دانشگاه آزاد اسلامی واحد تبریز

نام درس: داده کاوی

بخش: شروع داده کاوی

نام استاد: دکتر مسعود کارگر

Roadmap

- Data mining problems and data
- K-nearest Neighbor Classifier (KNN)
- Classification by K-nearest neighbor
- Weka System
- Discussion of the assignment 1

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Frequent Subgraph mining [Descriptive]
- …

Where do we get data? How do we evaluate?

Data Mining Data Repositories

- UCI Machine Learning Repositories
  - http://archive.ics.uci.edu/ml/datasets.html
  - 173 Data sets classified into Classification (114), Regression (12), Clustering (5), Other (44)
  - Also can be classified by
    - Attribute type,
    - Data Type,
    - Application Area
    - # of Attributes
    - # of Instances (Examples)
    - Format Type (Matrix or Non-Matrix)
- Data mining Competition

UCI MLR Data set example 1

- Abalone dataset
- The problem: Predicting the age of abalone from physical measurements. (x1, x2, …, x8) -→ Age?
  - The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task
  - Other measurements, which are easier to obtain, are used to predict the age
- No. of instances (4177)
- 8 Attribues/Features (Measurements)
  - Sex, Length, Diameter, Height, Wholeweight, Shucked weight, Viscera weight, Shell Weight, Rings

Attribute Properties

Sex / nominal / -- / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer / -- / +1.5 gives the age in years

Data examples

M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15
M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7
F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9
M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10
I,0.33,0.255,0.08,0.205,0.0895,0.0395,0.055,7
I,0.425,0.3,0.095,0.3515,0.141,0.0775,0.12,8
F,0.53,0.415,0.15,0.7775,0.237,0.1415,0.33,20
F,0.545,0.425,0.125,0.768,0.294,0.1495,0.26,16

How do we predict the rings based on the first 8 features?

Dataset example 2

- KDD Cup 2001 Competition dataset
- Problem: Prediction of Molecular Bioactivity for Drug Design -- Binding to Thrombin
- **The present training data set consists of 1909 compounds tested for their ability to bind to a target site on thrombin, a key receptor (protein) in blood clotting**
- **Of these compounds, 42 are active (bind well) and the others are inactive.**
- **Each compound is described by a single feature vector comprised of a class value (A for active, I for inactive) and 139,351 binary features, which describe three-dimensional properties of the molecule**

Summary of a data mining problem

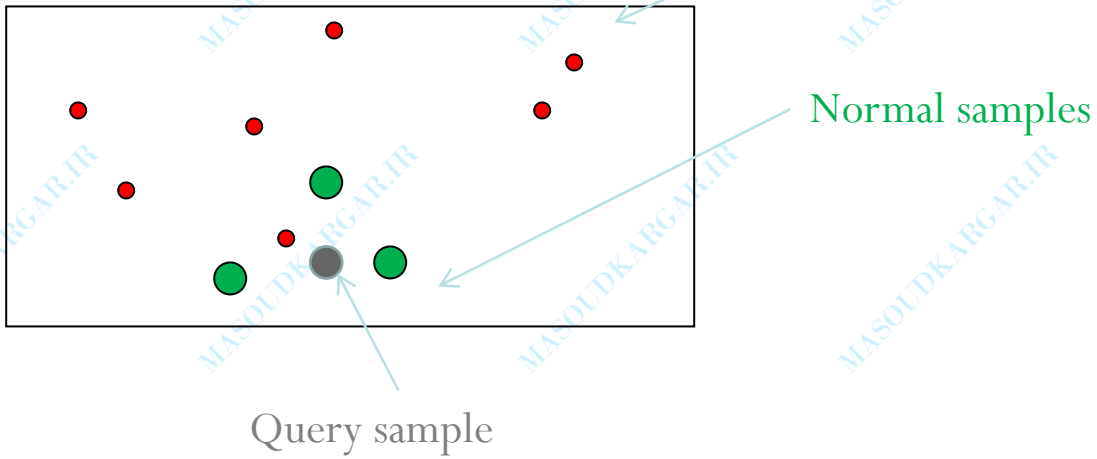
- Objective
  - Prediction of molecular bioactivity for drug design -- binding to Thrombin
- Data
  - Training: 1909 cases (42 positive), 139,351 binary features
  - Test: 634 cases
- Challenge
  - Highly imbalanced, high-dimensional, different distribution
- Approaches
  - Winners' Bayesian network predictive model

