クラスタ中心を再計算しない大規模データのための非階層的クラスタリング

中山 厚穂（首都大学東京大学院社会科学研究科経営学専攻）
出口 慎二（データエクスプローリング）
烏谷 正彦（カスタマー・コミュニケーションズ株式会社）

日本行動計量学会第44回大会　2016年8月31日　札幌学院大学

はじめに

- 大規模データに対して，非階層クラスタリングの代表的な手法であるK-means法を用いた際に効率的なクラスタリング実施するための方法について提案
  - 出口・中山・高崎(2014) (http://119.245.205.198/info/docs/out3.pdf)において提案したクラスタ中心を再計算しない大規模データに対する非階層的クラスタリング法を拡張した分析法

- クラスタ中心を再計算せずにクラスタリングを実行
  - クラスタ中心の移動は起こらず，各ケースの再分類は生じない
- クラスタリングする際の計算をシンプルにし使うリソースが少なくて済む
  - データをすべて読み込む必要がなく，データ全体はシーケンシャルに1回読めば実行でき，大きなメモリを必要としない

出口慎二・中山厚穂・高崎祐哉（2014）．大規模データのクラスタリング－クラスタ中心を再計算しない非階層的クラスタリング－．日本行動計量学会第42回大会抄録集，82–83.

出口・中山・高崎(2014)の分析手順

1. ID付POSデータから，「会員（レコード）」×「商品カテゴリ（フィールド）」のデータを作る．値は，ここでは「購買金額（値）」とする．
2. 「会員」×「商品カテゴリ」のデータから，ちいさな（具体的には，N=2,000 の）データを複数（ここでは，5つ）作る．これは，単純に，等間隔抽出で行う．
3. 複数の（5つの）サンプリングデータの階層的クラスタリングを行い，その結果から，全データのクラスタ構造を仮定する．
4. 仮定したクラスタ構造をもとに，クラスタ中心行列を作る．具体的には以下の通り．
    1. 複数のサンプリングデータ（5つ）それぞれを同じ数のクラスタ（ここでは6つ）に分ける．
    2. いずれも性質が同等と思われる複数のクラスタについて，算術平均によりクラスタ中心を計算する．
5. 作成したクラスタ中心行列を使って，あらためて，全データ（N=4,946,955）のクラスタリング（クラスタ中心の再計算を行わない）を行う．

本提案手法の分析手順

1. ID付POSデータから，「会員（レコード）」×「商品ないし商品カテゴリ（フィールド）」のデータを作る．値は，ここでは「購買金額」とする．
2. 会員×商品のデータから，小さな（階層的クラスタリングおよび結果の解釈を容易に行える程度に）データを複数作る．これは，単純に，会員を等間隔抽出して行う．
3. 複数のサンプリングデータをk-means法で想定されるクラスタ数よりも多めのクラスタ数(今回の分析ではクラスタ数は20)を指定して分析し，その結果得られたそれぞれの初期値を階層クラスター分析(Ward法) し，全データのクラスタ中心を作成する．
4. クラスタ中心行列を使って，あらためて，全データのクラスタリングを行う．もっともクラスタ中心が「近い」クラスタに，各ケースを振り分け，クラスタ中心の再計算は行わない．

はじめに

- k-means法は初期値に依存
  - 初期値を複数回ランダムに発生させて分析し，それらの中で分類の評価関数が最小となるものを選ぶ必要がある
  - 大規模データにおいては何度もk-means法で分析すると時間が非常にかかる場合もある
- サンプリングデータを利用することで最適な初期値を求めようとする研究も行われている (e.g. Bradley, 1998; Fahim, Salem, Torkey, Ramadan, Saake, 2009).
  - サンプリングデータから初期値を作成しているため全データにおける分析では反復は少なくて済む
- 表頭（変数）を階層クラスター法により分析し，その変数のクラスター数から表側（対象）のクラスター数に目星をつけることも行われている.
  - 変数のクラスタリングした結果に対して，対象の平均値を求めて初期値とすることも可能
  - 変数と対象が共通のクラスター構造となるという保証はない

Fahim, Salem, Torkey, Ramadan, Saake (2009)

1. 全データから，複数のサンプリングデータ抽出し，k-means法により同一のクラスタ数で分析する
2. 得られたそれぞれのクラスタ中心を，データとしてk-means法で分析する
3. その分析により得られたクラスタ中心を初期値として，全データに対してk-means法で分析する

提案手法の新規性

- k-means法を実施する際にクラスタ中心を再計算するかどうか
  - 本提案手法ではクラスタ中心を再計算しないため，既にクラスタ中心は定まっている
  - 分析結果はサンプリングデータのクラスタリングが示す通りとなる
  - クラスタ構造を見るだけなら必ずしも全てのデータを見る必要はない
  - 必ずしも特別なマシンスペックなどを要さずに，大量のデータを処理できる

