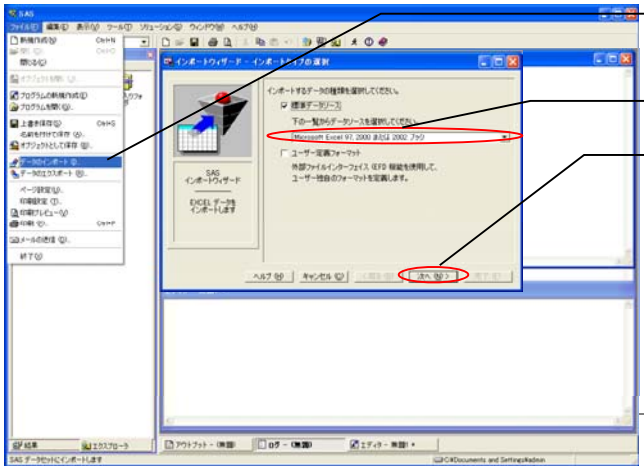
データマイニング　第5回
相関ルールの抽出（2）

総合政策学部　古谷知之

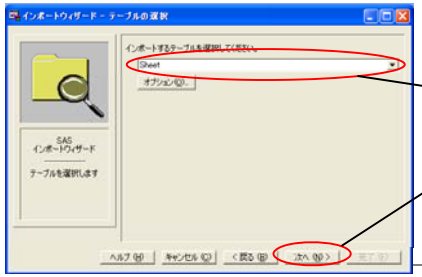
エクセルデータのインポート

SASは様々なファイル形式のデータを読み込み分析できる。
「ファイル（F）」から「データのインポート（I）」を選択。「Microsoft Excel」を選択し、「次へ（N）」をクリック

