RF-006

閾値関数による変化分析型アンサンブル学習を用いた商品・サービス普及予測

Prediction of Product/Service Diffusion based on Change Analysis Ensemble Learning

藤原 由希子† 富沢伸行† 井口浩人†
Yukiko Fujiwara Nobuyuki Tomizawa Hiroto Iguchi

1. はじめに

商品・サービスの普及予測は，費用対効果や調達資源を知るために重要である．普及とは，商品・サービスが，時間的経過の中で情報伝達を通して人々に採用される過程であり，累積採用者数は一般的に S 字曲線となる[1]．そのため，従来は，過去の普及実績を曲線へ適合することにより将来の普及予測が行なわれてきた[2]．しかし，普及初期段階の予測精度が低かった．

近年，顧客データは大量に蓄積され，顧客の多様な嗜好が分析されている．例えば，推薦システムとして，趣味の似た人の意見を用いた協調フィルタリングがある[3,4]．しかし，採用までの時系列変化は予測されていなかった．

我々は，データ加重による変化分析型決定木を用いた普及予測法を提案し，普及初期段階に曲線近似より高い予測精度となることを確認した[5]．しかし，普及開始直後の予測精度は低かった．

本稿では，閾値関数による変化分析型アンサンブル学習を提案する．普及における情報伝達は，互いに似ている同類的な人の間で最も繁盛に行なわれると指摘されている[1]．そのため，アンサンブル学習を用いて採用者らしさを示すスコアを出力するよう分類器を生成し，各未採用者は，スコアが分類閾値より高ければ同類的なため採用したとみなすこととした．提案法は，まず，過去の普及実績に適合するよう閾値を求め，普及率に対する閾値関数を推定する．次に，その閾値関数を用いて将来の普及の推移を予測する．提案法を携帯音楽機器 iPod の普及予測に適用し，手法の有効性を確認した．

2. 提案法

顧客データは，顧客の属性情報 X と，ある商品・サービスの採用情報 Y との組とする．X の各要素 x_{ij} は，i 番目の顧客の j 番目の属性であり，属性は，年齢，性別，他の商品・サービスの利用履歴など属性である．顧客数を M，属性数を N，i 番目の顧客の属性を $X_i=(x_{i1},x_{i2},\cdots,x_{iN})$とする．Y の各要素 y_{it} は，i 番目の顧客の時刻 t における採用の有無である．採用（正例）を 1，未採用（負例）を 0 で表すこととする．時刻 t の顧客全員の採用情報を $Y_t=(y_{1t},y_{2t},\cdots,y_{Mt})$とする．

提案法のアルゴリズムを図 1 に示す．

まず，過去の普及実績から，時刻 t の普及率 DR[t]を求め，次時刻での採用者増加数 Ntrue を求める．次に，時刻 t のデータから分類器 H を生成する．分類器生成は，アンサンブル学習である bagging with descriptor-sampling（BaggingDS）と呼ぶ手法を用いた[6]．アンサンブル学習は複数の分類器を生成して予測する手法であり，ノイズに頑強な手法として Bagging が有名である[7]．Bagging は，

† 日本電気（株），NEC Corporation

Algorithm：閾値関数による変化分析型アンサンブル学習
Input： X：顧客の属性情報
$Y_1,Y_2,\cdots,Y_T$：時刻 1 から T までの採用情報
w：分類閾値の探索間隔
Output： $Y_{T+1},Y_{T+2},\cdots$：時刻 T+1 以降の採用情報
Begin
For t=1,2,…,T-1
DR[t]=|{1≦i≦M: y_{it}=1}|/M
Ntrue = |{1≦i≦M: y_{it}=0 & y_{it+1}=1}|
(X,Y_t)を入力として分類器 H を生成
For thr=w,w×2,w×3,…,1
Npred(thr) = |{1≦i≦M: y_{it}=0, H (X_i)≧thr}|
Thres[t]=argmin_{thr} |Npred(thr)−Ntrue|
(DR[1],Thres[1]),…, (DR[T-1],Thres[T-1])の関数 f を推定
For t=T,T+1,…
DR[t]=|{1≦i≦M: y_{it}=1}|/M
(X,Y_t)を入力として分類器 H を生成
For i=1,2,…,M
y_{it+1}= 1 if H(X_i)≧f(DR[t])
y_{it} otherwize
End

図 1：閾値関数による変化分析型アンサンブル学習

データのランダムなリサンプリングを繰り返して複数の異なる分類器を生成し，予測ではこれらの分類器の予測結果の多数決を用いる．これに対し，BaggingDS は，データと同時に属性をランダムに選択して複数の分類器を生成する．BaggingDS は，Bagging より予測精度が概ね高く，また，他の学習法と比べ，正例数と負例数がアンバランスなデータにおける予測精度がやや高い[6]．普及予測では，特に普及開始直後に採用者である正例が極めて少ないため，本稿では学習法として BaggingDS を用いた．また，本稿では個々の分類器を決定木とし，ある時刻の決定木数を 500 個とした．