- 全データのクラスタ中心を複数のサンプリングデータから作成する際にk-means法を用いるのか階層クラスター分析を用いるか
  - 初期値を作成する際に，階層クラスター分析を用いる
  - いくつのクラスタ数を採用するかをその結合過程から判断することが可能

分析データ

- JICFSコード
  - 分析に用いたのはカスタマー・コミュニケーションズ株式会社からご提供いただいたTRUE DATAのドラッグストアのID付きPOSデータである(http://www.truedata.co.jp/about/)
  - JICFS分類についてレベル4（6桁）を使用. うち「2」まで（「食品」と「日用品」）を使用.
    1　食品
    2　日用品
    3　文化用品
    . . .
  - 商品コードの数（「食品」「日用品」のみ）は745個.
  - 実購買があるのは629個.

  →この629個のコードを使用.

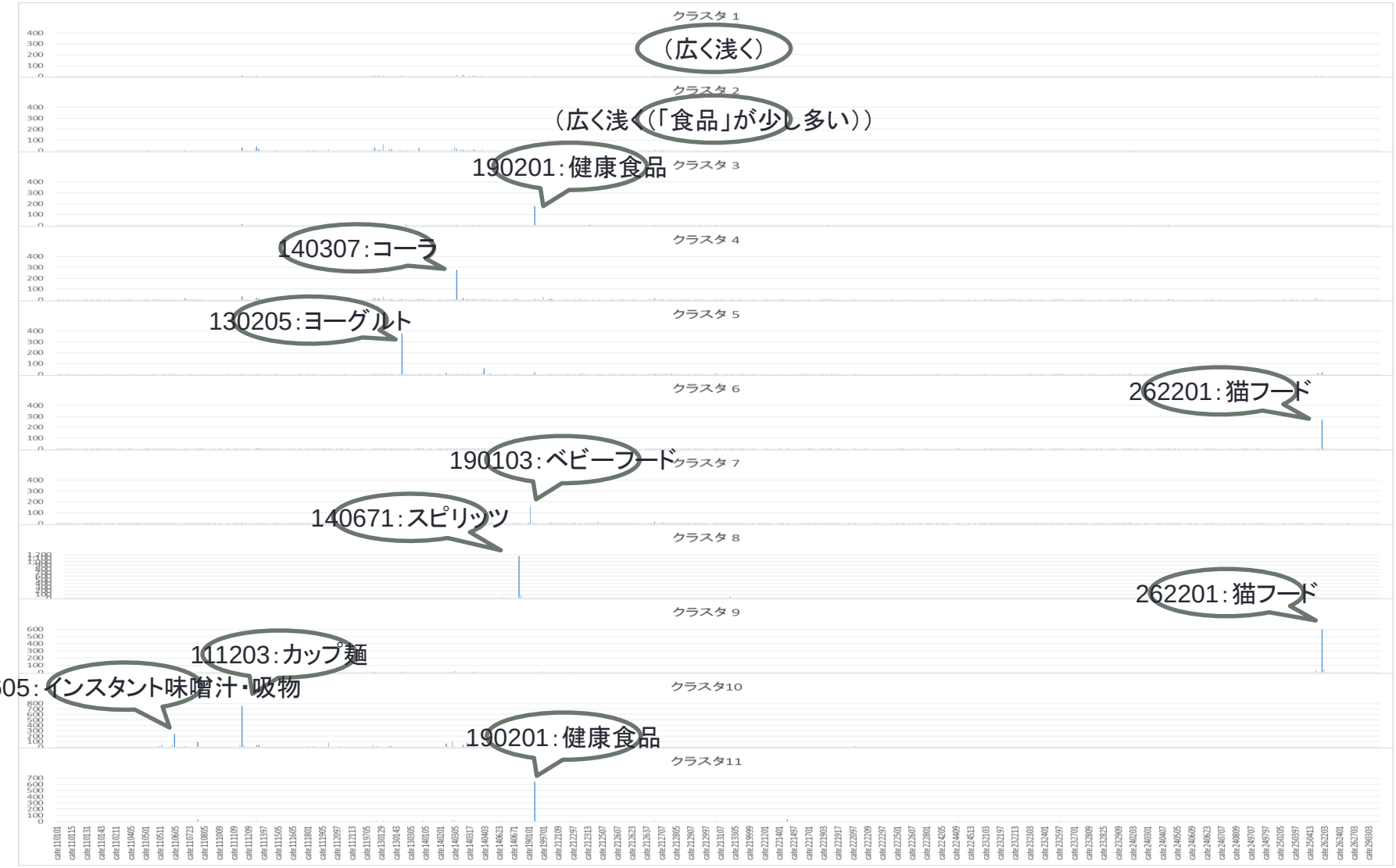
- 該当するレコード数
  - 全購買データ232,030,245レコードを属性ファイル上の会員番号でまとめると6,608,387レコード.
  - 属性ファイル上には5,821,992人分の情報.
  - 属性情報と購買情報の双方がある会員数は4,671,856レコード.

