定額制サービスにおける優良顧客の購買パターン分析

岩田具治† 斉藤和巳† 山田武士†

† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 619–0237 京都府相楽郡精華町光台 2–4
E-mail: †{iwata,saito,yamada}@cslab.kecl.ntt.co.jp

あらまし 近年，音楽や映像配信において定額制サービスが注目されている．利用回数に関わらず定額にすることにより，ユーザは一定の期間，料金を気にせず利用することができ，また，オンラインストアにとっては安定した収益が見込めるという利点がある．定額制サービスを提供するオンラインストアにとって，長期間契約してくれるユーザが優良顧客である．優良顧客に特徴的な購買パターンを分析することにより，配信コンテンツの選定や，優良顧客の特定に関する有益な情報が得られる．本稿では，生存時間解析を用いて購買パターンと契約期間の関係をモデル化し，大規模購買履歴データから高速に優良顧客に特徴的な購買パターンを抽出するためのアルゴリズムを提案する．人工データおよび携帯電話用漫画配信サイトの実ログデータを用い，提案法の有効性を示す．

キーワード 生存時間解析，特徴量選択，サブスクリプションサービス

Purchase Pattern Analysis of Loyal Customers in Subscription Services

Tomoharu IWATA†, Kazumi SAITO†, and Takeshi YAMADA†

† NTT Communication Science Laboratories
Hikaridai 2–4, Seika-cho, Soraku-gun, Kyoto, 619–0237 JAPAN
E-mail: †{iwata,saito,yamada}@cslab.kecl.ntt.co.jp

Abstract In recent years, subscription services are attracting attention in online music and movie distributions. Users are allowed unlimited use of their services with a fixed price, and online stores can expect constant revenue stream. For online stores providing subscription services, users that subscribe for long periods are loyal customers. To extract characteristic purchase patterns in loyal customers can help to select distribution contents, and to identify potential loyal customers. In this paper, we model subscription periods by purchase patterns using survival analysis, and propose an algorithm for extracting purchase patterns from a large-scale data set. We show the validity of our method with artificial and real data sets.

Key words survival analysis, feature selection, subscription service

1. ま え が き

近年，一定の期間の利用に対して課金する定額制サービス (サブスクリプションサービス) が注目されている．従来は，雑誌が中心であったが，音楽や映画，テレビ番組，ソフトウェア，携帯電話サービスなど様々な商品が定額制サービスで提供されている．利用回数に関わらず定額にすることにより，ユーザは一定の期間，料金を気にせず利用することができる．また，オンラインストアにとっては，各ユーザから毎月等決められた期間に対して一定額の収益が得られるため，予測可能な安定した収益が見込めるという利点がある．定額制サービスは，提供数が増えたとしてもあまりコストが増えないサービス，例えば音楽や映像のネット配信などでよく用いられる．音楽のネット配信ビジネスの場合，デジタルで録音された音楽の複製は容易であり，CD などを製造する必要もなく，インターネットを介して配信するため輸送費もかからない．通信回線の高速化と記憶装置の小型化，大容量化にともない，音楽や映像のネット配信は普及しつつあり，定額制サービスも今後ますます増加していくものと考えられる．

定額制サービスを提供しているオンラインストアにとって，長期間契約してくれるユーザが優良顧客である．優良顧客を増加させることは，収益増加に直結しているため，オンラインストアにとって重要な課題である．優良顧客に特徴的な購買パターンを分析することにより，優良顧客を増加させるための有益な情報が得られる．例えば，購買パターンから優良顧客になる可能性の高い顧客を特定することができ，それらの顧客に特

表 1　契約ログの例.
Table 1　An example subscription log.

ユーザ	契約状況	契約開始時刻	解約時刻
u_1	1	2004/8/16 11:50	2005/01/08 20:14
u_2	0	2004/8/16 18:01	
u_3	1	2004/8/17 16:10	2004/08/25 13:01
u_4	1	2004/8/17 21:39	2004/08/29 07:21
u_5	0	2004/8/18 01:44	
⋮	⋮	⋮	⋮
u_N	0	2005/10/28 23:10	

表 2　購買ログの例.
Table 2　An example purchase log.

ユーザ	商品	時刻
u_1	s_3	2004/8/16 12:06
u_1	s_1	2004/8/16 13:01
u_2	s_2	2004/8/16 18:51
u_1	s_6	2004/8/16 21:35
u_3	s_2	2004/8/17 16:42
⋮	⋮	⋮
u_N	s_{10}	2005/10/28 23:15

別なマーケティングをするなどの対策を打つことができる．また，優良顧客と同じような購買行動をリコメンドすることにより，優良顧客になる可能性を高めることが期待できる．さらに，優良顧客に多く購入されているという観点で提供する商品やサービスを選定することもできる．そこで本稿では，生存時間解析を用いて優良顧客の購買パターンの分析を行う．提案法では，購買パターンと契約期間の関係をモデル化し，契約期間に与える影響が強い重要な購買パターンを高速に抽出する．生存時間解析は，死亡や故障などあるイベントが発生するまでの期間を解析するための手法であり，ユーザの解約をイベントとすることで，定額制サービスにおける優良顧客分析に用いることが可能である．