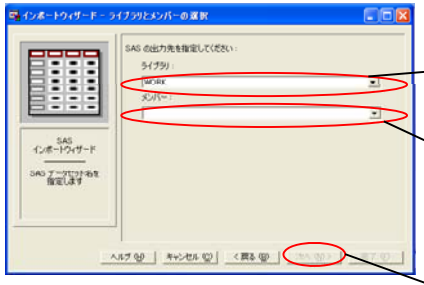


エクセルデータのインポート

「参照」でインポートするエクセルファイルを指定し、エクセルファイルのシートを指定、「次へ（N）」をクリック

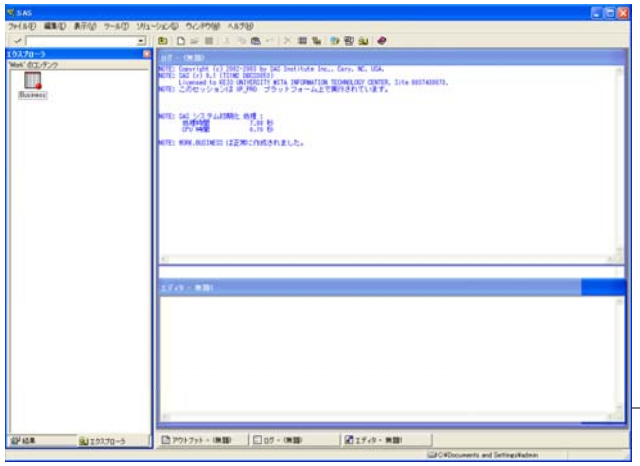


エクセルデータのインポート



エクセルデータのインポート

エクスプローラ部の中の「ライブラリ」をクリック、その中の「Work」をクリックすると「Business」というファイルができている



Workフォルダ

SAS上のフォルダ。
メモリに記憶されているので、一度SASを落としたらフォルダの中身は空になる。

相関ルールマイニングのデータ

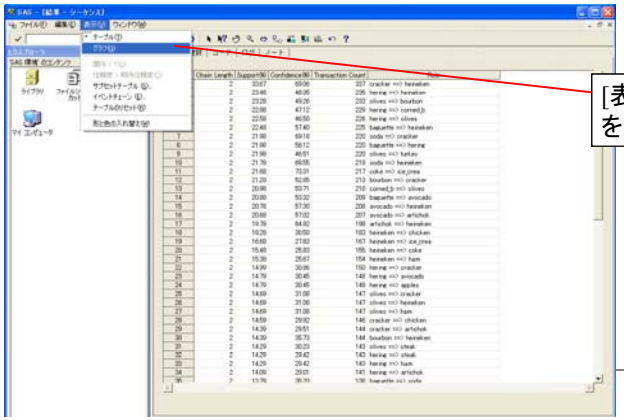
- http://fimi.cs.helsinki.fi/data/
- http://www.ics.uci.edu/~mlearn/MLRepository.html
  - The UCI KDD Archive の下のby application area の中の World Wide Web中のMicrosoft Anonymous Web Data
- http://www.cs.waikato.ac.nz/~ml/weka/index.html

相関ルールマイニング

- 通常の相関ルールマイニングでは，多すぎるほどの相関ルールが導出される．
  - その中から，専門家が有益と思うものを選択する
  - 数が多すぎると判断する作業もマイニング？？
    - 10000以上もルールが出てきてはさすがに困る
- 解決策
  1. 視覚化 (visualization)
  2. 得られた相関ルールをまとめる (Summarize & Aggregation)
     - 後処理(Post-Processing)

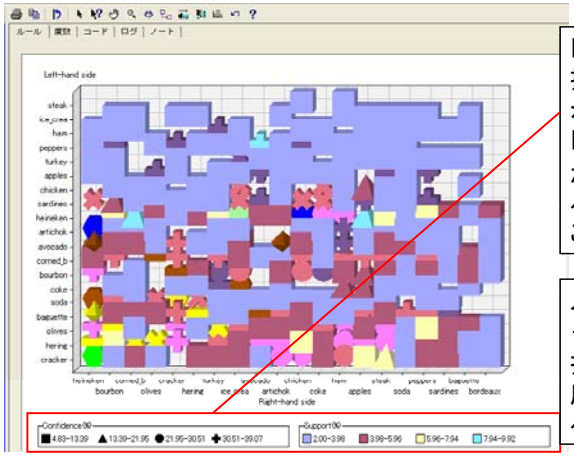
相関ルールの視覚化

□ 視覚化
[表示]から[グラフ(G)]を選択することで結果を視覚的にとらえることができる



相関ルールの視覚化

各ルールの支持度（Support）により、プロットの記号の色が決まる。信頼度（Confidence）の大きさにより、記号の形が決まる。



□ 信頼度（Confidence）、支持度（Support）の凡例が表示される
□ 枠の中を左クリックしながら上下にスクロールさせることで表示範囲を変えることができる

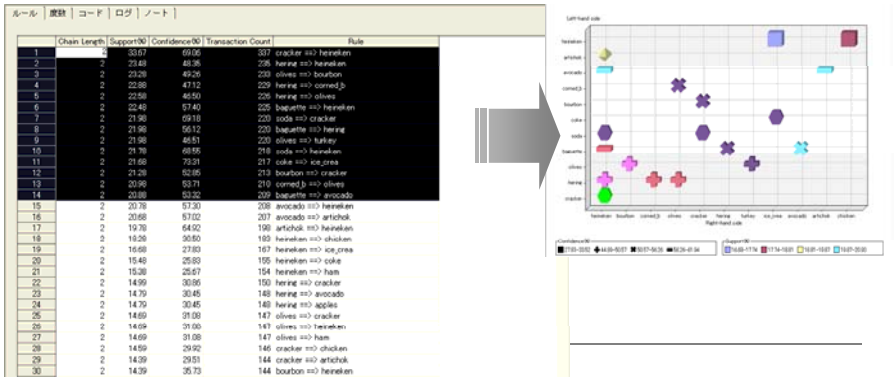
メインメニューから［表示］－［形と色の入れ替え］を選択することで支持度と信頼度を入れ替えることができる

