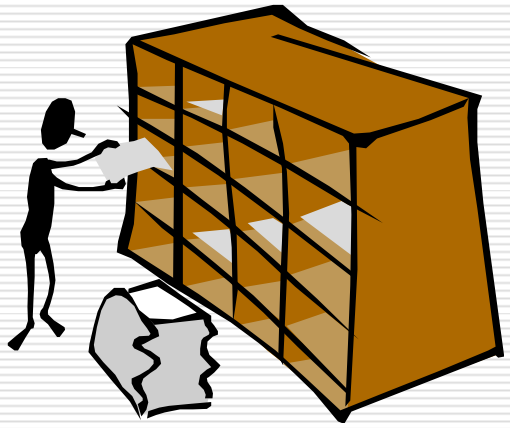
クラスタリング

定性データ分析入門第12回

担当：古川康一

クラスタリングとは

- 分類を目的とする手法の一つ
- データに基づいて分類対象をいくつかのクラスター(グループ)に分類する
- 似ているもの同士を同じクラスターに、似ていないもの同士は違うクラスターに分類される
- 類似性はOBS同士の距離から考える



クラスター分析の種類

- 排他的
  - 外的基準有
  - 外的基準無
    - 非階層的
    - 階層的
- 非排他的

Williams & Lance(1977)より

クラスタリングの事例

□Bank of Americaの優良顧客の発見

既存の住宅担保融資を利用する顧客にクラスタリングをかけた。探索的なデータマイニングを行い14のクラスタを作り、「企業家」という1つの有益なクラスタを発見した。

□アメリカ軍の女性兵士の服装

各兵士によく会う制服を提供する一方で、在庫の数を減らすことを目標とした。既存の洋服サイズはサイズが大きくなるごとに各部位のサイズも大きくなっていく。

詳細な兵士情報を分析しクラスタリングすることで「足が短く、ウエストが細く、胸が大きい女性」用の服などができた。

□遺伝子発現解析

各遺伝子が、いろいろな状況で蛋白質をどれくらい生成するかを調べる(遺伝子発現データ)。そのデータから、類似の振る舞いを示す遺伝子群を見つける。

決定木分析とクラスタリングの違い

- 決定木分析は、与えられたクラス情報に合うように分類基準を作る。すなわち、教師付き学習。
  - 各事例を分類するクラス、あるいはカテゴリーは分かっている。「環境情報学部の学生か、総合政策学部の学生か」、「定性データ分析入門を履修している、していない」、「下宿に住んでいる、住んでいない」など。それらのクラス情報が分かっている学生を訓練例として使う。
  - クラス情報のほかに、各学生の様々なデータが必要。出身地、興味、研究会、サークルなど。
  - クラス情報は分からないけど、その他の情報が分かっている学生に対して、そのクラス情報を当てる。
- 一方、クラスタリングは、分類すべきクラス自身が分からない、という事態を想定している。クラスタリングによって、どのようなクラスがあり得るのかを見出す。

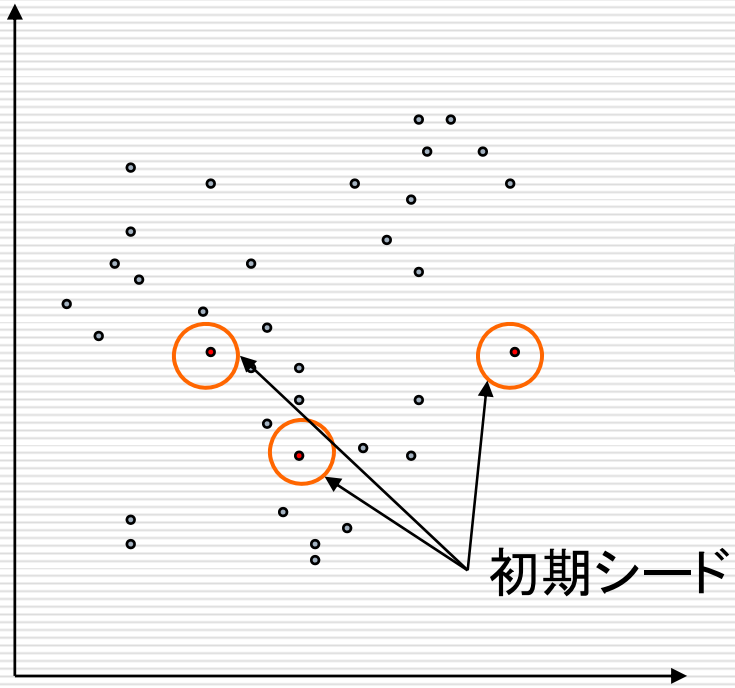
K-meansとは

- 非階層型の代表的クラスタリング手法の一つ
- 事前にクラスター数を指定しなくてはならない
- 以下のプロセスでクラスターを作成する

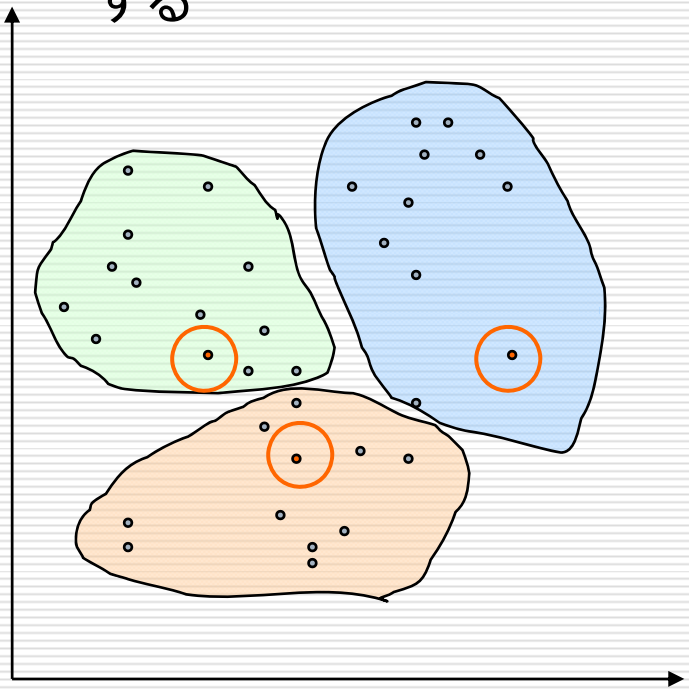
1. クラスタ数を指定
2. 各クラスターの重心（シード値）の座標を決める
    1. 乱数
    2. 個別のOBS
    3. 直接座標値指定
3. 各OBSを最も近いシードに分類
4. 分類されたメンバー内で新しいシード値を計算する
5. 旧シード≒新シードになるまで3&4を繰り返す