2. 関 連 研 究

生存時間解析を用いた解約予測や優良顧客の特定などの研究はこれまでに行われている [2], [7], [8], [10]．しかし，これらは性別や地域などのユーザの属性やサービス利用形態を共変量としており，購買パターンを共変量としたモデル化はされていないため購買パターン抽出のために用いることができない．以前我々は購買パターンを共変量としてモデル化を行っている [4] が，単純な購買パターンをのみを用いており，また，共変量選択を行っていない．生存時間解析において共変量選択手法は提案されているが [5]，大規模なデータ数，共変量数には対応していない．購買パターンは商品の組合せを考慮すると膨大な数にのぼり，従来法では実行が困難である．

3. 提　案　法

3.1 準　　備

定額制サービスを提供するオンラインストアが入手可能なログとして，契約ログと購買ログがある．契約ログとは，各ユーザの契約開始時刻，契約状況（契約中か解約済か），解約時刻のログである．解約時刻は解約済のユーザのみ得られる．表 1 に契約ログの例を示す．購買ログとは，各購買のユーザ，商品，時刻のログである．表 2 に購買ログの例を示す．

ユーザ u_n の契約期間を t_n，契約状況を e_n （解約済の場合 $e_n = 1$，契約中の場合 $e_n = 0$）とする．契約期間 t_n は，契約ログから得ることができる．ユーザ u_n の契約開始時刻を d_n^{start}，解約済の場合の解約時刻を d_n^{end}，ログの最終更新時刻を d_{end} とする．このとき契約期間は

表 3　入力データの例.
Table 3　An example input data.

ユーザ	契約状況	契約期間	購買商品 (購買時の契約期間)
u_1	1	145	$s_3(0)$, $s_1(0)$, $s_6(1)$, ...
u_2	0	438	$s_2(0)$, $s_8(3)$, $s_1(5)$, ...
u_3	1	8	$s_2(0)$, $s_{13}(7)$
u_4	1	12	$s_3(0)$, $s_1(2)$, $s_2(12)$
u_5	0	411	$s_5(0)$, $s_1(0)$, $s_8(2)$, ...
⋮	⋮	⋮	⋮
u_N	0	0	$s_{10}(0)$

$$t_n = \begin{cases} d_n^{end} - d_n^{start} & \text{if } e_n = 1, \\ d_{end} - d_n^{start} & \text{if } e_n = 0, \end{cases} \tag{1}$$

となる．契約中ユーザの場合，実際に解約した時刻はまだ分からないため，真の契約期間は未知である．このようなデータは打ち切りデータとよばれる．生存時間解析を用いることにより，未知の部分が含まれる打ち切りデータも適切に扱うことができる．

ユーザ u_n が k 番目に購入した商品を s_n^k，そのときの契約期間を t_n^k とする．購買商品 s_n^k は購買ログから，購買時の契約期間 t_n^k は購買ログおよび契約ログから得ることができる．ユーザ u_n が商品 s_n^k を購入した時刻を d_n^k とすると，購買時の契約期間は $t_n^k = d_n^k - d_n^{start}$ となる．提案法で必要となる入力データは各ユーザの契約状況 e_n，契約期間 t_n，購買商品 $\{s_n^k\}$，および，購買時の契約期間 $\{t_n^k\}$ である．表 3 に提案法の入力データの例を示す．

3.2 Cox 比例ハザードモデル

生存時間解析においては，生存時間は一般にハザード関数を用いてモデル化される．定額制サービスにおいては，契約期間 t が生存時間である．ハザード関数 $h(t)$ は，期間 t において契約しているユーザのなかで期間 t で解約するユーザの割合を表す．ハザード関数を用いると期間 t において契約している確率を表す生存関数 $S(t)$ は

$$S(t) = \exp\left(-\int_0^t h(\tau)d\tau\right), \tag{2}$$

で表される．ハザード関数の共変量として購買パターンを用いることにより，契約期間の長いユーザに特徴的な購買パターンを抽出することができる．本研究では，ハザード関数として生存時間解析で一般的に用いられる下式の Cox 比例ハザードモ

デル [3] を採用する．

$$h(t|u) = h_0(t)\exp(\sum_{f\in F}\beta_f x_f(u,t)), \tag{3}$$

ここで $h_0(t)$ はベースラインハザード関数，f は購買パターン，F は全購買パターンの集合，$\boldsymbol{\beta} = \{\beta_f\}$ は未知パラメータ，$x_f(u,t)$ はユーザ u の期間 t での購買履歴に購買パターン f があれば 1，なければ 0 となる関数，

$$x_f(u,t) = \begin{cases} 1 & \text{if user } u \text{ has purchase pattern } f \\ & \text{at subscription period } t, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

