データ解析実践演習

Wekaによるデータ分析と応用

データ解析の手順

- 収集したデータはそれぞれ形式が違う
- データを整形し，特徴を見極める.
- 例えば，Wekaに付属のデータは．．．

データ名	内容	特徴	正解情報
Breast-canser	乳がんの再発	ラベル	クラス（2値）
Contact-lenses	コンタクトレンズの推薦	ラベル	クラス（3値）
Cpu	CPUの性能評価	数値	数値
Credit-g	融資の審査	混合	クラス（2値）
diabetes	糖尿病の検査	数値	クラス（2値）
Iris	アヤメの分類	数値	クラス（3値）
Reuters-corn	記事分類	テキスト	クラス（2値）
Supermarket	スーパーの購買記録	ラベル	なし
Weather.nominal	ゴルフをする条件	ラベル	クラス（2値）
Weather.numeric	ゴルフをする条件	混合	クラス（2値）

ARFFファイルの復習

- CSVデータ　＋　ヘッダ情報　＝　ARFF形式
- ARFF形式（attribute-relation file format)

```
% 1. Title: Iris Plants Database
@RELATION iris
@ATTRIBUTE sepallength REAL
@ATTRIBUTE sepalwidth  REAL
@ATTRIBUTE petallength        REAL
@ATTRIBUTE petalwidth  REAL
@ATTRIBUTE class    {Iris-setosa,Iris-versicolor,Iris-virginica}
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
```

コメント

データセット名

特徴名と型

データ開始

学習データは1行に1事例

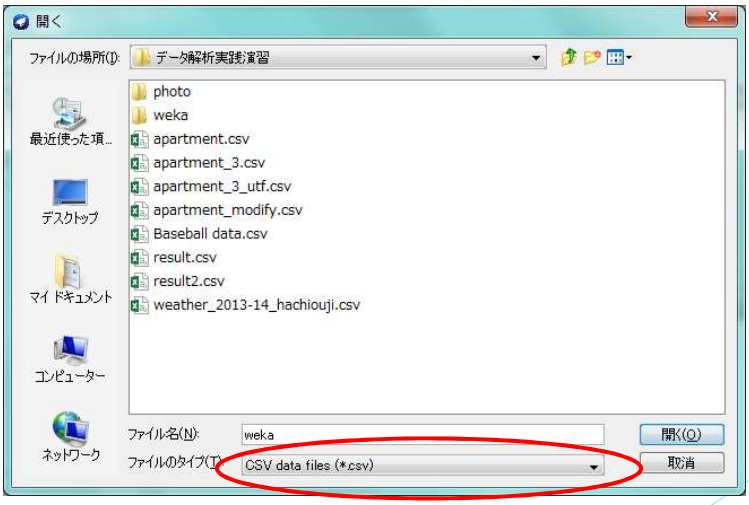
CSVファイルからARFFファイルへ

- Attributeを自分で書くのは大変.
- WekaはCSV形式のファイルを読み込むことができ，自動的にAttributeを埋めてくれる.
- 1行目に特徴名を入れると，自動的にattributeとして認識してくれる.

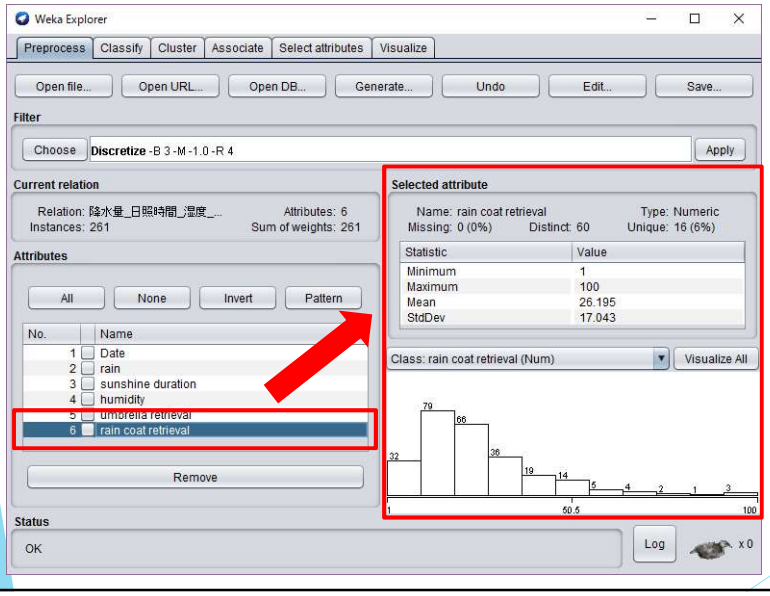
試してみよう

- Webサイトからファイルをダウンロードして Wekaで開いてみよう
  http://www.comp.sd.tmu.ac.jp/eri/weka/
- 降水量_日照時間_湿度_傘_レインコート.csv　を開いてみよう

CSVファイルを読み込む



データの確認



選択したデータの情報が確認できる.

Type：各データがどのような属性か確認
- Numeric（数値型）
- Nominal（名詞型）
- …etc

最近傍　K-NN

最近傍法（Nearest Neighbor）

- 降水量_日照時間_湿度_傘_レインコート.csvを対象として，K-NNを試してみる.
- 降水量_日照時間_湿度_傘_レインコート.csvには6個の属性が含まれている．どの属性を使うのが良いのだろうか？
  1. Date
  2. Rain
  3. Sunshine duration
  4. Humidity
  5. Umbrella retrieval
  6. Rain coat retrieval

勝敗を分類

- 6個の属性の中で Rain coat retrieval（レインコートの検索数，Googleトレンドより）　に着目
- どのデータを使って検索数を分類するか

  ↓

  何が検索数に関係しているのか？

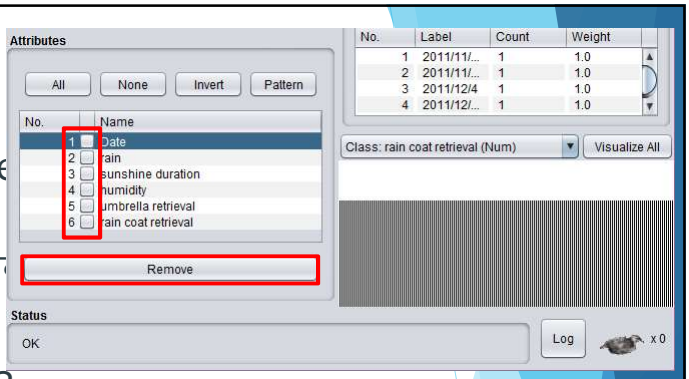
  使わない：Date, umbrella retrieval

  使う：rain, sunshine duration, humidity

- 前処理で使わない属性にチェックを入れて「Remove」

勝敗を分類

- 6個の属性の中で Rain coat re
  Googleトレンドより）　に着目
- どのデータを使って検索数を分類
  ↓
  何が検索数に関係しているのか？
  使わない：Date, umbrella retrieval
  使う：rain, sunshine duration, humidity



