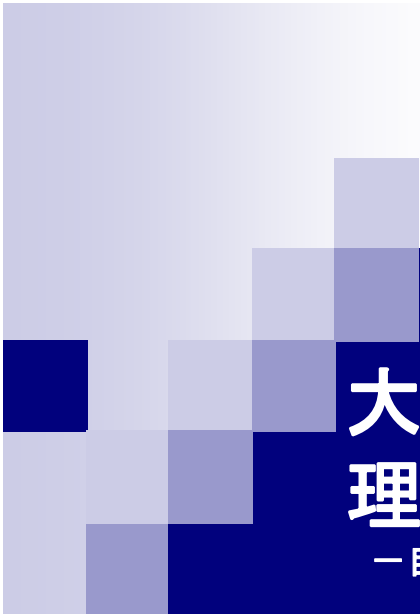




大規模データの中の多様性を理解する

ー自己組織化マップのサービス利用履歴への応用

群馬大学工学部情報工学科
関 庸一

at Splusユーザー会 2005.11.18

1



大規模データの特徴

■ 大規模データ

沢山の対象から、長期間に集められた多面的データ

= いろいろ混ざったデータ

■ 時間軸上での変化・対象の多様性が含まれている

□ 1つの対象が変化してゆく

- 時点ごとの様子の多様性
- 短期的/長期的変化

□ 対象の変化の構造 ... 変化の仕組みの多様性

■ データを仕分けして、説明する必要？

- 状態の様子のタイプ
- 変化の仕組みの違いの把握

2

多様性の捉え方

- 変化の姿
 - 振動変化...周期的変化
 - 傾向変化...時間的发展
- 変化の仕組みの多様性の姿
 - 全体を一つの共通モデルで説明できる？
 - 個体ごとにいろいろ ... 何もできなくなる
 - 個体を適当にグループ化すれば、共通のモデルで説明が可能
- 多様性をどう把握するか？
 - 主成分分析・双対尺度法(III類)など ... 線形な把握
 - クラスタ分析 ... 離散的な把握
 - SOM(Self-Organizing Map) ... 非線形な把握

3

SOM (Self-Organizing Map:自己組織化マップ)とは

Kohonenにより提案されたニューラルネットワークの一種

- 多次元データを分類、可視化する手法
- サンプルを低次元上の格子点(ノード)にクラスタリング
 - 入力：(サンプル×変量群)の行列データ
 - 出力：低次元(二次元)格子上のノード(クラスタ)ごとの特徴を表す参照ベクトル
 - マップ上において近隣ノードは類似する
 - サンプルは最も類似するノードに分類される
- k-means クラスタリング法との違い
 - K-means法：EMアルゴリズムの一種
 - 適当にクラスタ中心を用意したあと、以下を反復収束させる
 - M：個体を最も近い中心に配分
 - E：配分された個体の重心として、中心を再定義
 - サンプルを配置する空間が元データの空間でなく、新たに定義した、低次元空間：離散位相空間
 - 正方/六角などの指定した格子点(ノード)のみを利用
 - 各点に近傍が定義

4

SOMアルゴリズム

- あるランダムに選択された入力サンプル $z(t)$ と全ての参照ベクトル $m_i(t)$ とを比較し、ユークリッド距離が最小のノードを最整合ノード c とする

