



フリーウェアWekaによる DMアプリケーションの構築

阿部秀尚, 山口高平
静岡大学

hidenao@ks.cs.inf.shizuoka.ac.jp
yamaguti@cs.inf.shizuoka.ac.jp

2004/3/30

第19回 AIシンポジウム(SIG-J)

1

発表内容

■ Wekaの紹介

- Wekaの基本情報と構成
- DMプロセスとWekaの機能

■ 慢性ウィルス性肝炎データマイニングへのWekaの適用

- 属性構築: クラスタリングによる代表パターンの抽出
- マイニング: GUIを使った決定木の実行

2004/3/30

第19回 AIシンポジウム(SIG-J)

2

Wekaとは？

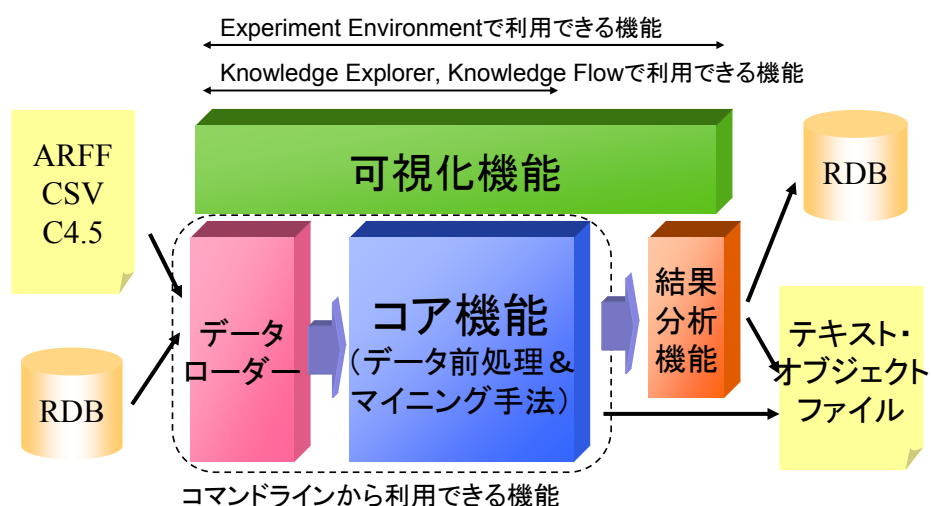
- 世界で最も使われているフリーのデータマイニングツール (by KDNuggets.com)
- オープンソース開発手法で開発が進められている
 - ワイカト大学(ニュージーランド)が中心に開発
 - Webを通じて無償で入手可能
 - <http://www.cs.waikato.ac.nz/ml/weka/>
 - 誰もがソースコードにアクセスでき、改変・再配布可能
- 最新版はVersion 3.4.1
- Java言語により実装(=マルチプラットフォーム)
 - Windows/MacOS X/JAR(ZIP)の各形式のパッケージにより配布

2004/3/30

第19回 AIシンポジウム(SIG-J)

3

Wekaの構成

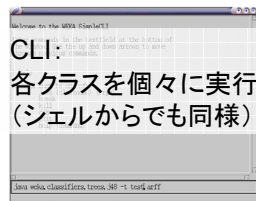


2004/3/30

第19回 AIシンポジウム(SIG-J)

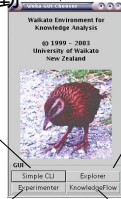
5

Wekaのインタフェース

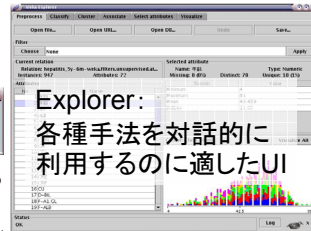
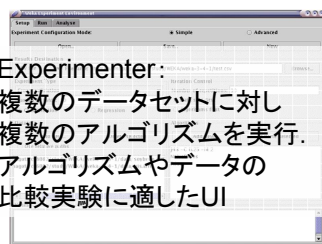


CLI:
各クラスを個々に実行
(シェルからでも同様)

コマンドラインやメニューから
GUIを起動



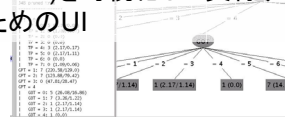
Experimenter:
複数のデータセットに対し
複数のアルゴリズムを実行.
アルゴリズムやデータの
比較実験に適したUI