複数の分類器を生成した後には，i 番目の顧客の採用者らしさを示すスコア $H(X_i)$を求める．スコアは，生成した全ての分類器が採用と予測するなら 1，全ての分類器が未採用と分類するなら 0 とする．よって，$H(X_i)\in[0,1]$となる．次の時刻の採用者を決めるためには，次の時刻の採用者が採用と予測されるような分類閾値を決めればよい．しかし，個々の顧客の採用時期はそれほど厳密でなく前後するため，閾値は，増加数の誤差が最小となるよう求めることとした．

このように時刻 t を変化させながら対応する閾値 Thres[t]を求め，普及率に対する関数として閾値関数を求める．ここで，閾値関数は，直前の普及の推移を重視するため，時刻 t における加重を $1\cdot r^{T-t}$ として加重最小二乗法で求めた．また，閾値は[0,1]となるため,普及率 100%における閾値が[0,1]の範囲となるようにした．

現在の時刻 T では，次の時刻の採用者は不明であるため，時刻 T の採用・未採用から分類器 H を生成し，推定された閾値関数を用いて，時刻 T の普及率から分類閾値を求め，分類閾値以上と予測される未採用者を採用者とみなす．時刻 T+1 では，みなされた採用により採用者数が増加するので，分類器や普及率を更新し，普及予測を続行する．

3．検証法

検証に用いたデータは，Web アンケート調査により収集した携帯音楽機器 iPod の普及データであり，[5]と同一である．調査は，対象地域を全国，男女はおよそ同数，年齢は 10 代から 40 代までとし，2007 年 9 月 7 日に開始し，約 1 日間で 2,458 件を回収した．質問は，携帯電話の利用履歴（15 種）やパソコンの利用履歴（7 種），先進性（8 種），メディアへの接触頻度（7 種）などである．

携帯電話やパソコンの利用履歴は，所持の有無や機種，様々なサービスの利用頻度である．なお，所持しない場合は，機種や利用頻度は欠損値とした．先進性は，デジタル機器一般に対して新しい商品・サービスの採用の相対的な早さや他人の行動に影響を与える度合いである．先進性の質問例を以下に示す．

- 携帯電話やパソコンなどの製品やサービスについてよく知っているほうだ
- 携帯電話やパソコンなどの製品やサービスについて人からよく聞かれるほうだ
- 周囲の人が新しい製品やサービスを利用していると気になるほうだ
- 自分の価値観に合わなければ周囲が全て利用していても，新製品を買ったり新サービスを利用したりしないほうだ

回答は，「あてはまる」，「ややあてはまる」，「あまりあてはまらない」，「あてはまらない」の 4 段階とした．メディアへの接触頻度は，新聞，テレビ，家族・友人との話，Web などで情報を得る頻度である．年齢，性別を加え，これら 39 種を顧客の属性 X とした．

普及実績を示す情報 Y もアンケートから求めた．iPod 発売から調査までの 6 年間で，採用者は 582 人であった．

このデータに対し，提案法（Change Analysis Ensemble Learning，CAEL）を用いて普及予測を行なった．また，比較するため，従来法として，S 字曲線への適合による普及予測を行なった．予測には，統計分析ソフト STATA（ver. 10）を用い，最大普及率 K を 0 から 100 の範囲で変化させ，誤差最小の S 字曲線を予測結果とした．さらに，変化分析型決定木（Change Analysis Decision Tree，CADT）[5]による普及予測も行なった．

4．実験結果

1，2，3，4，5 年目までの普及実績を用いて予測した結果を図 2（A）から（E）に示す．図に示すように，提案法 CAEL は，普及開始段階には従来の曲線への適合より高い精度で将来の実績を予測することができた．また，1 年目における提案法 CAEL，提案法 CAEL の[0,1]条件なし，および CADT の予測結果を図 2（A'）に示す．図に示すように，提案法 CAEL は，CADT や閾値が[0,1]の範囲にあることを考慮しない CAEL より優れた．