$$c = \arg \min_k \|z(t) - m_k(t)\|$$

- 以下の式で参照ベクトルを更新する

$$m_k(t+1) = m_k(t) + h_{ck}(t)[z(t) - m_k(t)]$$

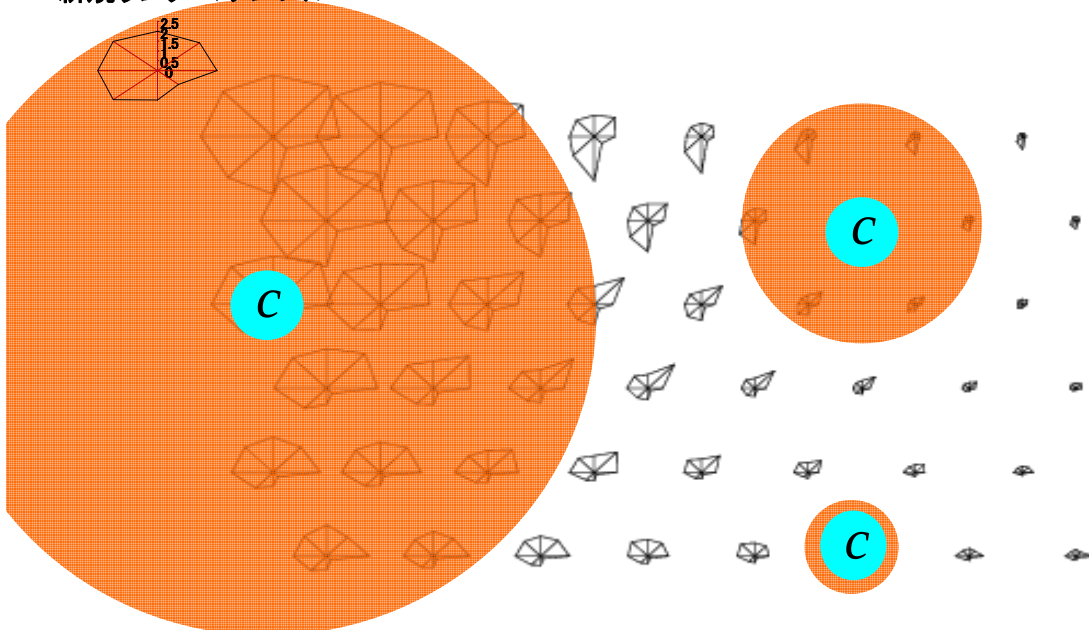
- $h_{ck}(t)$ は近傍関数と呼ばれ以下の式で与えられる
なお $N_c(t)$ は c の近傍集合範囲を表す減少関数

$$h_{ck}(t) = \begin{cases} \alpha(t) & k \in N_c(t) \quad (\alpha(t) : \text{学習率係数}) \\ 0 & k \notin N_c(t) \end{cases}$$

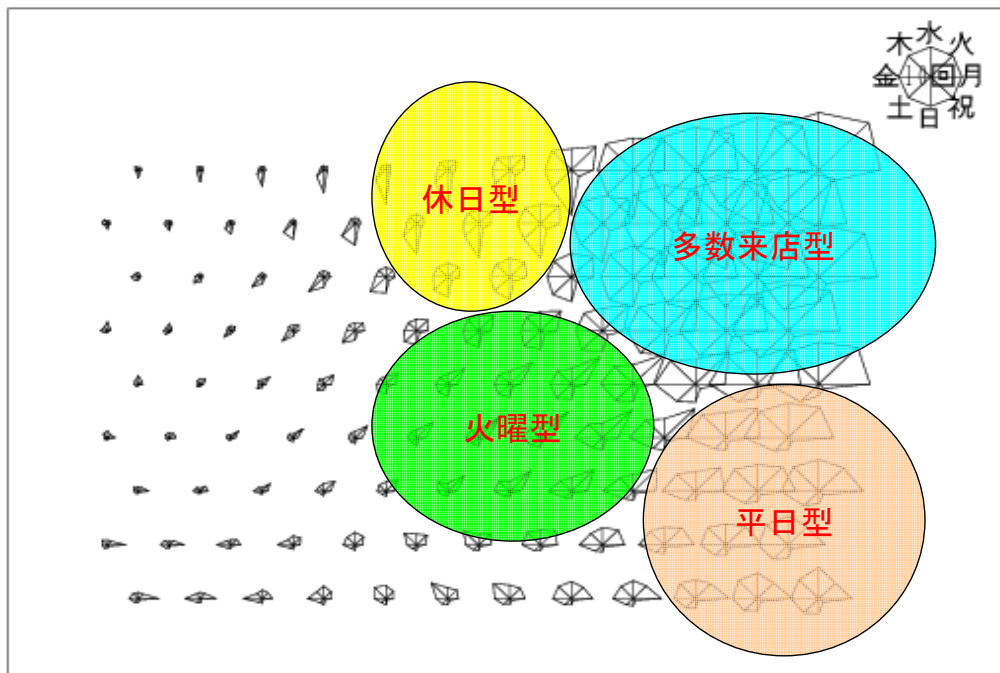
5

SOMアルゴリズム

新規ランダムサンプル



SOM(参照ベクトル収束結果例)