K-meansのプロセス

1. 事前に指定したクラスター数（この場合3）の重心（初期シード値）の座標を決める



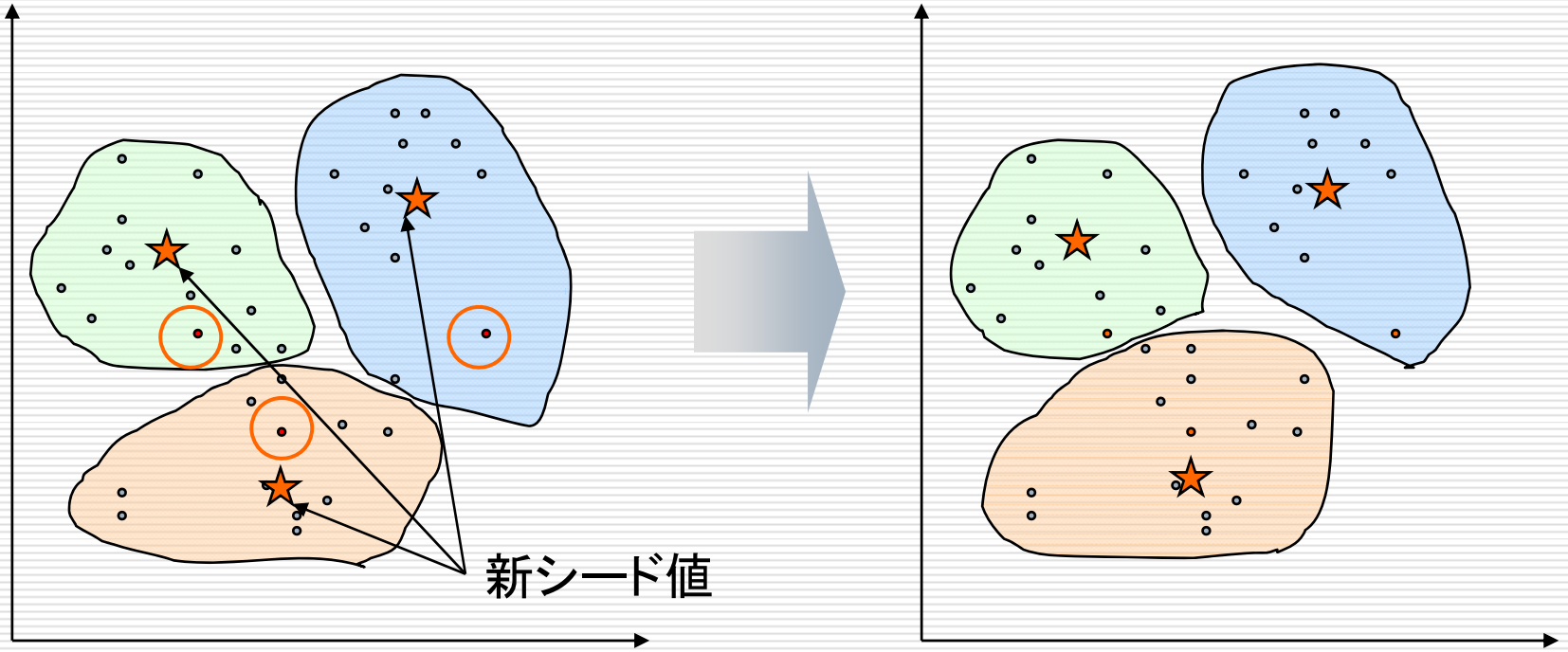
2. 初期シードとの距離で各クラスターがどの仮クラスターに所属するのか決定する



K-meansのプロセス

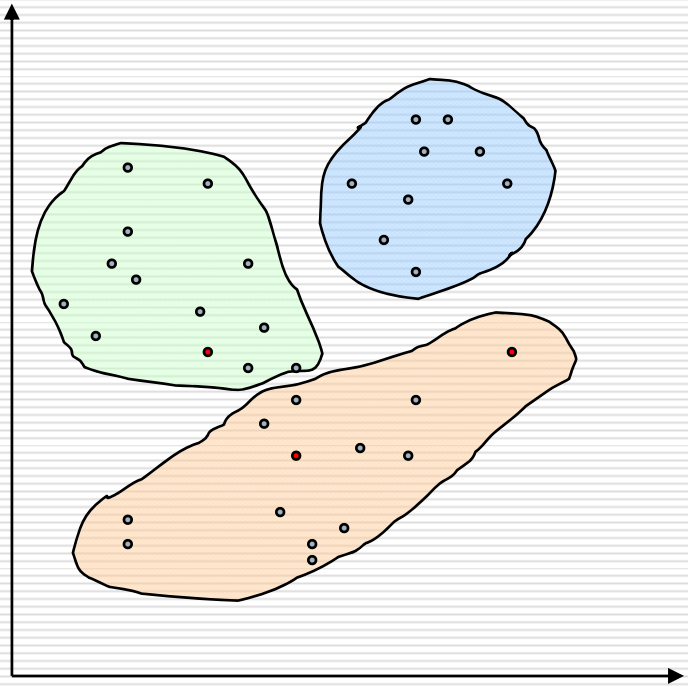
3．それぞれのクラスターで重心（新シード）を探す

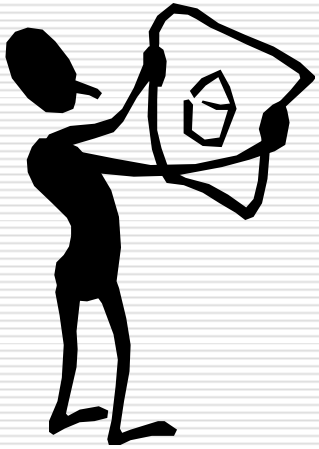
4．新シードからの距離を使い、クラスタリングをやり直す



K-meansのプロセス

5. 旧シードと新シードとの距離の移動がなくなるまで３＆４を繰り返す





多次元ベクトルデータ間の距離

- クラスタリングするデータが複数の属性を持つ場合、それらの各データは、多次元ベクトルとなる。
- 多次元ベクトル同士の距離は、多次元空間上でのユークリッド距離を考えればよい。