Explorer:
各種手法を対話的に
利用するのに適したUI



Knowledge Flow:
DMアプリケーションの流れ
(Flow)を可視化して実行する
ためのUI



2004/3/30

第19回 AIシンポジウム(SIG-J)

6

Wekaの入出力

■ 入力

- ☐ ARFFファイル
- ☐ CSV (Comma Separated Values)形式ファイル
- ☐ C4.5形式ファイル
- ☐ RDB (Relational Database)

■ 出力

- ☐ テキスト・オブジェクトファイルによる実行結果
- ☐ 可視化(グラフ, 2次元プロット)による実行結果

2004/3/30

第19回 AIシンポジウム(SIG-J)

7

ARFFの概要

```
@relation weather
@attribute outlook string
@attribute temperature numeric
@attribute humidity integer
@attribute windy {true, false}
@attribute time_s date "yyMMdd HH:mm"
@attribute play {yes, no}
```

```
@data
sunny,85,85,FALSE,010822 10:00,no
overcast,83,86,FALSE,010911 11:00,yes
rainy,70,?,FALSE,010921 10:30,yes
```

データセットの名前

属性名を縦に列挙する

数値属性を示す
(実数:real, 整数:integer)

名義属性は属性値
をカンマ区切りで示す

日付はdate+フォーマットで
定義する

クラスも他の名義属性と
同様に記述(実行時に指定)

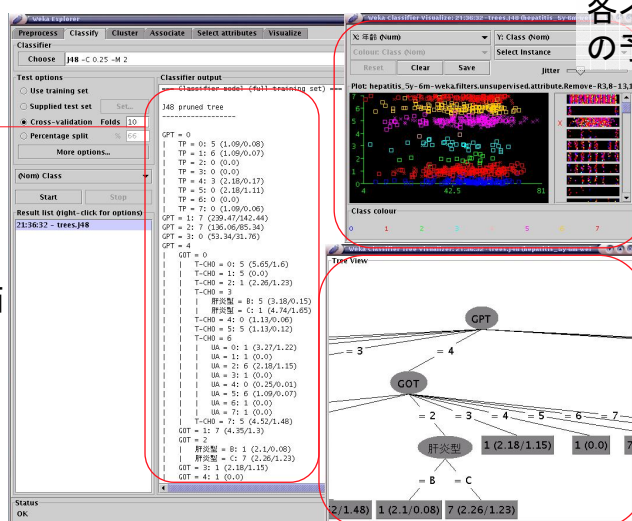
@data以降CSV形式の
データ

欠損値は'?'で表す

8

Wekaの出力: Explorerでの例

テキスト
による
実行結果.
以下に,
正解率など
の精度評価
が続く



各インスタンス の予測結果

グラフに
よる表示

9

Wekaの適用分野

- テキストマイニング
 - テキストクラスタリング
 - テキスト分類
- Webサーバログマイニング
- 各種表形式データに対するマイニング
 - RDBとの連携
 - POSデータ
 - DNAマイクロアレイ
 - 慢性肝炎データセット etc.

注: Wekalist, weka-jpで多く見られる分野

2004/3/30

第19回 AIシンポジウム(SIG-J)

10

DMプロセスとWekaの機能(1)

データの理解

データサイズ
外れ値、欠損値、
設定目標とデータの整合性



データの準備

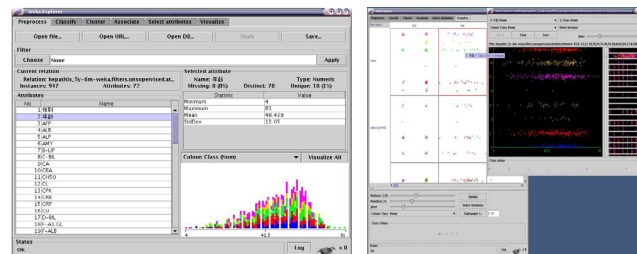
欠損値処理、離散化
サンプリング



モデル選択

クラスタリング、相関ルール、
決定木学習、ベイズ則による
分類器、線形回帰、
ニューラルネット...