- 前処理で使わない属性にチェックを入れて「Remove」

csv → arff

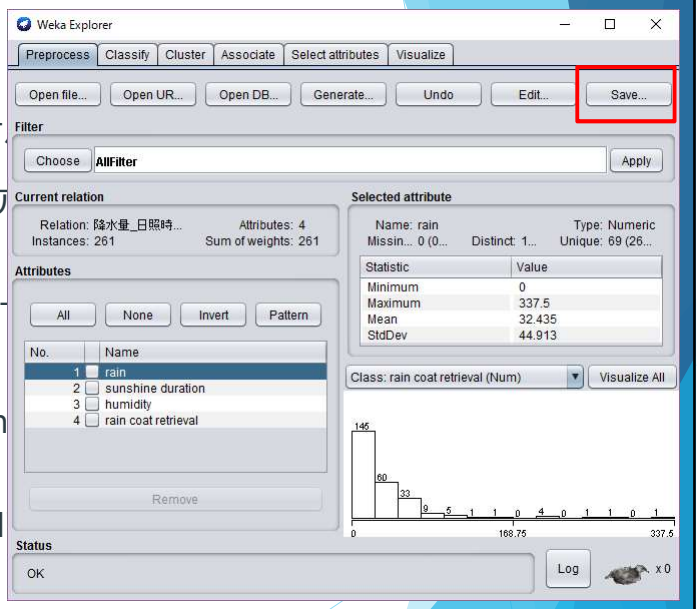
- 不要な属性を削除して，一度「保存」
- テキストエディタなどで保存したファイルを開いてみる.

- Attributeが自動的に追加されている.
  @attribute rain numeric
  @attribute 'sunshine duration' numeric
  @attribute humidity numeric
  @attribute 'rain coat retrieval' numeric

csv → arff

- 不要な属性を削除して，一度「
- テキストエディタなどで保存したフ

- Attributeが自動的に追加され
  - @attribute rain numeric
  - @attribute 'sunshine duration
  - @attribute humidity numeric
  - @attribute 'rain coat retrieval



K-NN

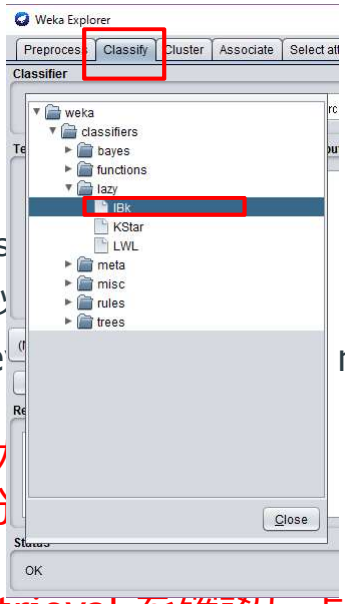
- 「Classify」タブに移動
- 分類器でIBkを選択する
  - Chooseボタン → Classifiers → lazy → Ibk
- Use training set (学習セット)を使用
- （Num）rain coat retrieval を選択して，Start (開始)

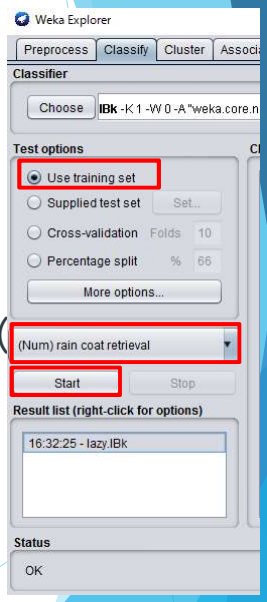
分類できていない・・・？
そもそも分類とは？

もう一度rain coat retrieval を確認しよう

K-NN

- 「Classify」タブに移動
- 分類器でIBkを選択する
  - Chooseボタン → Classifiers
- Use training set (学習セッ
- （Num）rain coat retrie rt (



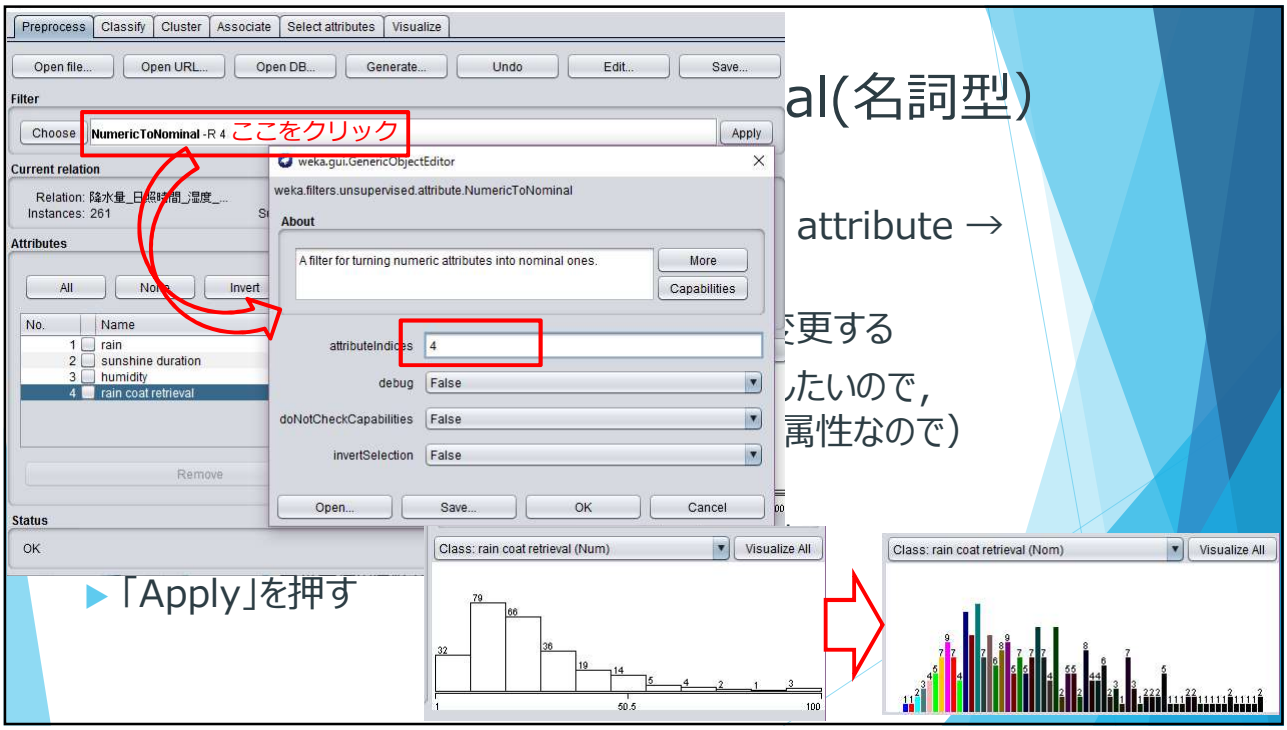


分類できてい
そもそも分

もう一度rain coat retrieval を確認しよう

Numeric(数値型) → Nominal(名詞型)