である．購買パターンの例として，商品 s_1 を購入したことがある，商品 s_1 かつ商品 s_2 を購入したことがあるなどが考えられる．購買履歴は期間 t によって変化するため，共変量 $x_f(u,t)$ は t に依存する時間依存共変量として扱う必要がある．ベースラインハザード関数 $h_0(t)$ はハザード関数の購買パターンに依存しない部分を表す．Cox 比例ハザードモデルでは購買パターンがハザード関数に与える影響は期間 t に依らず対数線形で一定であるとし，ベースラインハザード関数は期間 t に依存する．パラメータ β_f が低い ($\beta_f < 0$) 購買パターン f は，f があることによりハザード関数は低くなり解約されにくいことを表すため，契約期間の長いユーザに特徴的な購買パターンである．逆に，パラメータ β_f が高い ($\beta_f > 0$) 購買パターンは，契約期間の短いユーザに特徴的な購買パターンである．

期間 t で契約中であるユーザの集合を $R(t) = \{n|t_n \geqq t\}$ とし，ユーザ u が期間 t で解約したとする．契約中ユーザ集合 $R(t)$ のなかで解約するユーザが u である確率 $P(u|R(t))$ はハザード関数を用い下式で表され，

$$\begin{aligned} P(u|R(t)) &= \frac{h(t|u)}{\sum_{m\in R(t)} h(t|u_m)} \\ &= \frac{\exp(\sum_{f\in F}\beta_f x_f(u,t))}{\sum_{m\in R(t)}\exp(\sum_{f\in F}\beta_f x_f(u_m,t))}, \end{aligned} \tag{5}$$

Cox 比例ハザードモデルの場合，ベースラインハザード関数に依存しない．下式の部分尤度と呼ばれる与えられたデータに対する $P(u_n|R(t_n))$ を最大にすることで，Cox 比例ハザードモデルの未知パラメータ $\boldsymbol{\beta}$ を，ベースラインハザード関数を特定することなく推定できる．

$$\begin{aligned} L(\boldsymbol{\beta}) &= \log\prod_{n\in D} P(u_n|R(t_n)) \\ &= \sum_{n\in D}\Bigg(\sum_{f\in F}\beta_f x_f(u_n,t_n) \\ &\quad - \log\sum_{m\in R(t_n)}\exp(\sum_{f\in F}\beta_f x_f(u_m,t_n))\Bigg), \end{aligned} \tag{6}$$

ここで $D = \{n|e_n = 1\}$ は解約ユーザ集合を表す．部分尤度の 1 階微分，2 階微分はそれぞれ

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial\beta_f} &= \sum_{n\in D}\Bigg(x_f(u_n,t_n) \\ &\quad - \sum_{m\in R(t_n)} x_f(u_m,t_n)P(u_m|R(t_n))\Bigg), \end{aligned} \tag{7}$$

$$\begin{aligned} &\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_f\partial\beta_g} \\ &= -\sum_{n\in D}\Bigg(\sum_{m\in R(t_n)} x_f(u_m,t_n)x_g(u_m,t_n)P(u_m|R(t_n)) \\ &\quad + \sum_{m\in R(t_n)} x_f(u_m,t_n)P(u_m|R(t_n)) \\ &\quad \times \sum_{m\in R(t_n)} x_g(u_m,t_n)P(u_m|R(t_n))\Bigg), \end{aligned} \tag{8}$$

となり，未知パラメータに関して上に凸になるため，準ニュートン法 [6] などの最適化法により大域的最適解を求めることができる．同期間に複数のユーザが解約した場合は，Efron 近似 [5]

$$\begin{aligned} &L_{Efron}(\boldsymbol{\beta}) \\ &= \log\prod_t\Bigg(\prod_{n\in D(t)} h(t|u_n) \\ &\quad / \prod_{r=1}^{|D(t)|}\Big(\sum_{m\in R(t)} h(t|u_m) - \frac{r-1}{|D(t)|}\sum_{m\in D(t)} h(t|u_m)\Big)\Bigg) \\ &= \sum_t\Bigg(\sum_{n\in D(t)}\sum_{f\in F}\beta_f x_f(u_n,t) \\ &\quad - \sum_{r=1}^{|D(t)|}\log\Big(\sum_{m\in R(t)}\exp\big(\sum_{f\in F}\beta_f x_f(u_m,t)\big) \\ &\quad - \frac{r-1}{|D(t)|}\sum_{m\in D(t)}\exp\big(\sum_{f\in F}\beta_f x_f(u_m,t)\big)\Big)\Bigg), \end{aligned} \tag{9}$$

により部分尤度を計算することができる．ここで $D(t) = \{n|e_n = 1, t_n = t\}$ は期間 t で解約したユーザ集合，$|D(t)|$ はその要素数を表す．

3.3 共変量選択

Cox 比例ハザードモデルは対数線形モデルであるため，共変量として高次の購買パターン (商品 s_1 かつ商品 s_2 を購入したことがある，のような購買商品の組合せなど) を用いることにより，より記述力の高いモデルにすることができる．しかしながら，不必要な共変量を用いると学習データに対する尤度は高くなるが，過学習を起こしてしまい汎化性能が下がる可能性が高い．また，不必要な共変量が多数あると，どの購買パターンが重要であるか不鮮明になってしまうという問題がある．そこで，共変量の候補となる購買パターン集合から，契約期間に与える影響の大きい購買パターン集合を自動的に抽出する．

