知能と複雑系 109－8
（1997．8．30）

極値集合を用いたバスケット分析の電話回線データへの適用

松浦大樹*，鷲尾　隆，元田　浩
大阪大学　産業科学研究所

Abstract

標準的なバスケット分析は，閾値以上の支持度を持つ全アイテム集合に関する閾値以上の確信度を持つ全相関ルールを導出し，統計的に自明なルールを除去して結果を得る．この枠組みでは，統計的意外性や興味がある相関ルールが得られ，対象データに内在する規則性の分かりやすい説明知識が与えられる．一方，知識発見のもう1つの主眼は，対象の推定や予測に有用な知識の導出である．そこで筆者等は，極値集合の概念に基づき，一定の支持度と確信度の下で最小限の事実から最大限の推定や予測を行う相関ルールを導く基準と，それに基づく新たなバスケット分析を考案した．更に，それによる分析システムを構築し，マーケッティング分析として電話網呼のログデータの傾向推定問題への適用を試みた．

Basket Analysis Based on Extremal Set and Its application to Telecommunication Data

Hiroki MATSUURA, Takashi WASHIO and Hiroshi MOTODA
Institute of Scientific and Industrial Research, Osaka University
{washio,motoda}@sanken.osaka-u.ac.jp

Abstract

The standard Basket Analysis derives all item sets and all association rules having support and confidence levels greater than their thresholds, and filters out trivial rules in statistical sense. This framework gives comprehensive descriptions of regularities involved in the objective data. Another major purpose of knowledge discovery is to derive valuable knowledge for estimation and prediction on the objective system. We propose a novel criterion of association rules and a new Basket Analysis to provide maximal guesses from minimal facts based on extremal set principle. We applied our framework to an estimation problem in logging data of telecommunication network in terms of marketing analysis.

1　はじめに

データマイニング技術に基づく知識発見は，大きく分けて2つの目的を有すると考えられる[1]．1つは，対象データが含む何らかの規則性を解析者にとって分かりやすい，ないしは興味ある内容記述で与えることである．これは主にデータを通じた対象理解や意志決定を支援することに主眼を置いている．大量のトランザクションデータについて，アイテム間の相関ルールの導出を行う従来のバスケット分析は，基本的にはこの範疇に属する手法である．即ち，一定以上の支持度を持つ全アイテム集合に関する一定以上の確信度を持つ全相関ルールを導出し，統計的に自明なルールを除去して結果を得る[2, 3]．この枠組みでは，統計的意外性や興味がある相関ルールが得られ，対象データに内在する規則性の分かりやすい説明知識が与えられる．

一方，知識発見のもう1つの目的は，データが対象とする系に関する種々の推定や予測に有用な知識の導出である[1]．上述のバスケット分析で得られた相関ルールは，従来，この目的にも用いられて来た．しかしながら後述するように，相関ルール抽出の基準が必ずしも統計的に一様な信頼性の帰結を保証しないため，導かれる結論が如何なる信頼性を有するかが不明確となる．従って，信頼性を考慮することが重要な推定や予測にとって，不都合が生ずる．

*現在，富士通関西通信システム株式会社所属．

そこで筆者等は，相関ルールの生成及び抽出の原理として一定の支持度や確信度の下での極値集合の概念を導入することを考えた．そして，一定の支持度と確信度の下で最小限の事実から最大限の推定や予測を行う相関ルールを導く基準を構成し，それに基づく新たなバスケット分析の枠組み及びアルゴリズムを考案した．更に，本手法に基づくバスケット分析システムを構築し，我々の枠組みが対象データを通じた推定や予測問題において有効性を持つこと確認するため，電話網呼のログデータからのマーケッティングのための傾向推定問題への適用を試みた．

2 従来の相関ルール抽出基準と極値集合による基準

バスケット分析で対象とするデータは，1つ以上のアイテムの集合であるトランザクションで構成される．例えば，コンビニエンスストアのデータベースには各顧客について購入した商品が次のように保存されている．

$$\text{顧客}_1 = \{\text{食料品}_a, \text{食料品}_b, \text{日用品}_a, \ldots\}$$

$$\vdots$$

$$\text{顧客}_n = \{\text{お菓子}_a, \text{飲料水}_a, \text{日用品}_b, \ldots\}$$

顧客単位のデータをトランザクションと呼び，例えば食料品$_b$ が顧客$_1$によって購入された事が示されている．食料品$_b$のようにデータベース内で，ある事象を表わすために用いられる記号をアイテムと呼ぶ．