- 数値型を名詞型に強制的に変更する.
- Choose → filters → unsupervised → attribute → NumericToNominal を選択
- NumericToNominal をクリックして設定を変更する
- Rain coat retrieval だけを名詞型に変更したいので, attributeIndices に「4」を入力（4番目の属性なので）
  - 一番最後の属性なので「last」と入力してもOK
  - first – last だとすべての属性に対して名詞型にする
- 「Apply」を押す



al(名詞型)

attribute →

更する

たいので，

属性なので)

- 「Apply」を押す

もう一度　K-NN

- Classifyに戻る
- (Num) rain coat retrieval だったのが (Nom) rain coat retrieval となっているので，確かに数値型から名詞型になっている.
- Start
- Summary を　みると

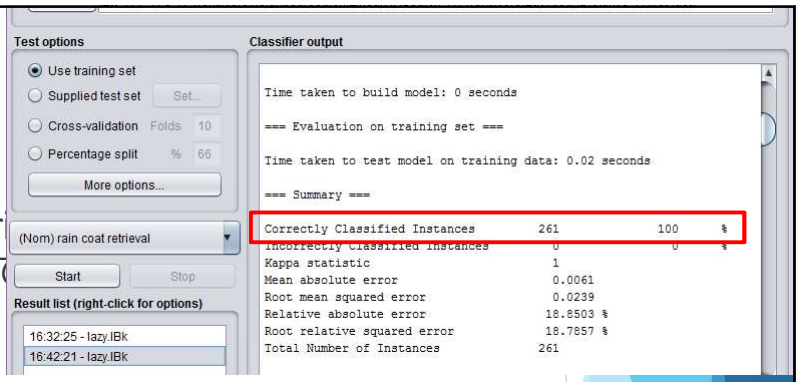
Correctly Classified Instances　261　100%

100%　大成功！

本当に？？？そもそも分類とは？
もう一度rain coat retrieval を確認しよう

もう一度　K-NN

- Classifyに戻る
- (Num) rain coat retri
  retrieval となっているので
- Start
- Summary を　みると



Correctly Classified Instances　261　100%

100%　大成功！

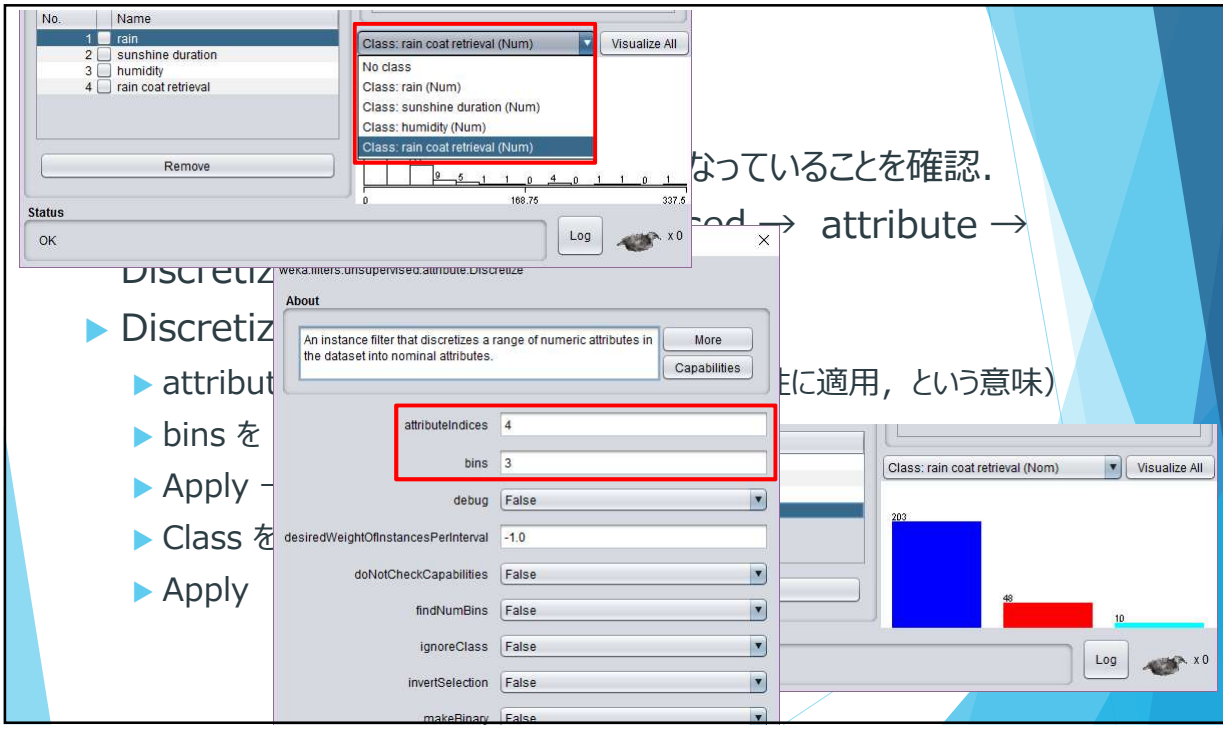
本当に？？？そもそも分類とは？
もう一度rain coat retrieval を確認しよう

クラスラベルを付ける（1/2）

- Rain coat retrieval の結果を，検索数が「多い」「多少ある」「少ない」の3段階にわけてラベルを付ける.
- ラベルのつけ方
  - Excelなどを利用して，IF文などでわけてあげる
  - Discretizes　を利用する.
- Discretizesを利用する
  - Numeric（数値型）にしか適用できない.
  - Nominal(名詞型)をNumeric(数値型)に変換するフィルターはない
    - 先ほど保存したファイルを読み込みし直す.
    - Preprocess　の　Undo　を試す