$$\mathbf{X} = (x_1, x_2, \ldots, x_n),$$

$$\mathbf{Y} = (y_1, y_2, \ldots, y_n)$$

のとき、両ベクトル間の距離は、

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

のようになる。

カテゴリカル変数の場合

- カテゴリカルな変数の場合、離散的な値で各属性を表現すればよい。
  - たとえば、企業の格付けでは、AAA,AA,A,BBB,BBのような値を取るが、それを9，8，7，6，5，…のような数値とすればよい。
  - アンケートで、5段階評価の場合、（とても強い、強い、普通、弱い、とても弱い）を（5，4，3，2，1）にする。
  - 健康状態のチェックで、過去に手術を受けたか受けていないか、現在治療中であるかないか、めまいがするかしないか、などのいくつかの項目にたいして、2つのうちどちらかを選ぶような場合、それらは0，1の2値で表現すればよい。

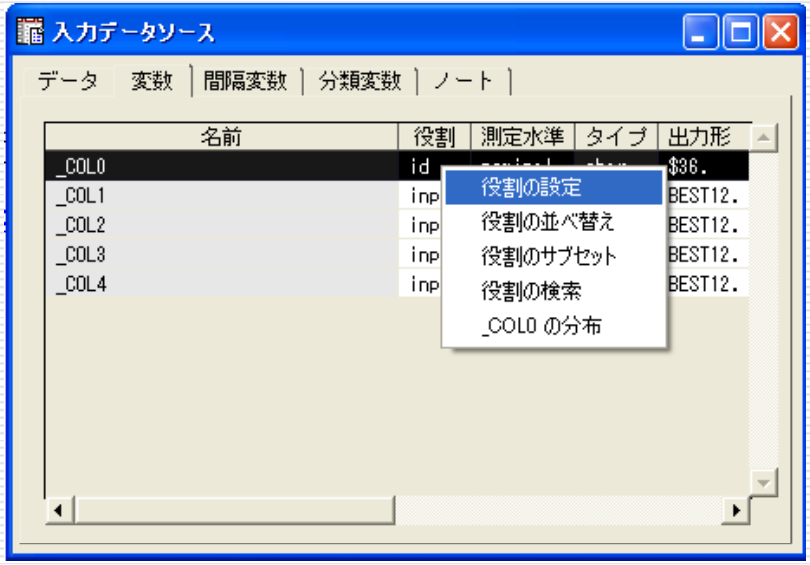
クラスタリングに使用するデータ

「clst.xls」
ブランドイメージ調査結果

変数名	役割	説明
ブランド名	id	ブランド名
実用性	input	企業に対する実用性のイメージを表す指標
先進性	input	企業に対する先進性のイメージを表す指標
かっこよさ	input	企業に対するかっこよさのイメージを示す指標
親しみ	input	企業に対する親しみのイメージを示す指標

クラスタリングノードの使い方

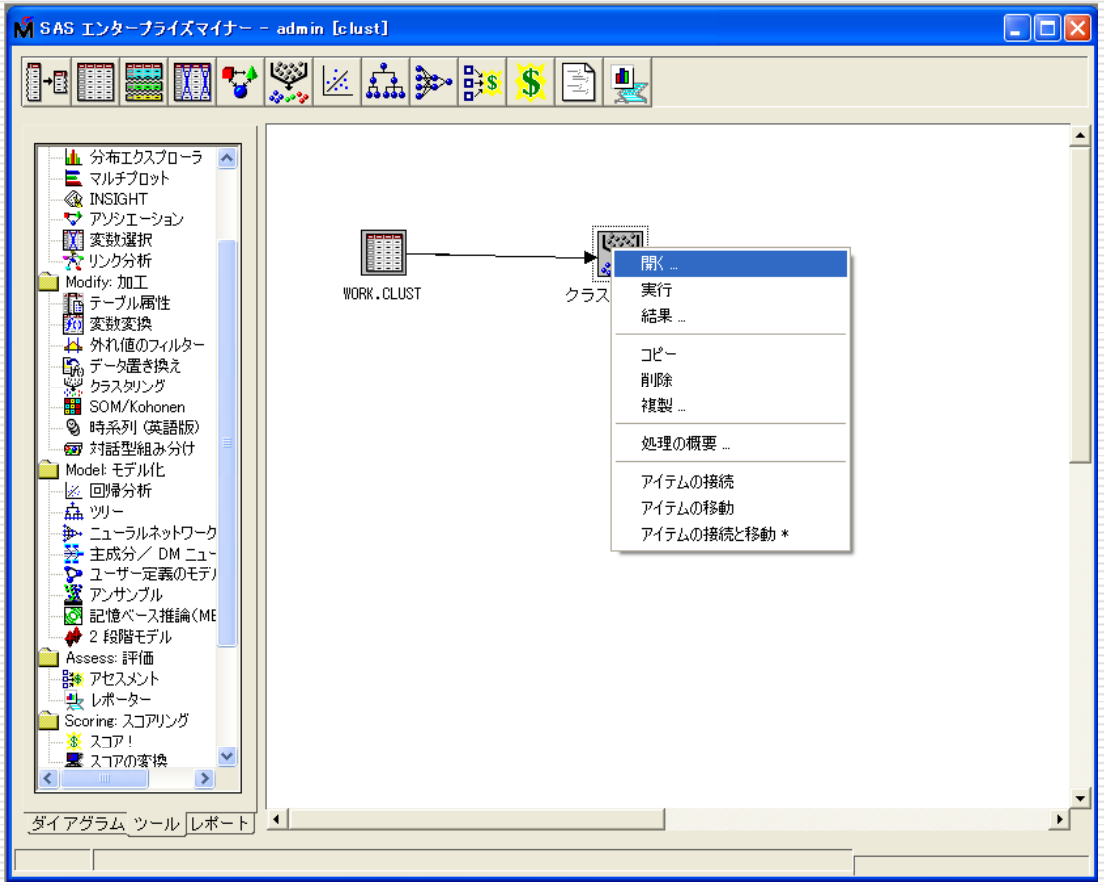
1）「clst.xls」をインポートしファイル名を「clst」とする。
2）EMを開き、入力データソースにインポートした「clst」を入れる
3）入力データソースの変数タブで「_COL0」の[役割]を[id]にする