■使用データ : ある食品スーパーの半年間の
顧客×日付(8変数)の来店頻度行列 (祝日を別の曜日と定義)

7

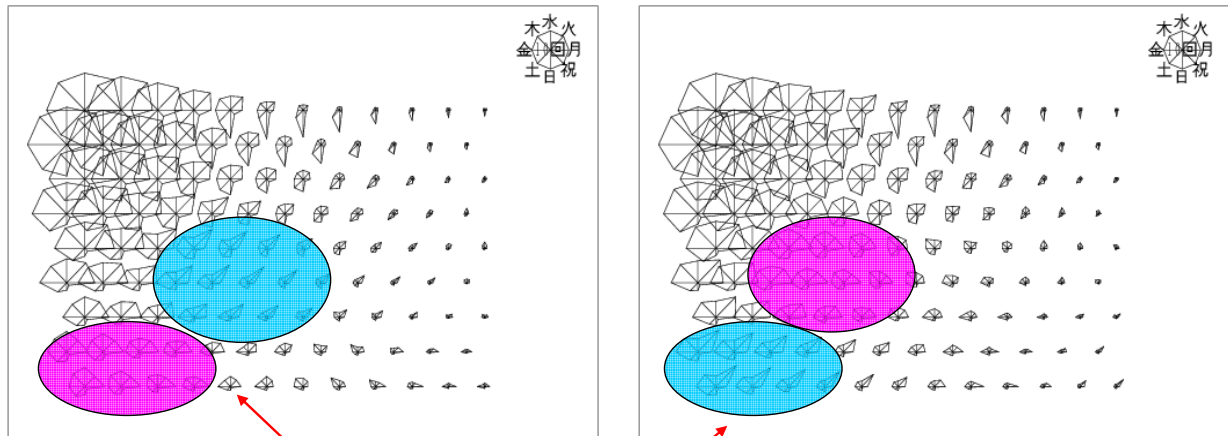
レーダーチャートを描画する関数

```
function(dat, x = 0, y = 0, names = NULL, color = rep(8, length(dat)), type = "b", llwd = 1, col = 1, lwd
= 1, axis = 2, dd = 0.1, cex = 1, r = 1, zero = 0, lty = 1, ddn = dd)
{ # (x,y) を中心として、dat をレーダーチャート描画する
  # type c: circle, s: stroke, b: both ... axis 軸をどの程度書くか
  if(!is.matrix(dat)) dat <- matrix(dat, nc = 1)
  if(!is.numeric(dat)) stop("dat is not numeric")
  n <- dim(dat)[1]; # 変数の数
  m <- dim(dat)[2]; # サンプルの数(m個のチャートを重ねがき)
  rad <- (0.5 - (0:(n - 1))/n * 2) * pi
  xx <- x + (dat - zero) * cos(rad) * r
  yy <- y + (dat - zero) * sin(rad) * r
  #----- axis 関連
  pretty.dat <- pretty(c(zero, dat), nint = 5)
  ddd <- (pretty.dat[2] - pretty.dat[1]) * dd * r
  Maxdat <- max(pretty.dat - zero) * r
  Mxx <- Maxdat * cos(rad)
  Myy <- Maxdat * sin(rad)
  for(i in 1:n) {
    if(axis) lines(c(x, x), c(y, y + Maxdat), lwd = 1, col = 8)
    if(axis > 1) text(x + 2 * ddd, y + (pretty.dat - zero) * r, pretty.dat, cex = cex, adj = 0)
    if(!is.null(names)) text(x + Mxx[i] * (1 + ddn), y + Myy[i] * (1 + ddn), names[i], cex = cex, srt =
      rad[i]/2/pi * 360, adj = 0)
  }
  if(axis) for(i in 1:length(pretty.dat)) lines(c(x, x + ddd), rep(y + (pretty.dat[i] - zero) * r, 2), lwd = 0.1, col = 8)
  #----- 本体
  if(n != length(color)) stop("Error:length of color is strange!")
  for(j in 1:m){
    for(i in 1:n) if(type == "s" | type == "b") lines(c(x, xx[i, j]), c(y, yy[i, j]), lwd = llwd, col = color[i])
    if(type == "c" | type == "b") lines(c(xx[, j], xx[1, j]), c(yy[, j], yy[1, j]), lwd = lwd, col = col, lty = lty)
  }
}
# plot(c(-10,10),c(-10,10),type="n",bty="n",xaxt="n",yaxt="n",xlab="",ylab="")
# pibar(1:6,0,0)
```