インターフェース: Explorer, Knowledge Flow
機能: Explorerの"preprocess"パネルでの表示
Explorerの"Visualize"パネルでの表示



提供されていない機能: 折れ線グラフ, 棒グラフ, 多次元グラフ

2004/3/30

第19回 AIシンポジウム(SIG-J)

11

DMプロセスとWekaの機能(2)

データの理解

データサイズ
外れ値, 欠損値,
設定目標とデータの整合性

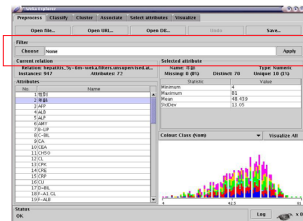
データの準備

欠損値処理, 離散化
サンプリング, 属性選択

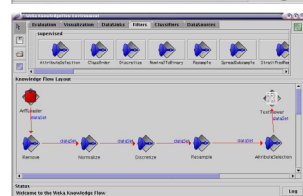
モデル選択

クラスタリング, 相関ルール,
決定木学習, ベイズ則による
分類器, 線形回帰,
ニューラルネット...

インタフェース: Explorer, KnowledgeFlow
コマンドライン



Explorerでは"preprocessパネル"
の"filters"から適用するフィルタを
選択して実行する(Undo可能)



Knowledge Flowでは"Filtersパネル"
からデータセットに適用する
フィルタを選択し, レイアウトする

提供されていない機能: 集計演算, 簡単な統計処理, データ結合

2004/3/30

第19回 AIシンポジウム(SIG-J)

12

Wekaのフィルタ(一部) 全種類: 39種類

- Supervised: 7種類

- Unsupervised: 35種類

Supervisedフィルタ	
AttributeSelection	属性選択アルゴリズムを適用して属性を選択する
Discretize	数値属性を指定した数の名義値に離散化する
NominalToBinary	名義値を複数の二値属性に変換
Resample	インスタンスをランダムにサンプリングする
SpreadSubSample	分散を考慮したランダムサンプリング
StratifiedRemoveFolds	交差検定で用いる訓練またはテストデータセットを作成
Unsupervisedフィルタ(Attribute)	
Add	指定した列に属性を追加する
Remove	指定した属性を除去する
AddCluster	インスタンスが属するクラス番号を属性として加える
Normalize	指定した数値属性の正規化を行う
Standardize	指定した数値属性の標準化を行う

2004/3/30

第19回 AIシンポジウム(SIG-J)

13

Wekaのフィルタ(一部)

Unsupervisedフィルタ(Attribute)	
ClusterMembersip	インスタンスが属するクラスタの確率を属性として加える
MakeIndicator	ある属性がある値であるか／ないかを属性として加える
ReplaceMissingValue	欠損値を指定した値に置き換える
NumericTransform	数値属性の値に指定した数式を適用する
StringToNominal	"string"と指定した名義属性のヘッダを名義値リストにする
SwapValues	名義属性の二つの値を交換する
Unsupervisedフィルタ(Instance)	
Randomize	データセット内のインスタンスを攪拌する
RemoveFolds	指定した番号の交差検定データセットを除去する
SparseToNonSparse	スパースなデータセットをスパースで無く加工する

2004/3/30

第19回 AIシンポジウム(SIG-J)

14

属性選択アルゴリズム(フィルタアプローチ)

フィルタアプローチ(8種類:探索法はRankerに固定)	
ReliefAttributeEval	Relief.Fアルゴリズムによる属性評価
InfoGainAttributeEval	情報利得による属性評価
GainRatioAttributeEval	情報利得比による属性評価
SymmetricalAttributeEval	クラスとの類似度／非類似度による属性評価
OneRAttributeEval	OneRアルゴリズムで使われる属性かどうかを評価
ChiSquareAttributeEval	χ^2 指標による属性評価
PrincipleComponents	主成分分析による属性評価
SVMAttributeEval	SVMによる属性評価

2004/3/30

第19回 AIシンポジウム(SIG-J)

15

属性選択アルゴリズム(ラッパーアプローチ)

