情報意味論(3) Wekaの紹介

Preslav Nakov (October 6, 2004)

http://www.sims.berkeley.edu/courses/is290-2/f04/lectures/lecture11.ppt

Eibe Frank
http://prdownloads.sourceforge.net/weka/weka.ppt

- ***WEKA: Explorer***
- WEKA: Experimenter
- WEKA: 使ってみよう

WEKA: 好奇心旺盛な飛べない鳥

Slide adapted from Eibe Frank's

WEKA: 用語

Weka で用いる単語は、多少、普通の用法と異なる:

- Attribute: 属性, feature (普通)
- Relation: 事例の集合, collection of examples
- Instance: その時点で使用中の事例集合
- Class: クラス、範疇、category

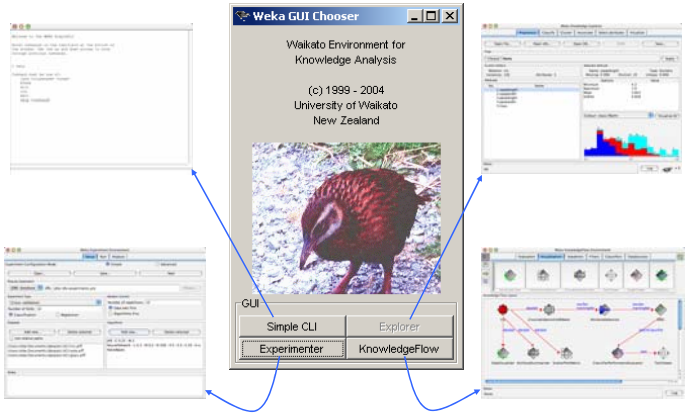
WEKA: The Software Toolkit

http://www.cs.waikato.ac.nz/ml/weka

- Java で書かれた、機械学習/データマイニングソフト
- GNU ライセンス
- 研究、教育、応用に用いられる
- Witten & Frank 著 “Data Mining” で使用
- 主な機能:
  - データ前処理ツール
  - 学習アルゴリズム
  - 評価手法
  - グラフィカルインタフェイス (データ可視化含む)
  - 学習アルゴリズムの比較

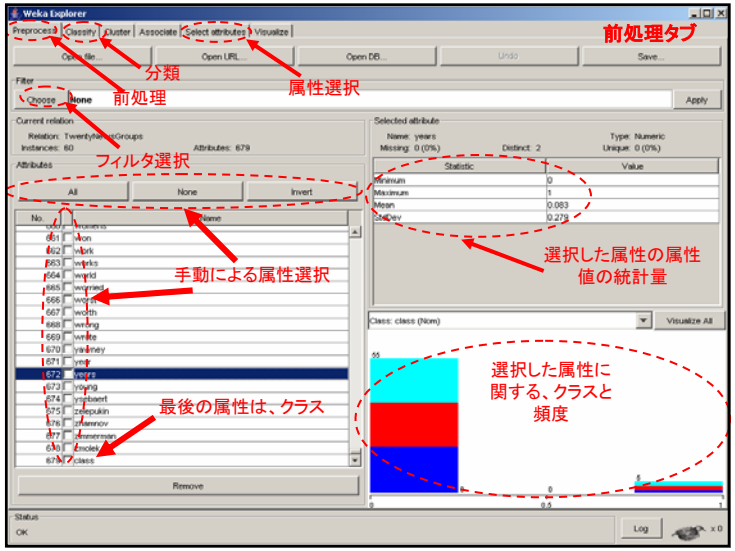
WEKA GUI Chooser

java -**Xmx1000M** -jar weka.jar



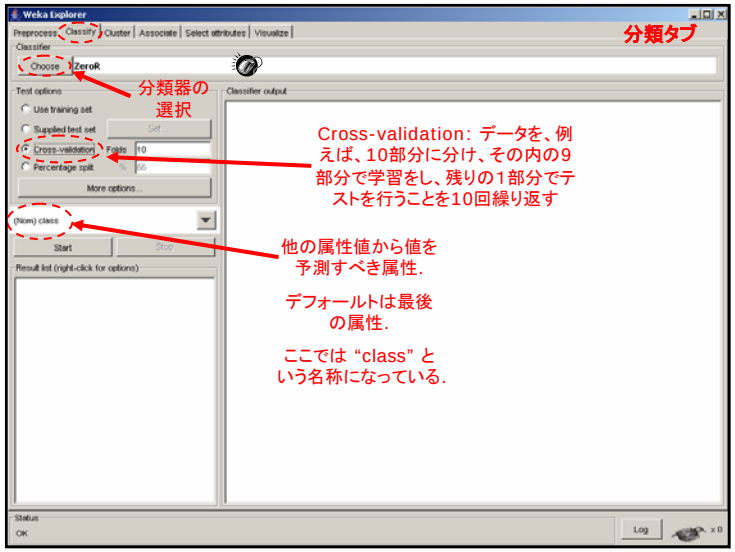
Explorer: データの前処理

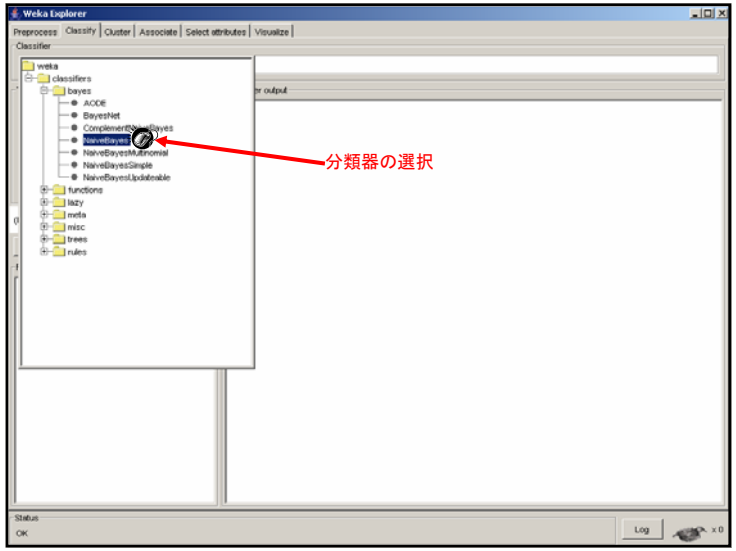
- WEKA はデータがインポートできる:
  - ファイルから: ARFF, CSV, C4.5, binary
  - URL から
  - SQL データベースから (JDBCを用いて)
- 前処理ツール(フィルター)が使われるのは:
  - 離散化 discretization, 正規化 normalization, resampling, 属性選択 attribute selection, 属性の変換と結合, etc.

