Dulmage-Mendelsohn分解を用いたマトリクスクラスタリングアルゴリズム

小股　正博†,　小林　学†, 坂下　喜彦†

†湘南工科大学　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．表1に簡単なPOSデータの例を示す．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．

本研究では，Dulmage-Mendelsohn分解[3]を用いてマトリクスクラスタリングを効果的に行うアルゴリズムを提案し，評価を行う．

表1　POSデータ例

	1	2	3	4	5	6	7
A	0	1	0	1	1	0	0
B	0	0	1	1	0	1	0
C	0	0	0	1	0	0	1
D	1	0	0	1	0	0	0
E	1	1	0	0	1	1	0
F	0	0	0	1	0	0	1
G	0	0	1	0	0	1	1

2. マトリクスクラスタリング [1,2]

マトリクスクラスタリングとは与えられた疎行列の行・列を入れ替えることにより，指定された面積(抽出された密行列の大きさ)以上で，指定された密度(部分行列における1の要素の比率)以上の密な部分行列を抽出することと定義する．さらに，アルゴリズムの評価をするため，面積と密度の一方を拘束条件，他方を評価値とすることにより指定された面積以上で密度が最大となる部分行列を見つけることを目的とする．

小柳らにより提案されたピンポン法[1]では，まず閾値より多く要素に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，これらの行に閾値より多くの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，閾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．閾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，高速に処理することが可能となる．また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．このアルゴリズムは適切な閾値を設定することにより，高速かつ良質な解を生成することが知られている．

3. Dulmage-Mendelsohn分解

Dulmage-Mendelsohn（DM）分解とは，非ゼロの係数が少ない場合の連立方程式の解法を求めるために開発された数学的技法であり，連立方程式の係数行列の行と列の並び替えによって行列のブロック三角化を与えるものである．

まず連立方程式の係数行列をAとし，行の集合をR，列の集合をCと現す．このとき係数行列Aに対する2部グラフは，$G(A)=(V,E)$,ただし$V=R\cup C$と表される．ここで枝の集合EはAの中で非ゼロの係数を持つ行と列に対する枝のみからなる集合である．このときDM分解は以下となる．

(Step1)　行列Aを2部グラフ$G(A)=(V, E)$, ($V=R\cup C$)として表し、$G(A)$の最大マッチングMを求める．

(Step2)　Step1で求めたMの辺を行集合Rと列集合Cの両方向へ向けて有効グラフを作る．さらにMに含まれないEの各辺をRからCへ向きをつけた有向グラフをつくり，その結果を$\tilde{G}=(V,\tilde{E})$とする．また，$\tilde{G}$において，RからMの端点を除いた頂点から有向道で到達できる頂点の集合をV_∞ ($\subseteq V$)とし，CからMの端点を除いた頂点から有向道で到達できる頂点の集合をV_0 ($\subseteq V$)とする．

(Step3)　$\tilde{G}$から$V_\infty\cup V_0$の頂点を除去したグラフ

$\tilde{G}$'の強連結成分分解を求める．ただし l<k のとき V_kから V_lへの有向道が存在しないように番号付ける．

(Step4) Step3 で求められた結果から行と列を並び替え，ブロック三角化する．

4. DM 分解を用いたマトリクスクラスタリング

DM 分解は行列の行と列を並び替え，ブロック三角化を与える．ただしそのままでは，大きなブロックが少数できてしまうケースも少なくない．ただし DM 分解の性質上，各ブロック内において非ゼロの成分は対角成分に集まる．そこで，DM 分解を用いた 2 段階のアルゴリズムを提案する．

(Step1) DM 分解を用いてブロック三角化し，対角ブロックを抽出する．

(Step2) Step1 で求めた対角ブロック内の対角成分を検索し，面積と密度の高い部分行列を取得する．

5. 人口データによる評価と考察

提案手法を評価するため，MUSASHI[4]の人口データを用いてピンポン法及び提案手法を用いてマトリックスクラスタリングを行った結果を表 3,4 に示す．なお人口データでは総トランザクション数が 3511．総アイテム数が 645 である．

表 3　ピンポン法

閾値	10	11
面積	6292	468
密度	0.46	0.77

表 4　提案手法

面積	49	49	112	162	180
密度	0.90	1.00	0.79	0.69	0.68

表 3 のピンポン法では閾値を 10 及び 11 とし，マーカ伝播の処理が終了したときの面積と密度を示している．提案手法では，Step1 の DM 分解において得られる行列の対角ブロックのサイズは，76×76，2×2，7×7 であった．またそれぞれの対角ブロックの密度は 0.123，1.0，1.0 である．ほとんどの非ゼロの要素はこのときすでに 76×76 のブロックに含まれている．従ってこのブロックに対して提案手法の Step2 の検索を行う．前節でも述べたように，DM 分解後は，対角ブロック中のさらに対角線近辺に非ゼロの要素が集まる．これは，DM分解において密な部分行列内の行と列の組み合わせは最大マッチングによって決まるためである．表 4 では対角ブロック内の対角線上を探索し，面積及び密度の高い部分を抽出している．面積及び密度が高いほど良質な解が得られていることになる．これを踏まえてピンポン法と提案手法を比較すると，提案手法はピンポン法に比べ，面積は小さいが密度の高い結果が得られる．

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．しかし提案手法では DM 分解後の行列の対角上に密な部分行列が集まることを利用しているため，多くの密な部分行列を高速に検索することが容易である．従って本手法はマトリクスクラスタリングに対して有効であると考える．

次に計算量の評価を行う．データベースのサイズを $n \times n$ と仮定し，その非ゼロの係数の密度を d とすると，ピンポン法の計算量は $O(d^2n^2)$ であることが知られている．また DM 分解の計算量は $O(n^2d\sqrt{n})$ である．大規模疎行列において n は d に比べ非常に大きい．そのためピンポン法のほうが 1 回の実行にかかる時間が短いことがわかる．ただし現在のコンピュータでは n がかなり大きくても DM 分解を問題なく行うことができる．また提案手法は 1 回の実行で複数個の良質な部分行列を効果的に抽出することができるため，提案手法は有効であると考える．

6.まとめ

本研究では DM 分解を用いたマトリクスクラスタリング手法を提案した．また従来有効とされているピンポン法と比較評価を行った．

本研究では人口データを使用して有効性の考察を行ったが，今後実際の POS データなどの実データを使用して，有効な情報を抽出できるか確認する必要がある．

また多次元行列，多値行列が取り扱えるようにアルゴリズムを拡張することも考えられる．

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，2001 年　Vol.42 No.8

[2]上原子　正利，小柳　滋，内積縮退：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)

[3]室田一雄，Dulmage-Mendelsohn 分解(DM 分解),http://www.misojiro.t.u-tokyo.ac.jp/~murota/lect-ouyousurigaku/dm050410.pdf

[4] 中原孝信, MUSASHI でデータマイニング, http://www.geocities.jp/dm_musashi/index.htm