.05	-100	1.05	60	0.112
3	980.1	79.7	1.05	-100	1.05	60	0.210
4	974.1	217.1	1.05	-100	1.05	60	0.224
5	927.2	-618.5	1.05	-100	1.05	60	0.158
...	...	...	...	...	...	...	...
2276	1090.4	-31.3	1.05	-100	1.05	60	0.157
...	...	...	...	...	...	...	...
4984	1090.2	-20.0	1.05	-100	1.05	60	0.158
...	...	...	...	...	...	...	...

Plot of database content



Source: *Automatic Learning techniques in Power Systems, L. Wehenkel*

How to measure information content

Entropy **H** is a measure of *Unpredictability*.

Defined as:

$$\square \; -\sum p_i \log p_i$$

Where

p_i is the probability of event i

Some examples

1. Flipping a coin
2. Rolling a 6 sided die
3. Rolling a loaded 6 sided die

Entropy in a Dataset

$$H_C(\boldsymbol{X}) \triangleq -\sum_{i=1,\ldots,m} P(\boldsymbol{C}_i|\boldsymbol{X}) \log P(\boldsymbol{C}_i|\boldsymbol{X}),$$

The classification Entropy, is the entropy related to the probability of a value x belonging to a class Ci

Or simply put, how difficult is it to guess which partition of U, i.e. Class Ci that an object o belongs to.

An example of classification entropy

Color	Size	Shape	Eadible?
Yellow	Small	Round	Yes
Yellow	Small	Round	No
Green	Small	Irregular	Yes
Green	Large	Irregular	No
Yellow	Large	Round	Yes
Yellow	Small	Round	Yes
Yellow	Small	Round	Yes
Yellow	Small	Round	Yes
Green	Small	Round	No
Yellow	Large	Round	No
Yellow	Large	Round	Yes
Yellow	Large	Round	No
Yellow	Large	Round	No
Yellow	Large	Round	No
Yellow	Small	Irregular	Yes
Yellow	Large	Irregular	Yes

Source: F. Aiolli - Sistemi Informativi University of Padova

Entropy example

Entropy for the example data set is calculated as:

$$I(all\_data) = -\left[\left(\frac{9}{16}\right)\log_2\left(\frac{9}{16}\right) + \left(\frac{7}{16}\right)\log_2\left(\frac{7}{16}\right)\right]$$

Giving: 0,9836

Is this reasonable?

Information Gain

The reduction in Entropy achieved by partitioning the dataset differently.

Lets separate for instance per the attribute Size.

Size

Small

Color	Size	Shape	Edible?
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Small	Irregular	+

Large

Color	Size	Shape	Edible?
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Irregular	+

Source: F. Aiolli - Sistemi Informativi University of Padova

Information Gain calculation

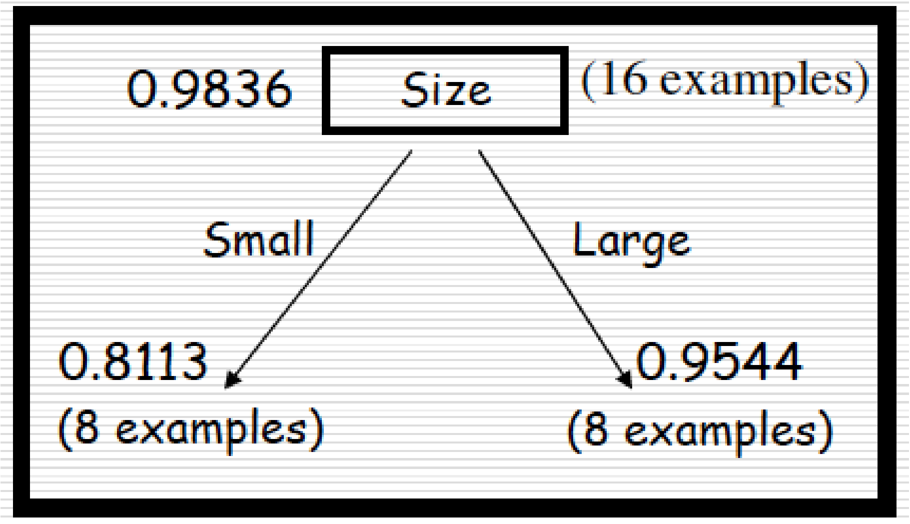
The two partitions has their own entropy value.