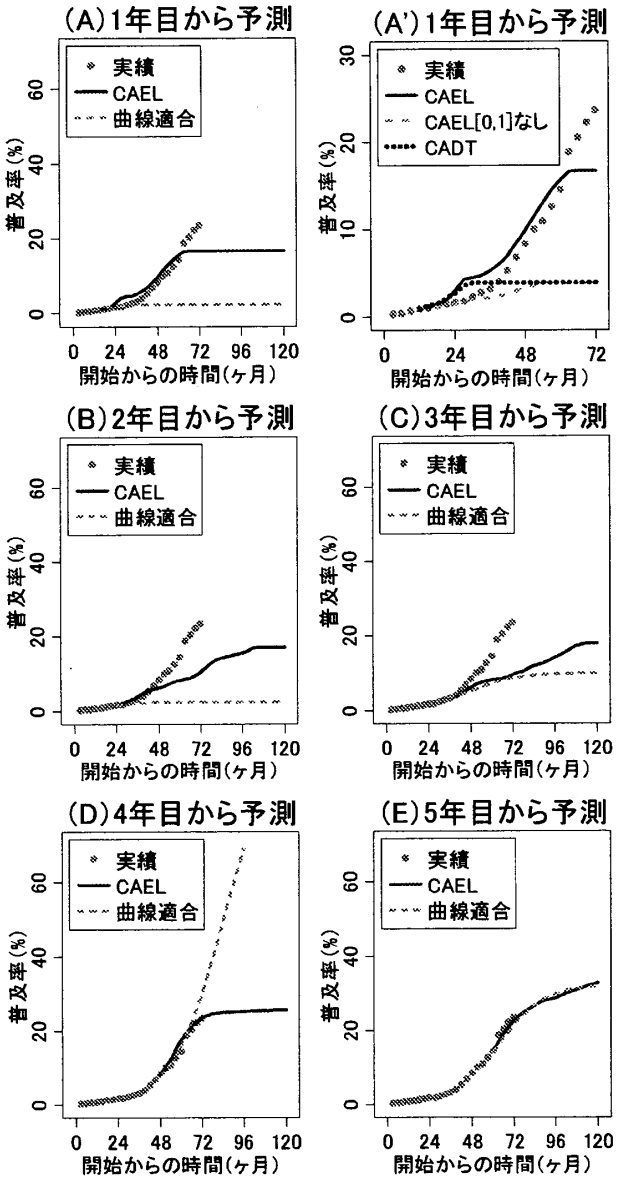

図 2　予測結果

5．考察

提案法は，普及開始直後から将来の普及の推移を予測することができた．1 年目の普及率は僅か 0.9%であり，普及開始直後に普及の推移がある程度決定されることが示唆された．また，提案法は，普及開始段階の予測では従来の S 次曲線への適合より優れた．なお，一般の時系列解析や曲線近似の場合は，曲線の形に制約がないため，S 次曲線への適合より予測が困難と考えられる．

閾値関数を説明するため，時刻に対する閾値を図 3（A），普及率に対する閾値を図 3（B）に示す．図に示すように，閾値は時系列では直線とならないが，普及率に対する閾値は概ね直線となった．従って，普及率に対する閾値関数という設定は有効と考えられる．また，1，2，3 年目の普及率は 0.009，0.02，0.03 であり，図 3（B）の普及率 0.009 以前の点，0.02 以前の点，あるいは 0.03 以前の点を用いて単純に推定すると普及率 100%で閾値 1 を

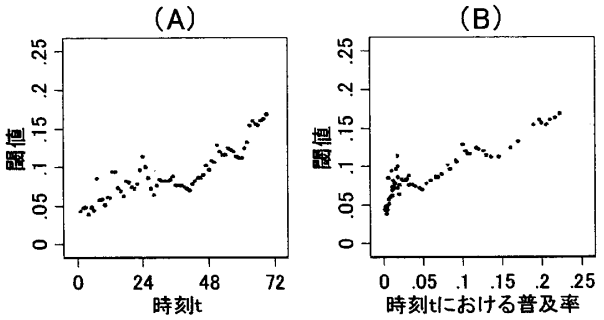

図 3（A）時刻に対する閾値，（B）普及率に対する閾値

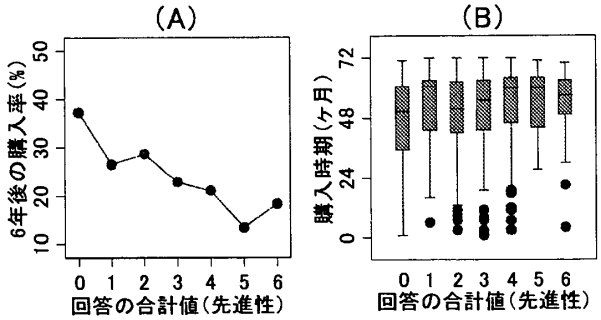

図 4（A）先進性と購入率．（B）先進性と購入時期

超える閾値関数となった．従って，閾値の範囲を考慮した提案法が，閾値の範囲を考慮しない CAEL や閾値自体を考慮しない CADT より優れたと考えられる．なお，提案法の 2，3 年目の予測精度が 1 年目に比べて低いのは，図 3（B）に示すように推定された閾値が高過ぎたためと思われる．