クラスタリングノードの使い方

4）クラスタリングノードをドラッグ &ドロップする

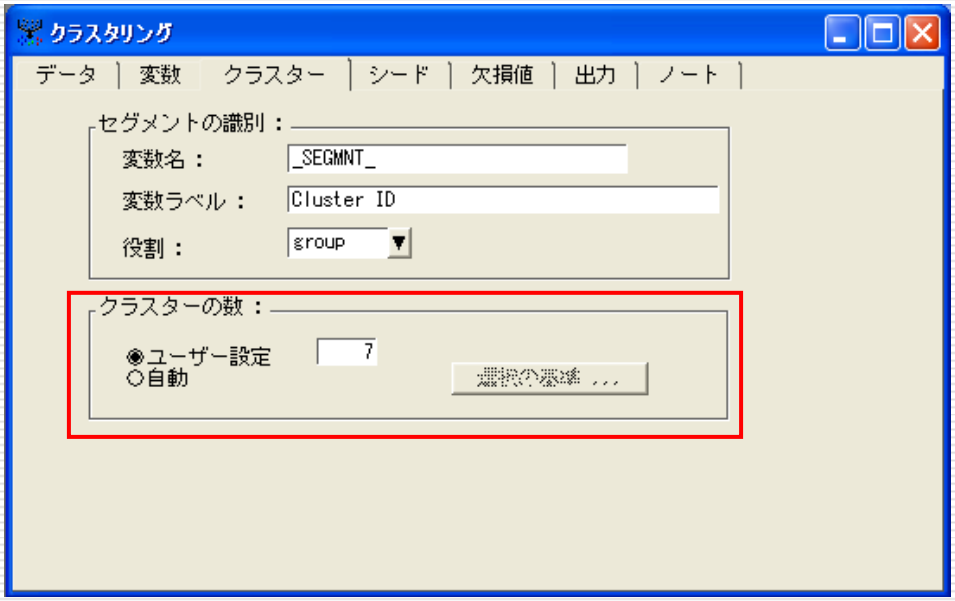
5）入力データソースから線を引き右クリックし「開く」を選択



クラスタリングノードの使い方

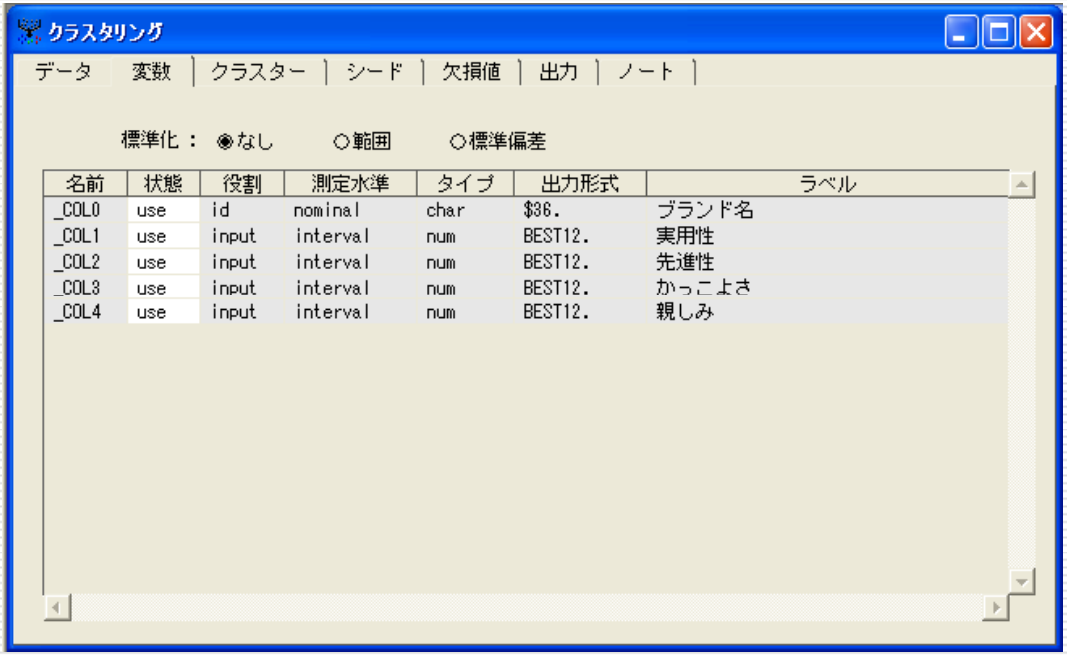
6）[クラスタータブ]でクラスターの数を設定する。[クラスターの数]をユーザー設定にし、（とりあえず）7つのクラスターを作る

7）クラスタリングノードを閉じて、実行する



クラスタリングノードの詳細設定

[変数タブ]
変数とその状態、モデルにおける役割、およびその他の属性を表示する。入力データソースで指定した度数変数とID変数とともに、すべての入力変数が表示される

