Hugin & Weka for Learning from Data

Using Bayesian Networks

Ashwin Patthi
University of South Carolina

Learning from Data?

Predicting Probability distribution of unknown variables from data/cases

Resulting learning model can be used as classifier

Knowledge discovery

Classification Task

Classifying variable y = x0 called the class variable given set of variables X=x1,x2...xk

called attribute variables

Classifier is learned from a dataset D consisting of samples over (x,y) on network Bx over probability distribution U

h: X --> y

Learning using Hugin

Uses EM (Estimation Maximization) Algorithm (Batch learning)

This is used only when structure is available

Experience table must be provided for nodes whose conditional probabilities are to inferred

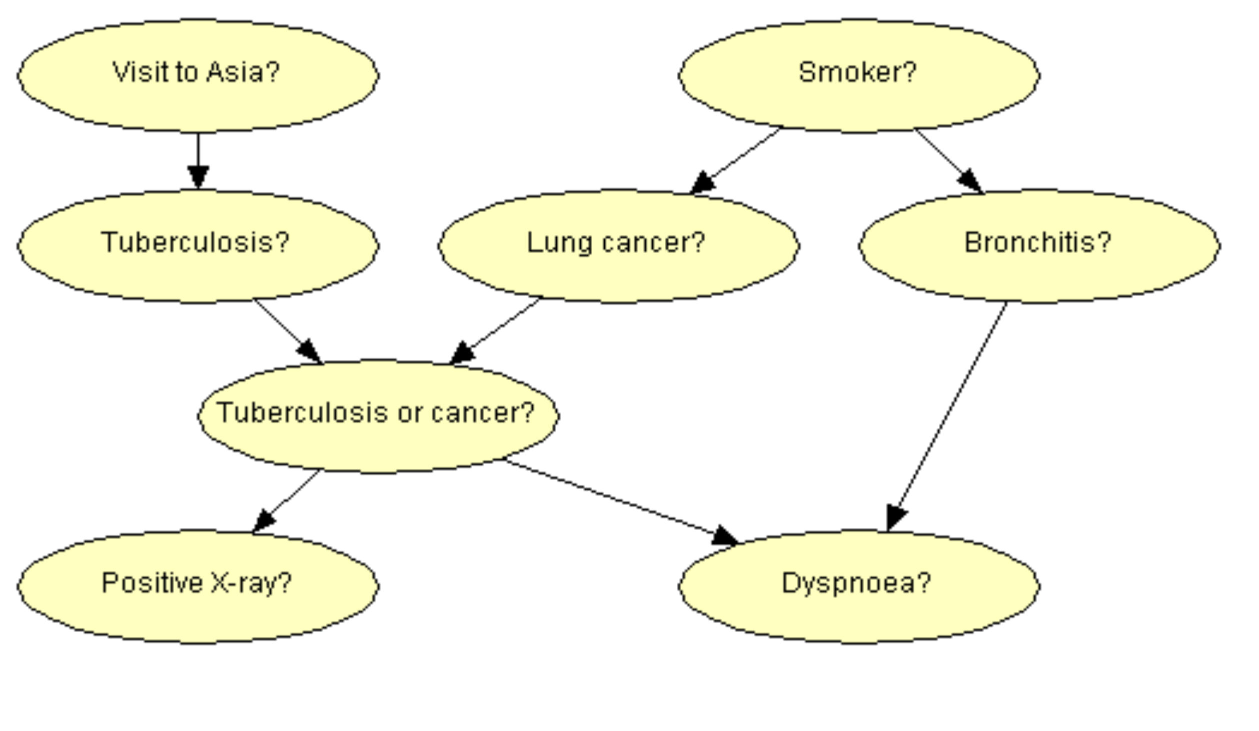
EM Algorithm

Performs number of iteration on cases

Computes log-likelihood and attempts to maximize it

Stops when two successive log-likelihood is less than tolerance

Learning on Chest Clinic data

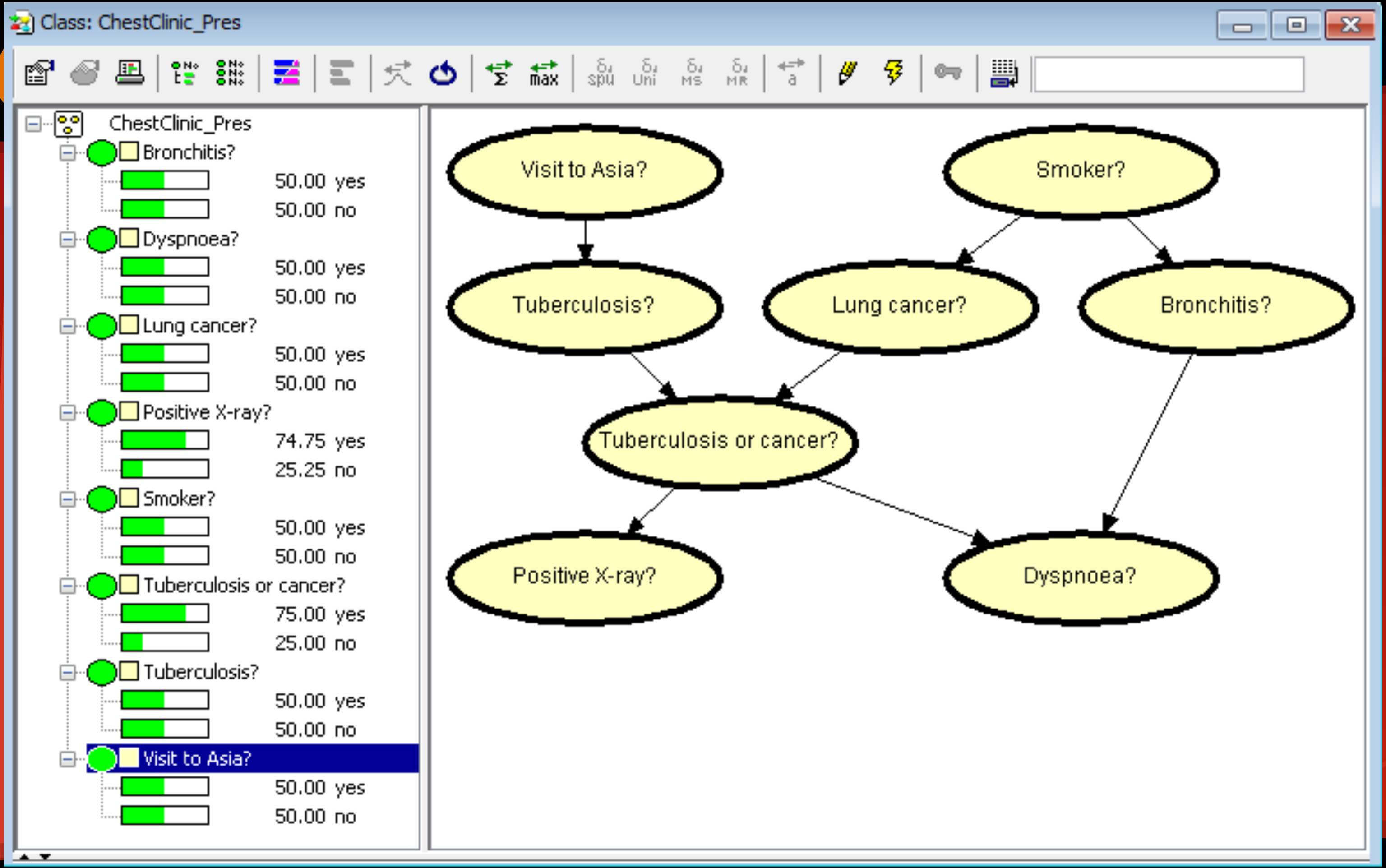


Assumptions

No prior knowledge on any distribution

Set probability distribution to 1 except for "Tuberculosis or Cancer"

Also experience value to 10 (or some low value)