Roadmap

- Data mining problems and data
- K-nearest Neighbor Classifier (KNN)
- Advanced KNN classifiers
- Weka System
- Discussion of the assignment 1

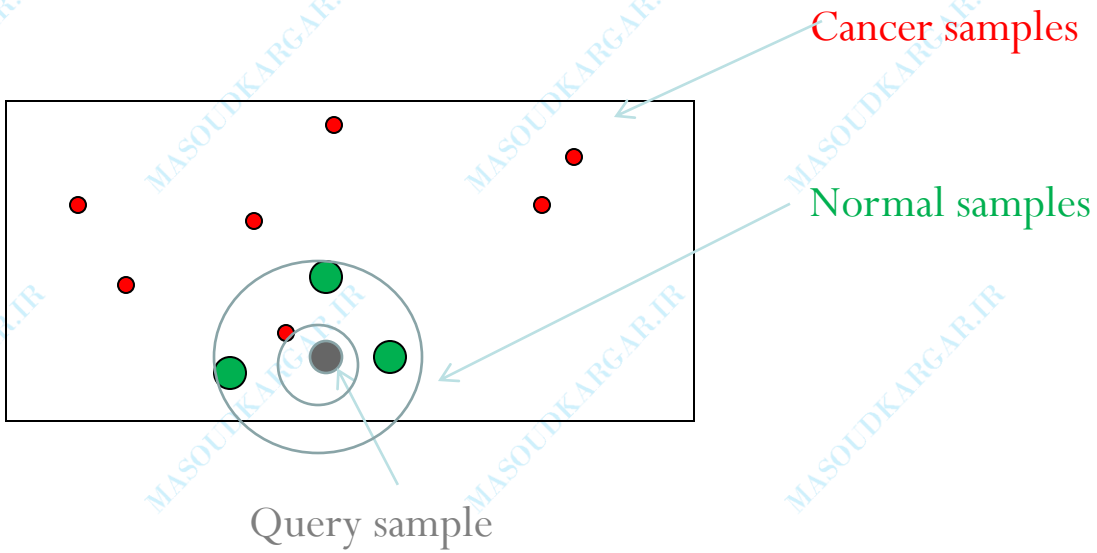
The Simplest Classifier: 1-NN

- Each sample is just a vector of numbers, each number is the expression level of a gene
  - (0.23, 0.34, 0.89, 0.99 …, 0.3435, 0) →Cancer (label)
  - (0.24, 0.33, 0.12, 0.56,…, 0.2, 0.5, 0.4)→ Normal (label)
- Given a test sample
  - Find the nearest (most similar) training sample, and predict its label as the class label for the test sample!



Build a Better KNN Classifier

- 1-NN is sensitive to outliers!
- K(4)-NN with voting works better
- Questions: How to determine optimal K?



