Übung 3

Alexander Hinneburg

Aufgabe 1

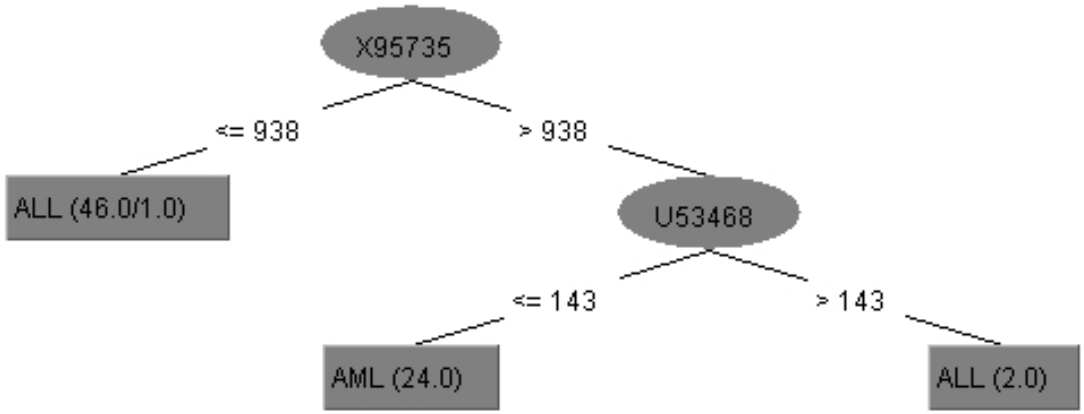
- Konvertierung von CSV nach Arff
  - Durch Klicken im Weka Explorer
  - Kommando Zeile

```
java -cp "c:\Programme\Weka-3-4\weka.jar"
    weka.core.converters.CSVLoader genes-leukemia.csv >genes-leukemia.arff
```

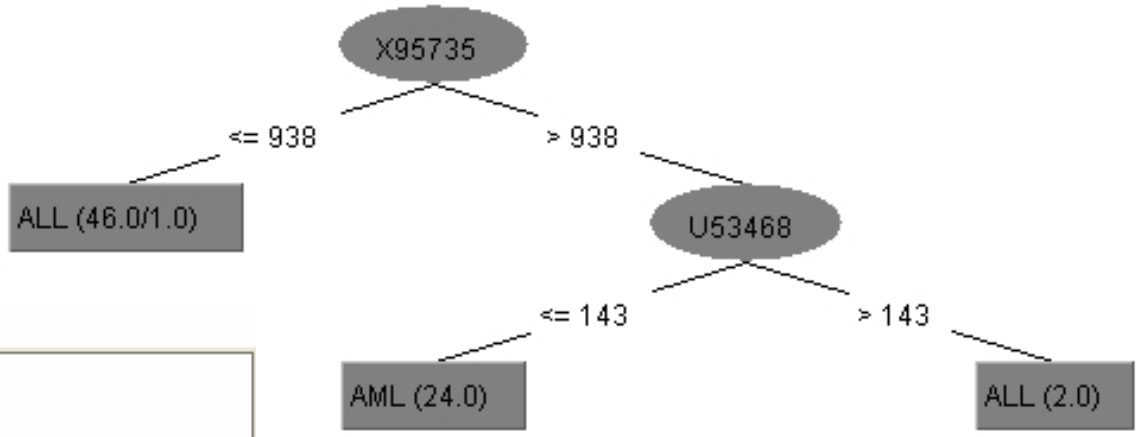
  - CLI ?
- Batch-Konvertierung von vielen Dateien
  - XXXLoader: lädt Datei im Format XXX und gibt Arff aus
  - XXXSaver: nimmt Arff und gibt XXX aus
  - Shell Skript: Performanz hängt von JRE ab
  - Java Programm: aufwendig zu schreiben