クラスラベルを付ける（2/2）

- Rain coat retrieval がnumeric になっていることを確認.
- Choose → filters → unsupervised → attribute → Discretize を選択
- Discretize をクリックして詳細設定画面へ
  - attributeIndicesに「４」を入力（4番目の属性に適用，という意味）
  - bins を「３」に設定 （3つのクラスにわける）
  - Apply → 変わらない！ → Class のリストで選択されていると適用できない
  - Class を No Classに
  - Apply



もう一度　K-NN

- Classifyに戻る
- Start
- Summary を　みると

  Correctly Classified Instances　　261　　100%

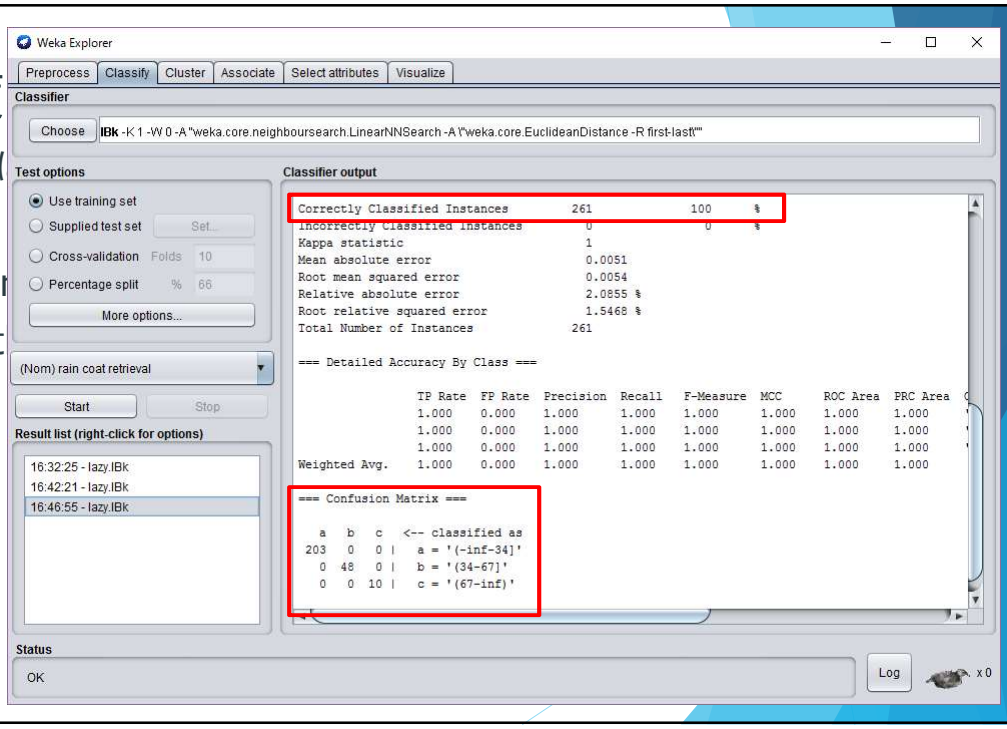
100%　大成功！

本当に？？？
交差検定をしよう！

もう一度

- Classify
- Start
- Summar

  Correct

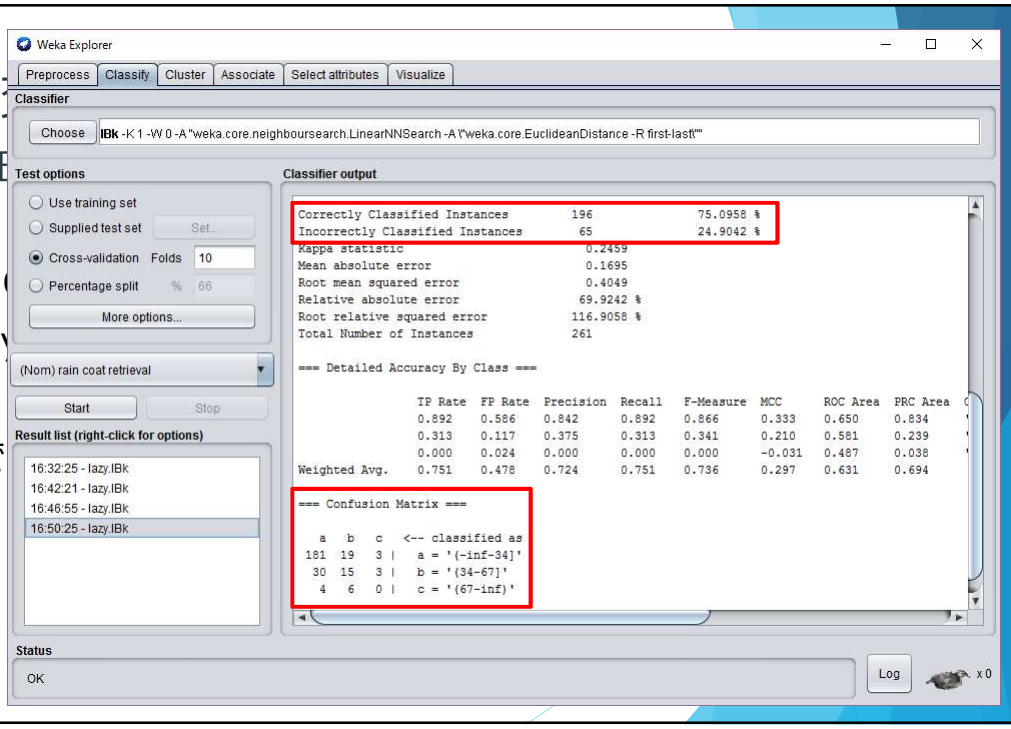


K-NN　交差検証

- 交差検証　フォールド　10

Correctly Classified Instances	196	75.0958 %
Incorrectly Classified Instances	65	24.9042 %

多少分類できそう．．．



パラメータ調整

- KNNの値を　3　に変更してみる
- 学習セット

Correctly Classified Instances	223	85.4406 %
Incorrectly Classified Instances	38	14.5594 %

- 交差検証　フォールド　10

Correctly Classified Instances	198	75.8621 %
Incorrectly Classified Instances	63	24.1379 %

どのデータがどのように分類されているのか確認してみる.