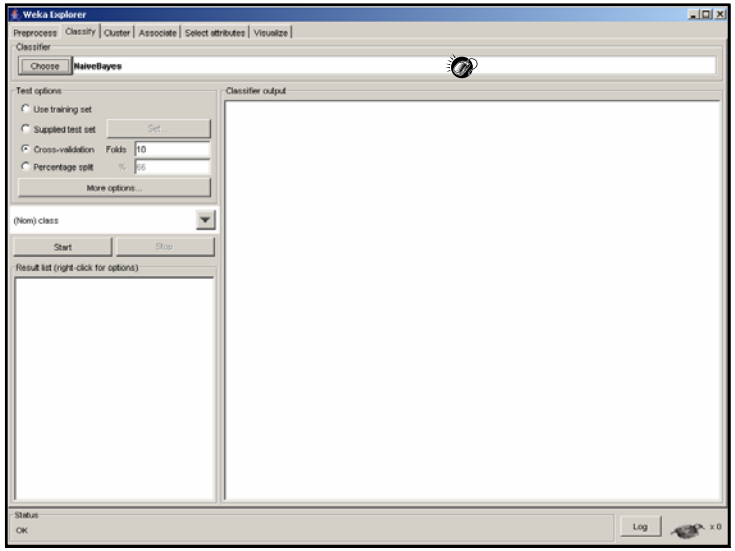


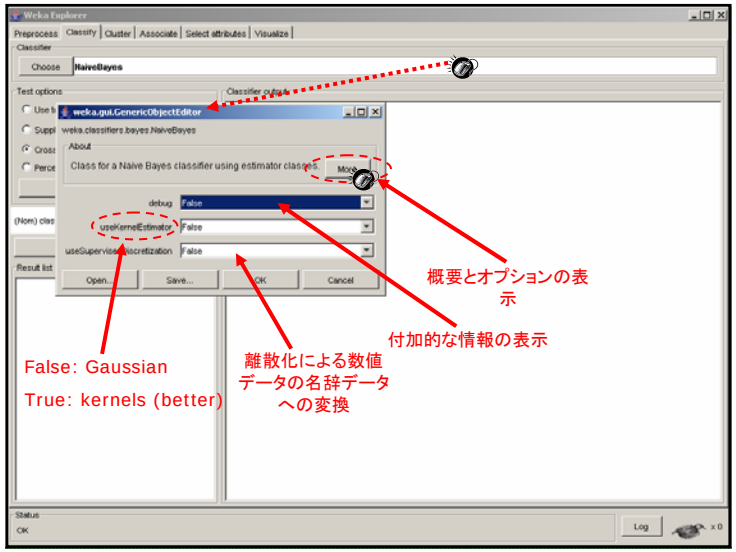
Explorer: 分類

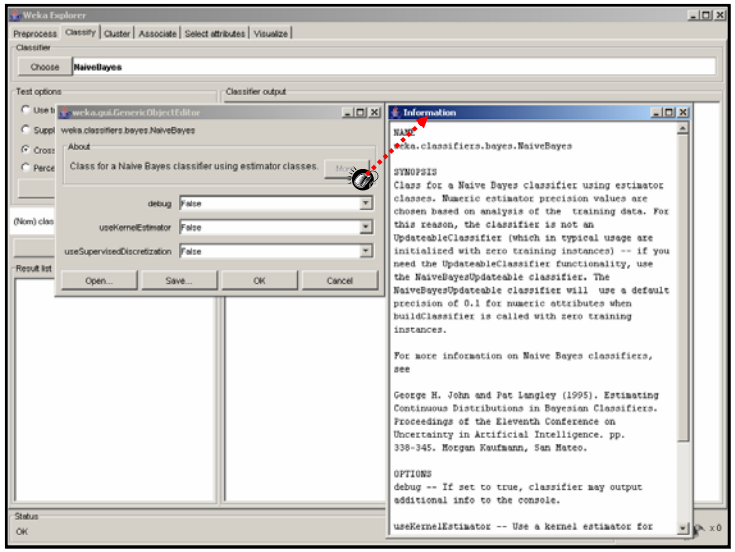
- Weka の *Classifiers* は:
  - 分類 classification (名辞で表すクラスの予測)
  - 回帰 regression (数値的な量の予測)
- 使用する学習アルゴリズム:
  - Naïve Bayes, 決定木 decision trees, kNN, support vector machines, ニューラルネットワーク, logistic regression, 等.
- メタ分類:
  - 単独で使うものではない
  - 学習アルゴリズムと組み合わせて使用
  - 例: boosting, bagging 等.

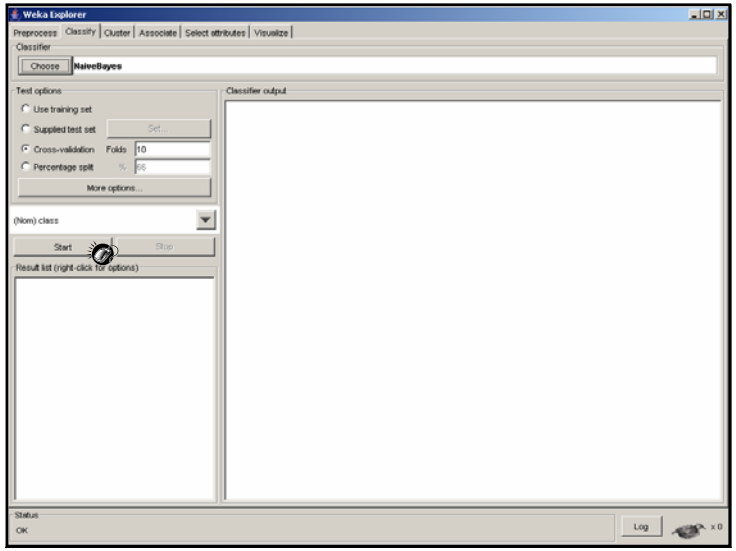


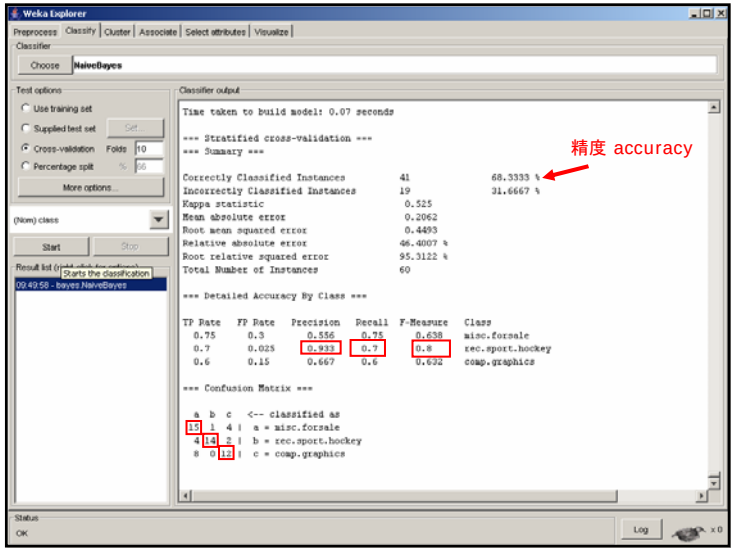





概要とオプションの表示
付加的な情報の表示
False: Gaussian
True: kernels (better)
離散化による数値データの名辞データへの変換