ラッパーアプローチ(評価法4種類)	
WrapperSubsetEval	交差検定による分類精度の評価
CfsSubsetEval	属性個々の分類性能による評価
ClassifierSubsetEval	分類器による精度の評価
ConsistencySubsetEval	各クラス値と属性値の関連度に基づく評価
ラッパーアプローチ(探索法7種類)	
BestFirst	前向き・後向き・両向き最良優先探索
ForwardSelection	前向き選択による貪欲探索
RaceSearch	複数の探索子によるレースによる探索
GeneticSearch	遺伝的アルゴリズムによる探索
RandomSearch	ランダム探索
ExhaustiveSearch	全探索
RankSearch	属性につけられた評価値のランクに基づく探索

2004/3/30

第19回 AIシンポジウム(SIG-J)

16

DMプロセスとWekaの機能(3)

データの理解

データサイズ
外れ値, 欠損値,
設定目標とデータの整合性

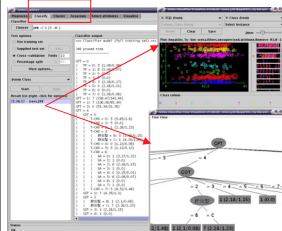
データの準備

欠損値処理, 離散化
サンプリング, 属性選択

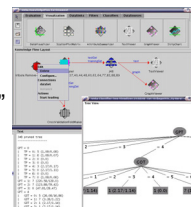
モデル選択

クラスタリング, 相関ルール,
決定木学習, ペイズ則による
分類器, 線形回帰,
ニューラルネット...

インタフェース: Explorer, Knowledge Flow
コマンドライン, Experimenter



Explorer(左図)では,
"Classify", "Cluster", "Associate"の各
パネルでアルゴリズムを選択して実行.
結果を可視化することが可能



Knowledge Flow(右図)では, "Classifiers"
パネルから分類・数値予測アルゴリズム
を選択し, 実行する. 結果は"Visualize"
パネルから可視化機能を選択し, 可視化.

提供されていない主な機能: 遺伝的アルゴリズム, C4.5rules

2004/3/30

第19回 AIシンポジウム(SIG-J)

17

分類学習・数値予測アルゴリズム(1)

ベイズ則に基づくアルゴリズム(8種類)	
AODE	
BayesNetk2, BayesNetB	K2/Bアルゴリズムによるベ이지アンネットワーク生成
ComplementNaiveBayes	Complement Classを使ったナイーブベイズ
NaiveBayes, NaiveBayesSimple	単純な確率モデルによる分類
NaiveBayesMultiNominal	単純な確率モデルによる数値予測
NaiveBayesUpdatable	ナイーブベイズの更新可能版
関数型アルゴリズム(12種類)	
SimpleLinearRegression, LinearRegression	線形回帰
SimpleLogistic, Logistic	ロジスティック回帰
SMO, SMOreg	サポートベクターマシンによる分類・数値予測
LeastMedSq	Least median squaredによる線形回帰の改良版
PaceRegression	Pace Regressionによる回帰モデルの導出
MultiLayerPerceptron	ニューラルネットワーク
VotedPerceptron	パーセプトロンの投票による分類器の学習
RBFnetwork	K-meansと線形回帰・ロジスティック回帰による分類器の生成
Winnow	WinnowとBalanced Winnowアルゴリズム

2004/3/30

第19回 AIシンポジウム(SIG-J)

18

分類学習・数値予測アルゴリズム(2)

インスタンスに基づく分類器(5種類)	
IB1, IBk	K-Nearest Neighboursアルゴリズムによる分類器
KStar	K*アルゴリズムによる分類器
LBR	怠惰なNaiveBayesによる分類器
LWL	局所重みによるNaiveBayes
決定木・モデル木学習アルゴリズム(10種類)	
Id3	ID3アルゴリズムによる決定木
J48	C4.5アルゴリズムによる決定木
DecisionStamp	1レベル決定木
ADTree	Boostingのような手法による決定木の学習
RandomForest	ランダムに生成した複数の決定木による分類
RandomTree	ランダムに生成した決定木による分類
REPTree	高速な決定木・モデル木の学習
LMT	ロジスティック回帰式を葉に持つモデル木
M5P	線形回帰式を葉に持つモデル木
UserClassifier	対話的に決定木・モデル木を構築

2004/3/30

第19回 AIシンポジウム(SIG-J)

19

分類学習・数値予測アルゴリズム(3)