8

複数の局所最適解

来店傾向がほぼ同じノードの位置が違う結果



火曜型と平日型のノードの位置がほとんど入れ替わっている
各サンプルと最近隣ノードとの差の和は変わらない

9

SOM収束実験

- 解の妥当性・収束を確かめる
 - 局所最小の可能性がある
 - ⇒ ランダムな初期解から複数回の分析
 - 最も良い解を選択

- 結果の評価に使用する指標

- 各ノードの参照ベクトルの更新量の二乗和

$$\sum_{k=1}^K \{m_k(t+1) - m_k(t)\}^2$$

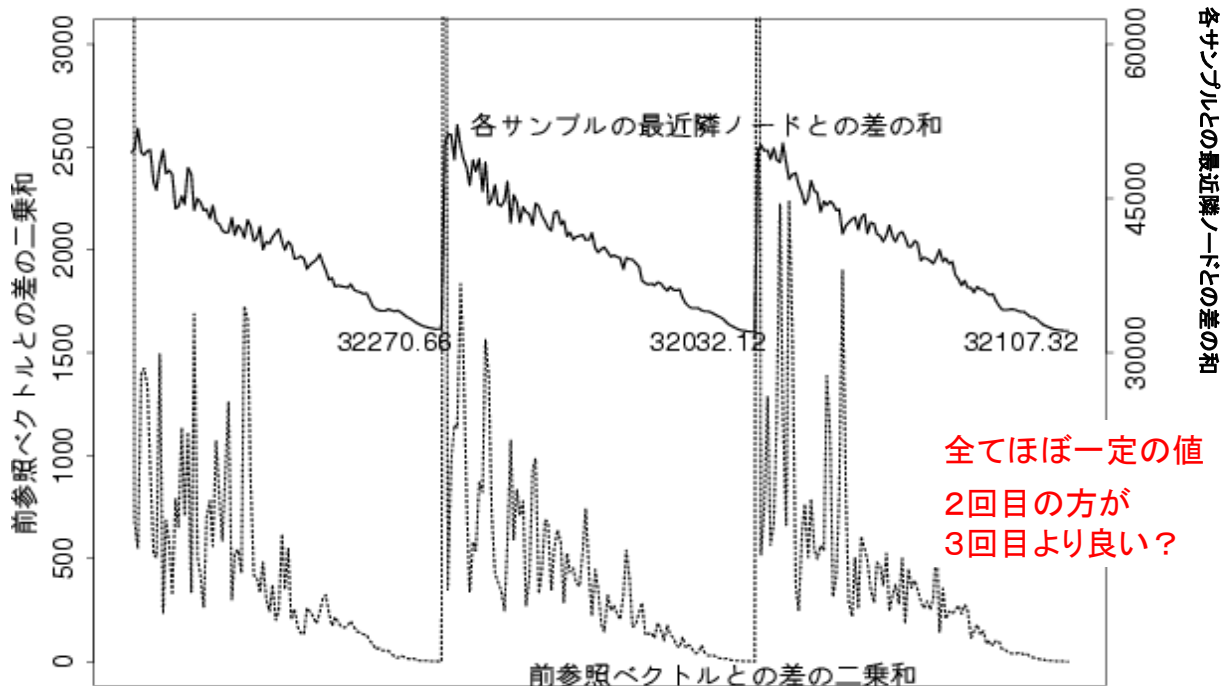
- 各サンプルと最近隣ノードとの差の和

$$\sum_{i=1}^n (\min_k \|z(t) - m_k(t)\|)$$

10

収束について

初期値: サンプル数多・反復数少・学習率係数大・近傍集合範囲広



11

事例: 百貨店3店舗での顧客買回りタイプの抽出

文献[4]

■ 顧客ごとの来店パターン → 買回りタイプ

3店舗のいずれに、何曜日に来店するか?

□ 原データ (平成15年度データ解析コンペティション提供)

対象2.5年間において、3店舗の何曜日に何回ずつ来たか?

49074人 × (8曜日 × 3店舗) の来店頻度行列

□ 各顧客を、2次元格子 (6 × 4 = 24個) のノードへ配分