Class: ChestClinic_Pres

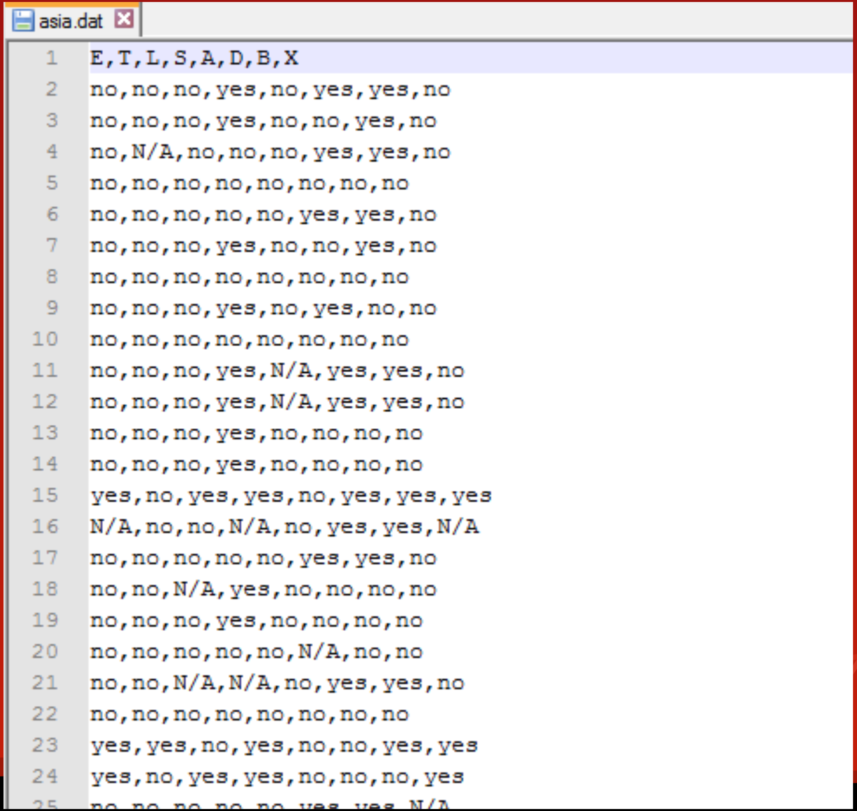
```
ChestClinic_Pres
  Bronchitis?
      50.00 yes
      50.00 no
  Dyspnoea?
      50.00 yes
      50.00 no
  Lung cancer?
      50.00 yes
      50.00 no
  Positive X-ray?
      74.75 yes
      25.25 no
  Smoker?
      50.00 yes
      50.00 no
  Tuberculosis or cancer?
      75.00 yes
      25.00 no
  Tuberculosis?
      50.00 yes
      50.00 no
  Visit to Asia?
      50.00 yes
      50.00 no
```

Data set (asia.data)