商品数が多い場合，全ての購買パターンの組合せを試すことは膨大な計算量が必要となる．そのため，共変量なしの Cox 比例ハザードモデルから購買パターンを 1 つずつ共変量として追加していくことにより，最適な共変量集合を探索する手法をとる．共変量候補集合のうち 1 つの購買パターンをモデルに追加

したとき，部分尤度が最も高くなる購買パターンを共変量として採用する．

$$\hat{g} = \arg\max_{g \in F_0} \left(\max_{\boldsymbol{\beta}} L_{F+g}(\boldsymbol{\beta}) \right), \tag{10}$$

ここで，$\hat{g}$ は追加する購買パターン，F_0 は共変量候補集合，F は現在抽出されている共変量集合，$L_{F+g}(\boldsymbol{\beta})$ は購買パターン g を追加したときの部分尤度を表す．上記の方法の場合，1 つの追加共変量を決定するために，各共変量候補集合について，$|F|+1$ 個の未知パラメータを推定する必要がある．ここで $|F|$ は共変量数を表す．より高速に実行するため，現在の共変量集合に対応するパラメータ $\boldsymbol{\beta}$ は新たな購買パターンを追加しても変化しないと考え，部分尤度を近似的に求める．このとき，追加パターンに対応する未知パラメータ β_g のみを推定すればよく，部分尤度は下式で表される．

$$\begin{aligned} & L_{F+g}(\beta_g) \\ &= \log \prod_{n \in D} \frac{P_F(u_n|R(t_n)) \exp\bigl(\beta_g x_g(u_n, t_n)\bigr)}{\sum_{m \in R(t_n)} P_F(u_m|R(t_n)) \exp\bigl(\beta_g x_g(u_m, t_n)\bigr)} \\ &= \sum_{n \in D} \Biggl(\log P_F(u_n|R(t_n)) + \beta_g x_g(u_n, t_n) \\ &\quad - \log \sum_{m \in R(t_n)} P_F(u_m|R(t_n)) \exp\bigl(\beta_g x_g(u_m, t_n)\bigr) \Biggr), \end{aligned} \tag{11}$$

ここで $P_F(u_n|R(t_n))$ は共変量集合が F の場合のユーザ集合 $R(t_n)$ のなかで解約するユーザが u_n である確率を表す．上式は β_g に関して上に凸になるため，ニュートン法により最大化することで大域的最適解を得ることができる．

共変量選択の手順をまとめると以下のようになる．

（1） $F = \phi$ とする．

（2） すべての $g \in F_0$ について $L_{F+g}(\beta_g)$ が最大となる β_g を推定する．

（3） $L_{F+g}(\beta_g)$ が最大となる g を F に追加し，F_0 から g を除く．

（4） F に対応する未知パラメータ $\boldsymbol{\beta}$ を推定する．

（5） 終了条件を満たさなければ Step 2 へ戻る．

終了条件として，共変量数が閾値以上になったとき，部分尤度の増加が閾値以下になったとき，最小記述長 (MDL) [9] が増加に転じたときなどが考えられる．なお MDL は次式で与えられる．

$$MDL = -L(\boldsymbol{\beta}) + \frac{1}{2}|F| \log |D|, \tag{12}$$

ここで $|F|$ は共変量数，$|D|$ は学習データ中の解約ユーザ数を表す．

4. 評価実験

4.1 人工データ

人工データを用い，提案法により契約期間に影響を与える購買パターンを抽出可能であるか評価した．表 4 に 1 ユーザの購買履歴および契約期間の作成手順を示す．ここで u は購買履歴，u_{+s_j} は商品 s_j 購入後の購買履歴，ϕ は空集合，$Bernoulli(\psi)$ は成功確率 ψ のベルヌーイ分布，$Multinomial(\boldsymbol{\theta})$ は j 番目の要素の成功確率が θ_j の試行回数 1 の多項分布を表す．まず，表 4 の行 3 から行 4 において，単位時間内でユーザが解約するかどうかを確率 $h(t|u)$ を用い決定する．ここでベースラインハザード関数は期間によらず一定 $h_0(t) = \lambda$ とした．次に行 7 から行 8 において，単位時間内で商品を購入するかどうかを確率 g を用い決定する．ここで g は期間 t によらず一定とした．購入商品は確率 $\boldsymbol{\theta}$ によって決定する (行 9)．購入確率は全商品で一様 $\theta_i = \frac{1}{V}$ とした．ここで V は全商品数である．商品は s_1 から s_{10} までの 10 商品とし，契約期間に影響の与える購買パターンとして商品 s_1 を購入したことがある，商品 s_2 を購入したことがある，商品 s_3 かつ商品 s_4 を購入したことがある，の 3 つとし，対応するパラメータはそれぞれ，

$$\beta_f = \begin{cases} -1 & \text{if } f = s_1, \\ 1 & \text{if } f = s_2, \\ -2 & \text{if } f = s_3, s_4, \\ 0 & \text{otherwise}, \end{cases} \tag{13}$$

とした．

表 4 ユーザシミュレーションアルゴリズム．
Table 4 a user simulation algorithm.