相関ルールの視覚化

□ ルールが多すぎてグラフが作成できない場合
1）テーブル内のルールの部分集合を選択する
または
2）拡大ツールで調整する

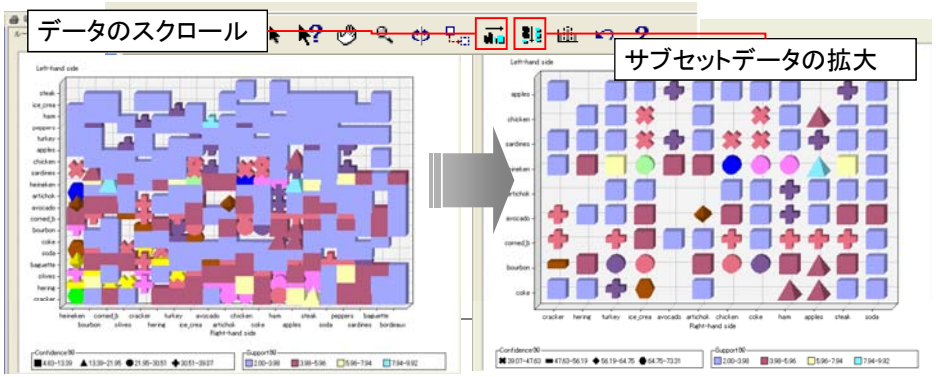
相関ルールの視覚化

□ テーブル内のルールの部分集合を選択する
1）テーブル内のルールを選択する
2）メインメニューから［表示］－［グラフ］を選択する



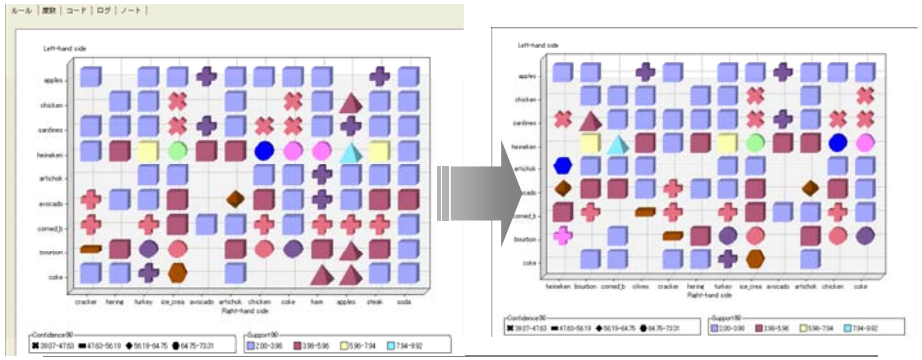
相関ルールの視覚化

□ 拡大ツールで調整する
1）アプリケーションツールバーの［サブセットデータの拡大］ツールを選択する
2）虫眼鏡のマウスカーソルを縦軸と横軸にスクロールさせ、グラフを拡大させる



相関ルールの視覚化

□ 拡大ツールで調整する
3）[データのスクロール]ツールを選択する
4）手の形のマウスカーソルを使い、グラフにあるほかの
　ルールを表示 させる。

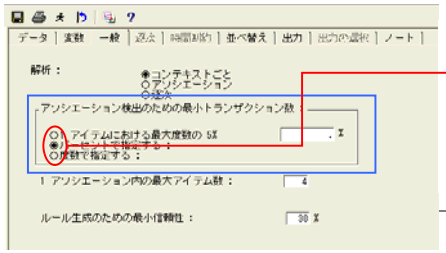


SASでの有効なルールの発見

- SASにおいて、支持度、信頼度、リフト値の3つの尺度で判断していく
- 3つの評価基準＋組み合わせの数を調整することで有用なルールを探し出す
- 調整には計算前の調整と計算後の調整がある
- 大量のルールを絞り込む場合
  - 考慮する組み合わせを少なくする
  - 支持度の水準を高くする
  - 信頼度の水準を高くする
  - リフト値の高いものを採用する
  - 4つを組み合わせる
- 他にも有効なものを探す手段として
  - 支持度が低いがリフト値・信頼度の高いものを探す
  - 組み合わせを増やした上で信頼度・支持度・リフト値を大きくする