First line is header

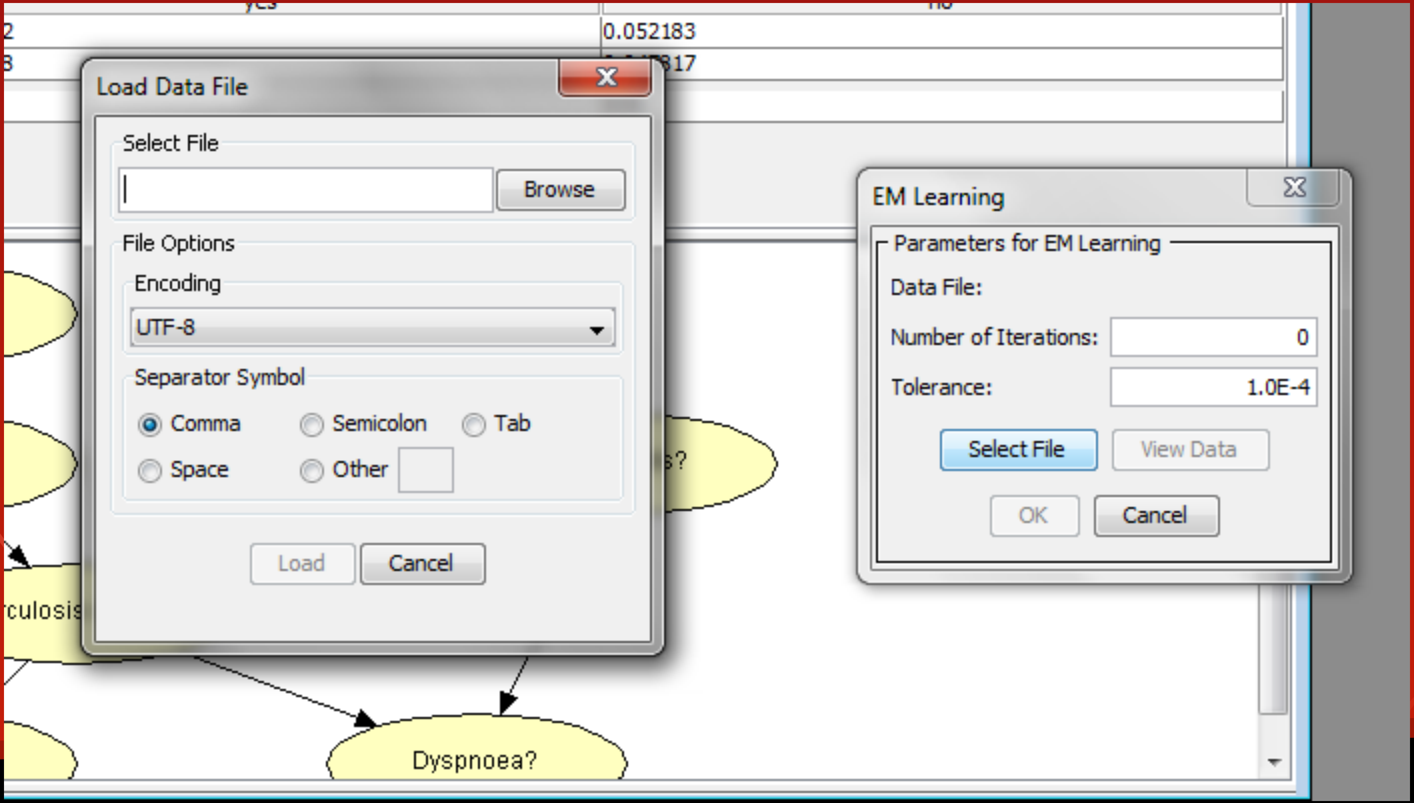
Each record is a case

N/A = Not available



```
E,T,L,S,A,D,B,X
no,no,no,yes,no,yes,yes,no
no,no,no,yes,no,no,yes,no
no,N/A,no,no,no,yes,yes,no
no,no,no,no,no,no,no,no
no,no,no,no,no,yes,yes,no
no,no,no,yes,no,no,yes,no
no,no,no,no,no,no,no,no
no,no,no,yes,no,yes,no,no
no,no,no,no,no,no,no,no
no,no,no,yes,N/A,yes,yes,no
no,no,no,yes,N/A,yes,yes,no
no,no,no,yes,no,no,no,no
no,no,no,yes,no,no,no,no
yes,no,yes,yes,no,yes,yes,yes
N/A,no,no,N/A,no,yes,yes,N/A
no,no,no,no,no,yes,yes,no
no,no,N/A,yes,no,no,no,no
no,no,no,yes,no,no,no,no
no,no,no,no,no,N/A,no,no
no,no,N/A,N/A,no,yes,yes,no
no,no,no,no,no,no,no,no
yes,yes,no,yes,no,no,yes,yes
yes,no,yes,yes,no,no,no,yes
no,no,no,no,no,yes,yes,N/A
```