- 各ノードは、ノードを代表する来店パターンをもつ。

(来店パターンは、近隣のノード間では類似するよう定められる。)

- 各顧客は、最も類似する来店パターンをもつノードに所属する。

12

買い回りタイプ分析 まとめ

1. 顧客の来店行動のタイプの識別
 - 三店舗の買い回りタイプ、来店曜日のタイプ
 - 大分には平日中心と土日中心の2タイプ. 土日は若年層
 - 別府タイプ: 別府近郊で女性の年輩層
 - わさだタイプ: 地域的には大分南部で若年層
 - 不活性タイプ: 遠方の顧客
2. タイプごとの属性・行動を集計: ex. 購入商品
 - それぞれ来店パタンの顧客は、どんな人々か?
 - どんな来店パタンの顧客が、何を購買しているか?
 - 大分・別府タイプは、婦人関連商品, セール品
 - 大分土日タイプは服飾雑貨を買い, セール品, 家庭用品を買わない
 - わさだタイプはベビー子供服, テナント
 - その他は、いろいろ満遍なく購入
3. ⇒ 何曜日にどの店舗では、何をすべきか?

13

事例: クレジットカードの利用履歴データ分析

文献[6]

■ データ概要

(平成16年度データ解析コンペティション提供)

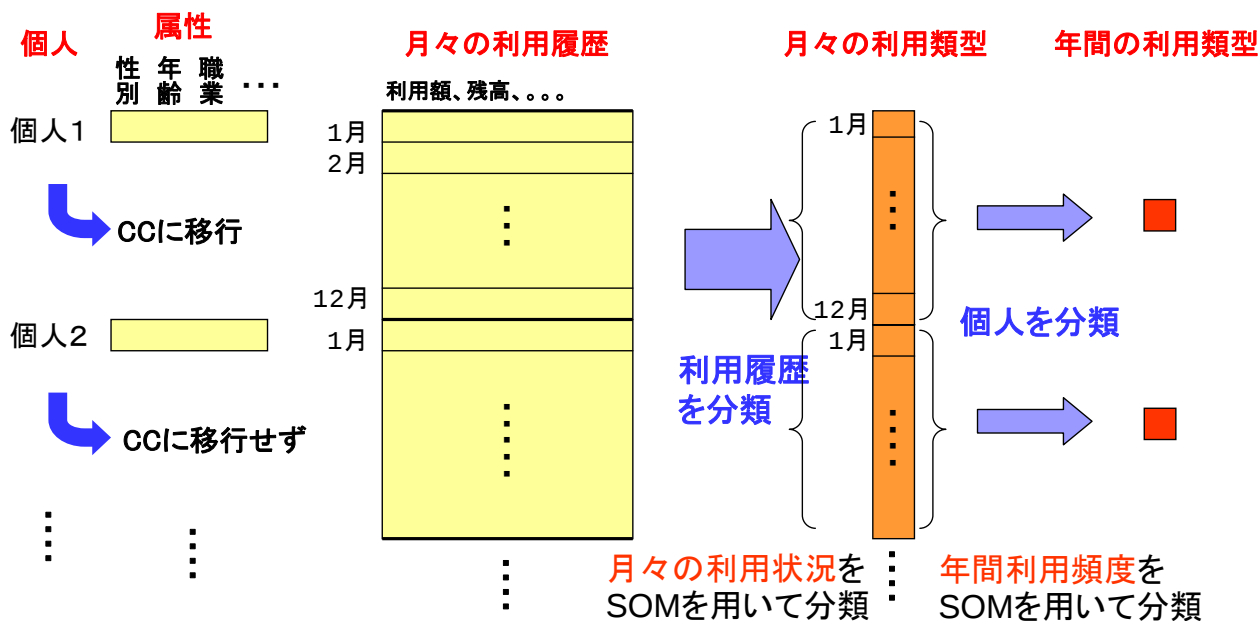
- 顧客の人数 54862人
- データの期間 2年間
 - 2002年1月から2003年12月