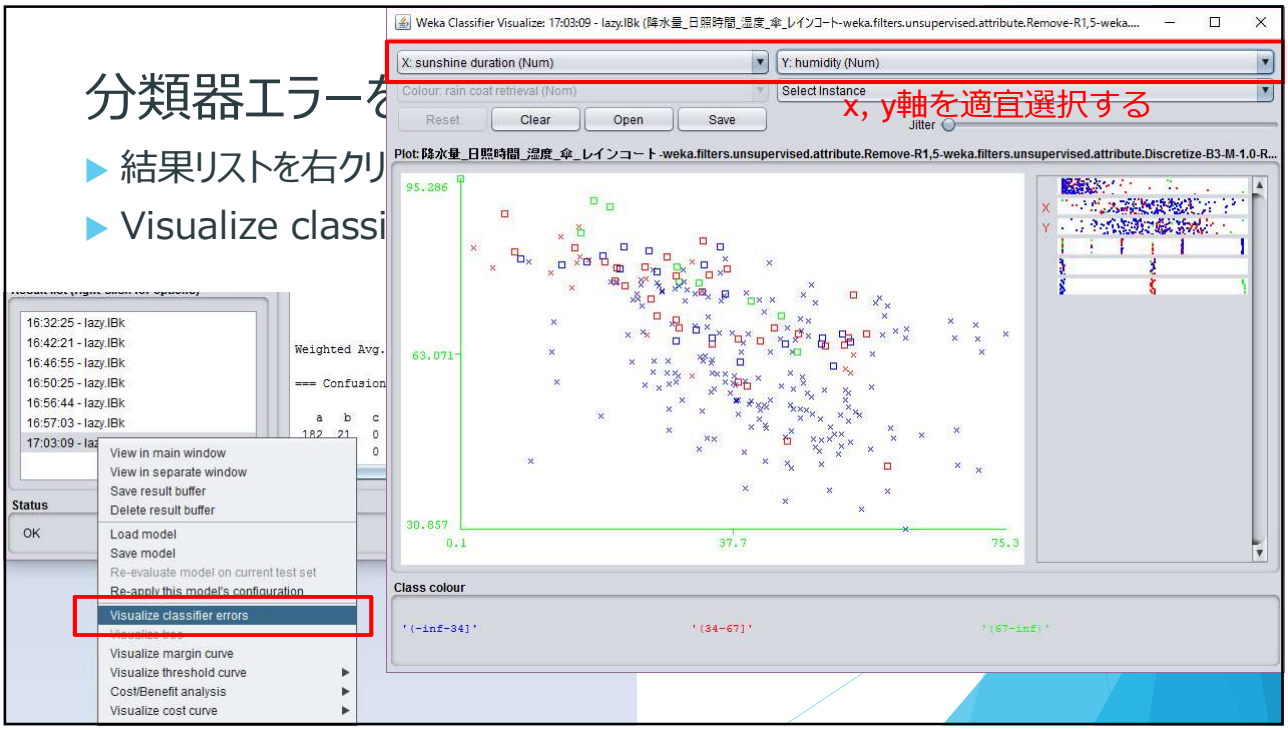
- 分類器をビジュアル化
- 予測結果を表示する

分類器エラーをビジュアル化

- 結果リストを右クリック
- Visualize classifier errorsを選択

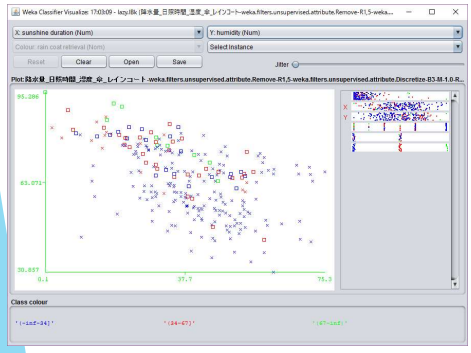
分類器エラーを

- 結果リストを右クリ
- Visualize classi



もっと結果を見やすくしたい

- 結果リスト → 右クリック → Save result buffer
- .txt や .xls などの拡張子をつけて保存

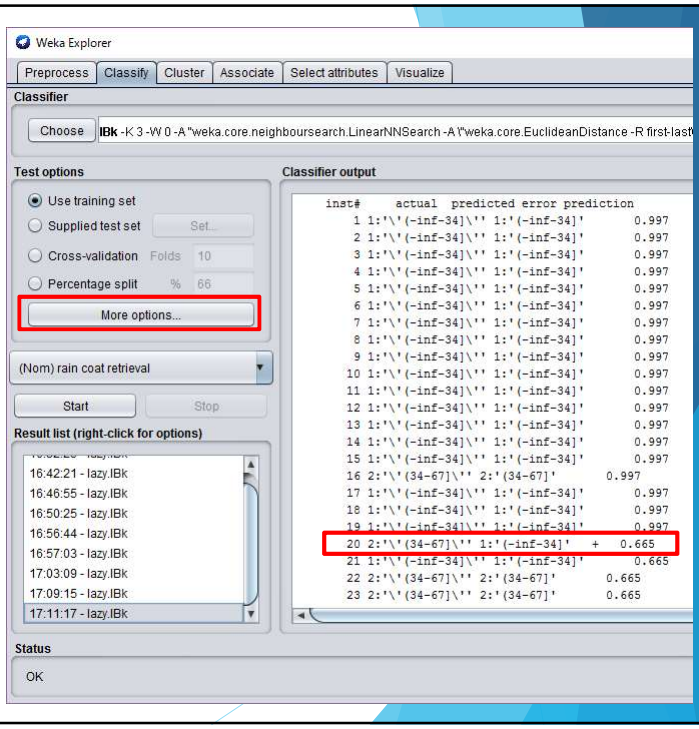


←この結果も，このまま論文に載せるのは．．．

「保存」をクリックして保存．
クラスタ情報が含まれた状態で保存される．

出力予測

More Options
↓
Output predictions
↓
Choose [PlainText]



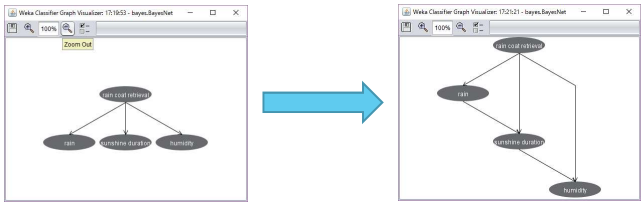
予測してみよう

- 分類ができるなら，予測にも使えるのでは？
- 保存したarffファイルを開いて，勝敗の列の一部を「？」に変更する
- もう一度実行してみる

BayesNet

BayesNet

- 「分類」タブで，分類器で「BayesNet」を選択
  - Classifiers → Bayes → BayesNet
- どのようなグラフになっているのか確認
  - 結果リスト → グラフをビジュアル化
  - 1 階層だけのグラフ → Search Algorithm → maxNrOfParents を 2 に変更してみる



サポートベクターマシン SVM

サポートベクターマシン

- Functions → SMO を選択
- オプションのKernelでカーネル関数を変更することができる.
  - Normalized Poly Kernel
    - Correctly Classified Instances 203 77.7778 %
    - Incorrectly Classified Instances 58 22.2222 %
  - RBF kernel, gamma=1.0E-4
    - Correctly Classified Instances 207 79.3103 %
    - Incorrectly Classified Instances 54 20.6897 %

ニューラルネットワーク

ニューラルネットワーク

- 分類器の選択 → functions → MultilayerPerceptron

補足
MultilayerPerceptron（多層パーセプトロン）は
フィードフォワード型のもっともオーソドック
スなニューラルネットワークです.