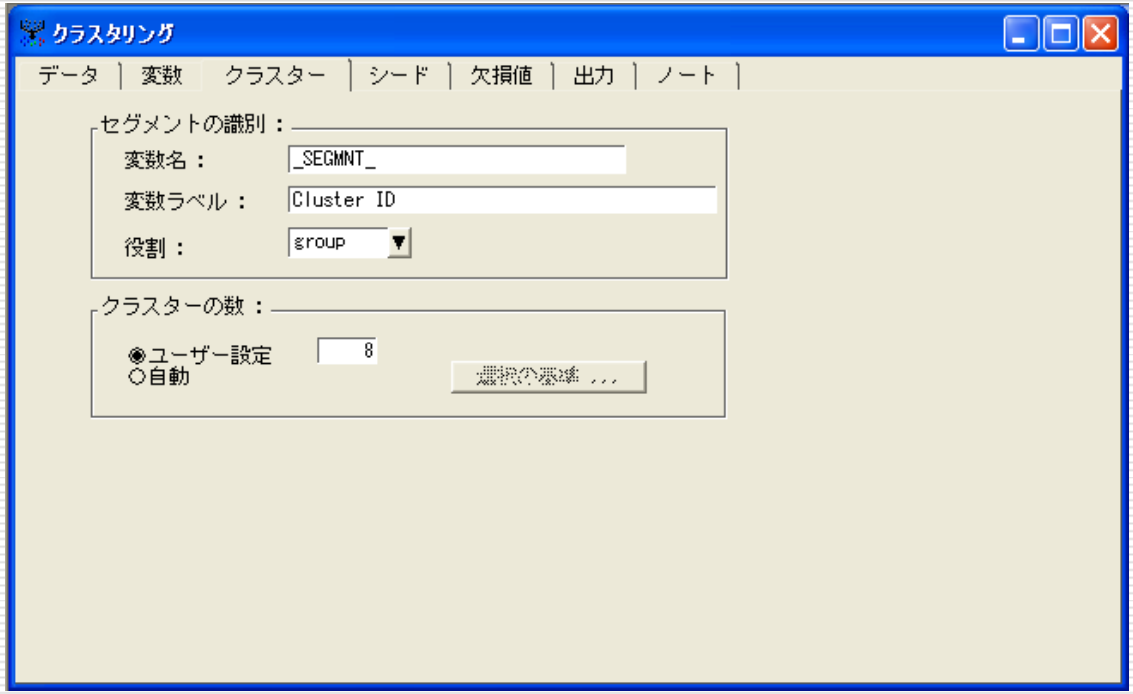


[標準化]
内部の標準化方法を設定できる

- [なし]（デフォルト）を指定すると、クラスタリングの前に変数の標準化を行わない
- [範囲]を指定すると、変数値が範囲によって除算される。平均値は減算されない
- [標準偏差]－変数値は、標準偏差によって除算される。平均値は減算されない。

