

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>

1

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

2

WEKA: 好奇心旺盛な飛べない鳥



Copyright: Martin Kramer (mkramer@wxs.nl), University of Waikato, New Zealand

Slide adapted from Eibe Frank's

3

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

4

WEKA: The Software Toolkit

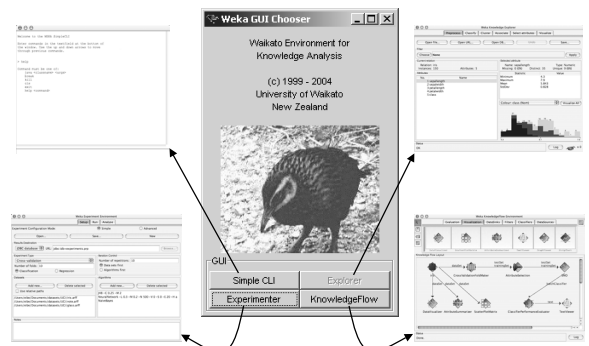
<http://www.cs.waikato.ac.nz/ml/weka>

- 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ース (データ可視化含む)
 - 学習アルゴリズムの比較

Slide adapted from Eibe Frank's

5

WEKA GUI Chooser `java -Xmx1000M -jar weka.jar`



Slide adapted from Eibe Frank's

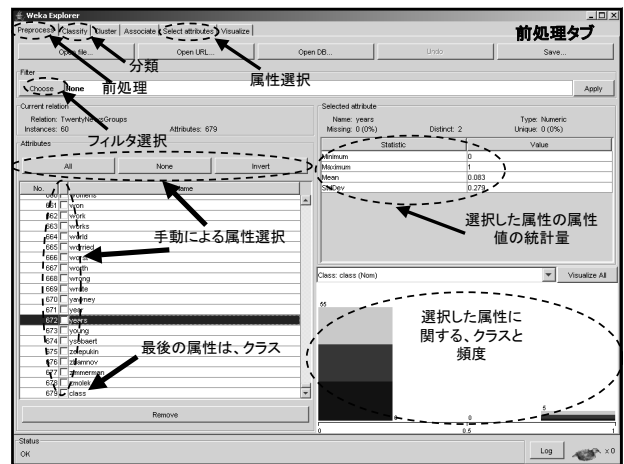
6

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.

Slide adapted from Elibe Frank's

7

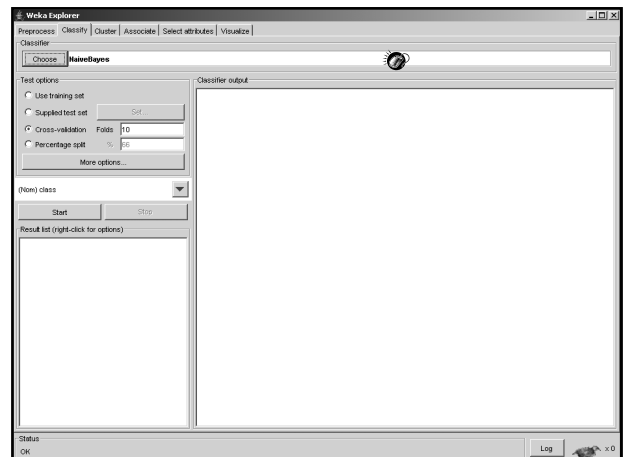
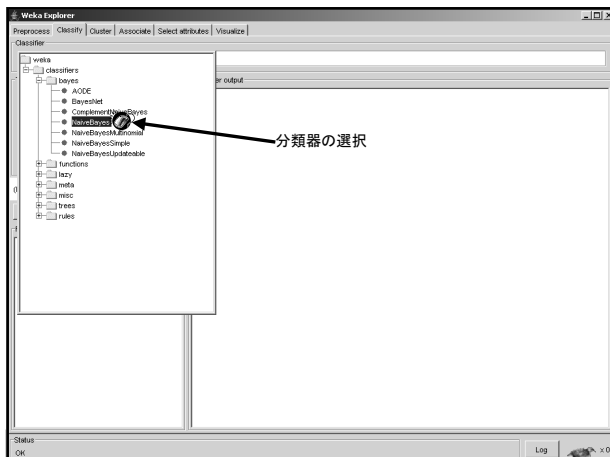
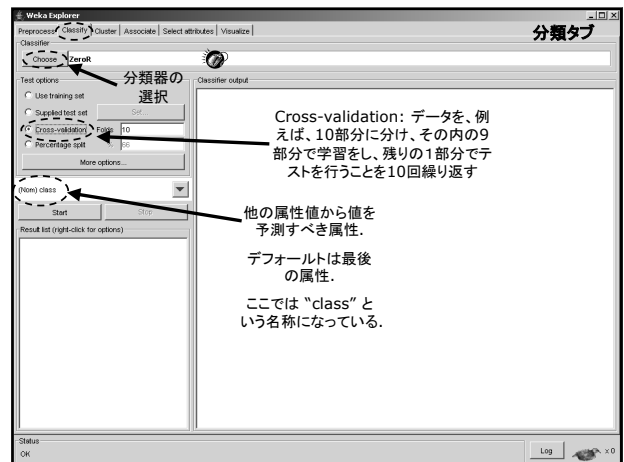