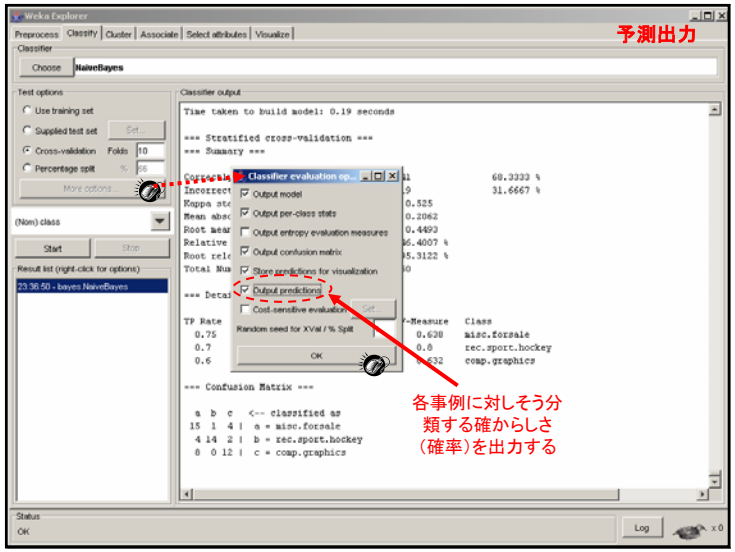


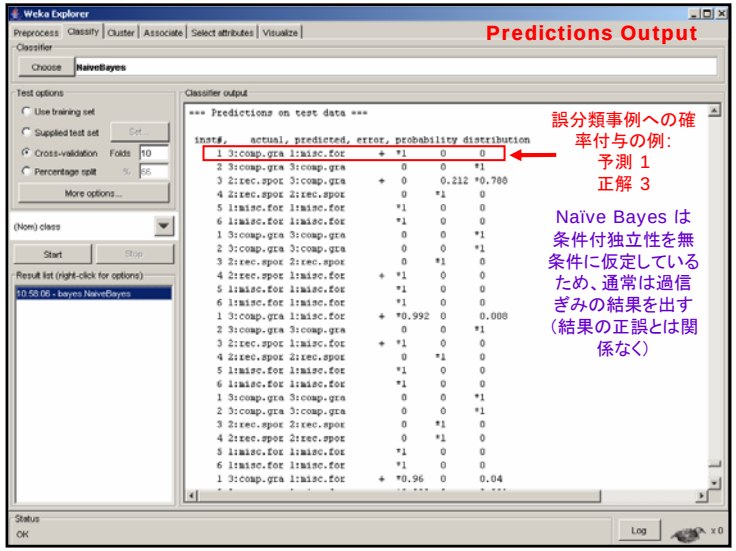

精度 accuracy

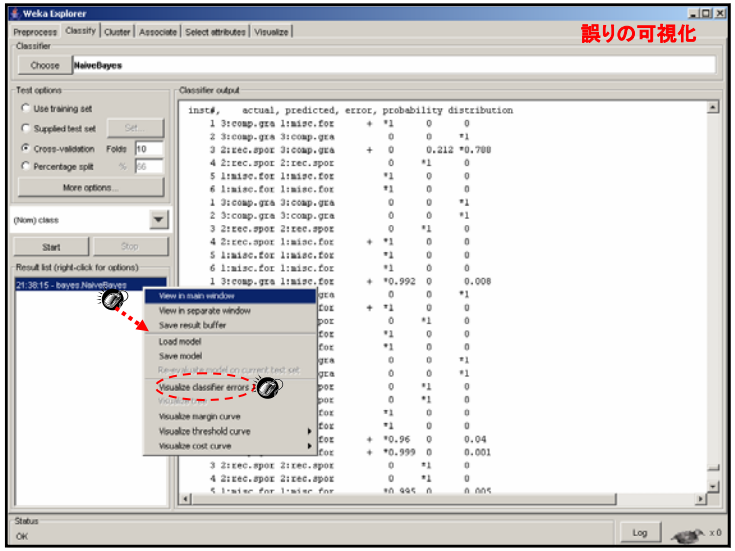
Confusion matrix

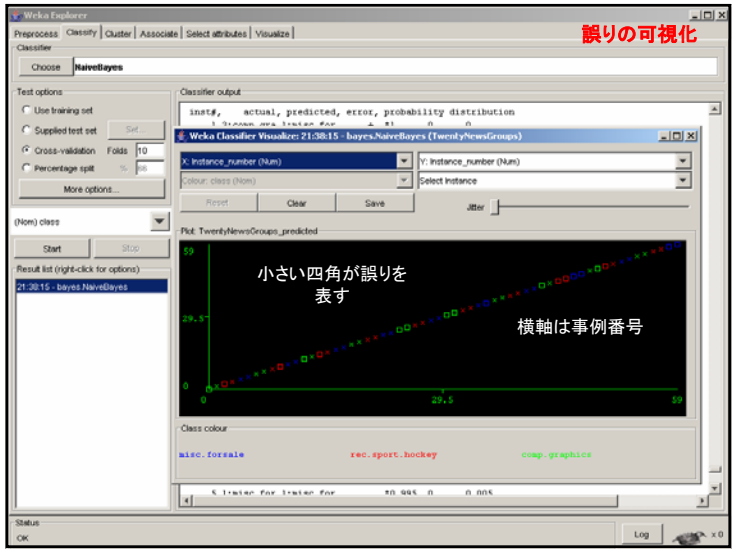
- 予測した分類と本当の分類との違いを数値化したもの
- これから様々な値が得られる:
  - accuracy: $(a+d)/(a+b+c+d)$
  - recall: $d/(c+d)$ => R
  - precision: $d/(b+d)$ => P
  - F-measure: 2PR/(P+R)
  - false positive (FP) rate: $b/(a+b)$
  - true negative (TN) rate: $a/(a+b)$
  - false negative (FN) rate: $c/(c+d)$
- これらは2クラス以上にも適用できる