Run the Learning Algo



Load Data File

Select File

File Options

Encoding: UTF-8

Separator Symbol: ● Comma ○ Semicolon ○ Tab ○ Space ○ Other

EM Learning

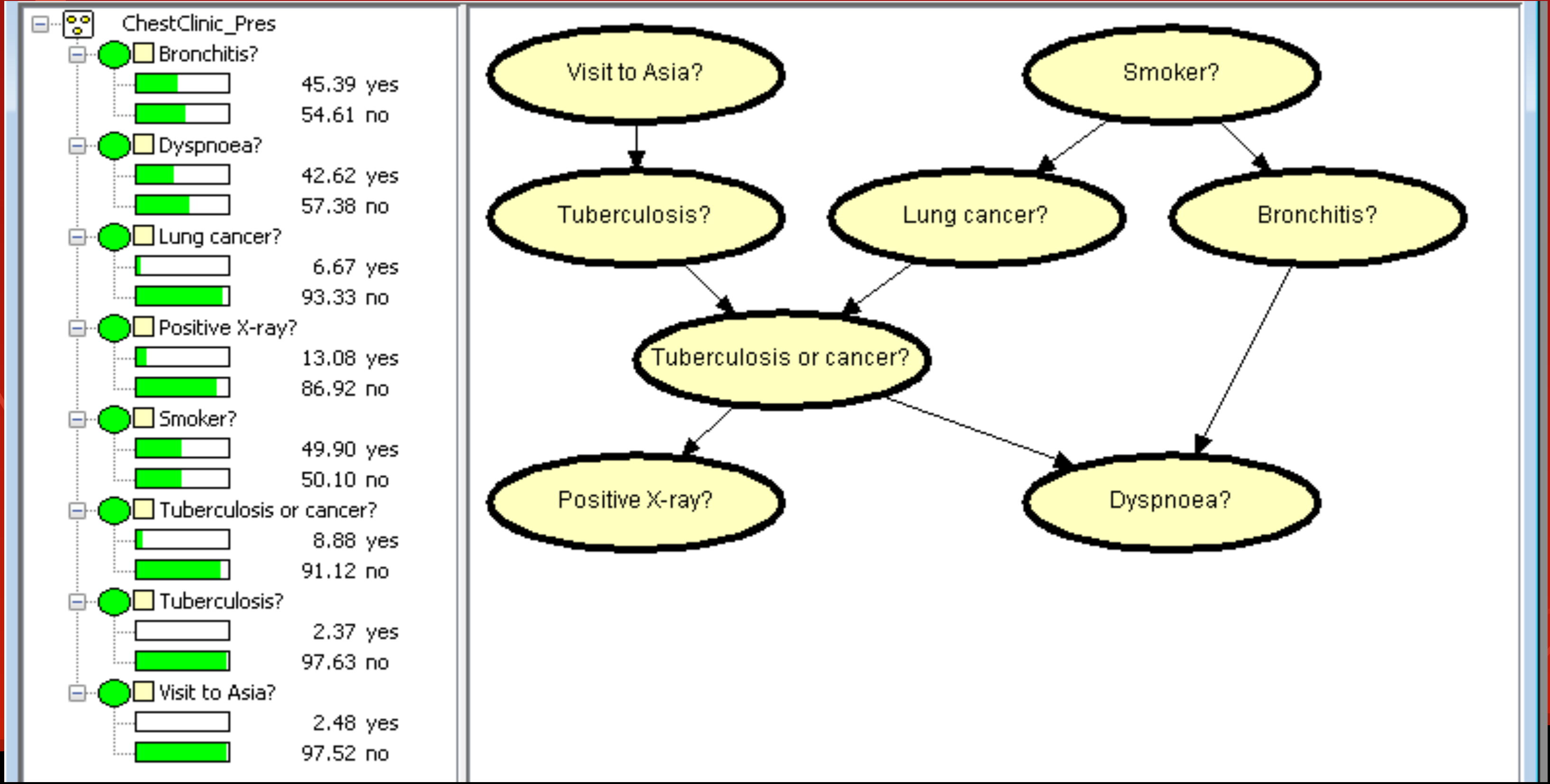
Parameters for EM Learning

Data File:

Number of Iterations: 0

Tolerance: 1.0E-4

Resulting Marginal Probabilities



```
ChestClinic_Pres
  Bronchitis?
      45.39 yes
      54.61 no
  Dyspnoea?
      42.62 yes
      57.38 no
  Lung cancer?
       6.67 yes
      93.33 no
  Positive X-ray?
      13.08 yes
      86.92 no
  Smoker?
      49.90 yes
      50.10 no
  Tuberculosis or cancer?
       8.88 yes
      91.12 no
  Tuberculosis?
       2.37 yes
      97.63 no
  Visit to Asia?
       2.48 yes
      97.52 no
```

Classifier

The resulting model with computed conditional probabilities can be used as classifier to predict the new unknown conditional probabilities

