# Part 1

## Chromosomes in the 100th generation

The attribute subsets apparently had equal fitness by the 92nd generation, so WEKA stopped there.

Generation: 92

merit scaled subset

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 2 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 2 5 6 7 9

0.97883 0 1 5 6 7 9

0.97883 0 1 5 6 7 9

## Weka’s list of the best tumor characteristics for classification

Selected attributes: 1,5,6,7,9 : 5

clump.thickness

single.epithelial.cell.size

bare.nuclei

bland.chromatin

mitoses

## Accuracy of classification and confusion matrixes

|  |  |
| --- | --- |
| **Naïve Bayes: All Attributes** | **Naïve Bayes: Fittest Attributes** |
| Correctly Classified Instances 432 97.2973 % | Correctly Classified Instances 435 97.973 % |
| a b <-- classified as  303 9 | a = 2  3 129 | b = 4 | a b <-- classified as  305 7 | a = 2  2 130 | b = 4 |

# Part 2

## Top 5 genes

Ranked attributes:

0.609 143 MPO from Human myeloperoxidase gene, exons 1-4./ntype=DNA /annot=exon

0.607 124 TALDO Transaldolase

0.559 145 Low-Mr GTP-binding protein (RAB31) mRNA

0.555 100 LRPAP1 Low density lipoprotein-related protein-associated protein 1 …

0.555 148 LPAP gene

## Comparison of accuracy and confusion matrixes

|  |  |
| --- | --- |
| **Naïve Bayes: All Genes** | **Naïve Bayes: Top 5 Genes** |
| Correctly Classified Instances 68 94.4444 % | Correctly Classified Instances 65 90.2778 % |
| a b <-- classified as  44 0 | a = ALL  4 24 | b = AML | a b <-- classified as  42 2 | a = ALL  5 23 | b = AML |
| True Positive rate = 100.0% (∆ +4.5%)  True Negative rate = 85.7% (∆ +3.6%) | True Positive rate = 95.5% (∆ -4.5%)  True Negative rate = 82.1% (∆ -3.6%) |
| False Negative rate = 0.0% (∆ -4.5%)  False Positive rate = 14.3% (∆ -3.6%) | False Negative rate = 4.5% (∆ +4.5%)  False Positive rate = 17.9% (∆ +3.6%) |

Classifying using Naïve Bayes with only the 5 most informative genes instead of all the genes decreases the accuracy by about 4.2% due to 3 additional misclassifications; 2 false negatives and 1 false positive.

# Breast Cancer

## Attribute Selection: Genetic Search

=== Run information ===

Evaluator: weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.bayes.NaiveBayes -F 5 -T 0.01 -R 1 -E DEFAULT --

Search: weka.attributeSelection.GeneticSearch -Z 20 -G 100 -C 0.6 -M 0.033 -R 20 -S 1

Relation: breast.cancer

Instances: 444

Attributes: 10

clump.thickness

uniformity.of.cell.size

uniformity.of.cell.shape

marginal.adhesion

single.epithelial.cell.size

bare.nuclei

bland.chromatin

normal.nuclei

mitoses

class

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Genetic search.

Start set: no attributes

Population size: 20

Number of generations: 100

Probability of crossover: 0.6

Probability of mutation: 0.033

Report frequency: 20

Random number seed: 1

Generation: …

Attribute Subset Evaluator (supervised, Class (nominal): 10 class):

Wrapper Subset Evaluator

Learning scheme: weka.classifiers.bayes.NaiveBayes

Scheme options:

Subset evaluation: classification accuracy

Number of folds for accuracy estimation: 5

Selected attributes: 1,5,6,7,9 : 5

clump.thickness

single.epithelial.cell.size

bare.nuclei

bland.chromatin

mitoses

## Naïve Bayes: All Attributes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: breast.cancer

Instances: 444

Attributes: 10

clump.thickness

uniformity.of.cell.size

uniformity.of.cell.shape

marginal.adhesion

single.epithelial.cell.size

bare.nuclei

bland.chromatin

normal.nuclei

mitoses

class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

…

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 432 97.2973 %

Incorrectly Classified Instances 12 2.7027 %

Kappa statistic 0.9362

Mean absolute error 0.0258

Root mean squared error 0.1539

Relative absolute error 6.1594 %

Root relative squared error 33.6744 %

Total Number of Instances 444

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.971 0.023 0.990 0.971 0.981 0.937 0.996 0.998 2

0.977 0.029 0.935 0.977 0.956 0.937 0.996 0.989 4

Weighted Avg. 0.973 0.025 0.974 0.973 0.973 0.937 0.996 0.996

=== Confusion Matrix ===

a b <-- classified as

303 9 | a = 2

3 129 | b = 4

## Naïve Bayes: Fittest Attributes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: breast.cancer-weka.filters.unsupervised.attribute.Remove-R2-4,8