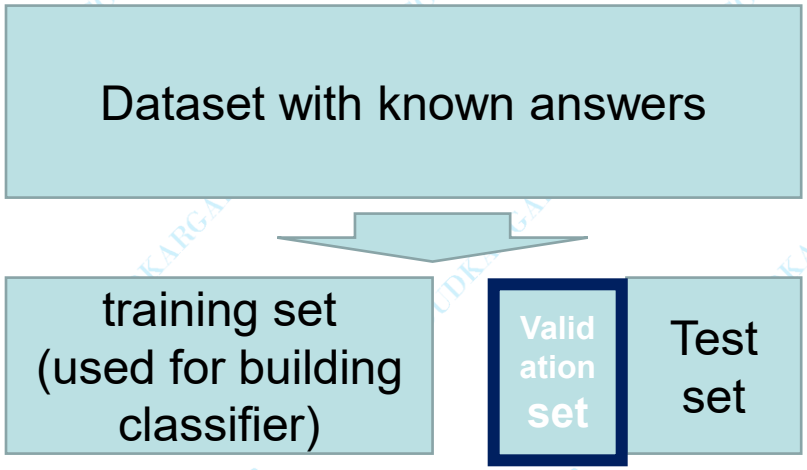
Implement KNN algorithm

- Algorithm Summary:
  - Step *1*: define your distance or similarity measure of two samples (instances)
  - Step 2: determine k (usually odd for easy voting)
  - Step 3: calculate the distances between the new input and all the training data
  - Step 4: sort the distance and determine k-nearest neighbors based on the k-th minimum distance.
  - Step 5: gather the class labels of those neighbors.
  - Step 6: determine the prediction label based on majority vote

How to Evaluate KNN Classifier

- Evaluation by Tests (exams!)
- To score the tests, you need to know the Correct Answers
- Two types of tests:
  - Tests on training set: 1-KNN should obtain 100% accuracy.
  - Tests on validation set (unseen examples by the classifier) to tune your parameters. E.g. to determine optimal K.
  - Tests on test set
- Define Accuracy:

$$\frac{\#\,correctly\ predicted\ labels}{total\ \#\ of\ testing\ samples}$$



Advanced KNN

- http://www.visionbib.com/bibliography/pattern621.html
- Speed-up by efficient indexing
  - JAGADISH, et al. 2005. iDistance: An Adaptive B+-Tree Based Indexing Method for Nearest Neighbor Search. ACM Transactions on Database Systems (TODS) ,30(2), 2005
  - Bin Zhang,Srihari, S.N. Fast k-nearest neighbor classification using cluster-based trees.IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004
- Learning a new metric/distance for KNN classification
- Bayesian KNN classification
- Zhan et al. 2004. Privacy Preserving K-nearest neighbor classification. International Journal of Network Security, Vol.1, No.1, PP.46–51, July 2005

Tricks to Tune KNN classifier

- Only consider partial attributes (when calculate distances, skip unimportant attributes)
- Determine the importance of attributes by comparing the distribution of its values for positive and negative classes
- Define different distance measures.

Roadmap

- Data mining problems and data
- K-nearest Neighbor Classifier (KNN)
- Advanced KNN classifiers
- Weka System
- Discussion of the homework1

WEKA:: Introduction

- http://www.cs.waikato.ac.nz/ml/weka/
- A collection of open source ML algorithms
  - pre-processing: feature selection, normalization, etc
  - classifiers
  - clustering
  - association rule
- Created by researchers at the University of Waikato in New Zealand
- Java based