SASでの有効なルールの発見

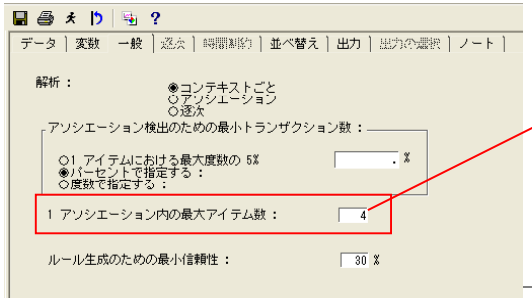
□ 支持度の調整方法
アソシエーションノードの[一般]タブで調整する。支持度を調整するには[アソシエーション検出のための最小トランザクション数]を調整する。
1）パーセント指定
または
2）度数指定
がある



ラジオボタンで選択する

SASでの有効なルールの発見

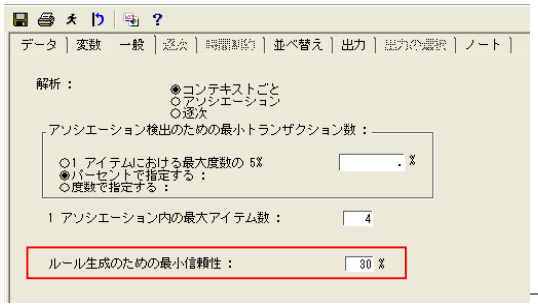
□ 組み合わせの数を調整する
アソシエーションノードの[一般]タブで調整する。[アソシエーション内の最大アイテム数]を調整する。
デフォルトでは最大4つのアイテムの組み合わせの相関ルールが検出される



数値を設定する

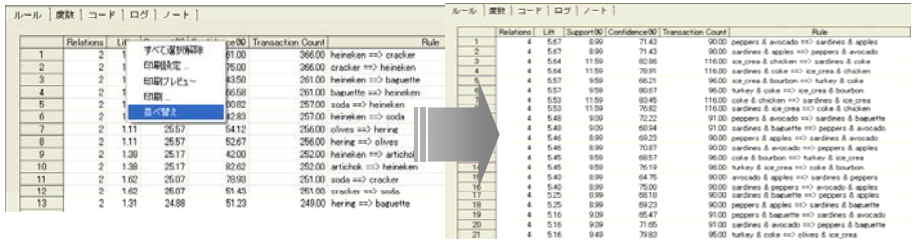
SASでの有効なルールの発見

□ 組み合わせの数を調整する
アソシエーションノードの[一般]タブで調整する。[ルール生成のための最小信頼性]を調整する。
デフォルトでは10%以上となっているが、信頼度の高いルールが必要なときは高い値を設定する



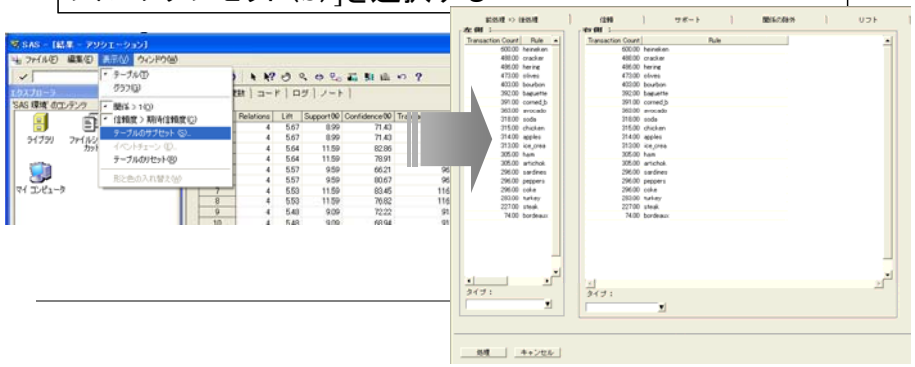
SASでの有効なルールの発見

□ リフト値で判断する
リフト値は計算前に調整することができない。アソシエーションルールの結果ウィンドウから、[Lift]のセル上で右クリック[並べ替え]を選択し、[Descending]を選ぶことにより並び替えられる