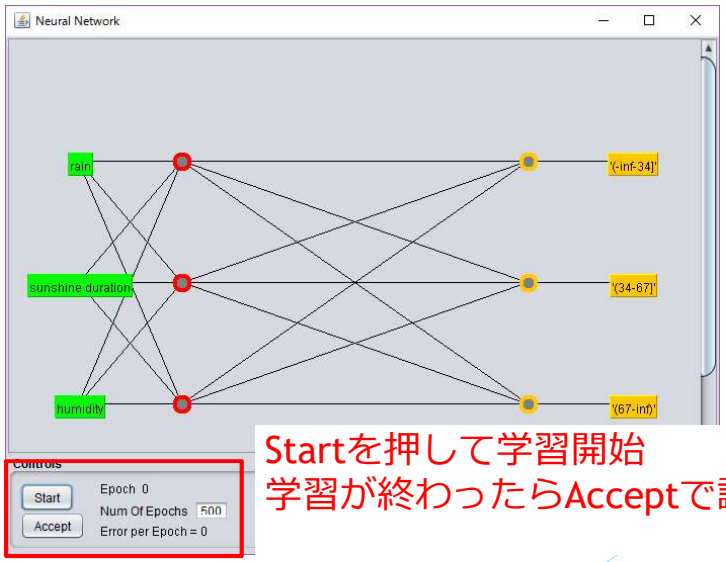
パラメータの設定

GUI：True
ネットワークの状態をグラフィカルに表示

HiddenLayers: a
隠れ層（中間層）のノード数を自動で決定

テストオプション　：　学習セットを使用

ニューラルネットワークの状態を確認



Startを押して学習開始
学習が終わったらAcceptで識別開始

結果を確認

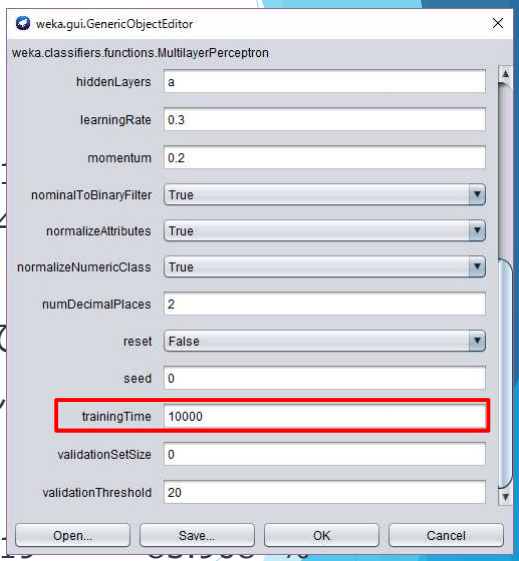
- 右側の出力結果を見ると
  - Correctly Classified Instances 214 81.9923 %
  - Incorrectly Classified Instances 47 18.0077 %

- 学習回数が多ければ，学習データに対して正答率が上がる.
- 学習回数が多いと，過学習が起こり，あいまいなデータに弱くなる…かも
  - 学習回数を 500回 から 10000回に変更
  - Correctly Classified Instances 219 83.908 %
  - Incorrectly Classified Instances 42 16.092 %

結果を確認

- 右側の出力結果を見ると
  - Correctly Classified Instances 21
  - Incorrectly Classified Instances 4

- 学習回数が多ければ，学習データに対して
- 学習回数が多いと，過学習が起こり，あい…かも
  - 学習回数を 500回 から 10000回に変更
  - Correctly Classified Instances 21
  - Incorrectly Classified Instances 42 16.092 %



交差検証して結果を比較しよう

- ただし，交差検証はパラメータ設定のGUIをFalseにすること.
- 交差検証は，学習データとテストデータを入れ替えて交差数分繰り返す.
- GUIだと，学習するたびに，Start，Acceptボタンを押す必要がある.
- 学習回数を500回に戻して実行.　（10000回のままでも構いませんが相当時間がかかります…）

交差検証の結果

- Correctly Classified Instances　202　77.3946 %
- Incorrectly Classified Instances　59　22.6054 %

- 学習セットを使用（学習データと識別データが同じ）に比べて，認識率が下がっている.

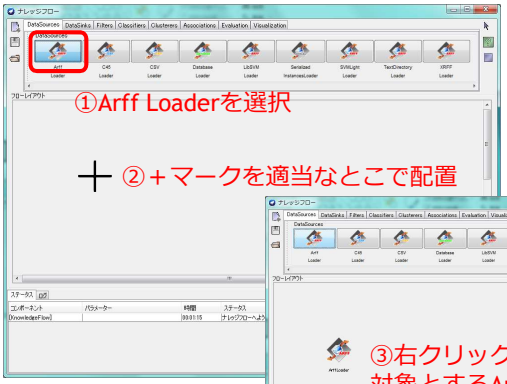
ナレッジフロー

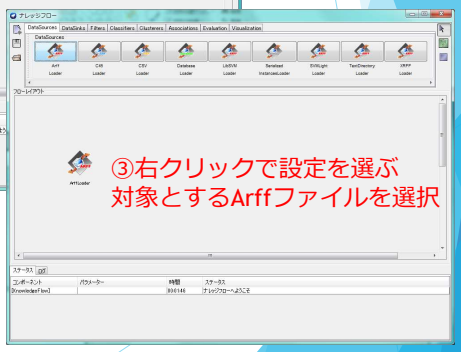
ナレッジフロー

- エクスプローラ
  - メリット
    - 様々な手法を試すことができる
    - 一つ一つの手順を確認しながら進められる
  - デメリット
    - 毎回手作業で手順を進めなければならない