		予測値	
		–	+
真	–	a	b
	+	c	d



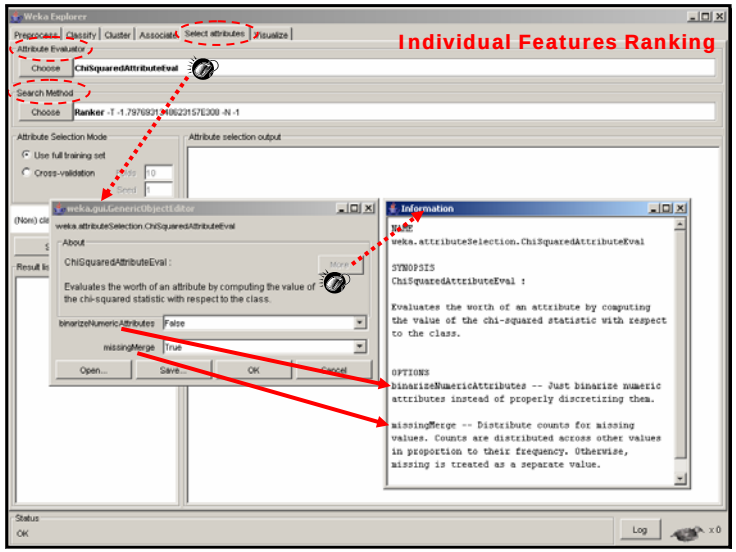


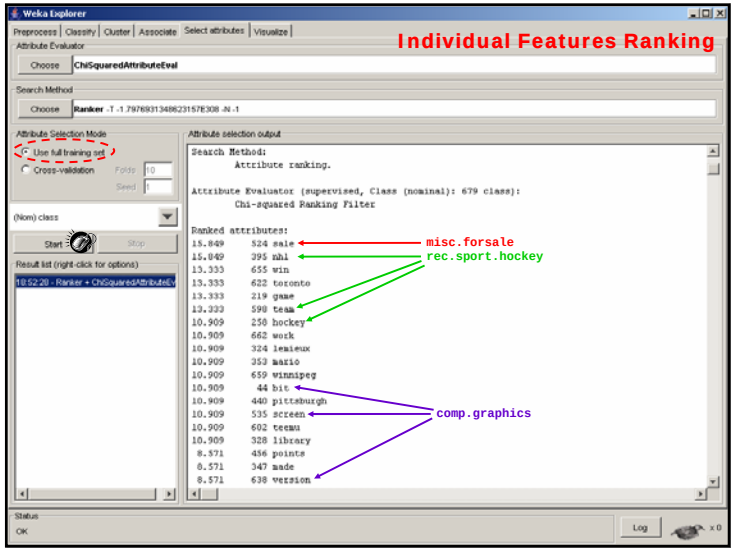


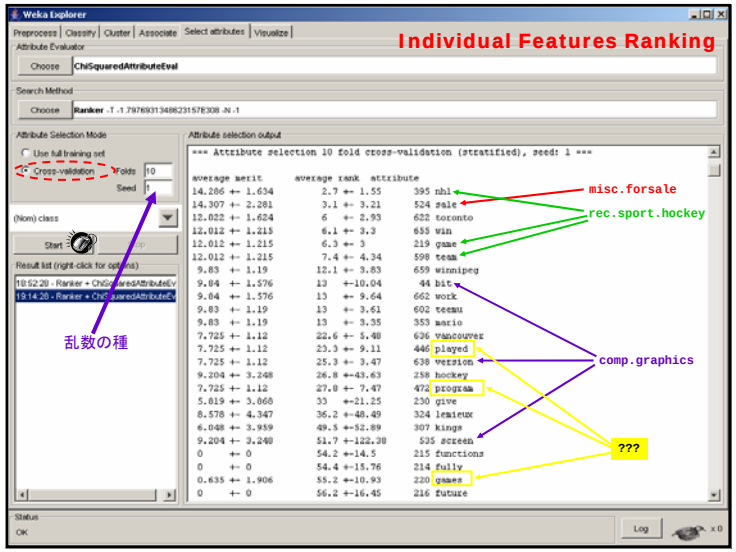


Explorer: 属性選択

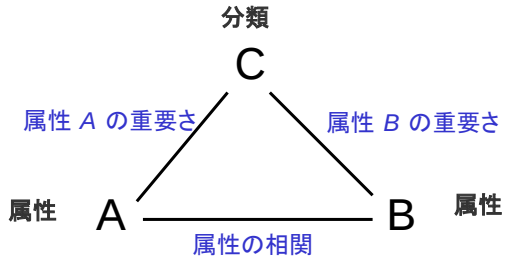
- 予測力の高い属性を探す
  - 予測力の低い属性を、予め、排除したい
- 2種類の方法:
  - 探索法:
    - best-first, forward selection, random, 網羅探索 exhaustive, 遺伝的アルゴリズム genetic algorithm, ranking
  - 評価法:
    - information gain, χ 自乗, etc.
- WEKA は (殆どの) 任意の組合せができる





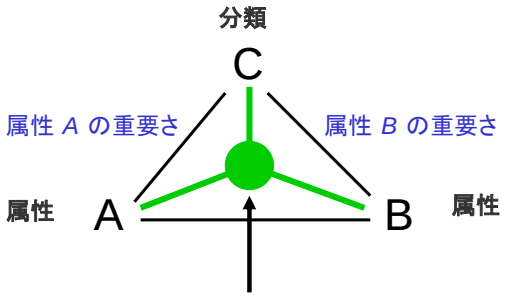


属性間の相互作用



2-方向 相互作用

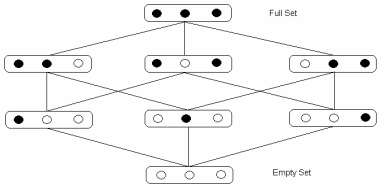
属性間の相互作用



3-方向 相互作用:
***A*, *B*, *C* に共通なもの;**
属性対からは推論できないもの.