バスケット分析では，支持度(Support)と確信度(Confidence)を基に，このようなデータベースから相関ルールを取り出す．取り出されるべきルールは次のように条件部である$Body : B$と結言部である$Head : H$によって表わされる．

$$B \Rightarrow H, \text{ ただし } B \subset H.$$

$Body : B$は例えば{食料品$_a \wedge$日用品$_a$}，$Head : H$は{食料品$_a \wedge$日用品$_a \wedge$食料品$_b$}のようなアイテムの集合であり，この場合，食料品$_a$と日用品$_a$を買う顧客は，それらに加えて食料品$_b$を購入しやすいことを表す．$Head : H$の支持度とは全トランザクションの中で，その$Head : H$を含むものの割合を示す値である。即ち，

$$sup(H) = \frac{H\text{を含むトランザクション数}}{\text{全トランザクション数}}.$$

$Body : B$の支持度についても同様に定義できる．確信度はルールの確からしさを表わし，$Body : B$を含むトランザクションの中で$Head : H$を含むトランザクションの数の割合で与えられる．即ち，

$$conf(B \Rightarrow H) = \frac{H\text{を含むトランザクション数}}{B\text{を含むトランザクション数}} = \frac{sup(H)}{sup(B)}.$$

バスケット分析では，これらの指標に対して最小支持度$(l - sup)$，最小確信度$(l - conf)$という閾値を設定し，それぞれより高い値を持つルールを取り出す。

また，更に興味深いルールのみを抽出するために，以下のようなフィルタリング操作を施す[3].

(A) $conf(B \Rightarrow BR) < sup(R)$ならば$B \Rightarrow BR$を削除.

(B) $conf(AB \Rightarrow ABR) < conf(B \Rightarrow BR)$ならば$AB \Rightarrow ABR$を削除.

(C) $conf(B \Rightarrow ABR) = conf(B \Rightarrow BR)$ならば$B \Rightarrow BR$を削除.

(D) $conf(B \Rightarrow BR) \times conf(B \Rightarrow AB) > conf(B \Rightarrow ABR)$ならば$B \Rightarrow ABR$を削除.

ただし，ここで$BR = B + R, ABR = A + B + R$である．

以上の過程で抽出される相関ルールが，一様な信頼性を有さないのは明らかである．たとえば，最小支持度や最小確信度を超えていれば，如何なる信頼性のルールも取り出される．また，上記フィルタ(A)のよ

うに，ルール自体が最小支持度や最小確信度の条件を満たしていても，削除されてしまう場合がある．このような抽出方法では，解析者に興味深い規則性の記述は得られるが，一様な信頼性を保証する推定や予測への適用は難しい．

そこで筆者等は，相関ルールを推定や予測問題へ適用することを眼目とし，その取り出しにおいて以下の条件を要請することとした．

1. 全ての相関ルールが指定されたほぼ同一の支持度と確信度を有すること．

2. 全ての相関ルールが極力少ない既知事実から極力特定的な結論を導くこと．

これら2つの要請が満たされれば，一定の信頼性を保証し，かつ推定や予測能力の高い相関ルールを得ることができる．

このような手法を構築するための準備として，極値集合の概念を基に以下の2つの定義の導入を行う．

極大支持集合：

ある指定支持度$s-sup$について，集合Hが

$$sup(H) \geq s-sup \text{ かつ } sup(S) < s-sup \quad \forall S \supset H$$

を満たす時，Hを指定支持度$s-sup$における極大支持集合という．

極小確信集合：

ある指定確信度$s-conf$について，ルール$B \Rightarrow H$が

$$conf(B \Rightarrow H) \geq s-conf \text{ かつ } conf(S \Rightarrow H) < s-conf \quad \forall S \subset B$$

を満たす時，Bを指定確信度$s-conf$の下で$Head:H$に関する極小確信集合という．

極大支持集合は，定義よりその任意の部分集合の支持度は$s-sup$より大きく，極小確信集合は，定義よりその任意の包含集合の確信度は$s-conf$より大きい．従って，極大支持集合Hを$Head$に持ち，その極小確信集合Bを$Body$に持つ$B \Rightarrow H$が上記2つの要請を満たす相関ルールとなる．

また，ルールのフィルタリング操作として，2つの異なる極大支持集合ABRとBCRが与えられ，それぞれより

$$AB \Rightarrow ABR \text{ 及び } B \Rightarrow BCR$$

という極小確信集合条件を満たすルールが得られた場合，前者より後者の方がより少ない事実BのみでRを推定できるため，前者を除去する．この操作を行っても残されたルールの支持度や確信度の一様性は確保される．