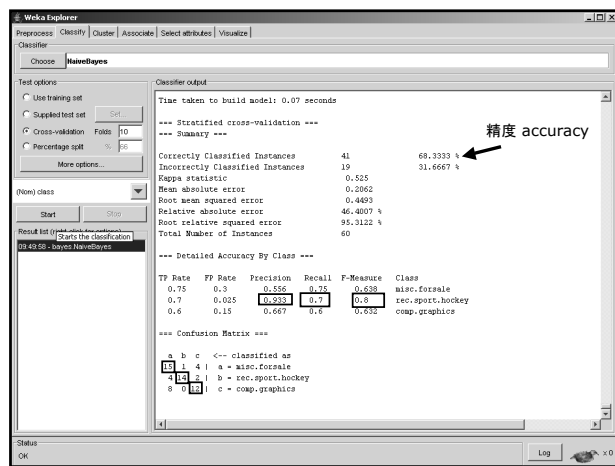
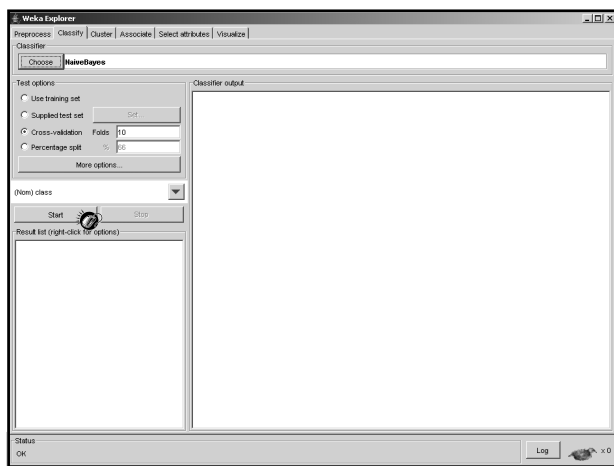
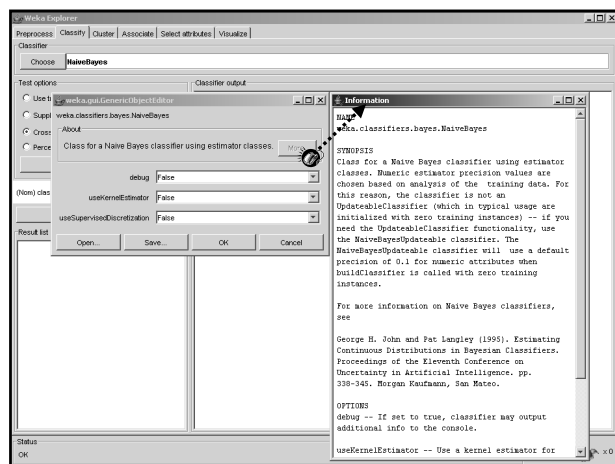
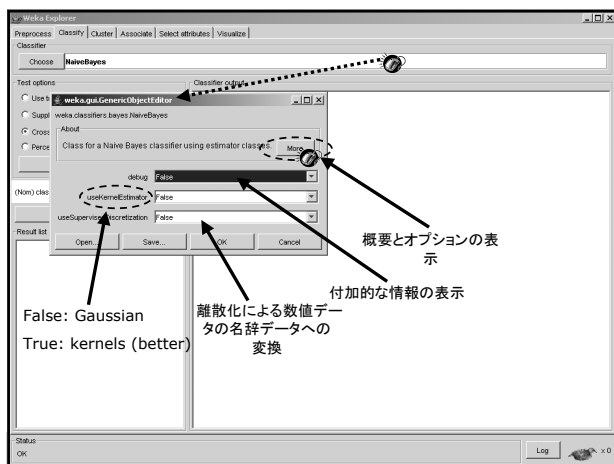
Explorer: 分類

