SAS ユーザー総会 2017

ID-POSだけで始める新しいRFM分析: データステップで実装するモンテカルロ法を用いて

片桐　智志[1]

([1]ネイチャーインサイト株式会社)

A New RFM Analysis Only with Scanner Data: Using Monte Carlo Simulation Implemented by a Data Step

KATAGIRI, Satoshi

Nature Insight, Co., Ltd.

要旨:

最近に提案された新しいRFM指標の推定方法では、従来のRFM分析の問題を解消したが、モンテカルロ法による計算が必要になる。我々はSASの基本的な機能であるデータステップを用いて高速に実行できることを説明する。

キーワード: RFM分析 マーケティング POSデータ MCMC モンテカルロ法 データステップ

目次

1. RFM分析の利点
2. 従来のRFM分析の問題点
3. 阿部 (2011) の方法
4. SAS での実装方法
5. 実験結果
6. 結論

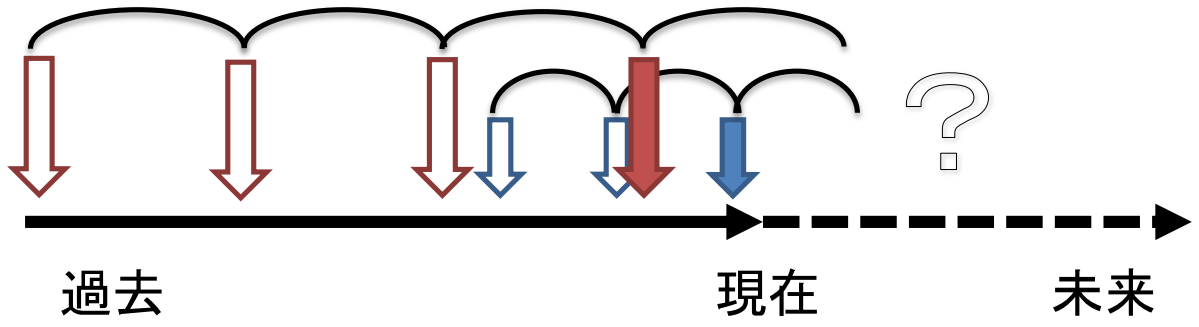
RFM分析の利点

- 単純な購買履歴データさえあれば可能
  - 購買者の個人ID、購買日、金額
  - ID-POS
  - オンラインショッピングのトランザクション
- シンプルで納得しやすい

RFM分析の問題点

- RFMのR (リセンシー) =顧客の離脱?
- 直近の購買日が離脱を表すとは限らない

青より赤のほうが離脱しやすい？



新しいRFM分析の方法

- 阿部(2011)の提案
- 購買回数と購買時期を同時に考慮した確率モデル
- 従来のRFM分析と同様のデータで実行可能
  - 追加データでさらに細かい分析も可能

阿部(2011)の提案する方法

- 最低限必要な顧客情報は以下

1. 最初・最後に購買した日
2. 観測開始日・終了日
3. 購買回数・平均購買額

阿部(2011)の提案する方法

- 顧客ごとに個別パラメータμ_i, λ_i, η_i を仮定する

1. 離脱までの期間は平均$1/\mu_i$ の指数分布
2. 各顧客の購買回数は平均λ_iのポアソン分布
3. 1回の購買額の対数は平均η_i の正規分布
4. 個別パラメータ μ_i, λ_i, η_i は対数正規分布

$$[\ln \mu_i \ln \lambda_i \ln \eta_i] \sim \mathcal{N}(\theta, \Sigma)$$

阿部(2011)の提案する方法

- 顧客情報を回帰することも可能:

$$\begin{bmatrix} \ln \mu_i \\ \ln \lambda_i \\ \ln \eta_i \end{bmatrix} \sim \boldsymbol{X_i d} + \boldsymbol{\varepsilon}_i$$

- 顧客の特徴がどのように購買に現れるのかを調べることができる

※今回はデータが存在しないため不可

阿部(2011)の提案する方法

- 個別パラメータ μ, λ, η から以下を計算できる
- 平均購買額
- ある期間内の期待購買回数
- ある期間内の期待購買額
- 一定期間後の離脱確率
- 生涯顧客価値