ルールベース分類器(10種類)	
PART	Partルール学習
Prism	PRISMアルゴリズムによるルール学習
OneR	1つの条件節のみによるルール集合生成
ConjunctiveRules	情報利得による分類・数値予測ルール集合学習
DecisionTable	決定表の作成
NNge	K-Nearest Neighborsのような方法でルールを学習
M5Rules	M5'モデル木によるルール集合
Ridor	Ripple-Downによるルール集合学習
JRip	記述長を制限したルール集合学習
ZeroR	デフォルトクラスによって分類する
その他のアルゴリズム(3種類)	
FLR	
HyperPipes	
VFI	

2004/3/30

第19回 AIシンポジウム(SIG-J)

20

メタスキーム(全23種類, 一部)

分類学習・数値予測アルゴリズムを内包し, 分類・予測精度を向上する

AdaBoostM1, LogitBoost	Boostingスキーム
Bagging	Baggingスキーム
Stacking, StackingC	Stackingスキーム
Vote	複数分類器集合による投票
ClassificationViaRegression	回帰式による分類
MultiClassClassifier	複数クラス値の問題を複数の2値クラス問題にして分類
etc...	

2004/3/30

第19回 AIシンポジウム(SIG-J)

21

クラスタリング・相関ルール学習 アルゴリズム

クラスタリング	
SimpleKMeans	距離尺度に基づく非階層的クラスタリング
EM	EMアルゴリズムを用いた非階層的クラスタリング
CobWeb	CobWebアルゴリズムによるクラスタリング
FarthestFirst	Farthest Firstアルゴリズムによるクラスタリング
MakeDensityBasedCluster	1つのクラスタリングアルゴリズムを複数回実行し、中間的なクラスターを得る
相関ルール学習	
Apriori	アプリアリアルゴリズムによる相関ルール学習
Tertius	Tertiusによる相関ルール学習

2004/3/30

第19回 AIシンポジウム(SIG-J)

22

発表内容

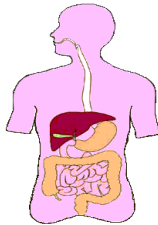
- Wekaの紹介
- 慢性ウィルス性肝炎データマイニングへのWekaの適用
 - 属性構築: クラスタリングによる代表パターンの抽出
 - マイニング: GUIを使った決定木の実行

2004/3/30

第19回 AIシンポジウム(SIG-J)

23

慢性ウィルス性肝炎データセット



慢性ウィルス性肝炎の
検査データセット

検査項目数: 957種類
患者数: 771名
レコード数: 約160万レコード

患者プロフィール: 患者ID, 性別, 生年月日

検体検査メタデータ: 血液検査, 尿検査の説明

検体検査の結果: 血液検査, 尿検査の結果(時系列データ)

肝生検の結果: 外科的な検査(病状を示す名義値)

インタフェロン投与情報: 特効薬の投与時期, 期間

2004/3/30

第19回 AIシンポジウム(SIG-J)

24

慢性ウィルス性肝炎データセット(一部)

[データベースリンク](#)
[Temp1 - Mozilla](#)

[ファイル\(E\)](#)
[編集\(E\)](#)
[表示\(V\)](#)
[Go](#)
[Bookmarks](#)
[ツール\(T\)](#)
[ウィンドウ\(W\)](#)
[ヘルプ\(H\)](#)

Menu	mid	test_date	item	results
● 一般DB操作 ○ 任意のテーブルの内容を表示する ○ 選択条件を式によって選択を行う ○ クロス集計を実行する ○ テーブルを結合(outer join)する	1	1981-02-19	ALB	
	1	1981-02-19	ALP	
	1	1981-02-19	CRE	
	1	1981-02-19	D-BIL	
	1	1981-02-19	F-A/G	
	1	1981-02-19	F-A1, GL	
	1	1981-02-19	F-A2, GL	
	1	1981-02-19	F-ALB	
	1	1981-02-19	F-B, GL	
	1	1981-02-19	G, GL	
● 時系列DB操作 ○ サブシーケンスを取り出す ○ サブシーケンス内のデータ問題を一定にする ○ データ間の線形補間を行う	1	1981-02-19	GOT	
	1	1981-02-19	GPT	
	1	1981-02-19	I-BIL	
	1	1981-02-19	LDH	
	1	1981-02-19	T-BIL	
	1	1981-02-19	T-CHO	
	1	1981-02-19	TP	
	1	1981-02-19	TTT	
	1	1981-02-19	UA	
	1	1981-02-19	UN	
● 時系列DMメソッド ○ 時系列クラスタを得る ○ 時系列クラスタの内容を表示	1	1981-02-19	ZTT	
	1	1981-03-26	ALB	
	1	1981-03-26	ALP	
	1	1981-03-26	CHE	
● ログアウト	1	1981-03-26	CRE	