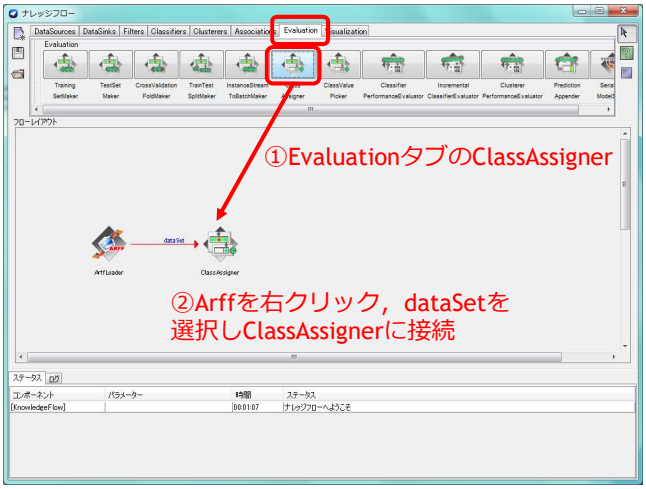
- ナレッジフロー
  - メリット
    - 一連の流れをフローでつないで，処理をつくることができる
  - デメリット
    - 一つ一つの手順を確認するのは面倒
    - ある程度の慣れが必要

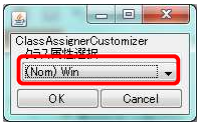
ナレッジフローでK-NN





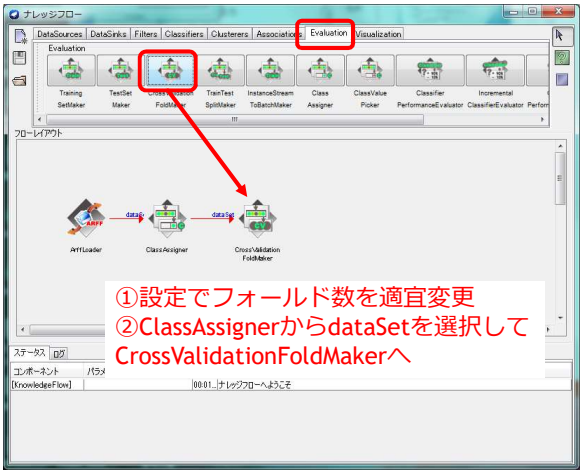
Class Assigner

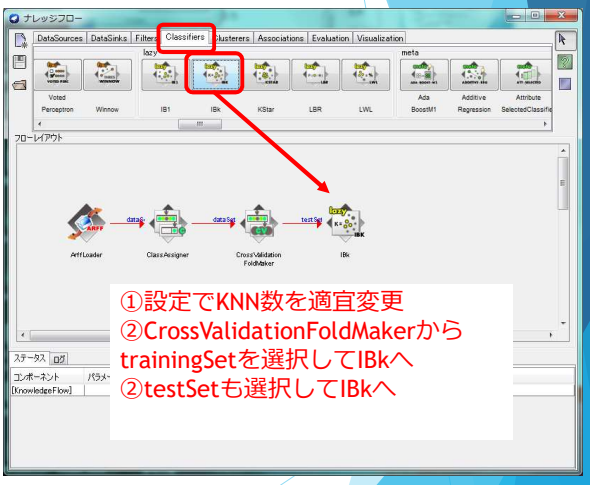




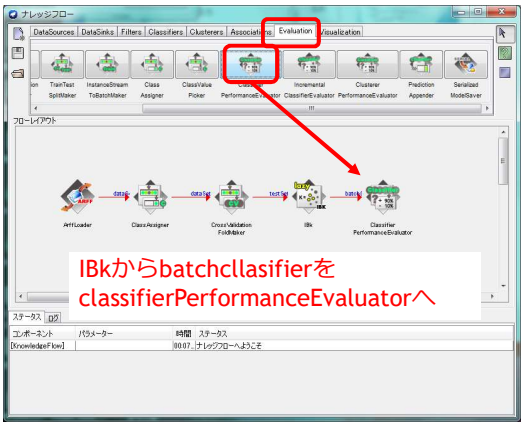
③ClassAssignerの設定で正解
データとするか指定

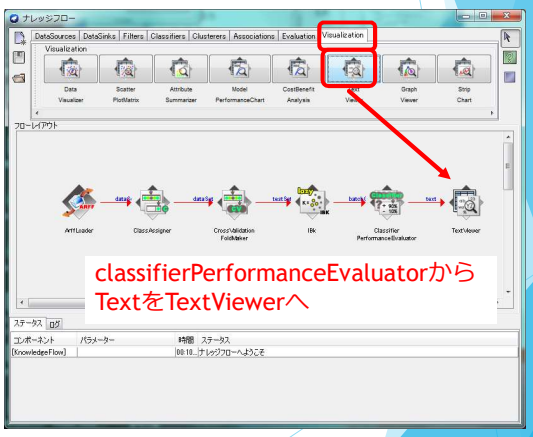
CrossValidationFoldMaker，IBk





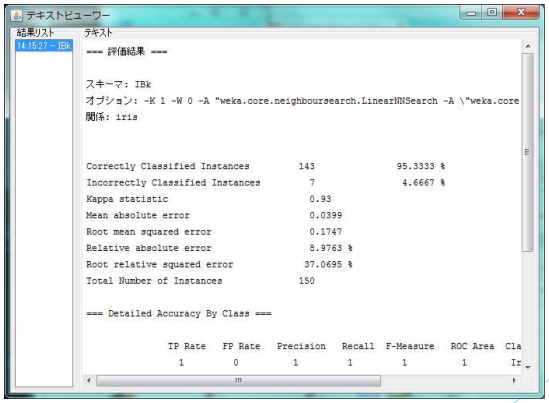
結果表示の準備
ClassifierPerformanceEvaluator, TextViewer