We can calculate for each possible attribute its expected entropy. This is the degree to which the entropy would change if partitioned based on this attribute.

To determine resulting entropy, you add the entropies of the two partitions, weighted by the proportion of examples from the parent node that ended up in that partition.

$$G(S,A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

Hence for the example



Entropy of left child is <u>0.8113</u>

$I(size=small) = 0.8113$

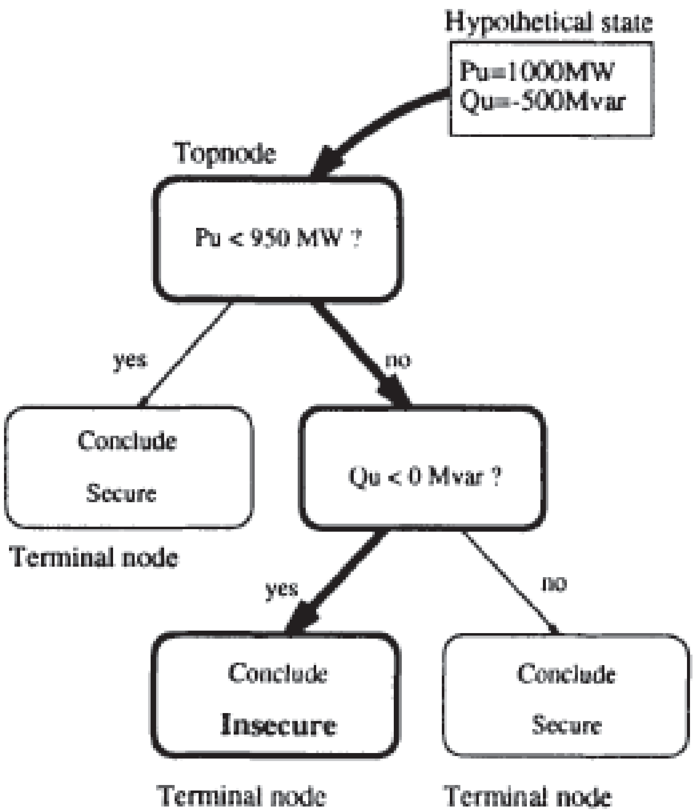
Entropy of right child is <u>0.9544</u>

$I(size=large) = 0.9544$

$$I(S_{Size}) = (8/16)^{*}.8113 + (8/16)^{*}.9544 = .8828$$

Contents

Back to our Power System example



Perhaps we can partition our dataset according to some attribute?

Lets try $P_u < 950$MW

Equivalent If-Then rules :

Rule 1 : If (Pu < 950MW) then Conclude Secure

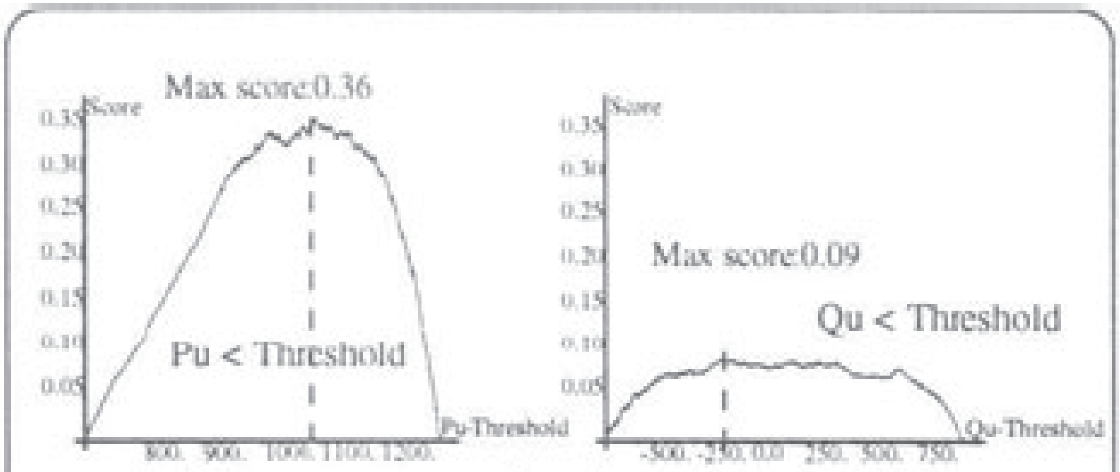
Rule 2 : If (Pu > 950MW) and (Qu < 0Mvar) then Conclude Insecure

Rule 3 : If (Pu > 950MW) and (Qu > 0Mvar) then Conclude Secure

Finding best partition.