1:	Set $t \leftarrow 0$, $u \leftarrow \phi$
2:	**loop**
3:	Sample $r_1 \sim Bernoulli(h(t\|u))$
4:	**if** r_1 is success **then**
5:	break
6:	**end if**
7:	Sample $r_2 \sim Bernoulli(g)$
8:	**if** r_2 is success **then**
9:	Sample $s_j \sim Multinomial(\boldsymbol{\theta})$
10:	Set $u \leftarrow u_{+s_j}$
11:	**end if**
12:	Set $t \leftarrow t + 1$
13:	**end loop**
14:	Output t, u

パラメータを $g = 0.1$，$\lambda = 0.01$ とし，上記アルゴリズムにより 1000 ユーザの購買履歴および契約期間を生成し，提案法を適用した．なお，共変量候補として商品集合のべき集合のうち要素数が 2 以下のものを用い，終了条件は MDL が増加に転じたときとした．共変量選択の結果，購買パターンとして商品 s_1 を購入したことがある，商品 s_2 を購入したことがある，商品 s_3 かつ商品 s_4 を購入したことがある，の真の契約期間に影響を与えるパターン 3 つが抽出された．対応するパラメータの推定値はそれぞれ，

$$\beta_f = \begin{cases} -0.981 & \text{if } f = s_1, \\ 0.918 & \text{if } f = s_2, \\ -2.106 & \text{if } f = s_3, s_4, \end{cases} \tag{14}$$

と真のパラメータの値と近く，提案法により購買パターンの契約期間への影響が適切に推定できることが示された．また，学習データとは別に 1000 ユーザの購買履歴および契約期間を生

表 5　人工データを用いた比較実験における部分尤度.
Table 5　Partial likelihoods for the artificial data.

	学習データ	テストデータ
提案法	-5.691	**-5.690**
共変量選択なし	-5.698	-5.744
履歴非依存	-5.957	-5.956

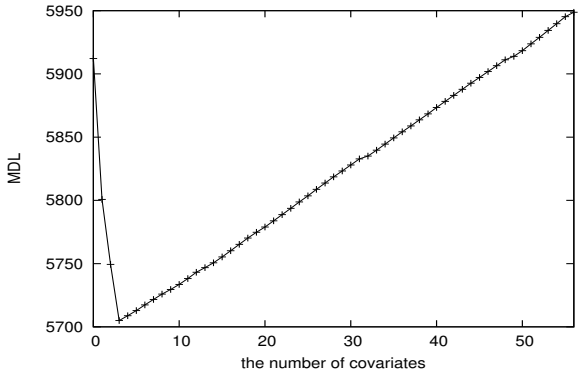


図 1　抽出共変量数を変えたときの MDL.
Fig. 1　The number of extracted features v.s. MDL.

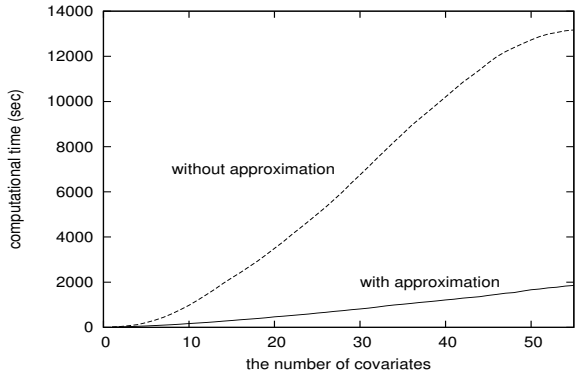


図 2　近似した場合と近似しない場合の抽出共変量数を変えたときの計算時間.
Fig. 2　Computational times with and without approximations.

成したものをテストデータとして用い，提案法を共変量選択を行わない Cox 比例ハザードモデル，ハザード関数が購買履歴に依存しないモデルと比較した結果，表 5 のようになった．ここで評価尺度として部分尤度を用いた．部分尤度が高いモデルは，契約中ユーザのなかから解約するユーザを予測する精度が高いことを表す．提案法のテストデータに対する部分尤度が最も高く，共変量選択をした Cox 比例ハザードモデルにより，解約ユーザを適切に予測できることを示唆する．

図 1 に抽出共変量数を変えたときの MDL を示す．真の共変量数である 3 のときに最も MDL が低く，MDL により適切な共変量数が選択できることを示唆する．

図 2 に近似した場合と近似しない場合の計算時間の比較結果を示す．近似した場合，候補である購買パターンを追加したときの部分尤度を計算する際に 1 つの未知パラメータのみの推定でよいため，近似しない場合に比べ計算量が少なくなり，また，抽出共変量数が多くなるにつれ計算量の差は大きくなる．

表 6　実データを用いた比較実験における部分尤度.
Table 6　Partial likelihoods for the real data.