[戻る](#)
[印刷](#)
[ヘルプ](#)
[検索](#)
[設定](#)
[ログアウト](#)

2004/3/30

第19回 AIシンポジウム(SIG-J)

25

システムの設計方針

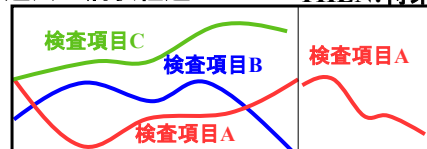


医療では、1つの検査結果だけではなく、
検査結果の時系列傾向で病状を把握する。
過去の病状から、将来の病状(予後)を
把握する必要がある。

属性: 複数の検査項目における、検査結果の**パターン**
クラス: 重要な1つの検査項目における、検査結果の**パターン**
システムは、
現在の**パターン**の**組合せ** ⇒ 予後の**パターン**の**推定**を行う。

IF:過去の病状経過

THEN:将来の病状経過



2004/3/30

第19回 AIシンポジウム(SIG-J)

26

慢性ウイルス性肝炎データセットからの 時系列ルール発見支援

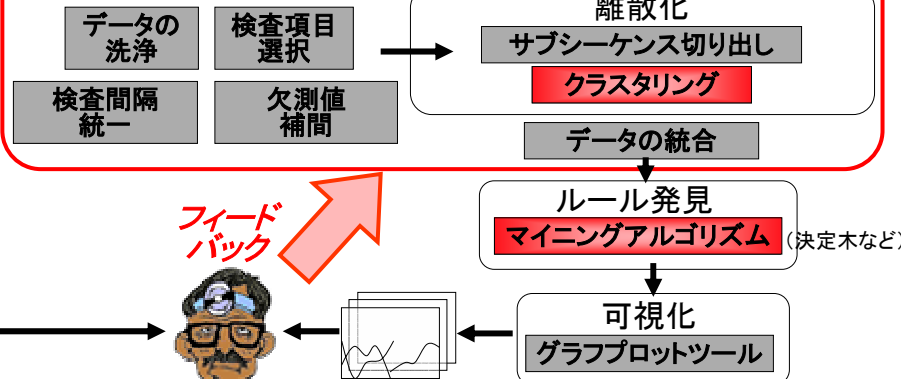
慢性ウイルス性肝炎
データセット

ドメイン依存

ドメイン独立

低レベルの前処理

高レベルの前処理

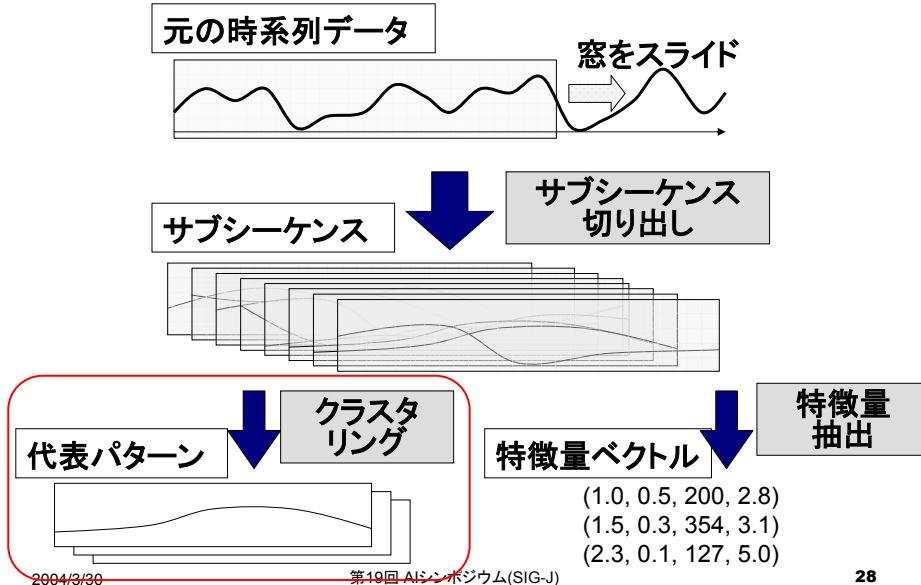


2004/3/30

第19回 AIシンポジウム(SIG-J)