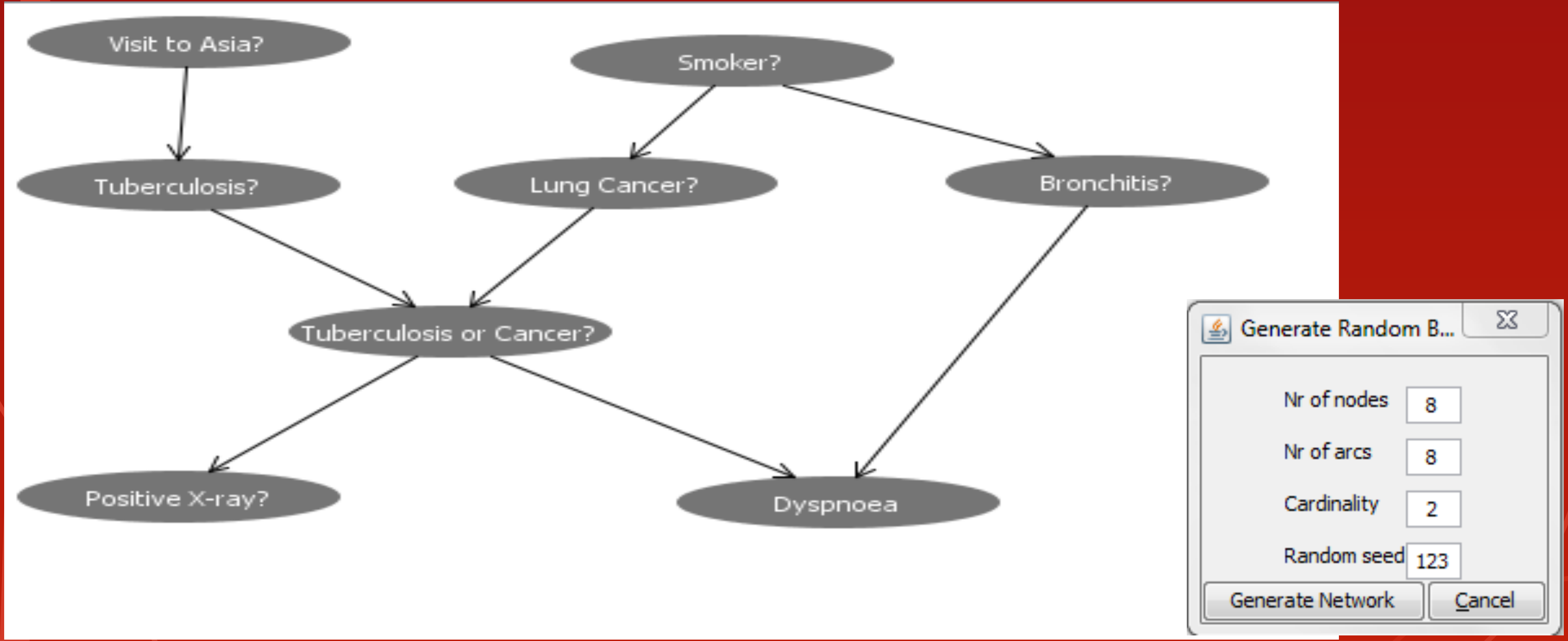
Learning using WEKA

Collection of Machine Learning Algorithms for data mining tasks

Contains tools for data pre-processing, classification, clustering, visualization etc.