3 極値集合によるアルゴリズムとその実装

従来のバスケット分析の枠組みにおいて，大量データを対象として相関ルールを導出する効率的アルゴリズムについては，種々の研究がなされている[2]．その中で，代表的なアルゴリズムとしてAprioriが挙げられる．これはアルゴリズム構造が比較的単純でメモリー消費量が少なく，かつ効率が高い特徴を有する．筆者等の提唱する極値集合に基づく相関ルール生成の枠組みにおいても，従来手法と同様に各アイテム集合について支持度や確信度の計算が要求されるため，Aprioriアルゴリズムをベースとした相関ルール導出アルゴリズムを用いた．

バスケット分析アルゴリズムは，大きく以下の2ステップから構成される．

1. 指定支持度以上を有する全てアイテム集合の導出．

2. 導出された各アイテム集合から指定確信度を有する相関ルールの導出．

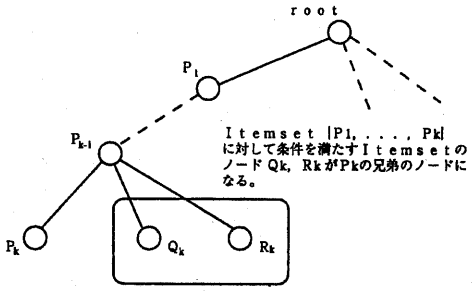

図 1: アイテム集合のトライ構造表現の利点

本枠組みにおいても，最初のステップは従来のAprioriアルゴリズムと同一のものを適用可能である．ここで，特に指定支持度以上を有するアイテム集合を大アイテム集合(Large Item Set)と呼ぶ．Aprioriアルゴリズムは，要素数1の大アイテム集合からはじめて，ボトムアップ的に要素数kの大アイテム集合から要素数k+1の大アイテム集合の候補を作り出すことを行う。要素数k+1のアイテム集合の各部分集合が指定以上の支持度を持たなければ，そのアイテム集合の支持度が指定値を超える可能性はない．従って，1つの要素以外が共通な要素数kの2つの大アイテム集合

$$P_k = \{item_1, item_2, \ldots, item_{k-1}, item_k\}$$

$$Q_k = \{item_1, item_2, \ldots, item_{k-1}, item'_k\}$$

が存在すれば，その和集合$P_k \cup Q_k$が要素数k+1の大アイテム集合の候補となる．このような各候補についてのみ，実際の支持度をデータから計算してチェックすることで効率的に大アイテム集合を求めることができる．

2番目のステップについては，従来の手法と異なり極大支持集合にのみ着目した処理を行う．即ち，はじめのステップで求めた大アイテム集合の間で，包含集合が存在しない要素数が極大な集合のみを残し，極大支持集合を得る．更に、各極大支持集合Hの中で指定確信度を満たす極小なBodyの集合Bを全て求め，相関ルール$B \Rightarrow H$を得る．

実装に用いたデータ構造を図1に示す．アイテム集合を一種のトライ構造で表現することで，上記のP_k, Q_kのような集合のマッチングを兄弟ノードとして効率的に処理する．また，最終的にトライの末端ノードが極大支持集合となる枝狩りも容易に行える．

以上のようなアルゴリズムとデータ構造に基づく相関ルール導出コードを作製し，実規模の推定・予測問題への適用を試みた．

4 電話網呼ログデータの傾向推定問題への適用

本枠組みの実規模の推定・予測問題への適用として，R.Kimball[4] のThe Data Warhouse Toolkitという書籍付属のCall Tracking Dataを用いた．これは実際の米国の電話会社のデータであり，1人の人物の1回の電話呼が1つのトランザクションに対応している．データ全体に含まれるトランザクション数は65,525個である．1つのトランザクションは各属性とその値のペアからなるアイテムを含むが，今回はその中から表1に示される内容の属性と値のペアを対象アイテムとした分析を行った．データ中に全部で221種類のアイテムが現れる．

表 1: 対象としたアイテムの内容

属性	値
Day Of Week	Sunday,Monday,Tuesday,Wednesday,Thursday,Friday,Saturday
Month	31-Oct,30-Nov,31-Dec
Calling City	Albany~Wilmington（９９都市）
Calling Marital	Married,Coresident,Single
Calling Sex	F,M
Called City	Albany~Wilmington（９８都市）
Called Marital	Married,Coresident,Single
Called Sex	F,M
Call Type Description	Voice,Busy,Data,Cellular