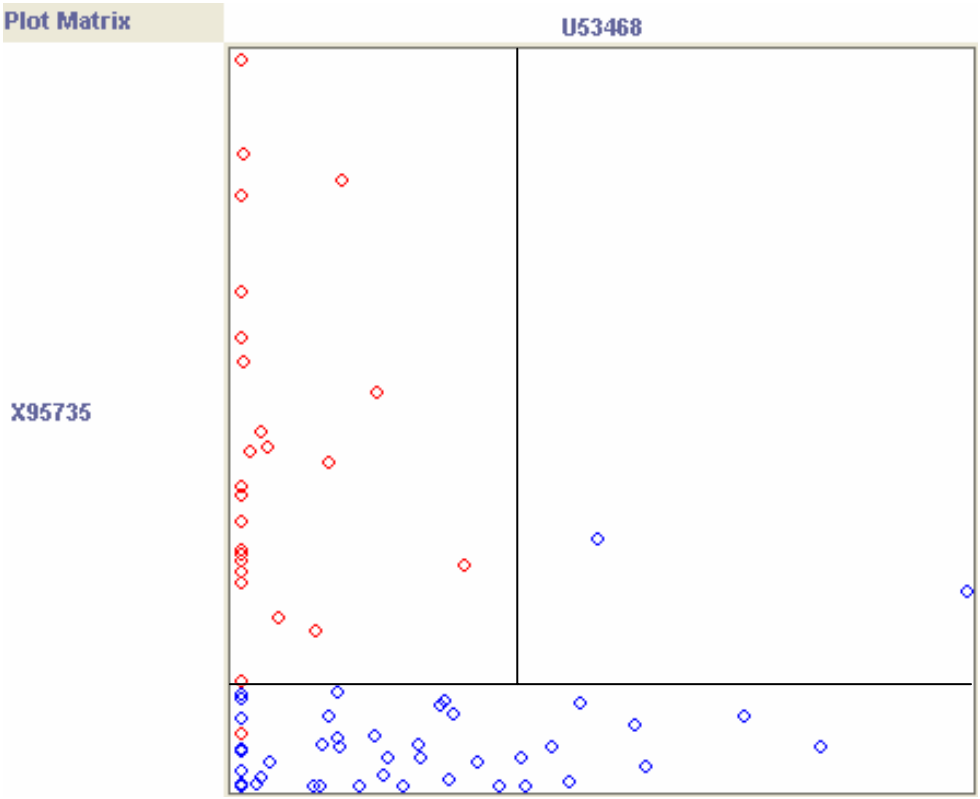
Aufgabe 2

- 72 Leukämie Patienten mit Expression für 40 Gene
- 2 Klassen, ALL (Acute Lymphoblastic leukemia),
  AML (Acute Myelogenous Leukemia)
- Klassifikator J48
  - Korrekt klassifizierte Instanzen: 71 (98.6111 %)
  - Kappa Statistik: 0.9691
    (Verbesserung gegenüber Zufallsklassifikator, 0%=Zufall)
  - TP Rate FP Rate Precision Recall F-Measure Class
    1 0.04 0.979 1 0.989 ALL
    0.96 0 1 0.96 0.98 AML
  - Confusion Matrix
    a b <-- classified as
    47 0 | a = ALL
    1 24 | b = AML



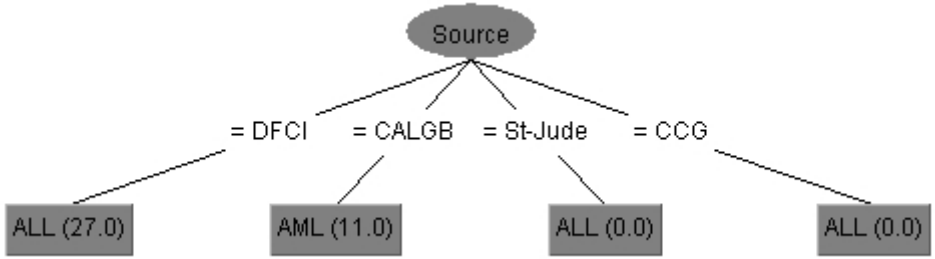
Aufgabe 2


X95735
<= 938
> 938
ALL (46.0/1.0)
U53468
<= 143
> 143
AML (24.0)
ALL (2.0)


Plot Matrix
U53468
X95735

Aufgabe 3

- Teilen der Daten mit Edit-Fkt von Weka
- Teil A
  - alles richtig, Kappa: 1
- Teil B



  - Korrekt klassifizierte Instanzen: 24 (70.5882 %)
  - Falsch klassifizierte Instanzen: 10 (29.4118 %)
  - Kappa: 0.32
- Diskussion
  - Source kennzeichnet Krankenhaus des Patienten
  - Als Erklärung für Krankheit wenig wahrscheinlich
  - Deshalb hoher Fehler auf Testmenge

Aufgabe 4

- Remove Source aus Trainings- und Test- Daten
  - Filter in Weka
- J 48 Klassifikator
  - Korrekt klassifizierte Instanzen: 31 (91.17 %)
  - Falsch klassifizierte Instanzen: 3 (8.82 %)
  - Kappa: 0.8198
  - Confusion Matrix

```
  a  b   <-- classified as
18  2 |  a = ALL
 1 13 |  b = AML
```

- Diskussion
  - Verwendung von Trainings und Testdaten verhindert sehr speziellen Test mit U53468
  - Fazit: um zwischen ALL und AML zu unterscheiden reicht X95735 aus

Zusatzaufgabe

- Bayes
  - Bayesnet 82.3529 %
  - Naive Bayes: 79.4118 %
  - Naive Bayes Update: 79.4118 %
- Functions
  - Logistic: 85.2941 %
  - RBFNetwork: 88.2353 %
  - SimpleLogistic: 82.3529 %
  - SMO: 85.2941 %
  - VotedPerceptron: 94.1176 %
- Lazy
  - IB1: 88.2353 %
  - LWL: 100%

Zusatzaufgabe

- Meta
  - AdaBoostM1: 100%
  - AttributeSelectedClassifier: 91.1765 %
  - Bagging: 58.8235 %
  - ClassificationViaRgerssion, CVParameterSelection, Grading, MultiScheme, …: 58.8235 %
  - Decorate: 82.3529 %
  - FilteredClassifier: 91.1765 %
  - LogitBoost: 97.0588 %
  - MultiBoostAB: 100 %
  - MultiClassClassifier: 85.2941 %
  - OrdinalClassClassifier: 91.1765 %
  - RandomCommittee: 76.4706 %
  - ThresholdSelector: 67.6471 %

Zusatzaufgabe

- Trees
  - ADTree: 91.1765 %
  - DecisionStump: 100%
  - J48: 91.1765 %
  - LMT: 82.3529 %
  - NBTree: 85.2941 %
  - RandomForest: 85.2941 %
  - RandomTree, REPTree: 58.8235 %
- Rules
  - Conjunktive Rule: 91.1765 %
  - Decision Table: 100%
  - JRip:70.5882 %
  - NNge: 88.2353 %
  - OneR, ZeroR: 58.8235 %
  - PART: 91.1765 %
  - Ridor: 91.1765 %