Bayesian Network classifier and editor is one of them

Creating Bayesian Network



Nr of nodes 8

Nr of arcs 8

Cardinality 2

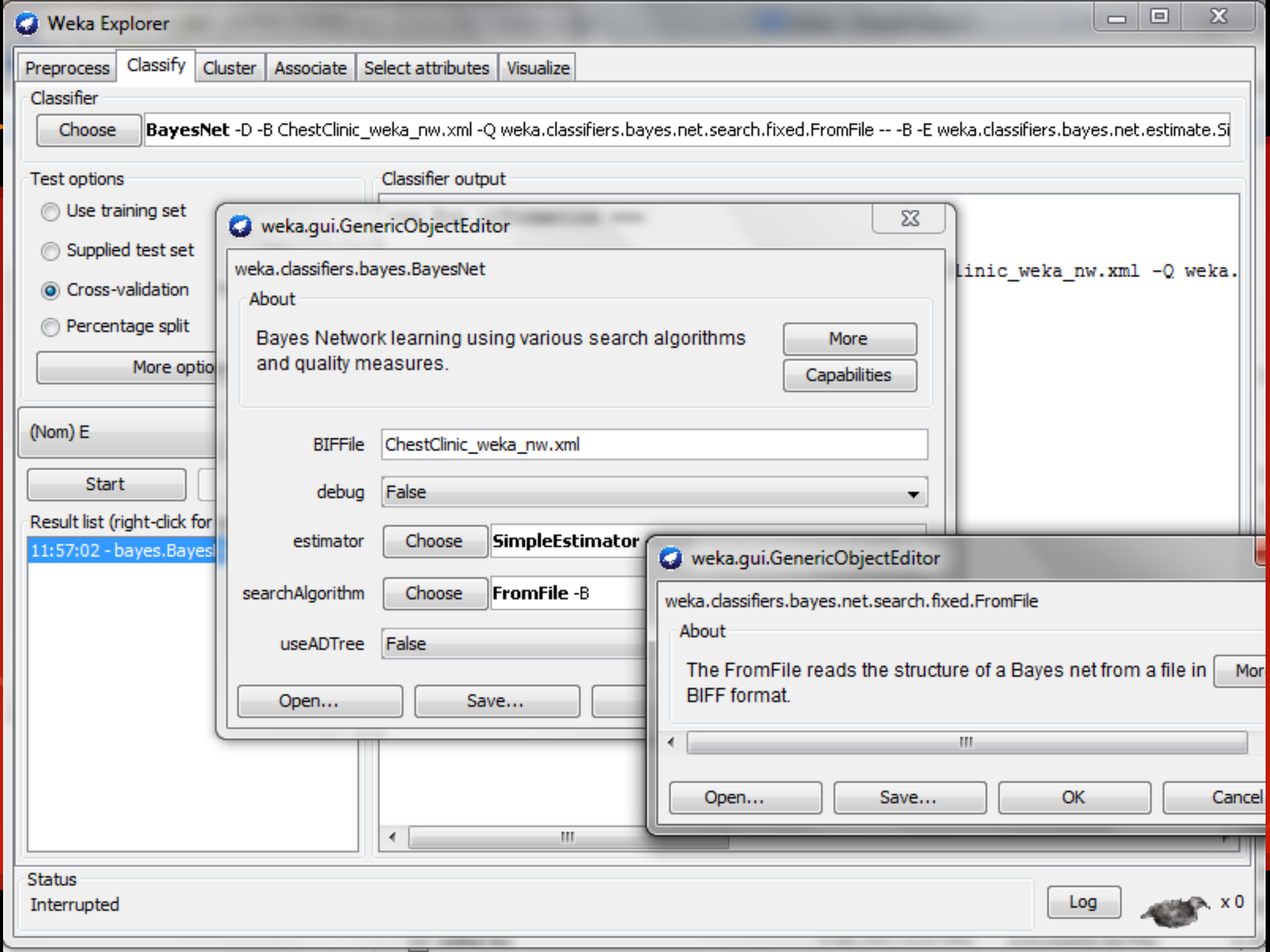
Random seed 123

Basic Assumptions

All variables are discrete finite variables

If continuous can convert to discrete using class filters.attribute.Discretize

No instances have missing values. If found can be filled by attribute.ReplaceMissingValues



BayesNet -D -B ChestClinic_weka_nw.xml -Q weka.classifiers.bayes.net.search.fixed.FromFile -- -B -E weka.classifiers.bayes.net.estimate.Si

○ Use training set
○ Supplied test set
● Cross-validation
○ Percentage split

(Nom) E

11:57:02 - bayes.Bayes

```
linic_weka_nw.xml -Q weka.
```

Status: Interrupted

weka.gui.GenericObjectEditor

weka.classifiers.bayes.BayesNet

About

Bayes Network learning using various search algorithms and quality measures.

BIFFile: ChestClinic_weka_nw.xml
debug: False
estimator: Choose SimpleEstimator
searchAlgorithm: Choose FromFile -B
useADTree: False