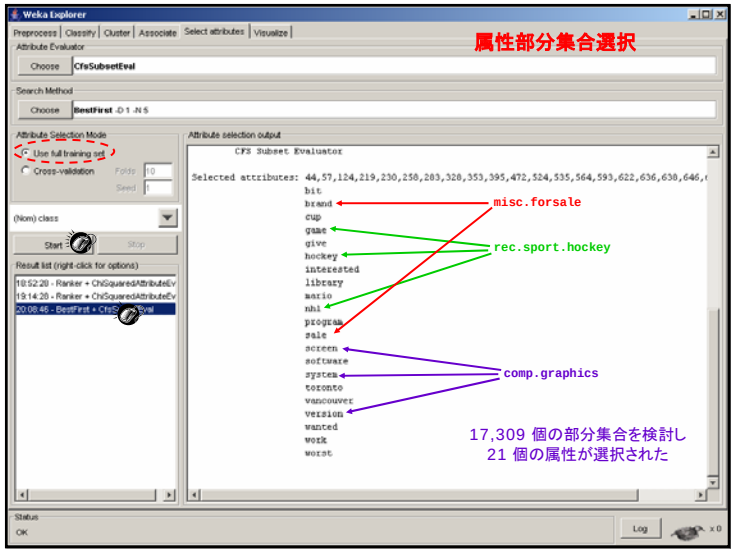
属性部分集合の選択

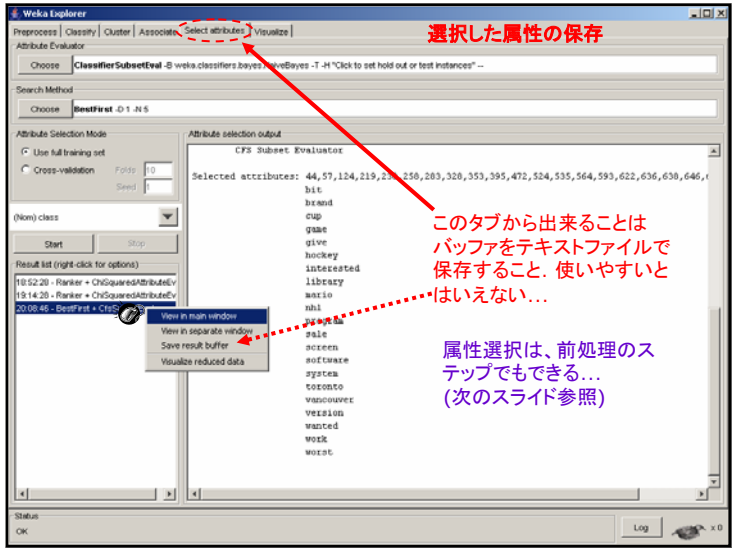
- 問題の例図
  - 全体集合 Full set
  - 空集合 Empty set
  - 数え上げ Enumeratic

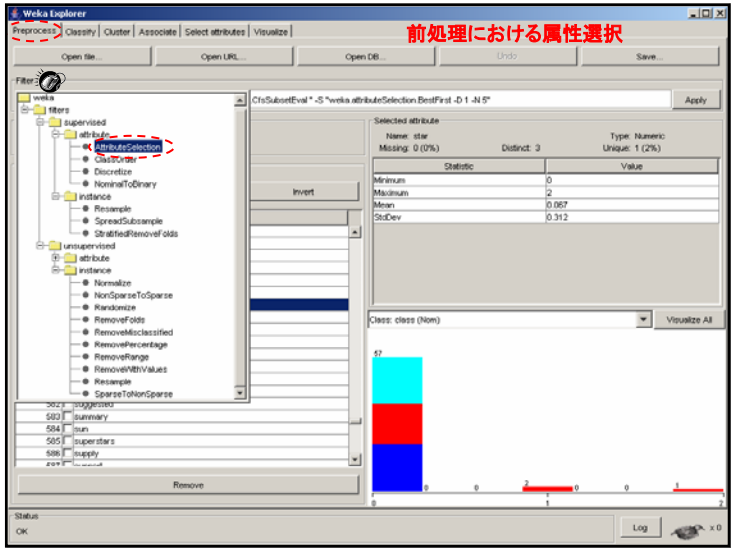


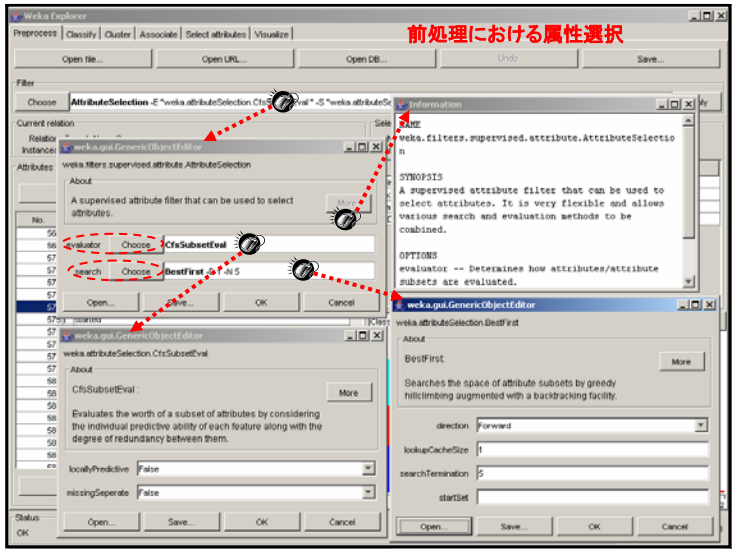
- 探索
  - 網羅的/完全 (数え上げ/branch&bounding)
  - ヒューリスティック (逐次的に 前進方向/後退方向)
  - 確率的 stochastic (生成/評価)
  - 個々の属性や部分集合の生成/評価

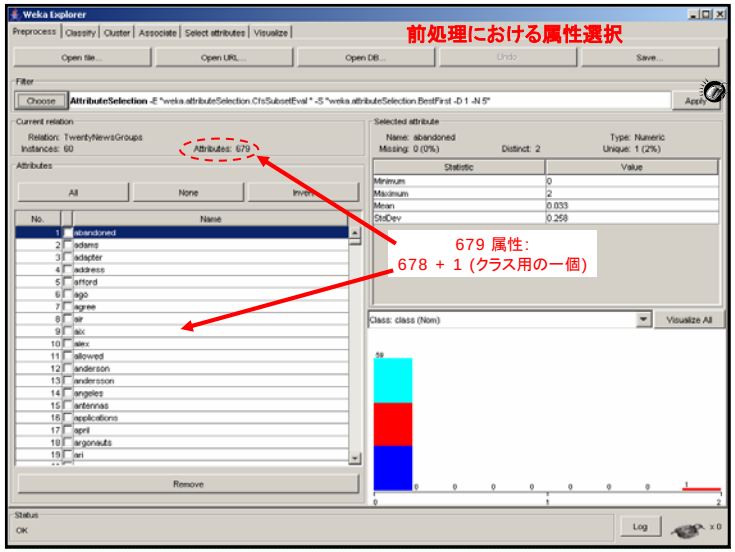

属性部分集合選択
CfsSubsetEval
BestFirst -D 1 -N 5
Information
Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.
OPTIONS
locallyPredictive -- Identify locally predictive attributes. Iteratively adds attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the attribute in question
missingSeperate -- Treat missing as a separate value. Otherwise, counts for missing values are distributed across other values in proportion to their frequency.