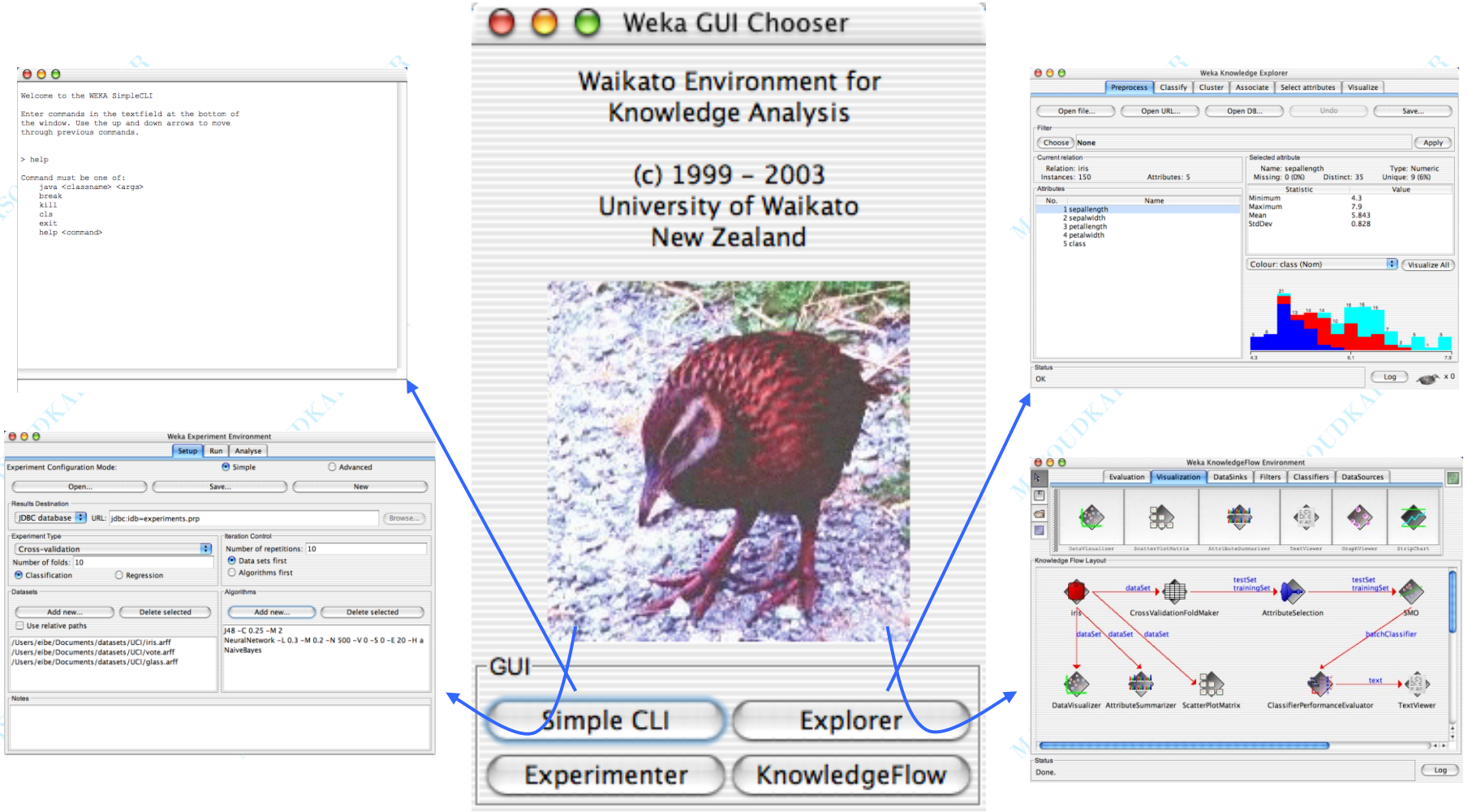
WEKA:: Installation

- Platform-independent: windows/Linux/Mac all ok!
- Download software from http://www.cs.waikato.ac.nz/ml/weka/
  - If you are interested in modifying/extending weka there is a developer version that includes the source code
- Set the weka environment variable for java
  - `setenv WEKAHOME /usr/local/weka/weka-3-4-13`
  - `setenv CLASSPATH $WEKAHOME/weka.jar:$CLASSPATH`
- Download some ML data from http://mlearn.ics.uci.edu/MLRepository.html

WEKA:: Introduction .contd

- Routines are implemented as classes and logically arranged in packages
- Comes with an extensive GUI interface
  - Weka routines can be used stand alone via the command line
    - `Eg. java weka.classifiers.j48.J48 -t $WEKAHOME/data/iris.arff`

WEKA:: Interface



WEKA:: Data format

- Uses flat text files to describe the data
- Can work with a wide variety of data files including its own “.arff” format and C4.5 file formats
- Data can be imported from a file in various formats:
  - ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)