はじめに，従来手法と本提案手法により生成される相関ルールの違いを検証した．その結果を表2に示す．フィルタリング前の生成ルール数については，本提案手法の方が従来手法に比較し少ない．特に指定支持度や指定確信度が低い場合，従来手法では種々の支持度や確信度を持つルールが生成され数が著しく増加するが，提案手法は極大支持集合及び極小確信集合の条件を満たすルールのみを生成するため，生成ルール数の抑制が顕著である．フィルタリング後の生成ルール数については，一概に大小は言えない．従来手法のフィルタに比べ，支持度や確信度を維持する本提案手法のフィルタの方が弱いので，フィルタリング前のルール数が同程度の場合には，抽出されたルール数は従来手法の方が少なくなる．ただし，指定支持度や指定確信度が低い場合には，本手法の方がフィルタリング前のルール数が少ないため，弱いフィルタを介しても最終的に抽出されるルール数は少なくなる．当然，従来手法で抽出される相関ルールの支持度や確信度は一様ではないのに対し，本手法で抽出されるルールの支持度及び確信度は均質であり，かつより大量のルールが生成される低い指定支持度や指定確信度の場合には，より縮約されたルール群が得られることが判る．

次に，指定支持度を2.0%，指定確信度を60.0%に設定して相関ルールの抽出を試みた．以下に，抽出された幾つかのルール例を挙げる．

＜例１＞

```
{ [Called_Marital]:Single  [Calling_Marital]:Coresident }==>{ [Calling_Sex]:F }
{ [Called_Marital]:Coresident  [Calling_Marital]:Single }==>{ [Called_Sex]:F }
```

この２つのルールから「同居者と未婚者の通話では，同居者は女性であることが多い.」と推定される．

＜例２＞

```
{ [Called_City]:Augusta }==>{ [Called_Marital]:Single }
{ [Calling_City]:Augusta }==>{ [Calling_Marital]:Single }
```

この２つのルールから「Augasta市では，未婚者が電話を利用することが多い.」と推定される．

＜例３＞

```
{ [Called_Sex]:M  [Calling_Sex]:M }
              ==>{ [Calling_Marital]:Married  [Called_Marital]:Married }
```

このルールから「男性同士の通話では，双方とも既婚者であることが多い.」と推定される．

このようにして得られたルールは，必ず一律に指定支持度を2.0%，指定確信度を60.0%を保証する傾向推定及び予測に利用することが可能である．

表 2: 相関ルールの生成数比較

ルールの抽出方式	指定支持度	指定確信度	トライのノードの数	取り出されたルールの数	フィルタ後のルールの数
従来方式	2.0 %	90.0%	1,257	1,292	4
極大極小方式				1,220	4
従来方式	2.0 %	70.0%	1,257	1,880	8
極大極小方式				941	24
従来方式	2.0 %	50.0%	1,257	2,980	71
極大極小方式				1,472	74
従来方式	2.0 %	30.0%	1,257	4,966	50
極大極小方式				1,386	52
従来方式	1.0 %	90.0%	3,079	3,695	301
極大極小方式				2,658	107
従来方式	1.0 %	70.0%	3,079	4,954	233
極大極小方式				2,046	129
従来方式	1.0 %	50.0%	3,079	7,470	302
極大極小方式				3,211	222
従来方式	1.0 %	30.0%	3,079	11,691	123
極大極小方式				2,915	115

5 おわりに

本研究では，標準的なバスケット分析の相関ルール導出規範が，対象データに内在する規則性の分かりやすい説明を与える一方，知識発見のもう1つの主眼である，対象の推定や予測に必ずしも適合しないことを指摘した．そこで一定の支持度と確信度の下で最小限の事実から最大限の推定や予測を行う相関ルールを導く基準と，それに基づく新たなバスケット分析の枠組みを考察した．その結果，極値集合の概念の導入と，従来のバスケット分析アルゴリズムの改変により，推定や予測問題に適合した手法を構築できた．

また，この手法を電話網呼のログデータの傾向推定問題への適用を試み，十分に縮約され，かつ均質な統計的信頼性をもたらす推定・予測ルールを導出可能であることを明らかにした．

参考文献

[1] U.Fayyad , G.Piatesky-Shapiro and P.Smyth. From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE FALL, pp.37-54, 1996.

[2] R.Agrwal and R.Srikant. First algorithms for mining association rules. In *Proceedings of the 20th VLDB Cnference*, pp.487-499, 1994.

[3] 野口 祐史，國藤 進．データベースからの知識発見法を用いた発想支援システムの研究．合同研究会" ＡＩシンポジウム' 96" (第7回)予稿集，pp.76-81, 1996.

[4] R.Kimball. The Data Warehouse Toolkit. Wiley Computer Publishing.