SASでの有効なルールの発見

□ 4つ尺度の計算後の調整
4つの尺度（信頼度（Confidence）、支持度（Support）、組み合わせの数（関係の数）、リフト値）のカットオフ範囲を指定することで、精密なルールを発見することができる。
1）アソシエーションの結果ウィンドウから[表示（V）]→[テーブルのサブセット（S）]を選択する

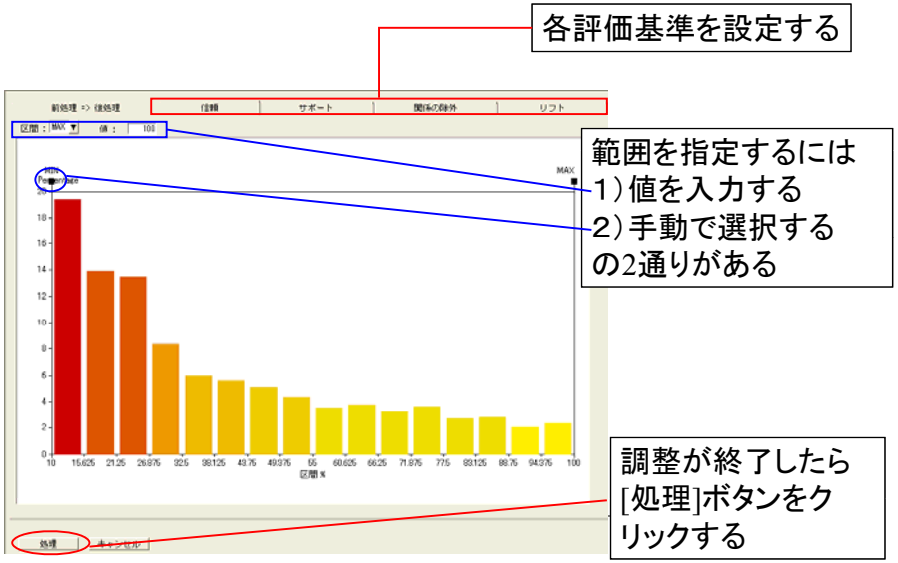


SASでの有効なルールの発見

各タブを選択することで、カットオフする範囲を指定できる
□ 最大値と最小値の指定方法
1）境界ラインの上部の黒いボックスを左クリックしたまま移動させる
または
2）[区間]ドロップダウンリストの矢印を選択し、MIN、MAX項目を選択し、[値]入力フィールドに値を入力する

上記の手法を他の評価基準でも設定し、設定終了後に[処理]をクリックする

SASでの有効なルールの発見



レポート課題

選択問題です。以下の1あるいは2を実行して、レポートとしてまとめてください。

1. WEBログデータに対して、相関ルールマイニングを行う。
    1. まず各自が前処理を行なう。
    2. そして、支持度、信頼度を色々変えて相関ルールを抽出する。
    3. そして、その相関ルールの中から意味のあるルールを探し出す。そのときに、支持度が小さく、リフト値が大きいルールを探す。
    4. そのときに、得られたルールを、対応表を使って、意味のあるルールに変換する。
    5. 相関ルールの表示も色々行なってみる。
2. 自分で問題を探して、データを用意し、相関ルールマイニングを行う。内容は、1に準じる。（ヒント：アンケート調査を行う。適当なアイテムを100個程度並べ、100人くらいの人にアンケートをとって、自分の好きなものを10個ずつ選んでもらう。各人のアンケート結果を一つずつのバスケットとして、相関ルールマイニングを行う）