27

時系列データ離散化

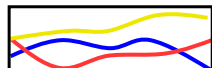


各検査項目の代表パターンを属性とした慢性肝炎データマイニングの例

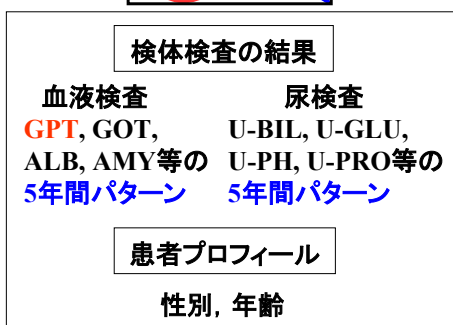
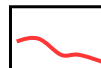


GPTは病状を把握するのに重要な検査項目。
病状の信頼性を確保する最低の事例数は10.

属性



クラス



予測

GPTの
6ヶ月パターン



2004/3/30

第19回 AIシンポジウム(SIG-J)

29

Knowledge Flowによるデモ

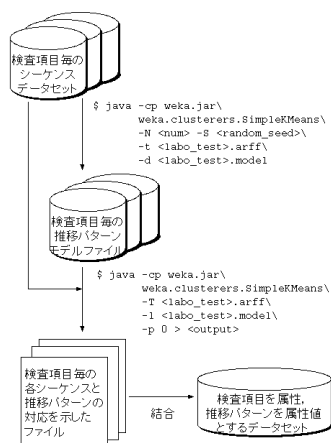
- 入力データセットは各患者の約5年間 (61×28日)のGPT値
- 出力はK-meansによるクラスタリングの結果
- AddClusterフィルターによるクラスタ付加を実行

2004/3/30

第19回 AIシンポジウム(SIG-J)

30

Wekaとスクリプト言語の連携による 代表パターンの抽出



```
foreach 検査項目{
  system ("java -cp weka.jar
    weka.clusterers.SimpleKMeans
    -N 8 -S 20 -t 検査項目.arff
    -d 検査項目.model");
  system ("java -cp weka.jar
    weka.clusterers.SimpleKMeans
    -T 検査項目.arff
    -l 検査項目.model
    -p 0 > 検査項目.prd");
}
Join すべての検査項目.prd
```

2004/3/30

第19回 AIシンポジウム(SIG-J)

31

Explorerによるデモ

- 入力データセットは
 - 属性: 約5年(61×28日)の検査結果の代表パターン, 年齢, 性別, 肝炎型
 - クラス: 約5年経過後, 6ヶ月のGPTの代表パターン
- 実行するアルゴリズム
 - J4.8: C4.5決定木学習アルゴリズムのJava実装版

2004/3/30

第19回 AIシンポジウム(SIG-J)

32

まとめ

- Wekaはデータマイニングプロセスに対し, 多くの手法を提供
 - 1入力に対して1出力が基本
- DMプロジェクトの立ち上げ期において強力なツールとなる
 - データの前処理では, RDBMS(SQLによる処理)やMUSASHI(意味を考慮した処理)などの外部ツールと連携
 - 複数のデータセットに対応するため, 必要に応じて周辺処理を追加

2004/3/30

第19回 AIシンポジウム(SIG-J)

33

最後に

- 公式Webページ&メーリングリスト
<http://www.cs.waikato.ac.jp/ml/weka/>
Wekalist@list.scms.waikato.ac.nz
- 書籍
I.H. Witten, E. Frank: “Data Mining”, MORGAN KAUFMANN
ISBN 1-55860-552-5
- 日本語による情報(非公式)
<http://panda.cs.inf.shizuoka.ac.jp/~hidenao/work/weka/>
weka-jp@ks.cs.inf.shizuoka.ac.jp
- 14:30よりの展示会に出展