岡山大学環境理工学部研究報告
第10巻，第1号，pp.41-47，2005年2月

分割表データにおけるApriori Alogrihtmを利用した変数選択手法

大野学*　　垂水共之†

Variable Selection Method using Apriori Algorithm on Contingency Table

Manabu Ohno*　　Tomoyuki Tarumi†

(Received December 28, 2004)

Summary

We proposed a new applied method for induction of variable selection on contingency table. This method is the application of Apriori algorithm on variable selection of contigency table with iteraction. We assume that variables are dichotomous variable. We confirm that can be select variable, when minimun support is low level by using AIC on variable selection criterion.

Key words: Apriori Algorithm, Varable Selection, Contigency Table, Interaction

1　はじめに

近年，情報技術に急速な技術革新によって，大量のデータが高速で安価に手に入る環境が整ってきた．そのような環境において、大量のデータから有用な知識を抽出するデータマイニングの研究が盛んに行われている．また，実際にデータマイニングを活用する，すなわちビジネスインテリジェンス (business intelligence) という形で，企業に存在する大量データを知識と利益に変える情報戦略が企業の意思決定において必要不可欠になりつつある．特に，顧客の購買情報を大量にもつ企業は，顧客のデータを通しての集中型マーケティング (focused marketing) が必須になっている．

顧客データとして代表的なものは，マーケットバスケットデータである．このデータの分析方法として，相関関係分析 (association analysis) がよく用いられる．それは，分析結果が直感的であり，また，そこから商品配置や商品のセット販売の施策に役立てられるからである．そのマーケットバスケット分析では，一般に「買った」,「買ってない」という2値の情報のうち,「買った」という一方のみを使用する．

マーケットバスケット分析では，サポート (support) という指標を用いて同時購買が行われている商品間の関連性の強さを示す．一般に，サポートが大きいアイテムの同時購買には，交互作用が作用していると考えることが出来る．アイテム i_1 とアイテム i_2 のアソシエーションルール $i_1 \rightarrow i_2$ のサポートが高ければ，i_2 の購買を i_1 の購買で予測することが出来る．しかし，i_2 の非購買を i_1 の非購買で予測するという問題に対しては，一般のアソシエーションルールのサポートの情報からは予測できない．その問題に対応するためには，2値型の変数の分布を考慮する必要がある．それは，目的変数と予測変数 (説明変数) の分割表データにおける，目的変数に有効な予測変数の変数選択を考えることに帰着する．本稿では，分割表データにおける高次元の予測変数 (高次の交互作用項) を考慮した効率的な

* 岡山大学大学院 自然科学研究科 資源管理科学専攻 〒700-8530 岡山市津島中 3-1-1 ohno@ems.okayama-u.ac.jp
† 岡山大学 環境理工学部 環境数理学科 〒700-8530 岡山市津島中 3-1-1 tarumi@ems.okayama-u.ac.jp

変数選択問題を考える．

いま，p個の予測変数があり，予測変数の主効果だけ仮定して目的変数を予測する場合，考えられる予測変数の個数は，たかだかp個である．しかし，予測変数に交互作用を仮定した場合，その個数は2^p-1個である．例えば，$p=15$といった場合でも，その数は32767個と膨大になる．よって，有効な変数を効率よく選択する必要がある．このような変数選択の必要性から，判別モデルの変数選択に対してアプリオリアルゴリズム [Agrawal et al., 1994] を利用することが [Ohno & Yamaguchi, 2003] で提唱された．

本稿は，[Ohno & Yamaguchi, 2003] を参考に，1) 変数の選択基準にAICを導入，2)`varialbe-delete`関数を提案し，不要な変数の大幅な削除の実現，の2項目の改良を行いその有効性を示す．

2 変数選択手法

本稿で想定しているデータを次に示す．いま，p+1個のカテゴリー変数$X_0,\cdots,X_p$がある．ここで，特に目的変数をX_0とし，それ以外の変数を予測変数とする．また，2値の値には優位性 (priority) はないと仮定する．よって，問題設定は，目的変数X_0を予測変数$X_1,\cdots,X_p$を使って予測するために有効な変数を選択する問題である．特に予測変数には高次までの交互作用項を考慮するものとする．この問題は前述したように，予測変数と目的変数の分割表データにおける予測変数の選択問題と等価である．