2段階のクラスタリング

- 1段階目：目標クラスタの作成
  - 初期的解析
    - 20のサンプリングデータ（1標本n=2,000）を作成.
    - 機械的に1データごとk-means法で20のクラスタを作成.
    - 各クラスタごとクラスタ平均を計算（20データ×20クラスタ＝400のクラスタ平均）.
    - 変数の数はいずれも632（629(変数)+3 (属性)）.

      →n=400のデータが1つできる
  - クラスタ中心の作成
    - 初期的解析で作ったn=400のデータを使い階層的クラスタリングを実行（平方距離+ward法）. 樹形図から11クラスタを採用. 11のクラスタ平均（クラスタ中心）を作成.（次頁図参照）

初期的解析後のクラスタ中心→目標クラスタ



最終クラスタ

- 2段階目：本解析
  - 11のクラスタ平均のうち，最も近いクラスタに6,000,000ほどの個々のデータを振り分ける．
    - 結果，8クラスタが採用された．→3クラスタは採用されなかった．

最終クラスタ

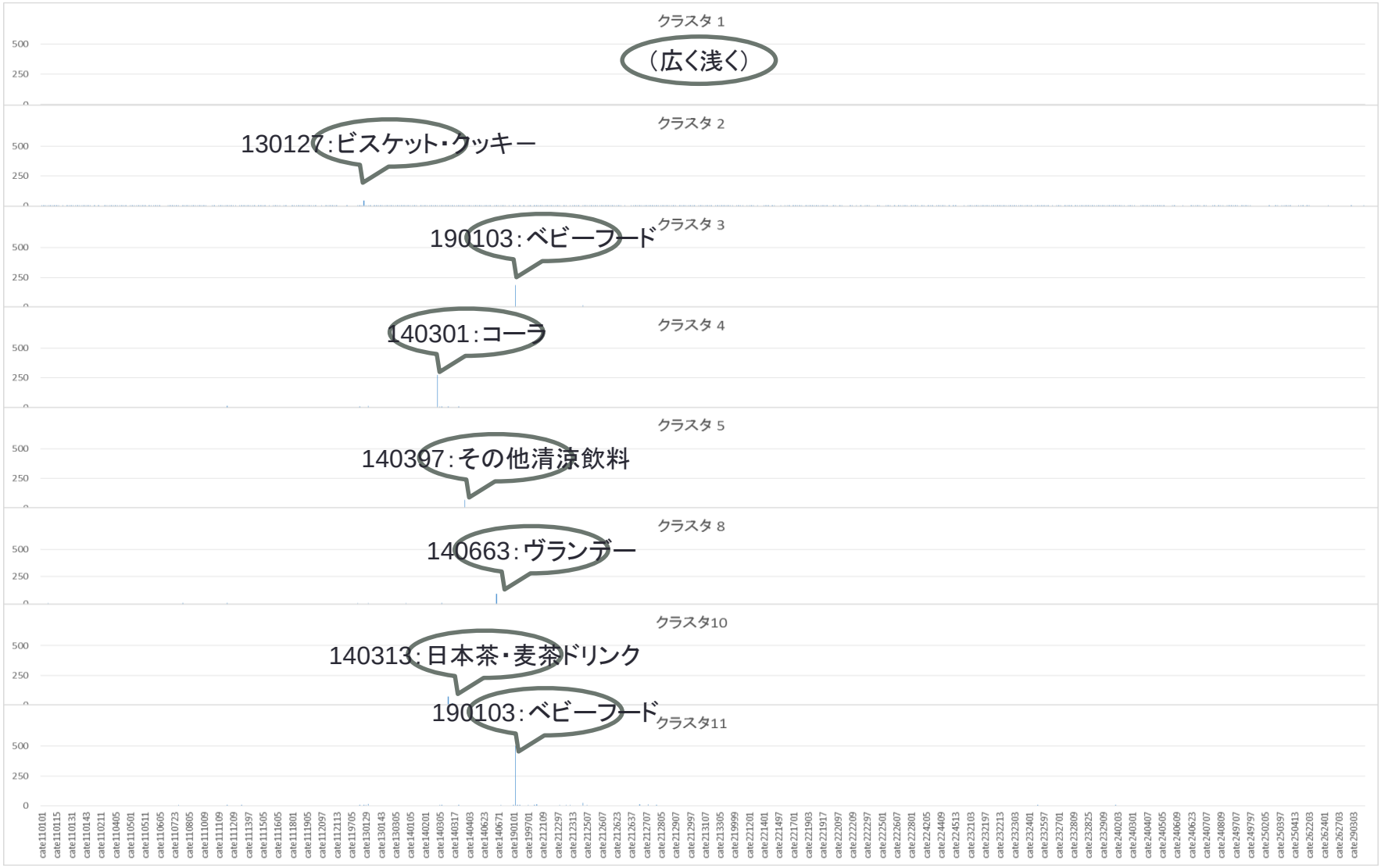
		度数	パーセント	有効パーセント	累積パーセント
有効	1	4668454	99.9	99.9	99.9
	2	277	.0	.0	99.9
	3	1735	.0	.0	100.0
	4	97	.0	.0	100.0
	5	453	.0	.0	100.0
	8	4	.0	.0	100.0
	10	747	.0	.0	100.0
	11	89	.0	.0	100.0
	合計	4671856	100.0	100.0	