クラスタリングノードの詳細設定

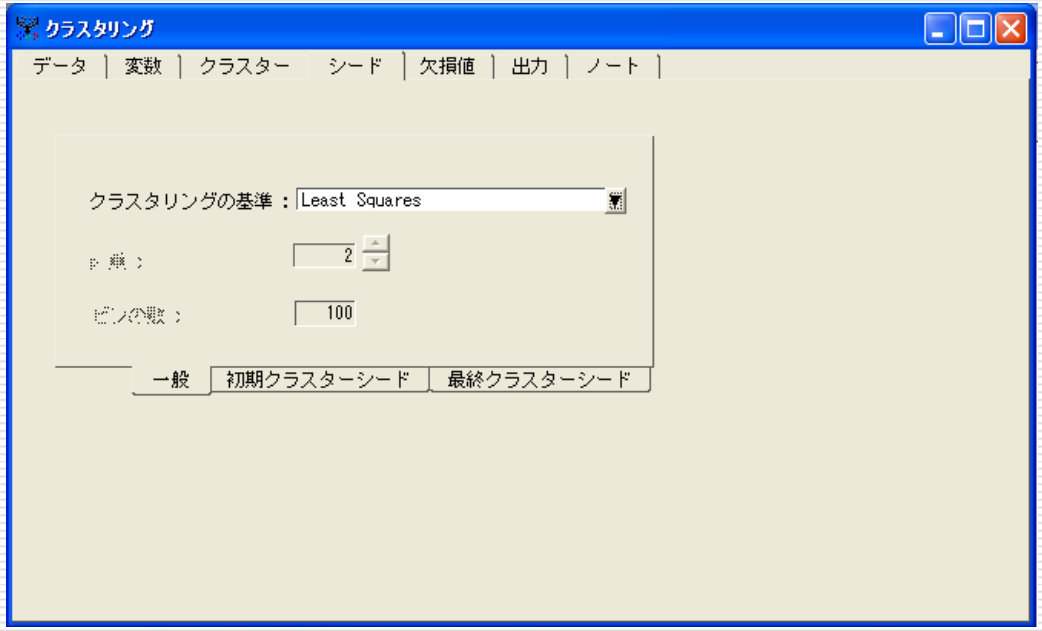
[クラスタータブ]
セグメント識別変数に関する指定とクラスター数の指定を行う。



クラスター数を[自動]で行う場合、[選択の基準]ボタンをクリックして表示されるダイアログボックスで、クラスター数が自動的に決定される方法を変更することができる。

クラスタリングノードの詳細設定

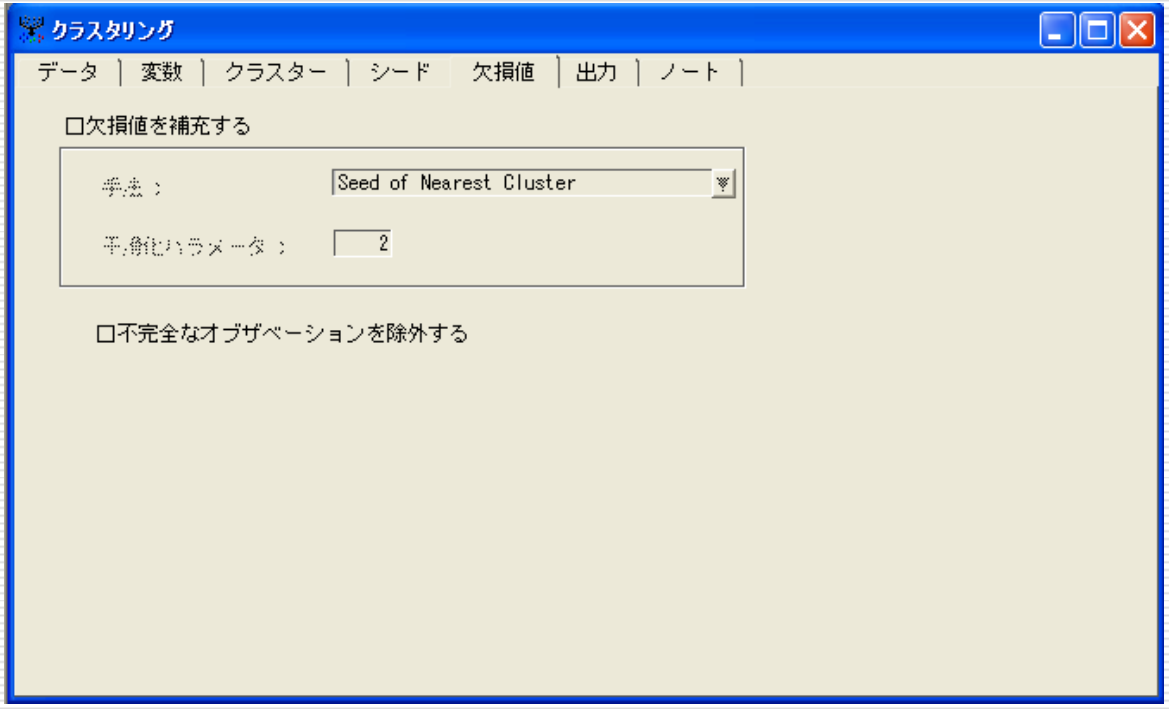
[シードタブ]
シードタブは[一般タブ][初期クラスターシードタブ][最終クラスターシードタブ]の3つのサブタブからなる



[一般タブ]
クラスター間の距離の基準等を決める
[初期クラスターシードタブ]
クラスターシードの初期化方法を指定できる
[最終クラスターシード]
反復を終了する条件を制御する