- 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.

Slide adapted from Elibe Frank's

9

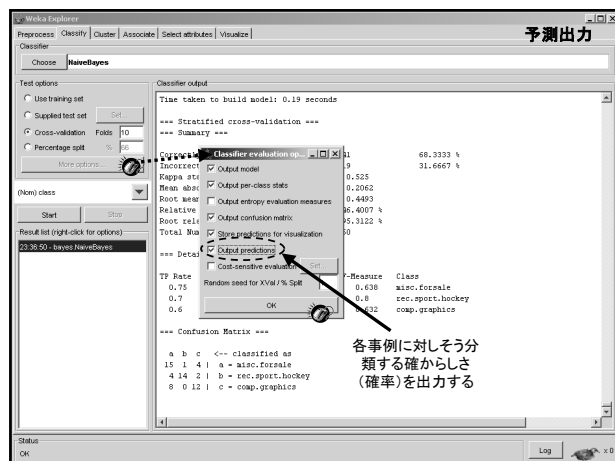




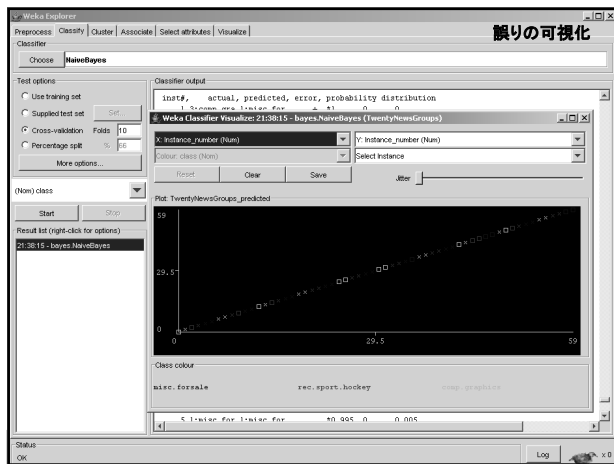
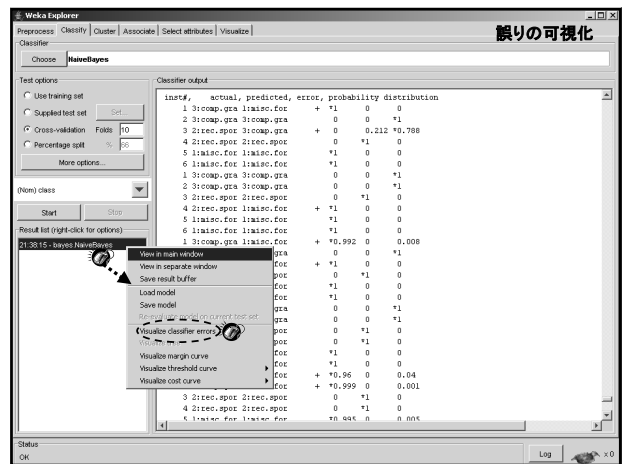
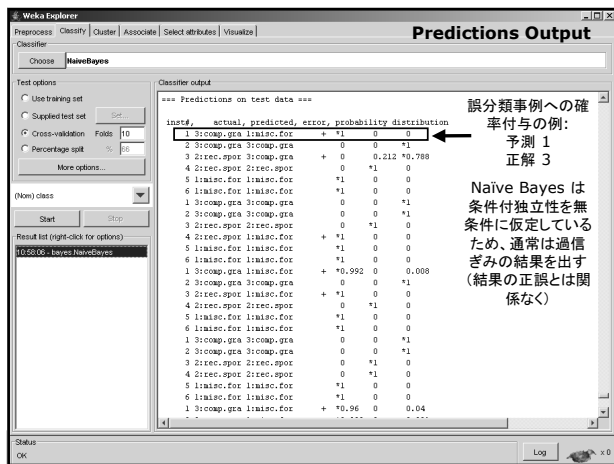
Confusion matrix