Starting with the candidate attributes (Pu and Qu) in our case

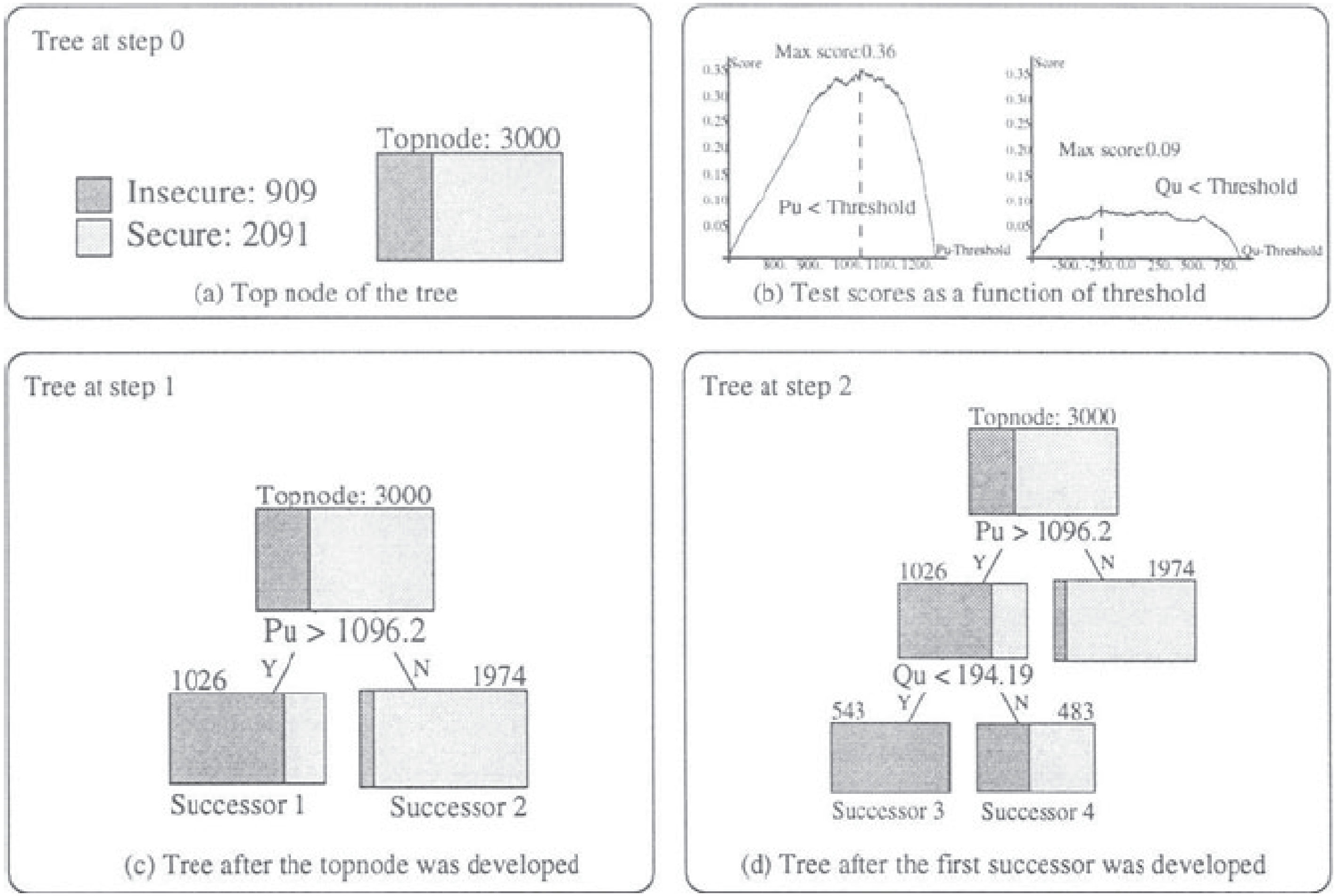
We check chich of the values for Pu and Qu that create the most valuable partition in terms of information gain.