2.1 apriori-gen 関数

ここで，本稿で用いる記号の定義を表1に示す．変数選択アルゴリズムの各ステップにおいて目的変数X_0を含んだサポートのみを考慮するため，予測変数は常に目的変数X_0との交互作用項の形で存在する．すなわち，下記のような交互作用項

$$\underbrace{X_0X_{i_1}\cdots X_{i_{k-1}}X_{i_k}}_{k+1}$$

である．しかし，実際に必要なのは予測変数の部分であるので，本稿では予測変数の部分のみを考慮し，その任意の予測変数をk変数とよぶ．

変数は変数選択アルゴリズムの各ステップにおいて，いずれかの集合の要素として存在する．k変数において，最小サポートを*minsup*と表し，*minsup*を満たす可能性のある変数の候補の集合を，候補変数C_kとする．このC_kの生成を，アプリオリアルゴリズムの`apriori-gen`関数によって生成される．その`apriori-gen`関数を図1に示す．

表1: 記号の定義

記号	定義
C_k	k変数の候補変数集合 (candidate variable set) とする．`apriori-gen()`関数によって生成された集合．
L_k	k変数のラージ変数集合 (large variable set) とする．C_kの中から*minsup*を満たす集合．
V	選択変数集合 (selected variable set) とする．$\bigcup_k L_k$の中から`variable-delete`関数によって選択された変数の集合．

この関数はアプリオリアルゴリズムの中核をなす関数である．この関数の1～4行を*join step*といい，5～8行の*prune step*という．本手法についても，この関数の*join step*を用いることによって生成する変数の数を減らすことを可能にした．

*join step*の動作を，候補アイテム集合のアイテム集合を用いて説明する．いま，k個のアイテムを要素とする集合をk-アイテム集合と呼び，k-アイテム集合pとqを以下のように表す．

$$\{p.item_1\ p.item_2\cdots p.item_{k-1}\ p.item_k\}$$

$$\{q.item_1\ q.item_2\cdots q.item_{k-1}\ q.item_k\}$$

このとき，*join step*は，以下のように最後のアイテム$p.item_k$と$q.item_k$だけが違うもの同士を結合させる機能をもつ (図1の2行目から4行目)．

$$\{p.item_1\ p.item_2\cdots p.item_{k-1}\ p.item_k, q.item_k\}$$

ただし，各アイテムは，あらかじめ決められた順序関係に従い，ソートされているものする．また，*prune step*では，*join step*で生成されたアイテム集合の部分集合が，L_kにないものを取り除いている．ここでは，本手法に対応して，アプリオリアルゴリズムで想定されている購買商品を表す*item*ではなく変数についての動作例を示す．

動作例

いま，L_3が$\{X_1X_2X_3\}$, $\{X_1X_2X_4\}$, $\{X_1X_3X_4\}$, $\{X_1X_3,X_5\}$であったとき，`apriori-gen`関数の*join step*では，$\{X_1X_2X_3X_4\}$, $\{X_1X_3X_4X_5\}$となり，*prune step*では，$\{X_1X_2X_3X_4\}$となり，C_4が生成される．

1) **insert into** C_k
2) **select** $p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}$
3) **from** $L_{k-1}\ p, L_{k-1}\ q$
4) **where** $p.item_1 = q.item_1, \ldots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$

Next, in the *prune* step, we delete all itemsets $c \in C_k$ such that some $(k-1)$-subset of c is not in L_{k-1}:

5) **forall** itemsets $c \in C_k$ **do**
6) **forall** $(k-1)$-subsets s of c **do**
7) **if** $(s \notin L_{k-1})$ **then**
8) **delete** c from C_k;

図 1: apriori-gen 関数

本手法では，目的変数を含んだ形での予測変数の組合せのみ考慮すればよいため，[大野，垂水, 2004] で提案されたapriori-gen 関数の *join step* を用いる．それは，*join step* の pass-2 の生成する変数集合において，$\{X_0\} \cup \{X_j\} \mid \forall j,\ j \neq 1$ のみを作るようにすれば，Pass-3 以降は，必ず $\{X_0\}$ を含んだアイテム集合しか生成されない．その動作を図 2 に示す．ただし，変数の並びは，X_0 を先頭にソートされているものと仮定する．目的変数 X_0 を含んだ変数しか生成しないことよって，無駄な変数の生成行わず，計算の高速化に寄与する．ただし，このとき，*prune step* は使えないのは自明である．