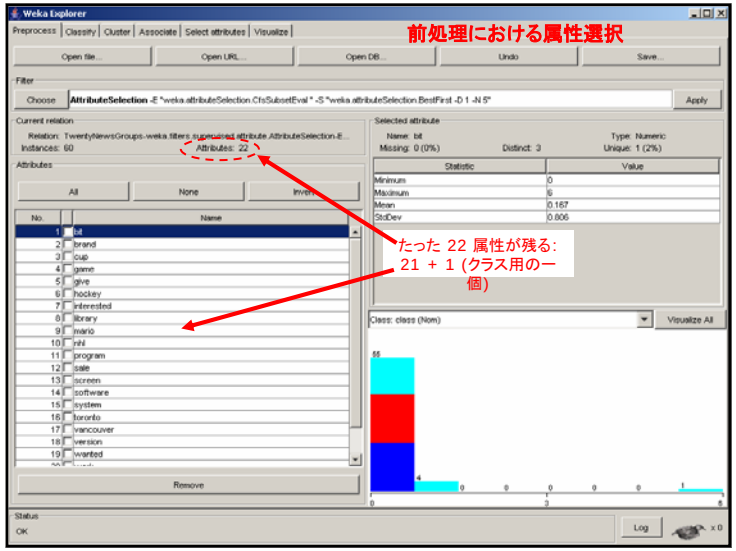

属性部分集合選択
misc.forsale
rec.sport.hockey
comp.graphics
17,309 個の部分集合を検討し
21 個の属性が選択された

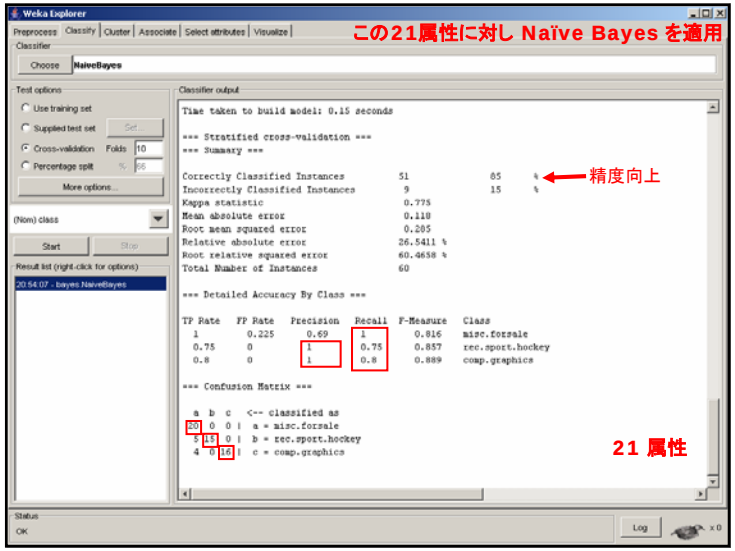

選択した属性の保存
このタブから出来ることは
バッファをテキストファイルで
保存すること．使いやすいと
はいえない…
属性選択は、前処理のス
テップでもできる…
(次のスライド参照)

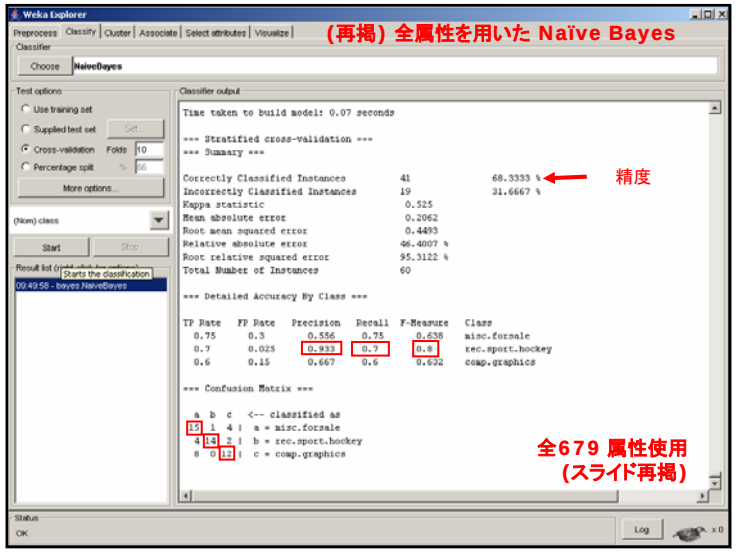

前処理における属性選択


前処理における属性選択


前処理における属性選択
679 属性:
678 + 1 (クラス用の一個)


前処理における属性選択
たった 22 属性が残る:
21 + 1 (クラス用の一個)


この21属性に対し Naïve Bayes を適用
精度向上
21 属性



重要なアルゴリズム

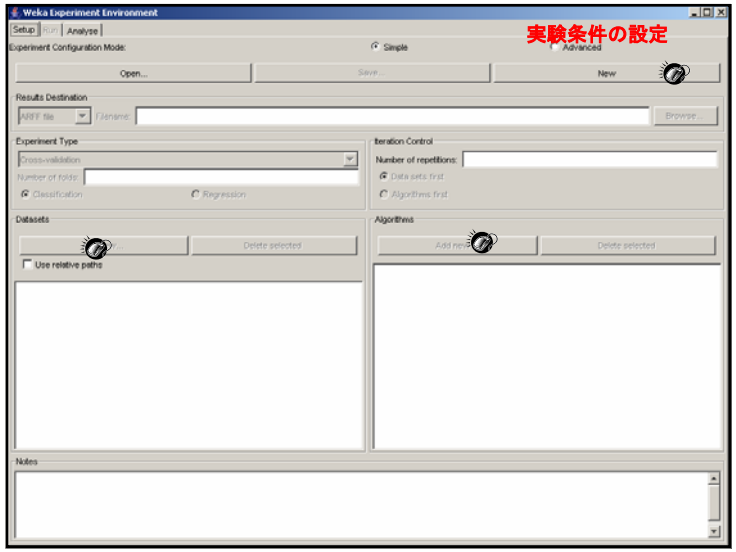
- WEKA はアルゴリズムに時々変わった名称をつけている
- 主なアルゴリズム:
  - Naïve Bayes: `weka.classifiers.bayes.NaiveBayes`
  - Perceptron: `weka.classifiers.functions.VotedPerceptron`
  - Winnow: `weka.classifiers.functions.winnow`
  - Decision tree: `weka.classifiers.trees.J48`
  - Support vector machines: `weka.classifiers.functions.SMO`
  - *k* nearest neighbor: `weka.classifiers.lazy.IBk`
- これらのものには、古典的なアルゴリズムではなくより洗練されたものがある
  - e.g. WEKA には古典的な Naïve Bayes は見つけられなかった (5 種類もの実装されているのに).