計算の流れ

- データ拡張法(Tanner & Wong,1987)を用いて計算

1. 顧客ごとに以下 (a) ~ (d) を計算
   a. λ_i, μ_iの暫定解から離脱の有無を乱数で判定
   b. (a) で離脱の場合、λ_i, μ_iの暫定解から生存期間を計算
   c. (a) , (b) より得た尤度でメトロポリス-ヘイスティングス法で λ_i, μ_i を新たにサンプリング
   d. ギブス・サンプラーで η_i を新たにサンプリング
2. 1.で得た λ_i, μ_i, η_iの平均と分散から共通パラメータ更新

SASデータステップでの実装

- 実は相性が良いデータステップとモンテカルロ法
  - 1オブザベーションごとの処理
  - 高速なループ処理

SASでの実装方法のポイント

1. 出力データはパラメータだけのワイド形式
2. データをハッシュオブジェクトとして読み込む
3. 確率の計算にFCMPプロシジャで登録した関数・サブルーチンを使う

実験

- Webで一般公開されているデータで実験する
- CDNOW のデータ
- Online Retail のデータ

CDNOWのデータ

- 配布元 http://www.brucehardie.com/
- CDNOW の販売履歴のデータ
- 1997~1998年の販売履歴うち39週間で推定

	N	平均	標準偏差	最小値	中央値	最大値
リセンシー(日数)	1000	180.48	77.10	1.00	206.00	272.00
購買期間(日数)	1000	48.51	75.90	0.00	0.00	269.00
観測期間(日数)	1000	228.98	24.09	34.00	229.00	272.00
フリクエンシー(回数)	1000	1.97	1.86	1.00	1.00	20.00
平均購買額($)	1000	31.49	25.60	3.99	23.54	259.96

CDNOWの推定結果の検証

- 39週間分で訓練しホールドアウト検証
- 相関係数と平均自乗誤差(MSE)を確認
- 1か月に短縮すると精度悪化

データ	指標	購買回数	購買平均額	購買総額
訓練	MSE	1.85	217.32	7550.70
検証	MSE	5.16	335.47	11750.00
訓練	相関	0.97	0.94	0.96
検証	相関	0.87	0.84	0.91

Online Retail のデータ

- 配布元 http://archive.ics.uci.edu/ml/datasets/Online+Retail
- 英国オンライン通販企業の販売履歴データ
- 2010~2011年の販売履歴うち6か月で推定

	N	平均	標準偏差	最小値	中央値	最大値
リセンシー(日数)	1000	64.51	51.65	1.00	52.00	181.00
購買期間(日数)	1000	46.18	58.50	0.00	0.00	179.00
観測期間(日数)	1000	110.69	53.76	1.00	116.00	181.00
フリクエンシー(回数)	1000	2.31	2.71	1.00	1.00	43.00
平均購買額(£)	1000	413.20	654.36	3.75	298.82	12864.20

Online Retail の推定結果の検証

- 6か月のデータと全体のデータでホールドアウト検証
- 相関係数と平均自乗誤差(MSE)を確認

データ	指標	購買回数	購買平均額	購買総額
訓練	MSE	2.68	287533.00	1930000.00
検証	MSE	21.23	249295.00	13100000.00
訓練	相関	0.99	0.76	0.86
検証	相関	0.93	0.74	0.87

計算速度の検証

- 1000人分のデータで12,000回サンプリング
- C++のプログラムと比して遜色ない計算速度

表：計算速度の比較(単位: 秒)

	N	平均	標準偏差	最小値	最大値
C++	100	257.025	2.324	251.63	261.89
データステップ	100	325.039	2.66	317.551	331.411

これからの課題

- 購買額の精度
- ハイパーパラメータ調整
  - 計算時間が早くとも、調整が必要で何度も繰り返し計算して確認する必要がある
- 収束の遅さ
  - バーンインを長く取る必要がある

結論

- 理論モデルに基づくRFM分析の手法
- 顧客ごとに様々な指標の計算が可能
- SAS の基本的な機能だけで計算可能
- C++と比べても遜色のない処理速度

参考文献

- Tanner, M. A. & Wong, W. H. (1987) “The Calculation of Posterior Distributions by Data Augmentation: Rejoinder,” *Journal of the American Statistical Association,* 82(398), pp. 548-550.
- 阿部 誠 (2011)『RFM指標と顧客生涯価値: 階層ベイズモデルを使った非契約型顧客関係管理における消費者行動の分析』, *日本統計学会誌,* 41巻1号, pp. 51-81