本手法は，アプリオリアルゴリズムを利用する手法であるので，そのアルゴリズムの特性上，本手法も *minsup* に強く依存している．それは，*minsup* を高く設定すれば，計算コストは軽減できるが，有効な変数が選択されない確率も同時に上がる．本手法では，与えれた *minsup* の中で目的変数に対して有効な予測変数を選択するという考え方をとるものとする．

1) **insert into** C_2
2) **select** $i_1.item_1, q.item_1$
3) **from** $L_1\ i_1, L_1\ q$
4) **where** $i_1.item_1 < q.item_1$

図 2: Pass-2 における Join Step

2.2 AIC を用いた分割表データの変数選択

AIC を用いた分割表データにおける変数選択については，[坂元, 1985] で述べられている．いま，目的変数を X_0 とし，予測変数を X_1 で表し，各変数は値 0 と 1 をとるとする．X_0 が値 $x_0 = 0, 1$ をとり，かつ，X_1 が値 $x_1 = 0, 1$ をとる確率を $p(i_0, i_1)$，対応する観測度数を $n(i_0, i_1)$ で表すことにすると，

$$\sum_{i_0=0}^{1} \sum_{i_1=0}^{1} p(i_0, i_1) = 1 \tag{2.1}$$

$$\sum_{i_0=0}^{1} \sum_{i_1=0}^{1} n(i_0, i_1) = n \tag{2.2}$$

である．ここで，n は標本数である．n が十分に大きいとして，確率の集合 $\{p(x_0, x_1)\}$，$x_0 = 0, 1$，$x_1 = 0, 1$ の下で観測度数の集合 $\{n(x_0, x_1)\}$ が得られる確率は多項分布，

$$M(\{n(x_0, x_1)\} \mid \{p(x_0, x_1)\})$$
$$= \frac{n!}{\prod_{x_0=0}^{1} \prod_{x_1=0}^{1} n(x_0, x_1)!} \prod_{i_0=0}^{1} \prod_{i_1=0}^{1} p(x_0, x_1)^{n(x_0, x_1))}$$

で求められる．上式を $\{n(x_0, x_1)\}$ が与えれれたときの $\{p(x_0, x_1)\}$ の尤度関数は，パラメータ $\{p(x_0, x_1)\}$ に無関係な定数項を無視すると，その対数尤度

1) L_1 := {Set of main effects those with minimum support }
2) **for**($k=2$; $L_{k-1} \neq \emptyset$; $k++$) **do begin**
3) C_{k+1} :=apriori-gen(L_k)
4) $L_k := \{c \in C_k \mid 0 \leq^{\exists} u \leq 2^k - 1; u \in \boldsymbol{Z}, c.support[u] \geq minsup\}$
5) **end**
6) V := variable-delete($\bigcup_k L_k$)
7) **end**
8) Answer=V;

図 3: 変数選択アルゴリズム

$l(\{p(x_0,x_1)\}$ は

$$l(\{p(x_0,x_1)\} = \sum_{x_0=0}^{1}\sum_{x_1=0}^{1} n(x_0,x_1)\log p(x_0,x_1) \quad (2.3)$$

これは2次元の分割表の場合であるが，多次元に拡張した場合においても同様の手続きによって，対数尤度を求めることができる．

いま，X_0 と k 個の変数 X_{i_1}, X_{i_2}, $\cdots$, X_{i_k} が関連があるとしたとき，それは X_{i_1}, X_{i_2}, $\cdots$, X_{i_k} が与えられれば，X_0 が定まると考えることができるので，そのモデルは

$$\mathrm{MODEL}(X_0; X_{i_1}, X_{i_2}, \cdots, X_{i_k}):$$
$$p(x_0 \mid x_{i_1}, x_{i_2}, \cdots, x_{i_k}) = a(x_0 \mid x_{i_1}, x_{i_2}, \cdots, x_{i_k}) \quad (2.4)$$

と書ける．$a(\ )$ は，$p(\ )$ の確率が $a(\ )$ の中の変数で決まる確率モデルであることを示す．このモデルを予測変数に関して k 次元に拡張した場合の対数尤度関数に代入し，制約条件，

$$\sum_{x_0=0}^{1} a(x_0 \mid x_{i_1}, x_{i_2}, \cdots, x_{i_k}) = 1 \quad (2.5)$$