WEKA:: ARRF file format

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal,
   atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

numeric attribute

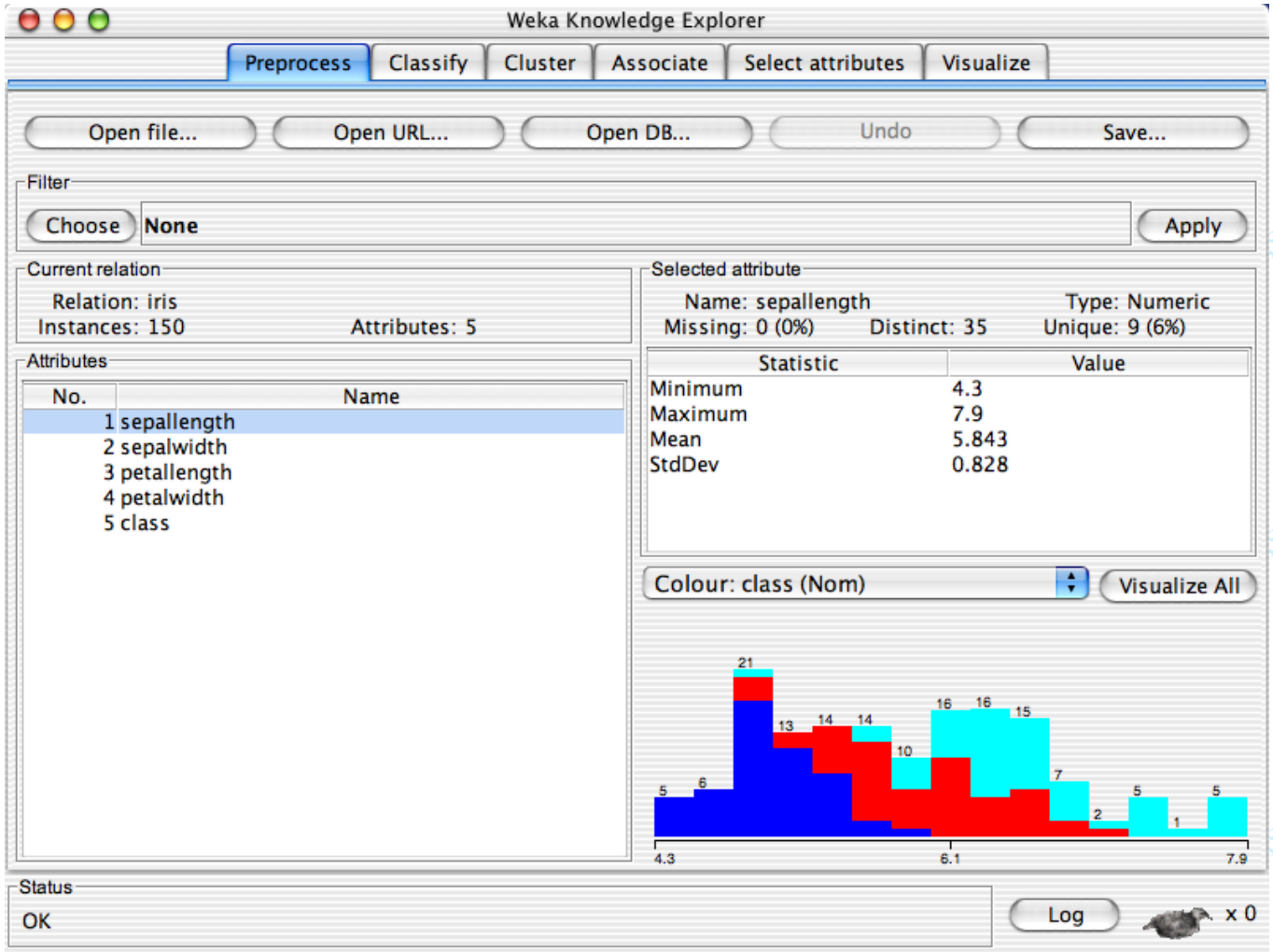
nominal attribute

A more thorough description is available here
http://www.cs.waikato.ac.nz/~ml/weka/arff.html

WEKA:: Explorer: Preprocessing

- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization,
  - normalization,
  - resampling,
  - attribute selection,
  - transforming,
  - combining attributes, etc

