5E-03 クラスタリング手法を用いたバスケット解析

5E－3

清水 周一　木村 雅彦　沼尾 雅之
日本アイ・ビー・エム株式会社 東京基礎研究所

1 はじめに

流通サービス業などの，売上記録や顧客情報などを蓄積するデータベースが，ますます肥大化し大規模なものになっている一方で，その大規模さゆえに，この蓄積されたデータから，そこに現れる性質について，トップダウン的に仮説を立てることが難しくなってきている．このような仮説検証型の問題点を解消するために，データの傾向や規則性を自動的に抽出するといったボトムアップ的な，発見型のデータ処理が強く求められている．

クラスタリングは，大量のデータを数え上げられるほどのクラスタに分割し，全体の見通しを良くするための手法である．われわれは，流通業の100万規模の売上データから，カゴを単位としてトランザクションを取り出し，カゴに含まれる商品の種類の偏りによって，カゴをいくつかのクラスタに分類するためのプロトタイプシステムを構築し，実験を行なった．

2 リレーショナル解析

リレーショナル解析[1]は，大量のオブジェクトの集合をいくつかのクラスタに自動分割するためのクラスタリング手法である．ここでは，オブジェクトの属性をビットベクトルで表現し，その属性ビットベクトルの間の類似度を定義することによって，二つのオブジェクトの類似度を計る．そして，似たオブジェクトは同じクラスタに入るように，また，異なった属性を持つオブジェクトは異なったクラスタに入るように，クラスタリング処理を行なう．

類似度は，例えば，以下のように定義できる．ここで，$11_{ii'}$ は，二つのオブジェクト i，i' の間で共通な属性の数を表し，$10_{ii'}$ は，オブジェクト i にのみ現れる属性の数を表す．

$$Sim_{ii'} = \frac{a11_{ii'}}{b11_{ii'} + c(10_{ii'} + 01_{ii'})} = \frac{A_{ii'}}{AM_{ii'}}$$

上式は，二つのオブジェクトに共通の属性の数と，少なくともどちらか一方にある属性の数との比をもって，一致の度合としている．この式に適切な係数 (a,b,c) を与えて，分子分母をそれぞれ，$A_{ii'}$，$AM_{ii'}$ とおくと，計算すべきクラスタ $C(i)$ は，以下の評価式を満たすものである．ここで，$C(i)$ は，オブジェクト i の属すクラスタを表す．

$$\max_{C} \frac{\sum_i \sum_{i' \in C(i)} A_{ii'} + \sum_i \sum_{i' \notin C(i)} (AM_{ii'} - A_{ii'})}{\sum_i \sum_{i'} AM_{ii'}}$$

分子第一項は，オブジェクト i，i' が同じクラスタに属すときの，属性が一致する数を表し，第二項は，異なるクラスタに属すときの不一致の数を表している．したがって，上式は，同じクラスタに属すオブジェクトの間では，属性の一致の数が最大に，また，異なるクラスタでは，不一致の数が最大になるような境界（クラスタ）を得るための評価式になっている．

上式にクラスタを表す配列 $x_{ii'}$ を導入すれば，以下の式となる．ここで，$x_{ii'}$ は，オブジェクト i，i' が同じクラスタに属すとき1となり，異なるクラスタのとき0となる二値配列である．

$$\max_{x} \sum_i \sum_{i'} (A_{ii'} - AM_{ii'}/2) \cdot x_{ii'}$$

反射則: $x_{ii} = 1$
対称則: $x_{ii'} = x_{i'i}$
遷移則: $x_{ii'} + x_{i'i''} - x_{ii''} \leq 1$

この式より，$A_{ii'} - AM_{ii'}/2$ が正のとき，すなわち，類似度 $Sim_{ii'} \geq 0.5$ のとき，$x_{ii'}$ の値が1になるように，逆に負のとき0になるように $x_{ii'}$ を決定すれば，上式を最大化するクラスタが得られる．ただし，$x_{ii'}$ は反射則，対称則，遷移則を満たす必要があるので，あまり単純ではない．

なお，類似度に関して $Sim_{ii'} \geq \alpha$ と一般化すれば，評価式は以下のようになる．

$$\max_{x} \sum_i \sum_{i'} (A_{ii'} - \alpha AM_{ii'}) \cdot x_{ii'}$$

Basket analysis based on an automatic clustering method
Shuichi SHIMIZU, Masahiko KIMURA, Masayuki NUMAO
IBM Japan Ltd., Tokyo Research Laboratory