	学習データ	テストデータ
提案法	-9.258	**-7.039**
共変量選択なし	-9.228	-7.055
共変量選択なし (パターン長 1)	-9.262	-7.048
履歴非依存	-9.396	-7.159

4.2 実データ

提案法を評価するため，携帯電話用漫画配信サイトにおける実ログデータに対し提案法を適用し，契約期間が長いユーザに特徴的な購買パターンを抽出した．用いたログは契約中ユーザ数 26953，解約済ユーザ数 8165，商品数 84 であり，500 ユーザ以上に現れた商品のべき集合を購買パターン共変量の候補とした．このとき共変量候補とする頻出パターンの発見には a priori アルゴリズム [1] を用いた．共変量候補数は 454，最長購買パターンは 5 商品を含むものであった．提案法を用いて最初に抽出された 3 つの購買パターンは，商品 s_3 を購入したことがある，商品 s_{10} を購入したことがある，商品 s_{25} かつ商品 s_{34} を購入したことがある，であり，このときのパラメータの値はそれぞれ，

$$\beta_f = \begin{cases} -0.672 & \text{if } f = s_3, \\ -0.435 & \text{if } f = s_{10}, \\ -0.680 & \text{if } f = s_{25}, s_{34}, \end{cases} \tag{15}$$

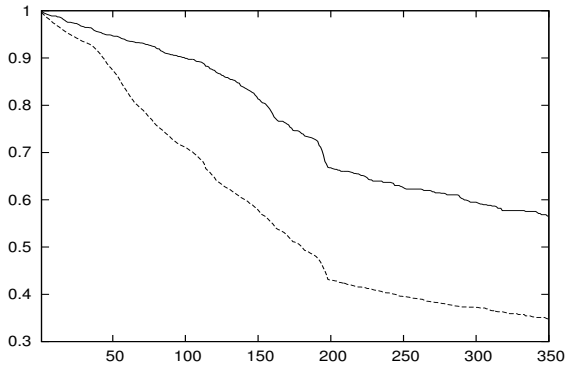
であった．この 3 つの購買パターンについて，購買パターンを購買履歴中に含むユーザと，含まないユーザの生存関数は図 3 のようになった．ここで，生存関数はノンパラメトリックモデルであるカプラン・マイヤー法

$$\hat{S}_{kaplan}(t) = \prod_{\tau<t} \left(1 - \frac{|D(\tau)|}{|R(\tau)|}\right), \tag{16}$$

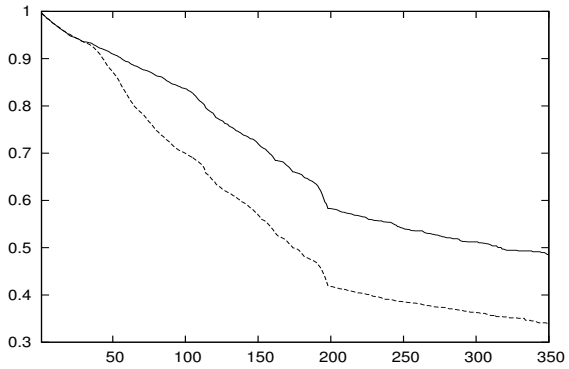
により推定したものである．ここで $\frac{|D(\tau)|}{|R(\tau)|}$ は期間 τ において契約中ユーザが解約した割合を表す．抽出されたパターンを持つユーザは生存関数が高く，契約期間が長いユーザに特徴的なパターンを抽出できたことを示す．

共変量選択の終了条件を MDL が増加に転じたときとしたとき，39 の共変量が抽出された．このときの提案法，共変量選択を行わない Cox 比例ハザードモデル，長さ 1 のみの購買パターンを用いた共変量選択を行わない Cox 比例ハザードモデル，ハザード関数が購買履歴に依存しないモデルを比較した結果，表 6 のようになった．提案法のテストデータに対する部分尤度が最も高く，予測精度が高いモデルであると言える．これは，購買パターンに含まれる契約期間に関する情報を適切に抽出できているためと考えられる．また，共変量選択なしの場合，学習データに対する部分尤度は高いが，テストデータに対する部分尤度は低い．これは過学習に陥っているためと考えられる．提案法では共変量選択を行い契約期間と相関の強い購買パターンのみを用いているため，汎化性能が向上していると考えられる．