- 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: $d/(c+d) \Rightarrow R$
 - precision: $d/(b+d) \Rightarrow P$
 - 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



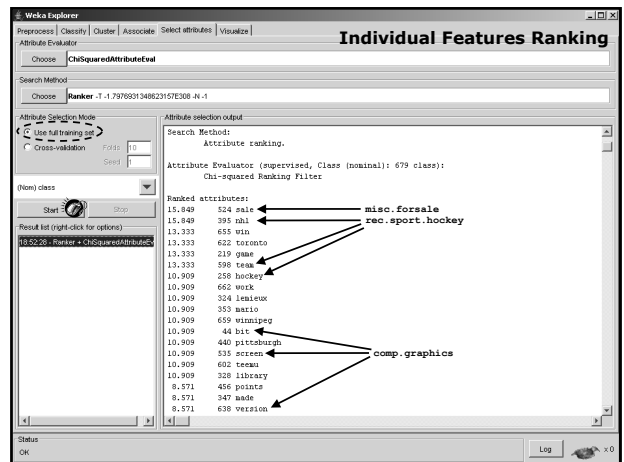
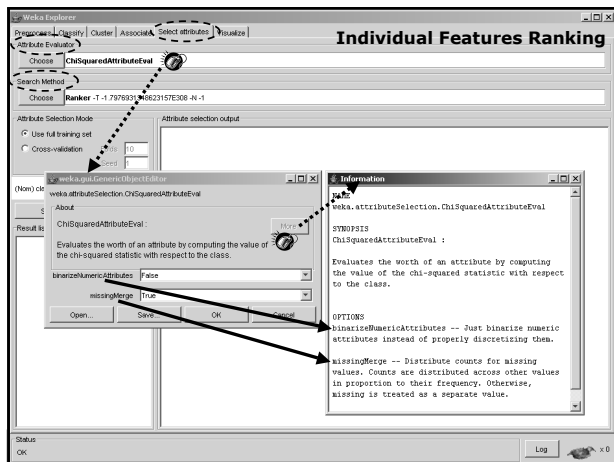
各事例に対しよう分
類する確からしさ
(確率)を出力する

