拡張アプリオリアルゴリズムに基づくマトリクスクラスタリング手法の提案

小股　正博†　　小林　学†　　坂下　喜彦†

†湘南工科大学大学院工学研究科電気情報工学専攻

1. はじめに

インターネットの急速な進展により電子商取引（EC）が広く普及し，各種商取引データのマーケティングへの応用が重要になっている．Customer Relationship Management （CRM）では，様々な販売チャンネルを通じた顧客の取引の履歴情報を一元管理し，個々の顧客に最適な対応を実施することにより顧客の維持率を高め，長期的な企業利益を高めることを目的とする．このとき顧客の購買履歴から顧客の特性を把握する上でデータマイニング技術が強力なツールとなりうるものと期待されている．

ここで顧客の購買履歴の表現方法として，顧客(トランザクション)を行に，商品(アイテム)を列に表し，購買されていれば要素が1，そうでなければ0とする疎行列を用いる．このデータベースに対し，なるべく1の多い行と列からなる部分行列を取り出す手法はマトリクスクラスタリングと呼ばれる[1,2]．これは上で述べた顧客のニーズを分析する上で役立つ手法であると考えられており，行・列置換法，ピンポン法などのアルゴリズムが提案されている[1,2]．

本研究では，アイテムとトランザクションの相関ルール抽出技術として広く利用されているアプリオリアルゴリズムを拡張し，指定した密度以上の部分行列を抽出するマトリクスクラスタリングアルゴリズムを提案し，評価を行う．

2. マトリクスクラスタリング

マトリクスクラスタリング[1,2]とは与えられた疎行列に対して，非零の要素の割合（密度）がしきい値以上で，指定した以上の大きさ（面積）を持つ部分行列を抽出することと定義する．小柳らにより提案されたピンポン法[1]では，まずしきい値より多く要素に1を持つ行を選び，その行を活性化する．次に，活性化された行についてのみ着目し，これらの行にしきい値より多くの 1 を持つ列を選び，その列を活性化する．この活性化の手順をマーカ伝播と呼び，行と列の間でマーカ伝播を繰り返すことにより，しきい値以下の行または列を枝刈りし，また閾値以上の行または列を活性化する．マーカ伝播を繰り返しても同一状態が続いた時，この行・列により構成される部分行列を結果として出力する．

Matrix Clustering Method using Extended Apriori Algorithm
Masahiro KOMATA†, Manabu KOBAYASHI† and Yoshihiko SAKASHITA†
†Graduate School of Engineering, Shonan Institute of Technology

しきい値の適切な設定により，活性化される行及び列の数をある程度に抑えることができ，高速かつ良質な解を生成することが知られている．

3. 準備

まず本稿で扱う記号を定義する．全アイテム集合を I，全トランザクション数を N と表す．

定義 1 アイテム集合 $X \subset I$ に対し，X の要素アイテムを全て購入しているトランザクションの集合を $F(X)$ と定義する．また X のサポート $S(X)$ を

$$S(X)=\frac{|F(X)|}{N} \quad (1)$$

と定義する． □

最小サポートを S_{min} と表すと，アプリオリアルゴリズムは $S(X) \geqq S_{min}$ となる全ての $X \subset I$ を効率よく求めるアルゴリズムである．アプリオリアルゴリズムを実行すると，$S(X) \geqq S_{min}$ を満足する X に対して X を列集合とし，$F(X)$ を行集合とする部分行列が得られる．この部分行列の密度は1であり，部分行列の面積は $|X|S(X)N$ である．

定義 2 アイテム集合 $X=\{x_1,x_2,\cdots,x_n\} \subset I$ に対し，$X^{(r)}$ を以下のように定義する．

$$X^{(r)}=\bigcup\{\{x_{i_1},x_{i_2},\cdots,x_{i_r}\} \mid 1 \leq i_1 \leq i_2 \leq \cdots \leq i_r \leq n\} \quad (2)$$

すなわち，$X^{(r)}$ は X の任意の r 個の要素を含んだ全てのアイテム集合からなる集合である． □

4. アプリオリアルゴリズムの拡張

部分行列の最小密度 d を指定し，$|X|=n$ の場合を考える．まず非負整数 m を $d \leq m/n$ を満足する最小の値とする．このとき本節では n に依存するしきい値 $S_{\min}^{(n)}$ に対し次式満足する X を求める手法を提案する．

$$S\left(\bigcup_{r=m}^{n} X^{(r)}\right) \geq S_{\min}^{(n)} \quad (3)$$

補題 1 アイテム集合 $X=\{x_1,x_2,\cdots,x_n\} \subset I$ に対し，

$$S\left(\bigcup_{r=m}^{n} X^{(r)}\right)=S\left(X^{(m)}\right)=\sum_{r=m}^{n}(-1)^{r-m}\binom{r}{m}\sum_{Y \in X^{(r)}} S(Y) \quad (4)$$

が成り立つ[4]． □

補題1を利用して，最小密度 d 以上でかつ式

(3)を満足するXを求める提案拡張アプリオリアルゴリズムを以下に示す.