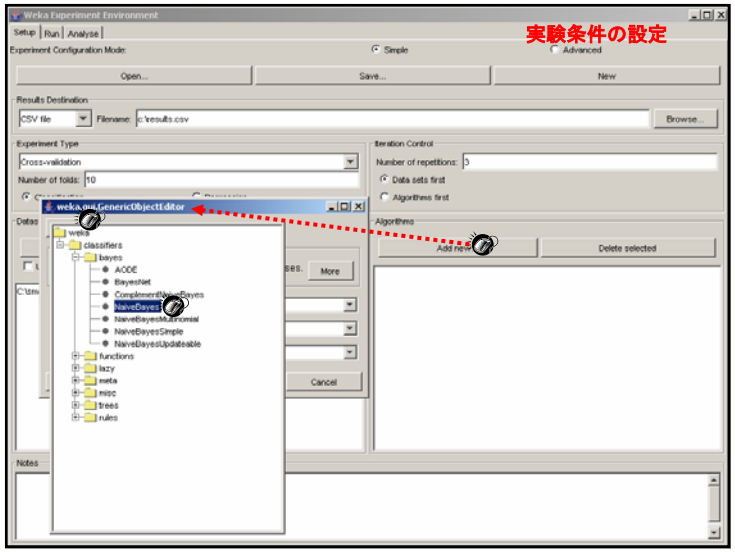
- WEKA: Explorer
- ***WEKA: Experimenter***
- WEKA: 使ってみよう

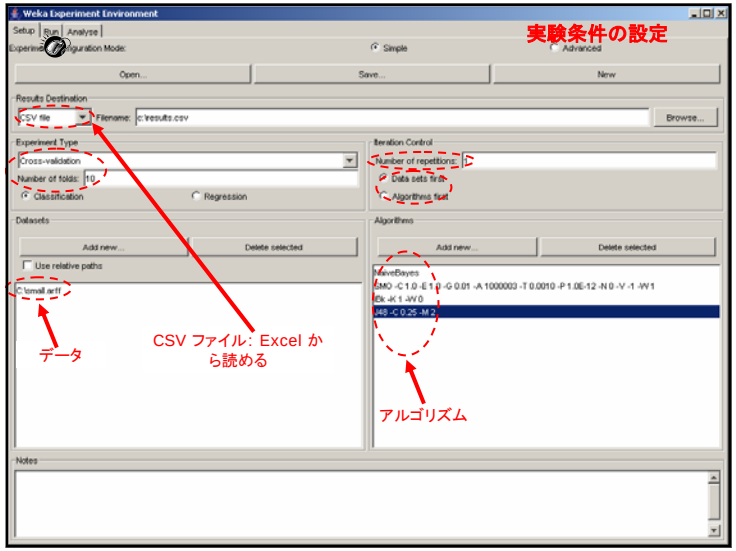
実験を行う

- *Experimenter* を用いると、異なる学習アルゴリズムを比較することが容易にできる
- 扱う問題:
  - 分類
  - 回帰
- 結果: ファイルに書き込まれる
- 評価方法の種類:
  - cross-validation
  - 学習曲線 learning curve
  - hold-out
- パラメータを変えて繰返すことができる
- 有意性検定が組み込み

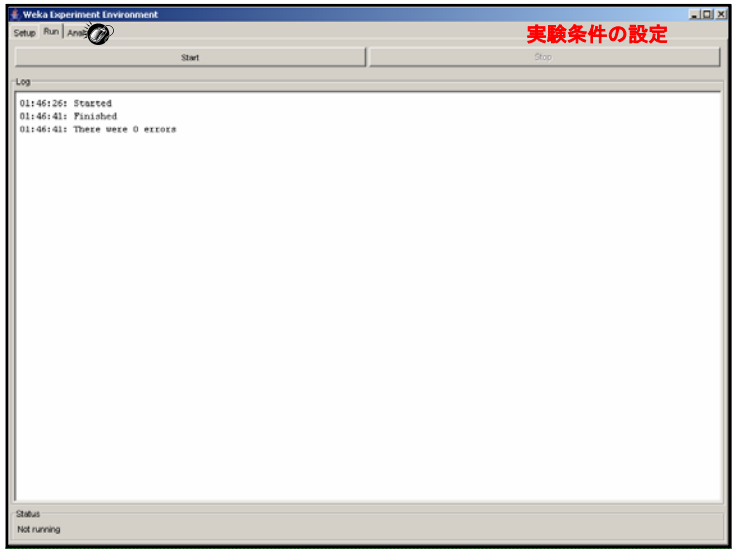
Slide adapted from Eibe Frank's

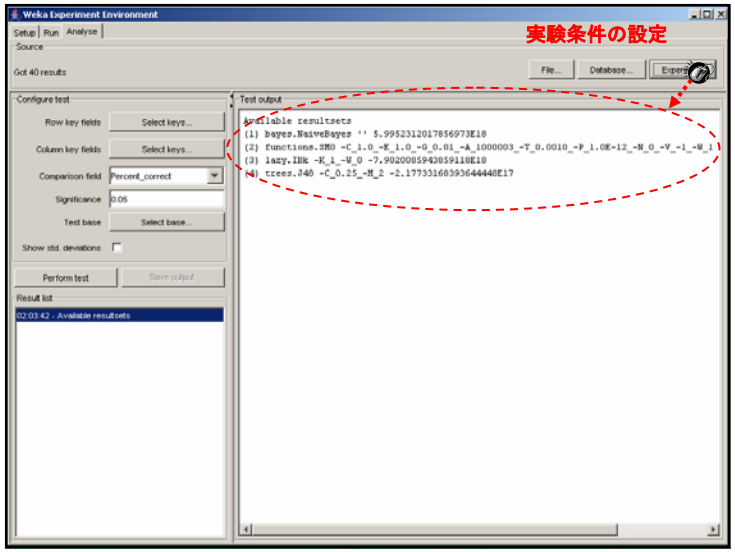

Weka Experiment Environment
実験条件の設定
Setup
Run
Analyse
Experiment Configuration Mode:
Simple
Advanced
Open...
Save...
New
Results Destination
ARFF file
Filename:
Browse...
Experiment Type
Cross-validation
Number of folds:
Classification
Regression
Iteration Control
Number of repetitions:
Data sets first
Algorithms first
Datasets
Add new...
Delete selected
Use relative paths
Algorithms
Notes

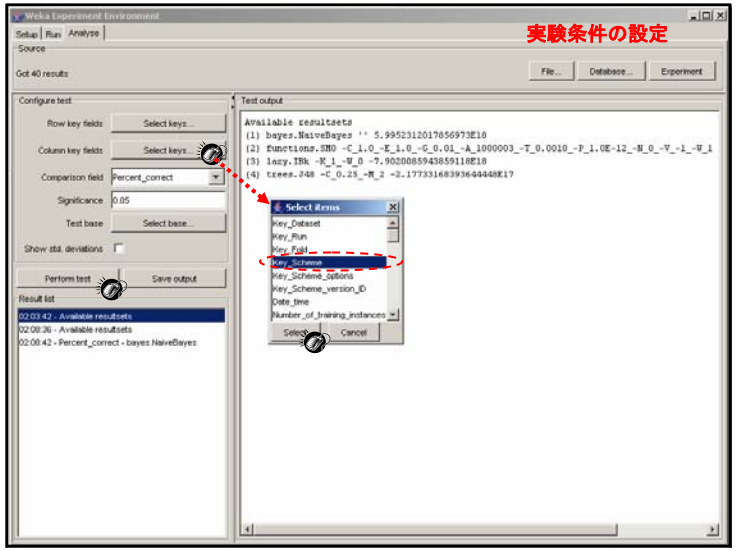

Weka Experiment Environment
実験条件の設定
weka.gui.GenericObjectEditor
weka
classifiers
bayes
AODE
BayesNet
ComplementNaiveBayes
NaiveBayes
NaiveBayesMultinomial
NaiveBayesSimple
NaiveBayesUpdateable
functions
lazy
meta
misc
trees
rules
More
Cancel