の下で，パラメータ $a(x_0 \mid x_{i_1}, x_{i_2}, \cdots, x_{i_k})$ の最尤推定量は

$$\tilde{a}(x_0 \mid x_{i_1}, x_{i_2}, \cdots, x_{i_k}) = \frac{n(x_0, x_1, \cdots, x_{i_k})}{n(x_1, \cdots, x_{i_k})} \quad (2.6)$$

で与えられる．よって，
$\mathrm{MODEL}(X_0; X_{i_1}, X_{i_2}, \cdots, X_{i_k})$ の AIC は

[坂元, 1985] より

$$\mathrm{AIC}(X_0; X_{i_1}, X_{i_2}, \cdots, X_{i_k}) = (-2) \sum_{x_{i_1}, x_{i_2}, \cdots, x_{i_k}} n(x_{i_0}, x_{i_1}, x_{i_2}, \cdots, x_{i_k}) \times \log \frac{n \cdot n(x_{i_0}, x_{i_1}, x_{i_2}, \cdots, x_{i_k})}{n(x_0) n(x_{i_1}, x_{i_2}, \cdots, x_{i_k})} + (2^{k+1} - 1) \quad (2.7)$$

分割表におけるこのモデルの $X_{i_1}, X_{i_2}, \cdots, X_{i_k}$ は k 変数の予測変数とみなせるので，予測変数の変数選択の評価基準としてこの AIC を用いる．

2.3 AIC に基づく変数の除去

$\bigcup_k L_k$ は $minsup$ を満たす変数の集合の集合である．目的変数の予測に有効な変数をより精練化するために，予測に有効でない変数を除去する．ここではそれを実行するための variable-delete 関数を図 4 に示す．この関数では，あらかじめ変数の AIC の低い値順にソートする．次に上位の変数が下位の変数に含まれているものを削除する．なぜなら，上位は変数が下位の変数に含まれている場合，その上位の変数の方がより目的変数を説明できる情報をもっており，それ以上の高次の変数を考える必要がないためである．

3 人工データによる数値実験

変数選択手法の有効性を次の人工データを用いた数値実験を通して確認する．

予測変数 X_i は $X_i \sim \mathrm{Bin}(1, u_i)$, $i = 1, \cdots, p$ に従う 2 項乱数を $n = 1000$ 個発生させる．u_i は互いに独立な一様分布 $\mathrm{U}[0,1]$ に従う乱数である．いま，目的変数は表 2 の各々のモデルを用いて，次の式で

1) **insert into** V_k
2) **from** $\bigcup_k L_k$　　// $\bigcup_k L_k$ is set of all-k large variable sets
3) **forall** $l \in \bigcup_k L_k$　　// l are kept sorted in small AIC order
4)　**if** ($l_i \subset l_j$　l_i.AIC $< l_j$.AIC) **then**
5)　　**delete** l_j from $\bigcup_k L_k$

図 4: `variable-delete` 関数

表 2: 交互作用モデル

Model Name	Model Formula: $f(\boldsymbol{X})$
Model 1	X_1X_2
Model 2	$X_1X_2X_3$
Model 3	$X_1X_2X_3X_4$

与えられる.

$$X_0 = \begin{cases} 1 & \text{if } \frac{20\times f(\boldsymbol{x})}{1+\exp f(\boldsymbol{x})} \geq u, \\ 0 & \text{if } \frac{20\times f(\boldsymbol{x})}{1+\exp f(\boldsymbol{x})} < u, \end{cases}$$

ただし, u は一様分布 U[0,1] に従う乱数である. また, 交互作用は次のような値,

$$\{X_{i_1}X_{i_2}\cdots X_{i_k}\} = \begin{cases} 1 & \text{if } \forall i; X_i = 1 \vee 0 \\ 0 & \text{その他} \end{cases}$$

をとる交互作用を仮定する. 各サポートごとに, 各モデルのもとで 100 セットのデータを前述した方法によって発生させ, 手法の選択精度を評価を行う.

4　結果と結論

実験結果の項目のは次の通りである.