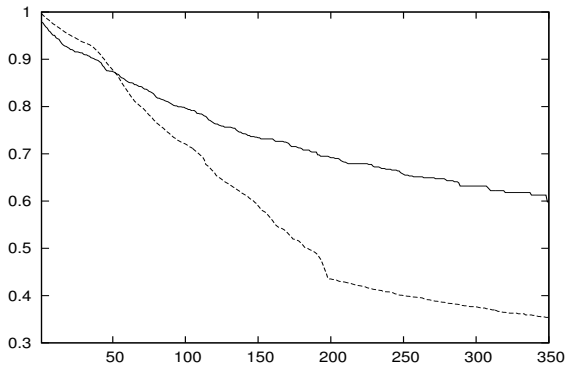
抽出共変量数を変えたときのテスト部分尤度および MDL を



(a) s_3



(b) s_{10}



(c) s_{25}, s_{34}

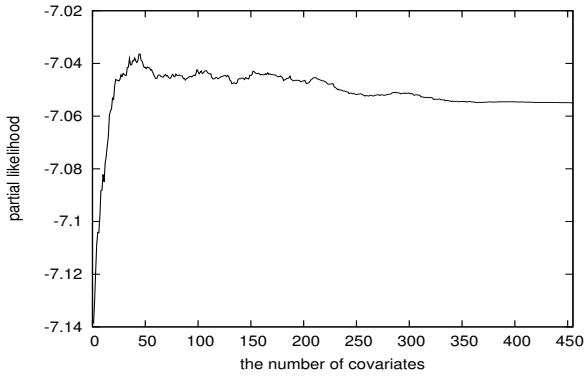
図 3　抽出された購買パターンを持つユーザの生存関数 $S(t)$．破線はそのパターンを持たないユーザの生存関数を表す．

Fig. 3　Survial functions $S(t)$ of users having extracted purchase patterns. Dot lines represent survival functions of users not having the patterns.

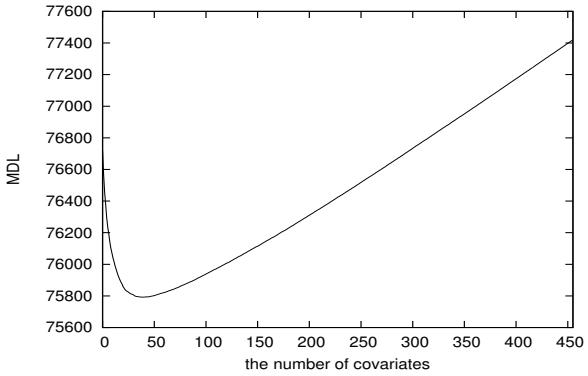
図 4 に示す．テスト部分尤度が最も高くなる共変量数は 45 であり，MDL が最も低くなるの共変量数は 39 であった．一致はしないものの両方ともに共変量数 40 前後で最良になる傾向があり，MDL によりテストデータに対する予測精度が高い共変量数を選択できることを示唆する．

4.3 議　　論

前節では，購買パターンを共変量とする Cox 比例ハザードモデルを用いることにより，購買パターン情報を用いない場合にくらべて解約ユーザを予測する精度が高くなることを示した．Cox 比例ハザードモデルによりモデル化することで，購買パターンのような時間依存共変量を含む場合でも比較的簡易にパ



(a) テスト部分尤度



(b) MDL

図 4　抽出共変量数を変えたときのテスト部分尤度 (a) および MDL(b).

Fig. 4　The number of extracted features v.s. test partial likelihoods(a) and MDL(b).

ラメータ推定が可能であり，また，ベースラインハザード関数を特定する必要がないという利点がある．しかしながら，Cox 比例ハザードモデルでなされる仮定が，常に定額制サービスにおけるログにおいて妥当であるわけではない．

Cox 比例ハザードモデルにおける生存関数は

$$S(t|u) = \exp\Big(-\exp\big(\sum_{f\in F}\beta_f x_f(u,t)\big)\int_0^t h_0(\tau)d\tau\Big), \quad (17)$$

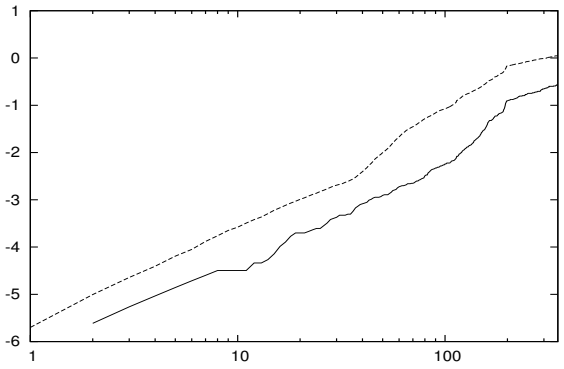
であり，両辺 2 重対数をとると，

$$\log\big(-\log S(t|u)\big) = \sum_{f\in F}\beta_f x_f(u,t) + \log\int_0^t h_0(\tau)d\tau, \quad (18)$$

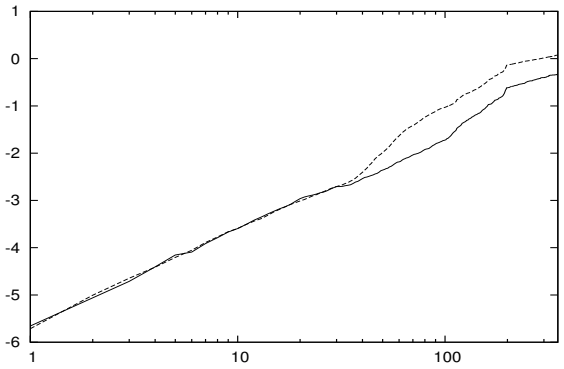
が成り立つ．共変量が x_f のみの場合を考え，購買パターン f を持つユーザを u_{+f}，持たないユーザを u_{-f} とすると，上式より，

$$\log\big(-\log S(t|u_{+f})\big) = \beta_f + \log\big(-\log S(t|u_{-f})\big), \quad (19)$$