- クレジットカード会社にとって、最大の収益はキャッシングによる返済金利
- キャッシングに移行する顧客を予測: 会社にとって有益
- しかし、顧客は多様なカード利用
→ 顧客を類型ごと分類して予測

- 1年目の利用履歴と個人属性などから、
2年目にキャッシングに移行する顧客を予測
➡ 1年目の顧客の特徴の把握?

14

利用履歴データの集約方法



個人での年間の利用履歴を
年間利用類型として集約し、顧客を層別

移行予測

15

クレジットカード まとめ

- 各種サービスの多様な利用履歴で顧客類型を構成
 - 月々の利用履歴から月間利用類型に集約
 - 月間利用の7つのタイプを把握
 - 2次元分布関数間距離を用いることで、月間利用類型から年間利用類型に集約
 - 顧客の1年間の利用の分類法を提案
 - CS、CC、CSリボの3タイプを把握
- 類型ごとの予測モデルを統合し、異なる予測が必要な類型群を構成
 - 顧客類型で用いなかった変数から、顧客類型群ごとのキャッシングの移行ロジスティックモデルを見出すことができた
 - 顧客類型群ごとに異なる説明変数が有効
 - 曜日、月初末、顧客属性

16

多様なデータの料理法

- 多面的観測の大規模データ ≠ 無作為抽出標本
対象間の独立性はいえても、同一分布でない。
対象の集団に構造が含まれている。
 - データの要約 ... 探索的データ解析
 - 対象の層別 ... 目的に適合した層の構成
 - 説明モデルの構築 ... 確証的データ解析
- 対象の構造を上手に見つけて、目的にあった分析を可能とする方法論！

cf. データの持つ情報密度

- S/N比 誤差の含まれる割合
- どこまで詳細に分析してよいか？

17

参考文献

1. Kohonen T.: ``Self-Organizing Maps, 3rd ed.“, Springer, 2001.
2. 関 庸一, ID付きPOSデータからの顧客行動パタンの抽出, オペレーションズ・リサーチ, 48, 2, 2003, pp.75-82.
3. 関 庸一, 小茂田 宏, 石原淳一郎, 事象系列のストリング分析 -- 百貨店における買回り行動の分析, オペレーションズ・リサーチ, 49, 2, 2004, pp.67-74.
4. 渡邊 亮, 北村裕人, 星野直人, 関 庸一, 買回りタイプによる顧客購買行動の理解, オペレーションズ・リサーチ, 50, 9, 2005, pp.42-51.
5. 関 庸一, 小茂田 宏, SOMノード群上での回帰モデル統合, 日本経営工学会春季大会予稿集, 150-151, 高崎経済大学, 2004年5月.
6. 関 庸一, 長井 歩, 渡邊 亮, 石原 淳一郎, クレジットカード利用履歴を利用したキャッシング移行予測, 経営情報学会2005年春季研究発表大会予稿集, pp.S56-S59, 早稲田大学, 2005年6月

18