Instances: 444

Attributes: 6

clump.thickness

single.epithelial.cell.size

bare.nuclei

bland.chromatin

mitoses

class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

…

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 435 97.973 %

Incorrectly Classified Instances 9 2.027 %

Kappa statistic 0.952

Mean absolute error 0.03

Root mean squared error 0.1463

Relative absolute error 7.1768 %

Root relative squared error 32.0178 %

Total Number of Instances 444

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.978 0.015 0.993 0.978 0.985 0.952 0.995 0.998 2

0.985 0.022 0.949 0.985 0.967 0.952 0.995 0.986 4

Weighted Avg. 0.980 0.017 0.980 0.980 0.980 0.952 0.995 0.994

=== Confusion Matrix ===

a b <-- classified as

305 7 | a = 2

2 130 | b = 4

# Leukemia

## Attribute Selection: Info Gain

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: leukemia gene expression data names changed

Instances: 72

Attributes: 151

[list of attributes omitted]

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 151 leukemia\_type):

Information Gain Ranking Filter

Ranked attributes:

0.609 143 MPO from Human myeloperoxidase gene, exons 1-4./ntype=DNA /annot=exon

0.607 124 TALDO Transaldolase

0.559 145 Low-Mr GTP-binding protein (RAB31) mRNA

0.555 100 LRPAP1 Low density lipoprotein-related protein-associated protein 1 (alpha-2-macroglobulin receptor-associated protein 1

0.555 148 LPAP gene

…

Selected attributes: 143,124,145,100,148,147,150,121,125,146,135,142,106,144,68,62,134,140,74,99,110,129,123,80,42,116,149,136,87,137,93,130,95,131,127,34,132,138,90,113,101,67,84,76,92,122,72,118,60,54,139,119,102,69,141,8,111,104,94,105,50,10,96,83,133,32,15,47,49,108,65,37,13,75,97,57,19,114,120,126,64,79,128,77,3,88,4,63,17,5,109,22,89,112,107,30,29,70,117,28,40,48,21,16,98,103,31,81,51,18,25,85,53,14,115,27,71,44,39,45,12,1,56,58,41,59,35,86,78,73,91,46,24,61,66,33,36,55,26,82,43,7,38,20,6,23,52,2,11,9 : 150

## Naïve Bayes: All Genes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: leukemia gene expression data names changed

Instances: 72

Attributes: 151

[list of attributes omitted]

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute ALL AML

(0.61) (0.39)

==========================================================================================================

…

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 68 94.4444 %

Incorrectly Classified Instances 4 5.5556 %

Kappa statistic 0.88

Mean absolute error 0.0556

Root mean squared error 0.2357

Relative absolute error 11.6559 %

Root relative squared error 48.2804 %

Total Number of Instances 72

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 0.143 0.917 1.000 0.957 0.886 0.942 0.933 ALL

0.857 0.000 1.000 0.857 0.923 0.886 0.977 0.974 AML

Weighted Avg. 0.944 0.087 0.949 0.944 0.944 0.886 0.956 0.949

=== Confusion Matrix ===

a b <-- classified as

44 0 | a = ALL

4 24 | b = AML

## Naïve Bayes: Top 5 Genes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: leukemia gene expression data names changed-weka.filters.unsupervised.attribute.Remove-R1-99,101-123,125-142,144,146-147,149-150

Instances: 72

Attributes: 6

LRPAP1 Low density lipoprotein-related protein-associated protein 1 (alpha-2-macroglobulin receptor-associated protein 1

TALDO Transaldolase

MPO from Human myeloperoxidase gene, exons 1-4./ntype=DNA /annot=exon

Low-Mr GTP-binding protein (RAB31) mRNA

LPAP gene

leukemia\_type

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute ALL AML

(0.61) (0.39)

==========================================================================================================

…

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 65 90.2778 %

Incorrectly Classified Instances 7 9.7222 %

Kappa statistic 0.7914

Mean absolute error 0.0993

Root mean squared error 0.3079

Relative absolute error 20.8388 %

Root relative squared error 63.0753 %

Total Number of Instances 72

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.955 0.179 0.894 0.955 0.923 0.795 0.952 0.955 ALL

0.821 0.045 0.920 0.821 0.868 0.795 0.943 0.920 AML

Weighted Avg. 0.903 0.127 0.904 0.903 0.902 0.795 0.948 0.941

=== Confusion Matrix ===

a b <-- classified as

42 2 | a = ALL

5 23 | b = AML