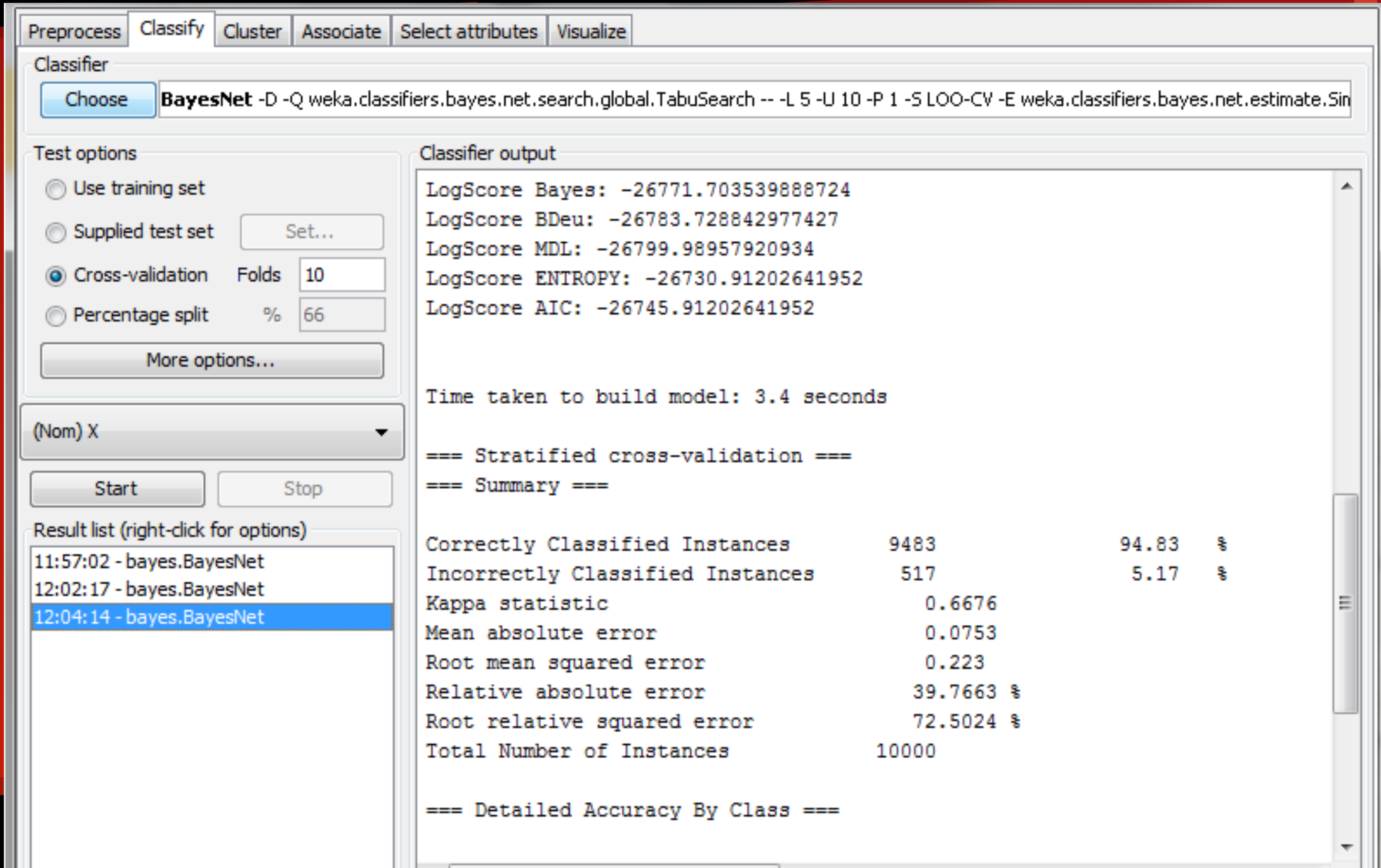
weka.gui.GenericObjectEditor

weka.classifiers.bayes.net.search.fixed.FromFile

About

The FromFile reads the structure of a Bayes net from a file in BIFF format.

Results



Classifier

BayesNet -D -Q weka.classifiers.bayes.net.search.global.TabuSearch -- -L 5 -U 10 -P 1 -S LOO-CV -E weka.classifiers.bayes.net.estimate.Sin

Test options

- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) X

Result list (right-click for options)

- 11:57:02 - bayes.BayesNet
- 12:02:17 - bayes.BayesNet
- 12:04:14 - bayes.BayesNet

Classifier output

```
LogScore Bayes: -26771.703539888724
LogScore BDeu: -26783.728842977427
LogScore MDL: -26799.98957920934
LogScore ENTROPY: -26730.91202641952
LogScore AIC: -26745.91202641952


Time taken to build model: 3.4 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        9483               94.83   %
Incorrectly Classified Instances       517                5.17   %
Kappa statistic                          0.6676
Mean absolute error                      0.0753
Root mean squared error                  0.223
Relative absolute error                 39.7663 %
Root relative squared error             72.5024 %
Total Number of Instances            10000

=== Detailed Accuracy By Class ===
```

Pros & Cons of Hugin

Pros:

Can generate data with missing values based on available network

Can predict prob. distribution for all variables

Cons:

No feature for handling missing values

No feature to predict class of all the cases at once

Pros and Cons of WEKA

Pros:

Feature to handle missing values

Can predict class of unknown data sets all at once

Cons:

Can’t generate data with missing values

Can’t predict probability distribution

Useful Links & References

1. Bouckaert, Remco R. Bayesian network classifiers in weka. Department of Computer Science, University of Waikato, 2004
2. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1
3. More information on Hugin tool
4. More information on WEKA tool

Questions ?