クラスタリングノードの詳細設定

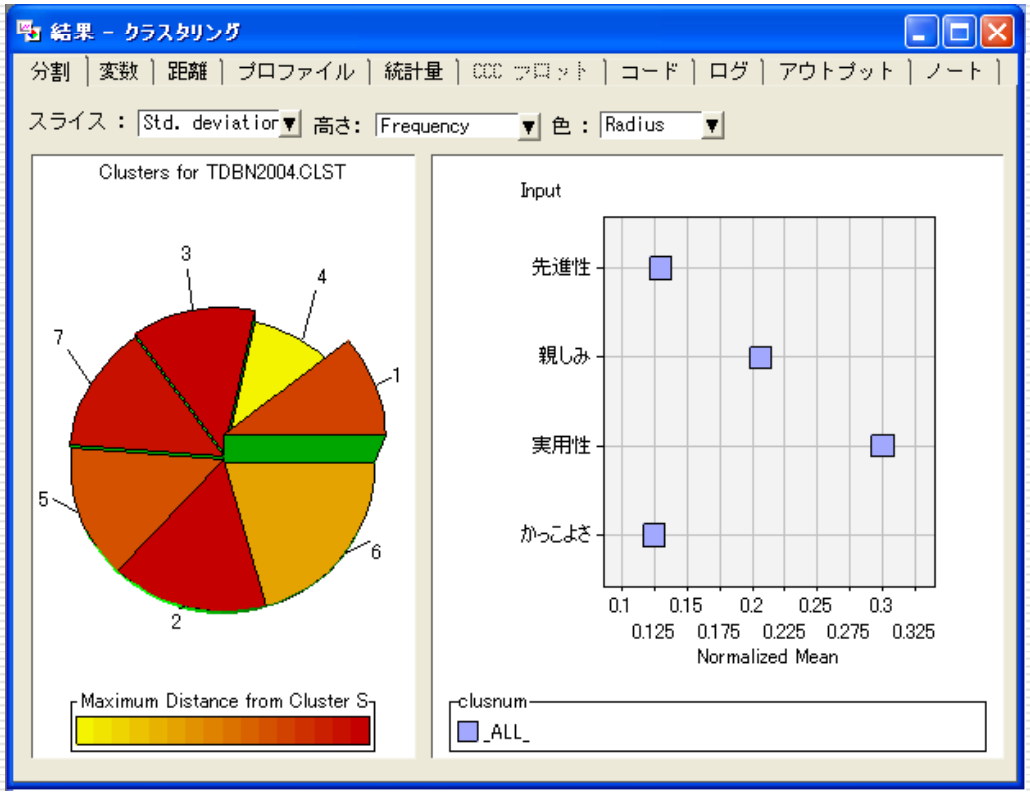
[欠損値タブ]
欠損値が含まれるオブザベーションの処理方法を指定する。



欠損値の処理には以下の二つの方法がある
1）クラスター分析のアルゴリズム実行時に欠損値を持つケースを除外する
2）出力データセット中の欠損値を置き換える

クラスタリングノードの結果

[分割タブ]
左側には立体的な円グラフが、右側にはクラスターセグメントの学習用データセット全体に対する入力平均が表示させられる

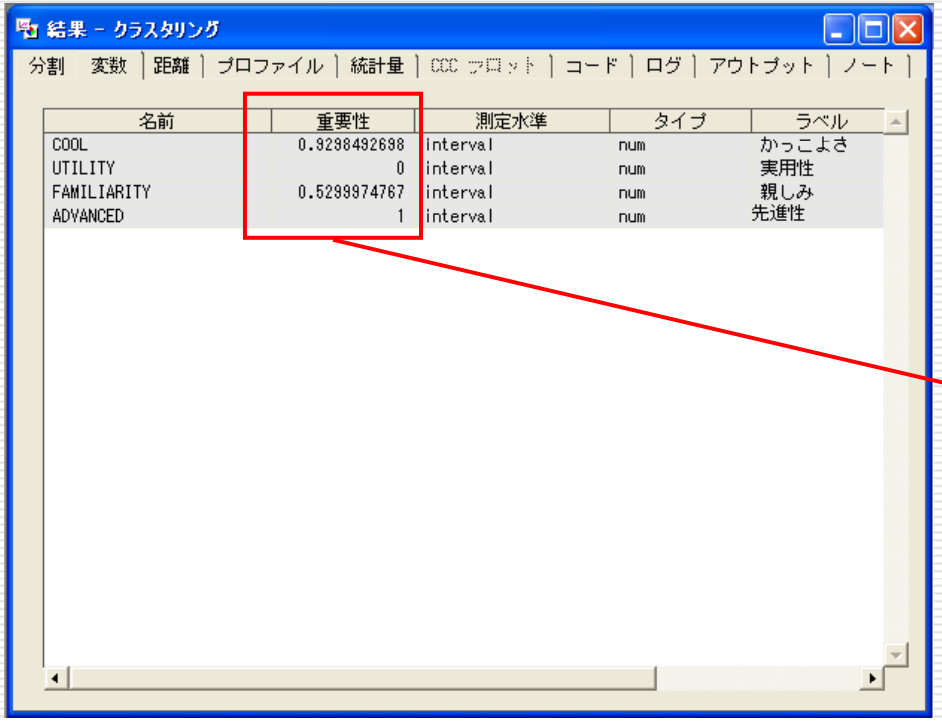


円グラフ表示の設定（スライス、高さ、色）を以下から選択できる

- 標準偏差：クラスター内のオブザベーションのバラつき
- 度数：クラスターに含まれるオブザベーションの数
- 半径：クラスターから最も遠いクラスターメンバーまでの距離

クラスタリングノードの結果

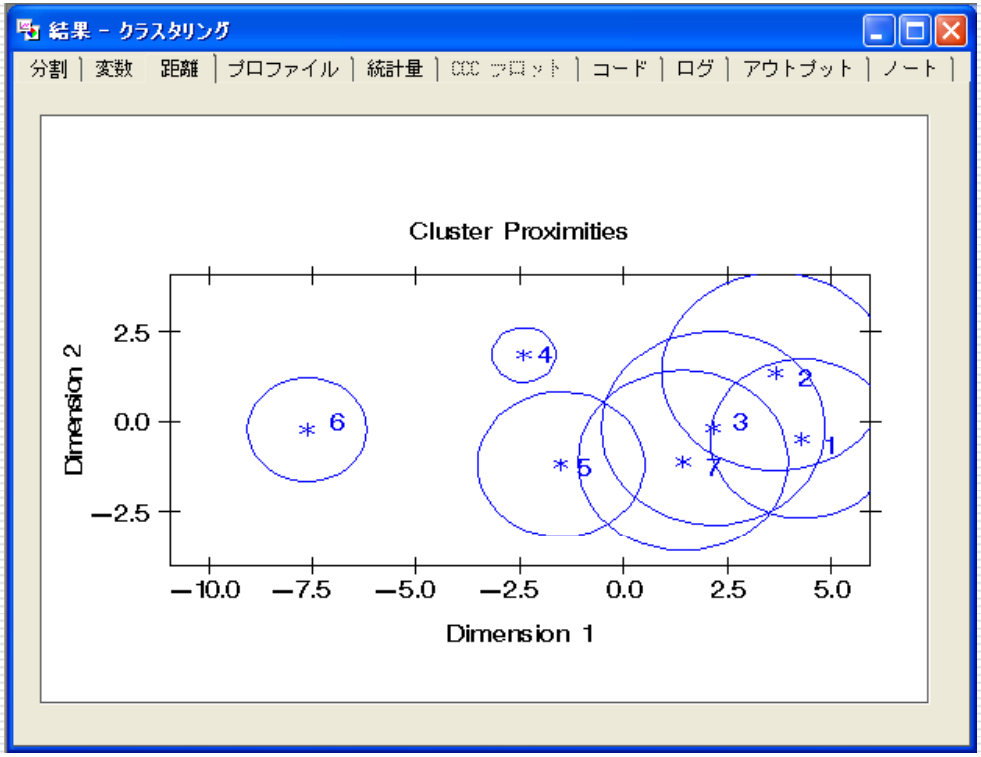
[変数タブ]
クラスター分析に使用される入力変数、測定水準、タイプ、およびラベルが表示される。「重要度」には各変数のクラスター形成に対する変数の重要度（0～1）が示される。