本稿で推定した閾値関数は，iPod の普及速度を規定しており，他の商品・サービスに直接適用できない．他の商品・サービスの普及を予測する場合には，提案法を再度適用して，閾値関数を求める必要がある．しかし，提案法は，閾値関数の推定が自動的であり，他の商品・サービスへも活用できると考える．

図 2（A）（D）に示すように，提案法の 1 年目や 4 年目からの予測結果は，5 年目頃からの普及が横ばいになった．原因の一つは，提案法が新商品販売などによる顧客価値の変動には対応していないためである．提案法は，価値変動に追随するため，直前の普及の推移を重視する加重を用いたが，直後に変動がある場合に予測できない．今後，新商品販売の効果を追加する手法の開発が必要と考える．また，手法的に，採用者が増加しないと予測されると，分類器や普及率が変化せず，普及停止と予測されるという問題もある．本稿ではサンプリングのための乱数の種を固定していたが，偶然増加数が 0 となる場合を考慮し，乱数の種を変化させるなどと手法を改善したい．

デジタル機器一般に対する先進性と特定商品 iPod の購入とは関係があると考えられる．そこで，まず，先進性に関する 8 種の回答結果を，「あてはまる」を 0 点，「あてはまらない」を 3 点の 4 段階として，信頼性係数（クロンバックのα係数）を調べた．すると，2 つの質問「よく知っている」，「よく聞かれる」の場合が，係数 0.83 と最も高く，内容的にも先進性を最も示すと考えられた．そのため，これら 2 質問の回答の合計値を先進性として，購入との関係を調べた．先進性と購入率を図 4（A），購入者における先進性と購入時期との関係を図 4（B）に示す．図 4（A）に示すように，合計値の低い先進的な顧客ほど購入者の比率が高かったが，決定的ではなかった．また，図 4（B）に示すように，先進的な顧客ほど購入が早い傾向はあったが，バラツキが大きかった．従って，単純に先進性が高いから早期に購入したと予測できるわけではなかった．

採用者の特徴を調べるため，1 年目からの予測により，2 年目の時点で予測された採用情報を用いて決定木を生成した．その結果，採用者と未採用者を分類する最も重要な属性は，「新しいものが好き」に「あてはまる」か否かであった．一方，2 年目の実際の採用情報に基づいて生成した決定木でも最も重要な属性は同じであった．類似の特徴を示す属性が多いために決定木の細部は異なるものの，提案法は，採用・未採用を分類するのに重要な属性を自動的に抽出しつつ予測を行なっており，先進性の度合いだけでなく，メディアへの接触頻度や音楽の利用頻度が高いと採用と予測されていた．

提案法は，個別の顧客の採用の有無を予測する手法である．顧客に対しては常に多くの商品・サービスを推薦すると効果的でなくなると指摘されており[8]，適切な時期に個別広告するなどの活用へ発展させていきたい．

6. まとめ

本稿では，閾値関数による変化分析型アンサンブル学習を用いた普及予測法を提案した．提案法を携帯音楽機器 iPod の普及予測に適用し，手法の有効性を確認した．

参考文献

[1] Rogers, E. M.: Diffusion of Innovations, 5th Edition, New York, Free Press, 2003.

[2] Mahajan, V., Muller, E., Bass, F. M.: New Product Diffusion Models in Marketing: A Review and Directions for Research, Journal of Marketing, Vol. 54, pp.1-26, 1990.

[3] Goldberg, D., Nichols, D. Oki, B. M., Terry, D.: Using Collaborative Filtering to Weave an Information Tapestry, ACM Communications, Vol. 35, No. 12, pp.61-70, 1992.

[4] Felfernig, A., Friedrich, G., Schmidt-Thieme, L.: Guest Editors' Introduction: Recommender Systems, IEEE Intelligent Systems, pp.18-21, 2007.

[5] 藤原由希子，富沢伸行，井口浩人：変化分析型決定木を用いた製品・サービスの普及予測，第 22 回人工知能学会全国大会，2008.

[6] Fujiwara, Y., Yamashita, Y., Osoda, T., Asogawa, M., Fukushima, C., Asao, M., Shimadzu, H., Nakao, k., Shimizu, R.: Virtual Screening System for Finding Structurally Diverse Hits by Active Learning, J. Chem. Inf. Modeling, Vol. 48, No. 4, 930-940, 2008.

[7] Breiman, L.: Bagging Predictors, Machine Learning, Vol. 24, No. 2, 123-140, 1996.

[8] Leskovec, J., Adamic, L. A., Huberman, B. A.: The Dynamics of Viral Marketing, ACM Transactions on the Web, Vol. 1, No. 1, 2007.