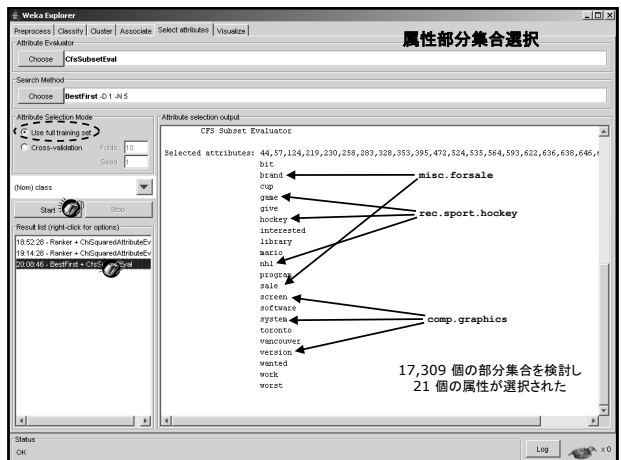
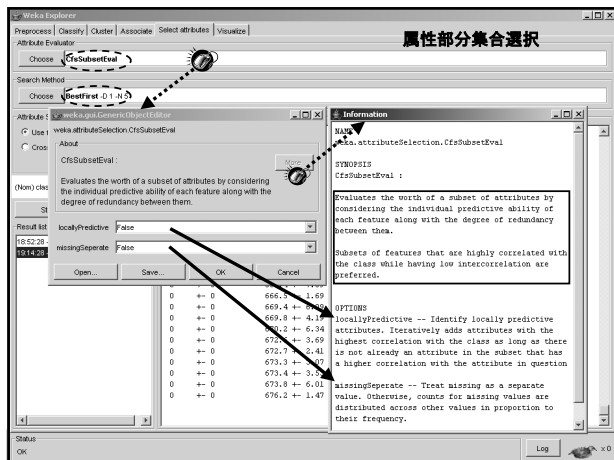
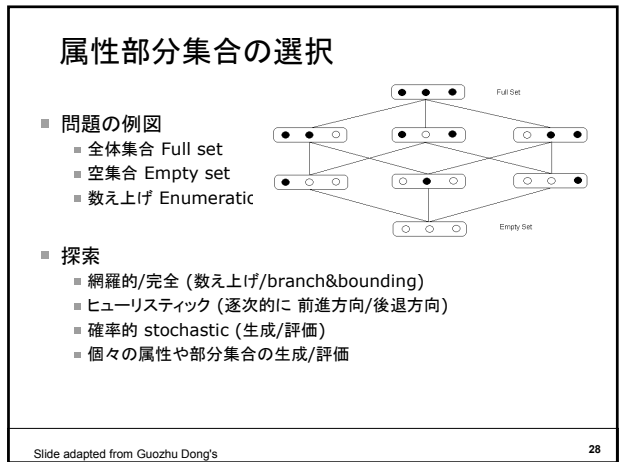
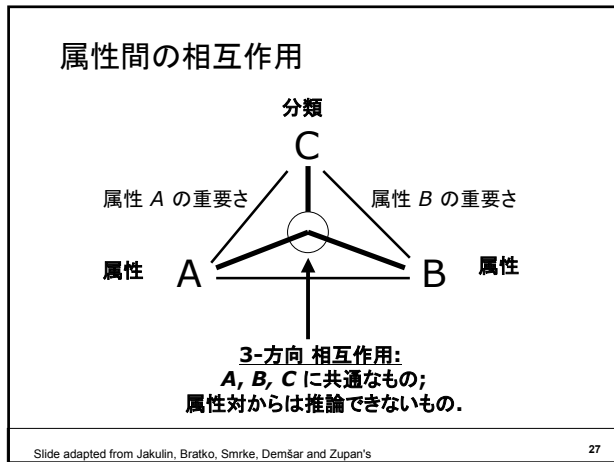
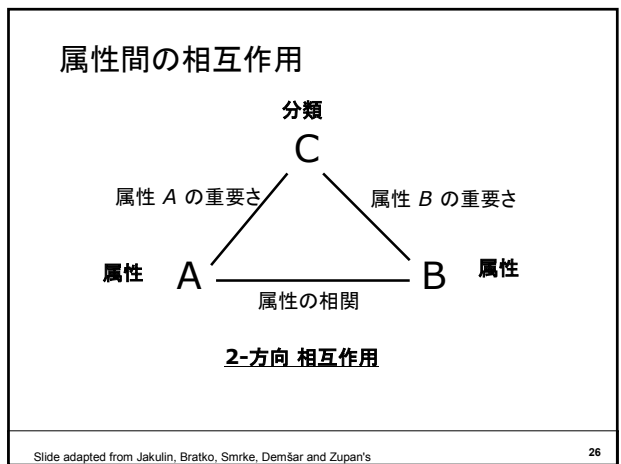
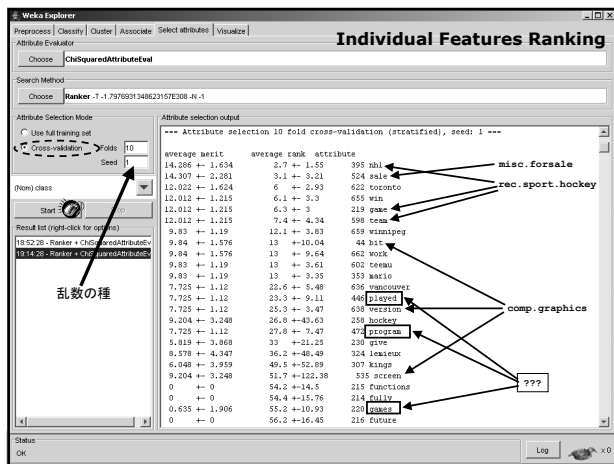