クラスターの形成に対する変数の重要度（価値の指標）を表示する。

クラスタリングノードの結果

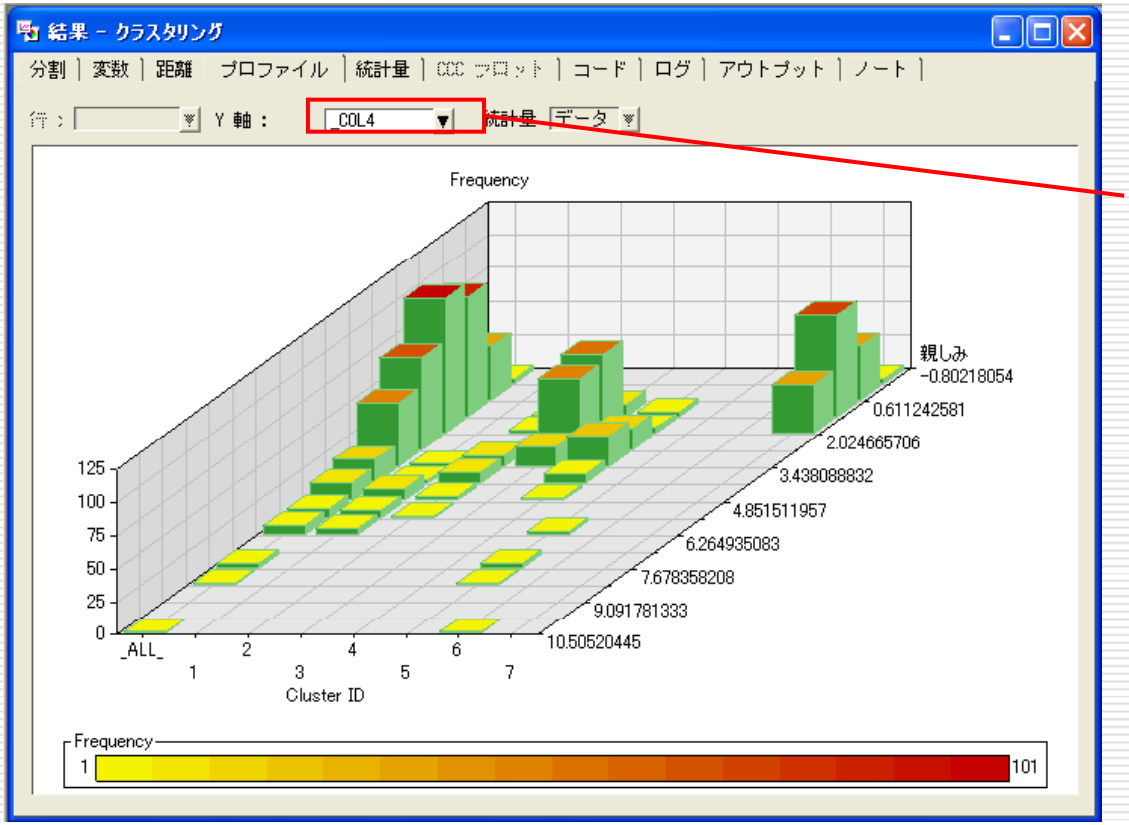
[距離タブ]
各クラスターのサイズおよびクラスター間の関係が表示される。



◆軸は、クラスター平均間の距離行列を入力に使用した多次元尺度法により決定される。
◆アスタリスクはクラスターの中心、円はクラスターの半径を表す。
◆表で見たい場合は、[プリファレンス]⇒[距離テーブル]を選択する

クラスタリングノードの結果

[プロファイルタブ]
各クラスターの変数がグラフ化して表示されている。

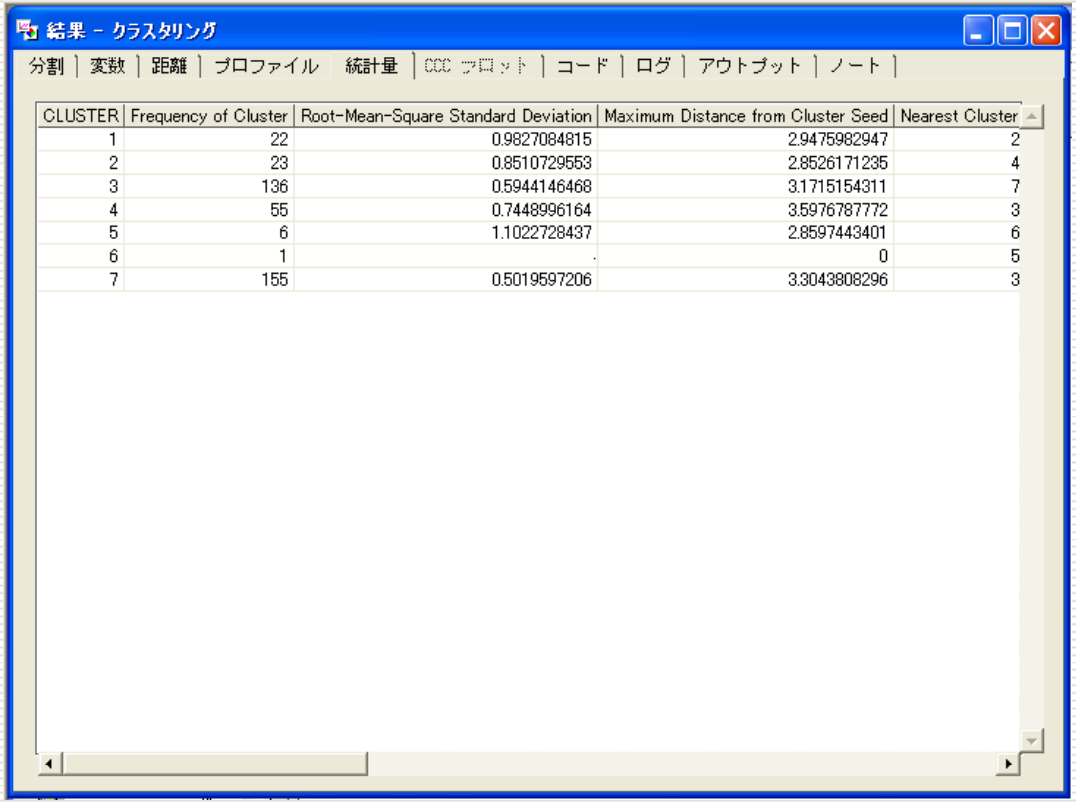


◆Y軸を選択することで、その変数がどのような分布になっているのかがわかる

◆この場合、間隔変数だけだが、カテゴリカル変数がある場合、[表示]⇒[カテゴリカル変数]で変えられる

クラスタリングノードの結果

[プロファイルタブ]
各クラスターに関する情報が表形式で表示される。



CLUSTER	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster
1	22	0.9827084815	2.9475982947	2
2	23	0.8510729553	2.8526171235	4
3	136	0.5944146468	3.1715154311	7
4	55	0.7448996164	3.5976787772	3
5	6	1.1022728437	2.8597443401	6
6	1	.	0	5
7	155	0.5019597206	3.3043808296	3

表示される情報は以下の通り
- ◆各クラスターのケースの数
- ◆自乗平均の平方根標準偏差
- ◆クラスターシードからの最大距離
- ◆最も近いクラスター
- ◆最も近いクラスターまでの距離
- ◆各入力変数の平均

クラスタリングノードの結果

クラスタープロファイルツリー
[表示]⇒[クラスターのプロフィールツリー]を選択すると、クラスター分割のルールをツリー状にして表示される。