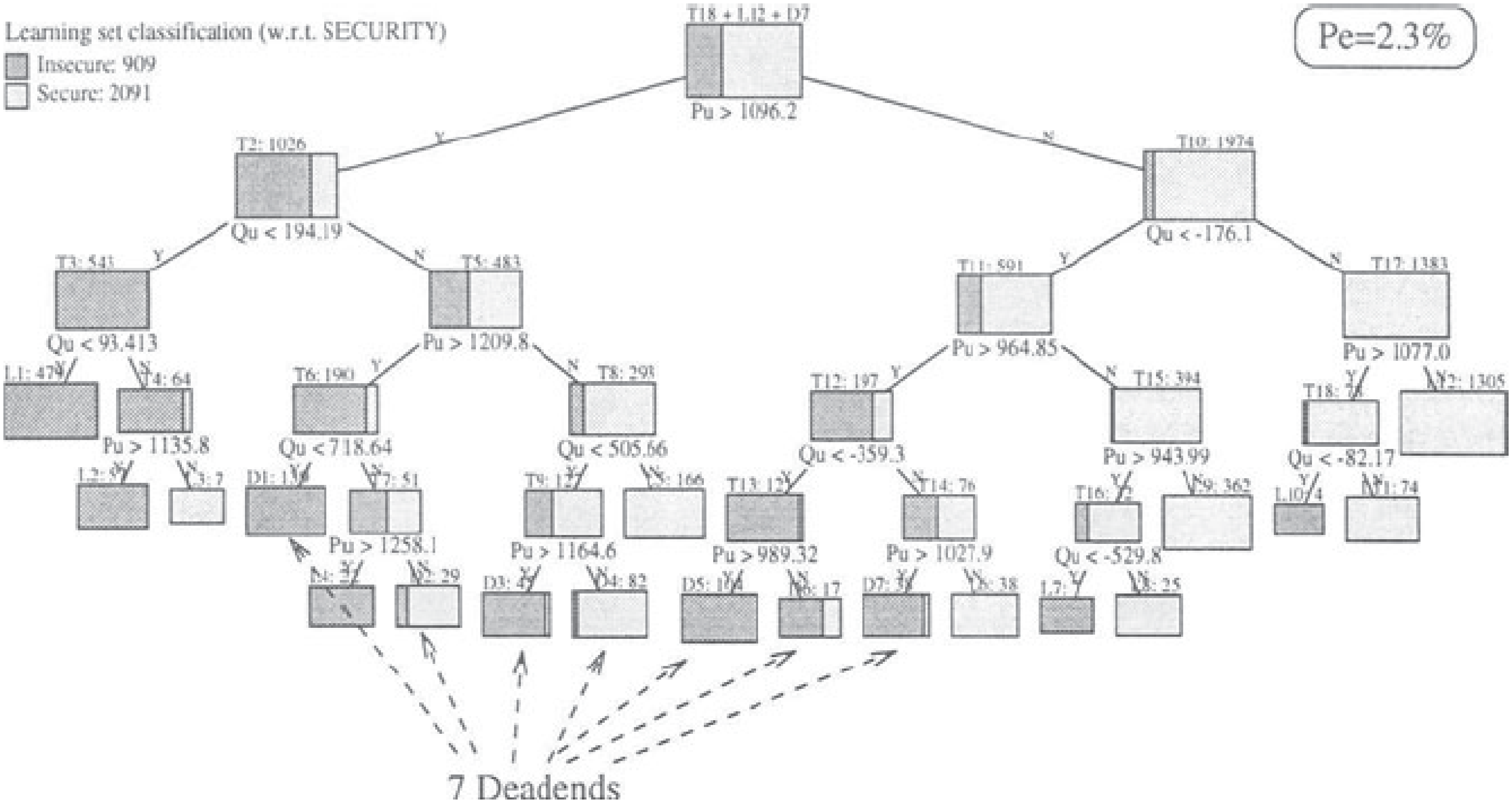
(b) Test scores as a function of threshold

Pu > 1096,2 MW is the best partition

Gradual expansion of the Decision Tree



Complete Decision Tree



How to stop?

The splitting of data sets continues until either:

A perfect partition is reached – i.e. One which perfecly explains the content of the class – a *leaf*

One where no infomration is gained no matter how the data set is split. – a *deadend*.

Validation of the Decision Tree

By using the Test Set (2000 samples) we can calidate the Decision tree.

By testing for each Object in the Test Set, we determine if the Decision tree provides the right answer for the Object.

In this particular example, the probability o error can be determined to 2,3. I.e. Of the 2000 samples 46 were classififed to the wrong class.