Explorer: 属性選択

- 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

Slide adapted from Elibe Frank's





選択した属性の保存

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Choose **ClassifierSubsetEval** -> Weka classifiers bayes, naiveBayes, 1-H "Click to set hold out or test instances" --

Search Method Choose **BestFirst** -> D 1-N 5

Attribute Selection Mode ☒ Use full training set ☐ Cross-validation

Attribute selection output CFS Subset Evaluator

Selected attributes: 44,57,124,219,220,258,283,328,353,395,472,524,535,564,593,622,636,638,646,651,652,653,654,655,656,657,658,659,660,661,662,663,664,665,666,667,668,669,670,671,672,673,674,675,676,677,678,679,680,681,682,683,684,685,686,687,688,689,690,691,692,693,694,695,696,697,698,699,700,701,702,703,704,705,706,707,708,709,710,711,712,713,714,715,716,717,718,719,720,721,722,723,724,725,726,727,728,729,730,731,732,733,734,735,736,737,738,739,740,741,742,743,744,745,746,747,748,749,750,751,752,753,754,755,756,757,758,759,760,761,762,763,764,765,766,767,768,769,770,771,772,773,774,775,776,777,778,779,780,781,782,783,784,785,786,787,788,789,790,791,792,793,794,795,796,797,798,799,800,801,802,803,804,805,806,807,808,809,810,811,812,813,814,815,816,817,818,819,820,821,822,823,824,825,826,827,828,829,830,831,832,833,834,835,836,837,838,839,840,841,842,843,844,845,846,847,848,849,850,851,852,853,854,855,856,857,858,859,860,861,862,863,864,865,866,867,868,869,870,871,872,873,874,875,876,877,878,879,880,881,882,883,884,885,886,887,888,889,890,891,892,893,894,895,896,897,898,899,900,901,902,903,904,905,906,907,908,909,910,911,912,913,914,915,916,917,918,919,920,921,922,923,924,925,926,927,928,929,930,931,932,933,934,935,936,937,938,939,940,941,942,943,944,945,946,947,948,949,950,951,952,953,954,955,956,957,958,959,960,961,962,963,964,965,966,967,968,969,970,971,972,973,974,975,976,977,978,979,980,981,982,983,984,985,986,987,988,989,990,991,992,993,994,995,996,997,998,999,1000

Result list (right-click for options)

18:22:28 - Filterer - CHCappedAttributeS...