[拡張アプリオリアルゴリズム]

1) $C_1=\{\{r\}|r\in I\}$;
2) for(n=1; $C_n\neq\Phi$; n++){
3) $m=\arg\min\left\{m'\middle|{}^{m'}/_{n}\geq d\right\}$
4) Obtain $S(X)$ for all $X\in C_n$;
5) Calculate $S(X^{(m)})$ from Lemma1 for all $X\in C_n$;
6) $L_n=\{X\in C_n\,|S(X^{(m)})\geqq S^{(n)}{}_{\min}\}$;
7) $C_{n+1}=\{Y\subset I\ |\ |Y|=n+1$ and $Y^{(n)}\subset L_n\}$;
8) }
9) Output $\cup_n L_n$; □

ここでd=1と設定し，$S^{(n)}_{\min}=S_{\min}, n=1,2,\cdots$，と設定すると通常のアプリオリアルゴリズムとなる．従って，上記のアルゴリズムはアプリオリアルゴリズムの拡張となっている.

5. 新聞のテキストデータによる評価と考察

提案手法を評価するため，1994年の毎日新聞のデータを対象とし，経済欄の7066記事を用いて実験を行った．まずこれらの記事すべてに対して茶筅を用いて名詞を抽出し，各名詞をアイテムとした．アイテム数は19106である．また各記事がトランザクションである．これらのデータを用いてピンポン法及び拡張アプリオリアルゴリズムによるマトリクスクラスタリングの結果をそれぞれ表1,2に示す．なお拡張アプリオリアルゴリズムでは表2よりも数多くの結果が同時に得られているが，代表的なもののみを示した．また拡張アプリオリアルゴリズムにおいて用いたしきい値を表3に示す.

表1　ピンポン法の結果

しきい値(行－列)	5－1100	6－900	7－800	8－700
面積	40625	148896	196252	223941
密度	0.51	0.31	0.28	0.26

表2　拡張アプリオリアルゴリズムの結果

Item数	7	8	8	9	9
面積	8204	12360	3464	8100	2700
密度	0.79	0.72	1.00	0.76	0.89

表3　拡張アプリオリアルゴリズムのしきい値

n	1	2	3	4	5
$S^{(n)}_{\min}N$	100	200	300	300	300
n	6	7	8	9	
$S^{(n)}_{\min}N$	300	300	200	300	

表1において，ピンポン法のしきい値をそれぞれ「行のしきい値－列のしきい値」として表示している．なお各トランザクションにおけるアイテム数はさほど大きくないため，行のしきい値は列のしきい値よりも大変小さな値となる．なおピンポン法において，まず行のしきい値を固定し，最も密度が大きくなる列のしきい値を求め，その結果を表1に示している．表2の拡張アプリオリアルゴリズムではd=0.6と設定した．さらに式(3)を満たすXに対し，その部分行列内に面積が$S^{(n)}_{\min}|X|N$で密度が0.8以上の部分行列が存在するXをラージアイテム集合とした．表1及び表2において面積が大きく密度が高いほど良質な解を得られていることになる．表1,2の結果から，提案した拡張アプリオリアルゴリズムはピンポン法に比べて面積は小さいが密度の高い結果が得られることが分かる.

ピンポン法はパラメータ設定による実効時間の違いが少なく，行と列を対称的に取り扱うため，大きな行列に対して適用可能である．しかし，解を求める際に面積や最小密度を設定することができないという欠点を持つ．また1回の実行に対して1種類の部分行列しか求めることができない．一方提案手法はピンポン法に比べ計算量は非常に大きいが，数多くの良質な解をまとめて生成することができ，指定した密度以上の良質な解を求めることが可能である．従って多少時間をかけてでも類似した顧客の動向を解析したり，商品間の傾向を分析するために大いに有効であると考える.

6. まとめと今後の課題

本研究ではアプリオリアルゴリズムを拡張することにより，指定した密度以上の部分行列を抽出するマトリクスクラスタリングアルゴリズムを提案した．また従来法のピンポン法と比較評価を行い，有効性を示した.

提案手法では，着目するアイテム数nの値によって式(3)の左辺の値が非常に大きい場合があり，しきい値による判定が意味を成さなくなることがある．そのため，nの値によって有効な$S^{(n)}_{\min}$の値を決定する方法を検討する必要がある.

参考文献

[1] 小柳滋，久保田和人，仲瀬明彦，" Matrix Clustering:CRM向けの新しいデータマイニング手法"，Vol.42 No.8,2001.
[2] 上原子正利，小柳滋，" 内積縮退MC：類似行の検出と類似列の検出を組み合わせたマトリクスクラスタリングアルゴリズム"，情報処理学会論文誌 Vol,45 No.SIG 7(TOD22),2004.
[3] CD-毎日新聞94'データ集，日外アソシエーツ，1995.
[4] ホール，組合せ理論，吉岡書店,1971.