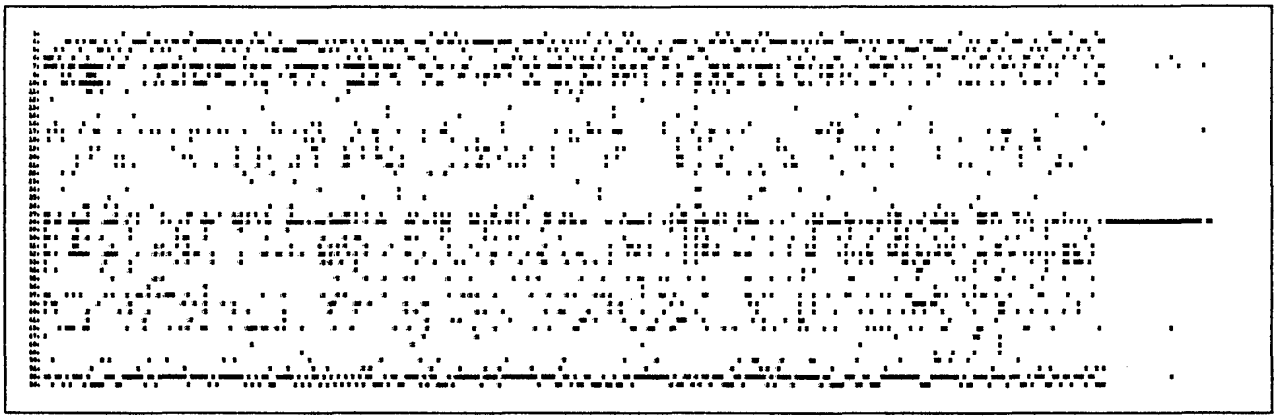

図 1: 購入商品のジャンル，顧客情報などを属性にした購買履歴

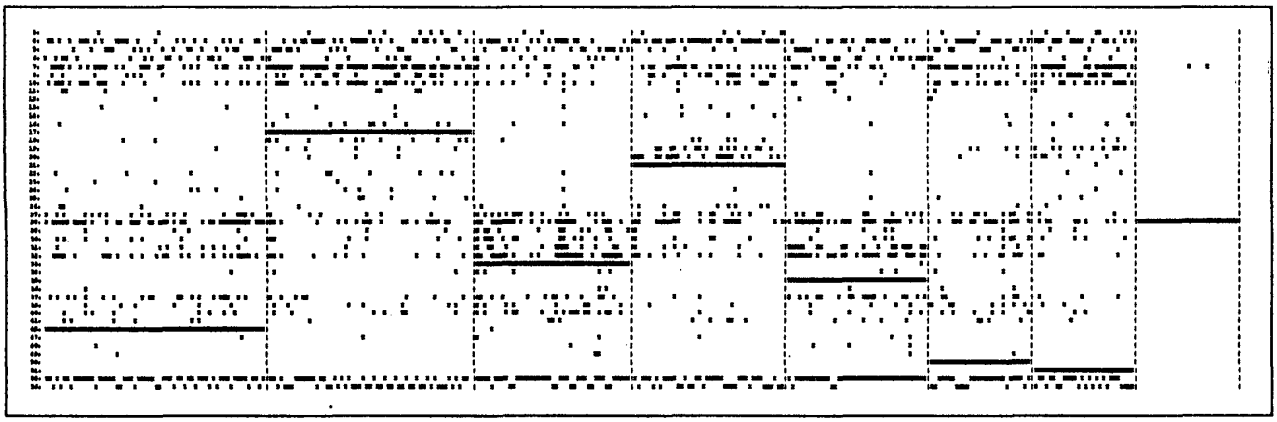

図 2: クラスタリングの結果 (一部)

3 実験例

図 1 に，13,523 人の購買履歴をもとに，購入商品のジャンルや個人の性別，年齢などを 56 の属性としてプロットしたデータの一部を示す．ここで，縦軸の上部 4 行は年代，最下部 2 行は性別，残りはジャンルである．

図 2 に，クラスタリング処理の結果を示す．各クラスタでの傾向がわかるように，縦列 (個人) を並び変えて，クラスタの大きい順にソートした．また，総クラスタ数 79 のうち，33 から 40 番目のクラスタのみ示した．

4 統計的手法との比較

リレーショナル解析では，属性ビットベクトルの分布に偏りがあれば，それがクラスタとなって現れる．同様のクラスタリングは，Nearest-Neighbor 法などの統計的手法を使って行なうことができるが，リレーショナル解析の場合には，初期値としてクラスタ数やクラスタの種を与える必要がないので，分割結果がそのような初期値に大きく依存するという問題もなく，非常に扱いやすい．なお，クラスタの数は，前述の α を変えることにより調整できる．

分割結果のクラスタに含まれる属性は，相関の強いものが集まっているように観測される．一般に商品や個人属性の間の相関を計るには，相互情報量 [2] や共起の頻度を計算する統計的な手法もあるが，組合せを数えるにあたって，属性の分類の大きさ (粒度) を適切に調整しないと，組合せとしてなかなかヒットせず絶対数としての閾値を越えられないという問題がある．一方，リレーショナル解析では，粒度が小さい場合でも，クラスタが構成されるように観測される．また，図 2の最右のクラスタのように，単独で現れる属性についても分割結果に現れる．

参考文献

[1] Chantal Bédécarrax et al.: "A New Methodology for Systematic Exploitation of Technology Databases", Information Processing & Management, Vol.30, No.3, pp.407–418, 1994

[2] 宮川: "情報理論", コロナ社, 1979