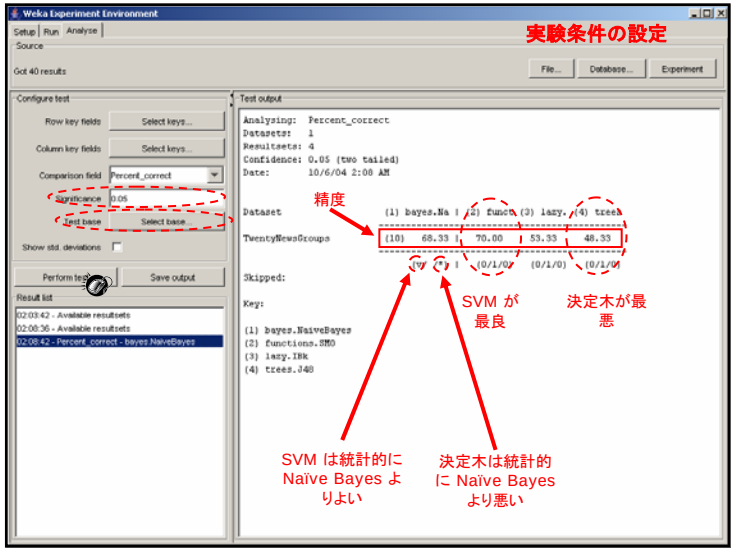

Weka Experiment Environment
実験条件の設定
CSV file
Filename: c:\results.csv
Cross-validation
Number of folds: 10
Number of repetitions: 10
Data sets first
Algorithms first
Classification
Regression
Use relative paths
C:\small.arff
NaiveBayes
J48 -C 0.25 -M 2
データ
CSV ファイル: Excel から読める
アルゴリズム

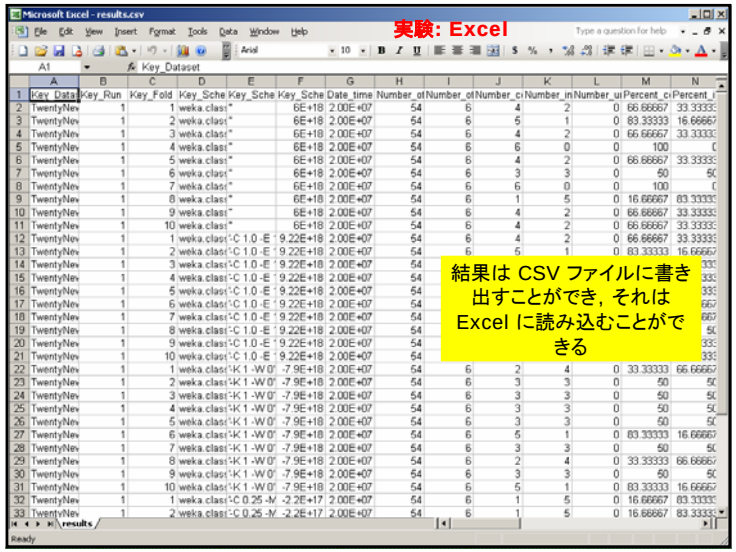

Weka Experiment Environment
実験条件の設定
Start
Stop
Log
Status
Not running


Weka Experiment Environment
Setup
Run
Analyse
Start
Stop
Log
01:46:26: Started
01:46:41: Finished
01:46:41: There were 0 errors
Status
Not running
実験条件の設定


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 40 results
File...
Database...
Experiment
Configure test
Row key fields
Select keys...
Column key fields
Select keys...
Comparison field
Percent_correct
Significance
0.05
Test base
Select base...
Show std. deviations
Perform test
Save output
Result list
02:03:42 - Available resultsets
Test output
Available resultsets
(1) bayes.NaiveBayes '' 5.9952312017856973E18
(2) functions.SMO -C_1.0_-K_1.0_-G_0.01_-A_1000003_-T_0.0010_-P_1.0E-12_-N_0_-V_-1_-W_1
(3) lazy.IBk -K_1_-W_0 -7.902008594385911 8E18
(4) trees.J48 -C_0.25_-M_2 -2.1773316839364444 8E17
実験条件の設定


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 40 results
File...
Database...
Experiment
Configure test
Row key fields
Select keys...
Column key fields
Select keys...
Comparison field
Percent_correct
Significance
0.05
Test base
Select base...
Show std. deviations
Perform test
Save output
Result list
02:03:42 - Available resultsets
02:08:36 - Available resultsets
02:08:42 - Percent_correct - bayes.NaiveBayes
Test output
Available resultsets
(1) bayes.NaiveBayes '' 5.9952312017856973E18
(2) functions.SMO -C_1.0_-K_1.0_-G_0.01_-A_1000003_-T_0.0010_-P_1.0E-12_-N_0_-V_-1_-W_1
(3) lazy.IBk -K_1_-W_0 -7.902008594385911 8E18
(4) trees.J48 -C_0.25_-M_2 -2.1773316839364444 8E17
Select items
Key_Dataset
Key_Run
Key_Fold
Key_Scheme
Key_Scheme_options
Key_Scheme_version_ID
Date_time
Number_of_training_instances
Select
Cancel
実験条件の設定


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 40 results
File...
Database...
Experiment
Configure test
Row key fields
Select keys...
Column key fields
Select keys...
Comparison field
Percent_correct
Significance
0.05
Test base
Select base...
Show std. deviations
Perform test
Save output
Result list
02:03:42 - Available resultsets
02:08:36 - Available resultsets
02:08:42 - Percent_correct - bayes.NaiveBayes
Test output
Analysing: Percent_correct
Datasets: 1
Resultsets: 4
Confidence: 0.05 (two tailed)
Date: 10/6/04 2:08 AM
Dataset
(1) bayes.Na
(2) funct
(3) lazy.
(4) tree
TwentyNewsGroups
(10) 68.33
70.00
53.33
48.33
(0/ 0)
(0/1/0)
(0/1/0)
(0/1/0)
Skipped:
Key:
(1) bayes.NaiveBayes
(2) functions.SMO
(3) lazy.IBk
(4) trees.J48
精度
SVM が最良
決定木が最悪
SVM は統計的に Naïve Bayes よりよい
決定木は統計的に Naïve Bayes より悪い
実験条件の設定


Microsoft Excel - results.csv
実験: Excel
結果は CSV ファイルに書き出すことができ，それは Excel に読み込むことができる