各クラスタの男女比・平均年齢

clus	全体	男女	年齢
クラスタ 1	4,668,454	1.83	50.28
クラスタ 2	277	1.73	58.99
クラスタ 3	1,735	1.89	40.11
クラスタ 4	97	1.63	52.62
クラスタ 5	453	1.77	58.10
クラスタ 8	4	1.25	59.25
クラスタ10	747	1.65	57.12
クラスタ11	89	1.87	42.21

※「男女」は女性を「2」男性を「1」とした場合の平均値．「年齢」は平均値．

最終クラスタ


クラスタ 1
（広く浅く）
クラスタ 2
130127:ビスケット・クッキー
クラスタ 3
190103:ベビーフード
クラスタ 4
140301:コーラ
クラスタ 5
140397:その他清涼飲料
クラスタ 8
140663:ヴランデー
クラスタ10
140313:日本茶・麦茶ドリンク
クラスタ11
190103:ベビーフード
500
250
0

結果

- 最終クラスタは8クラスタに分かれた(うち1つは度数が1桁).
- 初期解は11クラスタだったが, 3つのクラスタには最も近いデータはなかった.
  - 無かったクラスタ
    (カッコ内は最頻カテゴリ)
    - 6(262201:猫フード)
    - 7(190103:ベビーフード)
    - 9(262201:猫フード)
- 目標クラスタと最終クラスタは似ていた.

各クラスタの「男女」と平均「年齢」			
clus	全体	男女	年齢
クラスタ 1	4,668,454	1.83	50.28
クラスタ 2	277	1.73	58.99
クラスタ 3	1,735	1.89	40.11
クラスタ 4	97	1.63	52.62
クラスタ 5	453	1.77	58.10
クラスタ 8	4	1.25	59.25
クラスタ10	747	1.65	57.12
クラスタ11	89	1.87	42.21

目標クラスタ	
code	
-	広く浅く
130127	ビスケット・クッキー
190103	ベビーフード
140301	コーラ
140397	その他清涼飲料
140663	ブランデー
140313	日本茶・麦茶ドリンク
190103	ベビーフード

最終クラスタ	
code	
-	広く浅く
「1」番台	「食品」が少し多い
190201	健康食品
140307	コーラ
130205	ヨーグルト
140671	スピリッツ
110605, 111203	インスタント味噌汁・吸い物, カップ麺
190201	健康食品

考察

- ランダム初期解の是非
  - 出口・中山・高崎(2014)(http://119.245.205.198/info/docs/out3.pdf) と同様，解釈的にサンプリングデータから予想されるクラスタ数/クラスタ平均を計算して初期解を求めることもできる.
  - 今回は629変数と解釈するには変数が多かったため，機械的な処理を検討.

→ただしSPSSのディフォルトでは「平均値」の値が望ましくない.

- ランダム初期解の是非
  - 選ばれなかった解もあった.（クラスタ6，7，9）
    - 初期解の中に近い解が無ければ選ばれない
  - クラスタ4や11は，当てはまりが悪かった

	各クラスタの中心からの距離の平均値	度数	標準偏差
クラスタ 1	319	4,668,454	56,530
クラスタ 2	175	277	191
クラスタ 3	1,408	1,735	2,052
クラスタ 4	3,502	97	4,879
クラスタ 5	378	453	1,700
クラスタ 8	551	4	216
クラスタ10	621	747	463
クラスタ11	11,476	89	20,416

謝辞

- 本研究は，JSPS科研費基盤研究（C）（課題番号16K00052）の助成を受けて実施いたしました.

- また，ID付きPOSデータにつきましてはカスタマー・コミュニケーションズ株式会社よりTRUE DATAをご提供頂きました.ここに記して厚く御礼申し上げます.