20:08:46 - BestFirst - CHC...

New instance window
View in separate window
Save result buffer
Visualize reduced data

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。使いやすといえない...

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

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Save...

Filter Choose **AttributeSelection** -> "S" weka attributeSelection BestFirst -> D 1-N 5

Current relation: supervised

Selected attribute: Name: star, Missing: 0 (0%), Distinct: 3, Type: Numeric, Unique: 1 (2%)

Statistic	Value
Minimum	0
Maximum	2
Mean	0.667
StdDev	0.312

Class: class (Nom)

Visualize All

Status: OK

前処理における属性選択

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Save...

Filter Choose **AttributeSelection** -> "S" weka attributeSelection BestFirst -> D 1-N 5

Current relation: supervised

Selected attribute: Name: star, Missing: 0 (0%), Distinct: 3, Type: Numeric, Unique: 1 (2%)

Statistic	Value
Minimum	0
Maximum	2
Mean	0.667
StdDev	0.312

Class: class (Nom)

Visualize All

Status: OK

前処理における属性選択

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Save...

Filter Choose **AttributeSelection** -> "S" weka attributeSelection BestFirst -> D 1-N 5

Current relation: supervised

Selected attribute: Name: star, Missing: 0 (0%), Distinct: 3, Type: Numeric, Unique: 1 (2%)

Statistic	Value
Minimum	0
Maximum	2
Mean	0.667
StdDev	0.312

Class: class (Nom)

Visualize All

Status: OK

前処理における属性選択

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Save...

Filter Choose **AttributeSelection** -> "S" weka attributeSelection BestFirst -> D 1-N 5

Current relation: supervised

Selected attribute: Name: star, Missing: 0 (0%), Distinct: 3, Type: Numeric, Unique: 1 (2%)

Statistic	Value
Minimum	0
Maximum	2
Mean	0.667
StdDev	0.312

Class: class (Nom)

Visualize All

Status: OK

この21属性に対し Naïve Bayes を適用

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier Choose **NaiveBayes**

Test options ☒ Use training set ☐ Supplied test set ☐ Cross-validation ☐ Percentage split

Classifier output

Time taken to build model: 0.15 seconds

--- Stratified cross-validation ---

Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances
51	9	0.775	0.118	0.285	26.541 %	66.458 %	60

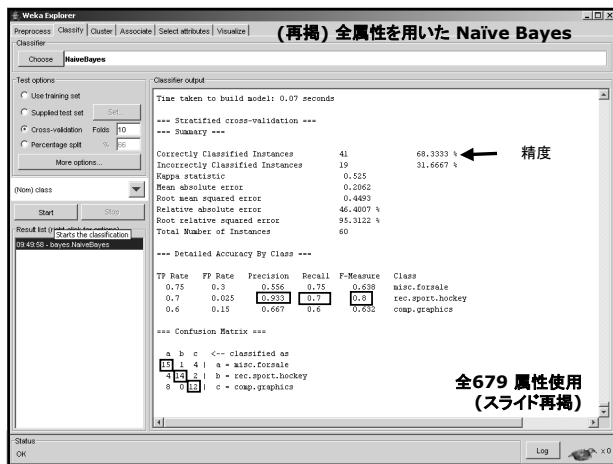
--- Detailed Accuracy by Class ---

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.225	0.69	1	0.816	0.816	misc,forensic
0.75	0	0.75	0.857	0.857	rec.sport.hockey
0.8	0	1	0.8	0.889	comp.graphics

--- Confusion Matrix ---

a \ b	c <= classified as	a = misc,forensic	a = rec.sport.hockey	a = comp.graphics
b 0	51	0	0	0
b 1	0	0	0	0
b 2	0	0	0	0

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 は見つけれなかった (5 種類もの実装されているのに).

38

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

39

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

40