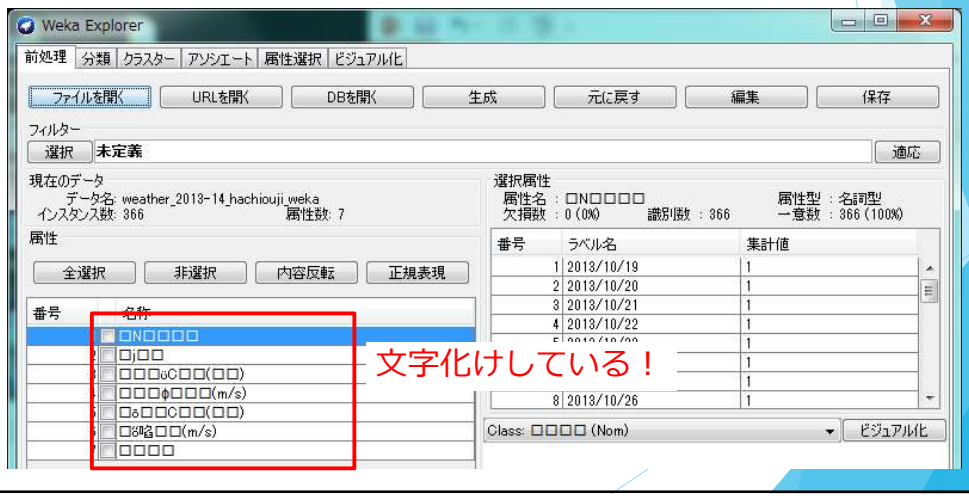
実行と結果確認

- ArffLoaderの「読み込み開始」をクリック
- ステータスがすべて「完了」したら
- TextViewerの「Show results」をクリック

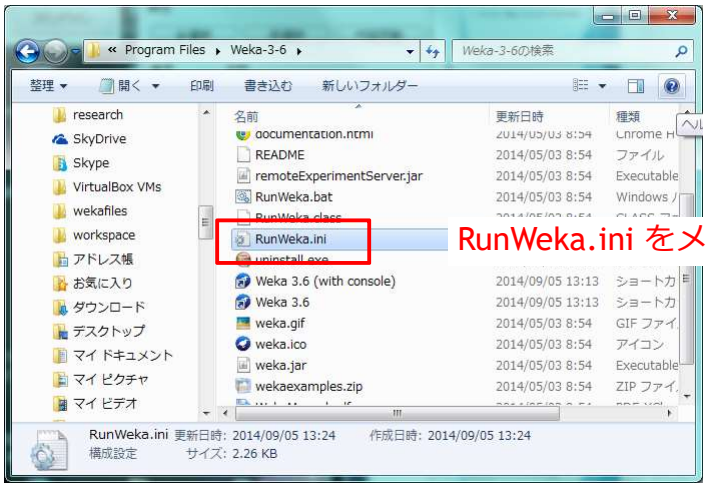


日本語の含まれたファイルを開く

- soccer.csv　をダウンロードして開いてみよう

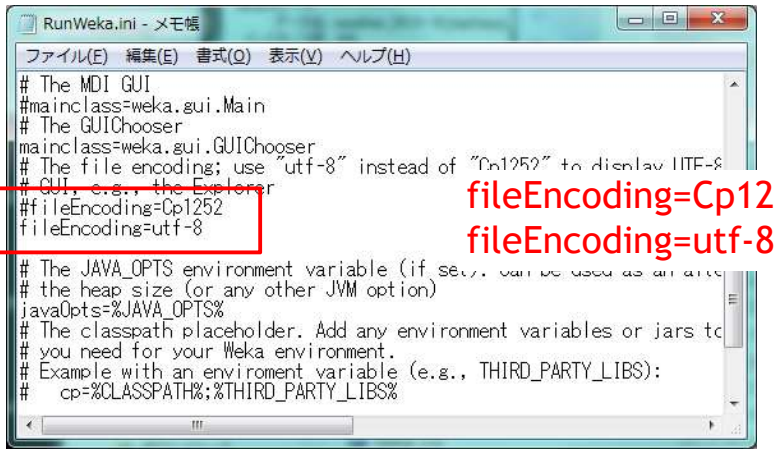


日本語が文字化けしている時の対処法



RunWeka.ini をメモ帳などで開く

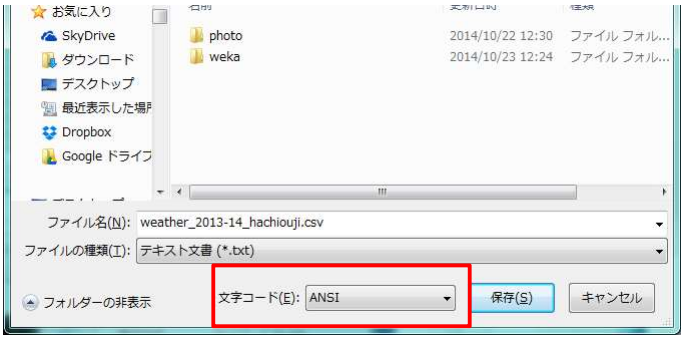
RunWeka.ini の修正



fileEncoding=Cp1252 をコメントアウト
fileEncoding=utf-8を追記

文字コードはUTF-8

- CSVファイルをメモ帳で開き，「名前をつけて保存」を選択.
- 文字コード欄からUTF-8を選択して保存



UTF-8を選択

Macの場合

- 設定ファイルの変更は不要
- csvファイルをテキストエディットで開いて，複製を選択.
- 保存する時に，エンコーディングを日本語(Mac OS)にする.

エラーが出てファイルが開けない

- 文字コードの問題ではない場合，次の点を確認してください．

- , = “ ‘ *+ -% が含まれているとエラーが起きるかも
- 誤った情報でファイルが上書きされてしまった

目的のアルゴリズムが適用できない

- 目的のアルゴリズムが，グレーになっていて選択できないことがある．アルゴリズムによって扱える次元数や，属性型に制限があるため．

- 対処方法
  - 次元数を減らしてみる：不要なデータを削除してみる．
  - 属性型を変更する：数値型から名詞型に変えてみる．

Wekaをプログラムに

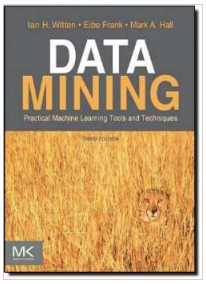
- WekaのGUIだけでなく，自身のプログラムに取り込みたい
- Wekaのインストールディレクトリ内，weka.jar を利用してweka APIを使ってJavaのプログラムを作成することが可能.
- 詳細はwekaインストールディレクトリ内のDocumentation.htmlを参照

データ収集

- CSV形式で提供されている統計データを利用
  - 総務省統計局：http://www.stat.go.jp/
  - 政府統計の総合窓口：http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do
  - 気象庁：http://www.jma.go.jp/jma/index.html
- 各種発表データから整形する
- 自分で収集する

参考文献

- フリーソフトではじめる機械学習入門，荒木雅弘著，森北出版
- Data Mining: Practical Machine Learning Tools and Techniques, Ian H. Witten, Eibe Frank, Mark A. Hall



グループワーク

- グループ分け（１グループ　３－５名程度）
  - 目的や対象データを決める
  - 作業分担
- 10月27日の予定
  - 14:40 – 15:10 各班まとめ作業30分程度
  - 15:30 – 発表会
    - 発表10分　＋　質疑5分
    - データ解析の目的
    - 用いた手法
    - 評価結果