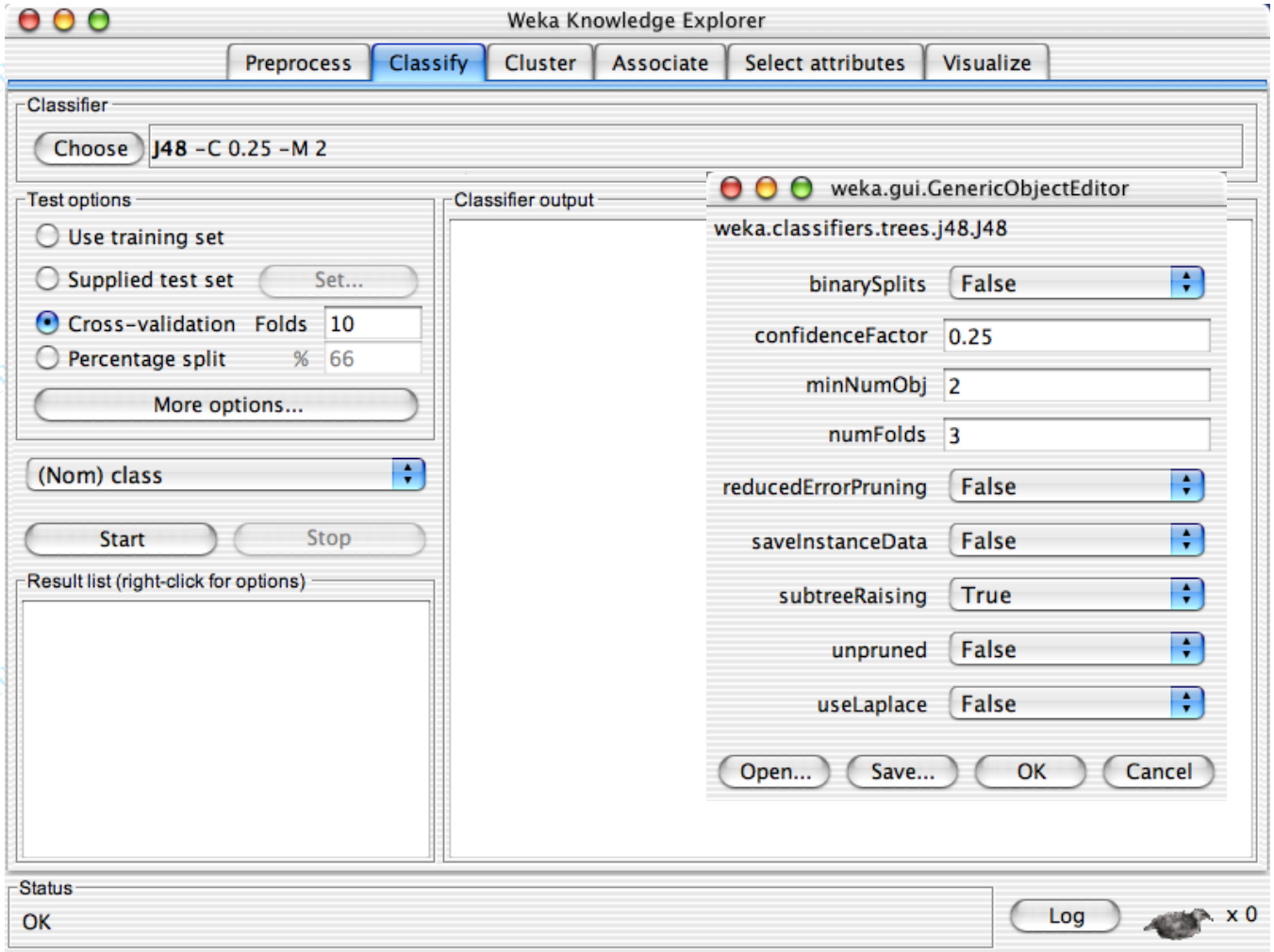
Weka environment



WEKA:: Explorer: building "classifiers"

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, …
- "Meta"-classifiers include:
  - Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, …

Weka environment



قدردانی

- Dr. Jianjun Hu
  http://mleg.cse.sc.edu/edu/csce822/
- University of South Carolina
- Department of Computer Science and Engineering