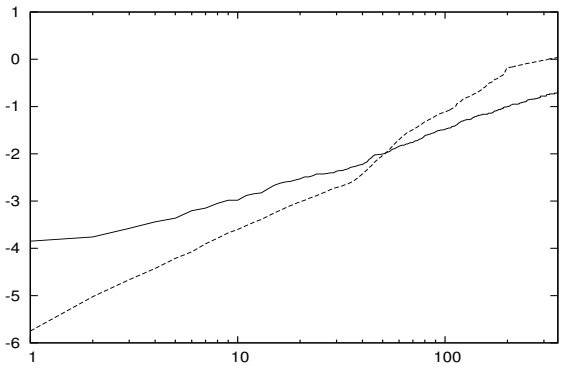
となる．つまり，Cox 比例ハザードモデルでは，購買パターン f を持つユーザと持たないユーザの生存関数の 2 重対数 $\log(-\log S(t))$ は，平行になるはずであり，この性質は比例ハザード性と呼ばれる．実データから抽出された購買パターンを持つユーザと持たないユーザの生存関数の 2 重対数プロットは図 5 のようになった．ここで図 3 と同じく生存関数はカプラン・マイヤー法により求めた．図 5(a) では，各線が平行になっ



(a) s_3



(b) s_{10}



(c) s_{25}, s_{34}

図 5　抽出された購買パターンを持つユーザの生存関数の 2 重対数プロット ($\log t$ v.s. $\log(-\log S(t))$)．破線はそのパターンを持たないユーザのものを表す．

Fig. 5　$\log t$ v.s. $\log(-\log S(t))$ of users having extacted purchase patterns. Dot lines represent those of users not having the patterns.

ており，購買パターン s_3 に関しては比例ハザードモデルが適切であると言える．しかしながら，図 5(b)(c) では，平行になっておらず，比例ハザード性は成立していない．これは Cox 比例ハザードモデルの限界を示しており，今後，比例ハザード性が成立しない共変量も適切に扱うことができるモデルの検討が必要である．

5. むすび

本稿では，定額制サービスにおける優良顧客である長期間契約しているユーザに特徴的な購買パターンの解析を行った．提案法では，生存時間解析を用いて購買パターンと契約期間の関係をモデル化し，契約期間に与える影響の強い購買パターンを抽出する．人工データおよび携帯電話用漫画配信サイトにおけるログデータを用いて提案法の有効性を確認した．

今後の課題として，前節で述べた Cox 比例ハザードモデルの限界に対応するため，生存時間解析の分野での成果を参考に，モデルを改良する必要がある．また，今回は購買パターンとして購買商品集合を用いたが，順序情報を採り入れた購買系列パターンなど，他のパターンの解析も重要であると考えられる．

今回は定額制サービスを解析の対象とした．しかしながら，ユーザが他のサービスに乗り換えることを解約することととらえると，購入した商品に応じて課金する従量制サービスにも応用可能であると考えられる．従量制においても，長期間サービスを利用してくれるユーザは優良顧客である．従量制では，定額制の場合と異なり，解約時刻に関するデータが通常得られないという課題があるが，ある一定期間の利用がなければ解約とみなしたりすることで対処可能であろう．

また，優良顧客の購買パターンの定性的特徴の分析も必要である．例えば，優良顧客が多く購入する商品は，幅広い層に人気ある商品であるのか，または，少数の特定のユーザだけに人気が高い商品であるのか，関連商品が多い商品であるのか，などの意味付けを試みたい．

文　献

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of 20th International Conference Very Large Data Bases*, pp. 487–499, 1994.

[2] Wai Ho Au, Keith C. C. Chan, and Xin Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Trans. Evolutionary Computation*, Vol. 7, No. 6, pp. 532–545, 2003.

[3] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, Vol. 34, No. 2, pp. 187–220, 1972.

[4] Tomoharu Iwata, Kazumi Saito, and Takeshi Yamada. Recommendation method for extending subscription periods. In *Proceedings of the Twelfth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006.

[5] Elisa T. Lee and John Wenyu Wang. *Statistical methods for survival data analysis, Third Edition.* Wiley, 2003.

[6] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Programming*, Vol. 45, No. 3, pp. 503–528, 1989.

[7] J. Lu. Modeling customer lifetime value using survival analysis - an application in the telecommunication industry. In *Proceedings of SUGI 28*, pp. 120–128, 2003.

[8] D. R. Mani, James Drew, Andrew Betz, and Piew Datta. Statistics and data mining techniques for lifetime value modeling. In *Proceedings of the Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 94–103, 1999.

[9] Jorma Rissanen. *Stochastic complexity in statistical inquiry.* World Scientific, 1989.

[10] Yuji Shono, Yohei Takada, Norihisa Komoda, Hriaki Oiso, Ayako Hiramatsu, and Kiyoyuki Fukuda. Customer analysis of monthly-charged mobile content aiming at prolonging subscription period. In *Proceedings of IEEE Conference on Computational Cybernetics*, pp. 279–284, 2004.