- 1st Acc.　モデルの変数が選択変数集合 V の中でもっとも AIC が低い値をもつ 100 回のシミュレーションにおける出現率.
- Overall Acc.　モデルの変数の選択変数集合 V における 100 回のシミュレーションにおける出現率.
- NS. Ave.　選択変数集合 V 中の 100 回のシミュレーションにおける変数の個数の平均.
- NS. Var.　選択変数集合 V 中の 100 回のシミュレーションにおける変数の個数の分散.
- DR. Ave.　`variable-delete` 関数よる 100 回のシミュレーションにおける削除率の平均. すなわち, $1-(|V|/|\bigcup_k L_k|)$ の 100 回の平均.
- DR. Var.　`variable-delete` 関数よる 100 回のシミュレーションにおける削除率の分散. すなわち, $1-(|V|/|\bigcup_k L_k|)$ の 100 回の分散.

表 4 と表 5 より, Model2 と Model3 において, サポートが高い場合は, 変数が正確に選択されない. これは, モデルに含まれる変数のサポートが *minsup* より低いため, ラージ変数集合として選択されなかったためである. これは, 前述したように, *minsup* に強く依存したアルゴリズムであるという本手法の特性上, 必然的な結果である. このとき, 他の変数が選択変数集合 V に含まているわけであるが, その AIC は正しく変数が選択された状態に比べて, 各変数間であまり差がない状態になる.

各モデルにおいて, ラージ変数集合に正しく変数が含まれた場合, 高い精度で選択できていることがわかる. そのとき, サポートをさらに低くした場合においても, 選択精度が極端に悪くなることなく, 緩やかに選択精度が減少している. これは, 前述したとおり, 変数の選択規準に AIC の値が変数の次数を考慮した規準であり, それによって無意味な高次の変数をなるべく選択しないように働く.

次に, `variable-delete` 関数の削除率であるが, 表 3,4,5 の DR.Ave. と 1st Acc. より, この関数によって不必要な変数を正確に削除していることがわかり, それはこの関数の有効性を示すものである. この関数の動作によって, サポートを低くした場合においても, 選択する変数の個数を 20 個前後に収めることが可能になっている. 以上の結果より, 結論として

- 分割表データの変数選択に対して, 変数選択規準に AIC を用いることにより, 低いサポートおいても, その有効性を確認した.
- `variable-delete` 関数によって, $\bigcup_k L_k$ 中の不必要な変数を大幅に削減でき, その有効性を確認した.

表 3: Model1 の数値実験結果

Model 1 : X_1X_2						
minsup	1st. Acc	Overall. Acc.	NS. Ave.	NS. Var.	DR. Ave.	DR Var.
200	1.00	1.00	20.81	2.84	0.52	2.00×10^{-2}
175	1.00	1.00	22.49	5.64	0.86	4.79×10^{-4}
150	0.99	0.99	21.95	4.59	0.88	1.28×10^{-4}
125	0.82	0.82	21.92	5.81	0.88	1.56×10^{-4}
100	0.71	0.71	21.06	5.95	0.95	1.52×10^{-4}
75	0.71	0.71	20.53	6.85	0.98	5.48×10^{-6}
50	0.73	0.74	19.79	5.37	0.99	2.26×10^{-6}
25	0.73	0.73	18.8	4.66	0.99	1.03×10^{-7}

表 4: Model2 の数値実験結果

Model 2 : $X_1X_2X_3$						
minsup	1st. Acc	Overall. Acc.	NS. Ave.	NS. Var.	DR. Ave.	DR Var.
200	0.00	0.00	18.64	0.23	0.04	4.04×10^{-3}
175	0.00	0.00	18.96	1.87	0.33	2.35×10^{-2}
150	0.00	0.00	22.18	4.87	0.84	1.20×10^{-3}
125	0.56	0.56	21.91	4.64	0.88	1.29×10^{-4}
100	1.00	1.00	20.71	5.11	0.90	1.74×10^{-4}
75	1.00	1.00	20.71	4.87	0.97	2.25×10^{-5}
50	0.83	0.83	19.1	4.57	0.98	5.12×10^{-6}
25	0.77	0.77	18.8	4.78	0.99	1.55×10^{-5}

といえる.

参考文献

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data.* 207-216.

[Agrawal et al., 1994] Agrawal, R., Imielsinki, T. & Swami, A. (1994). Fast Algorithm for mining association rules. *The Proceedings of International Conference on Very Large Data Bases.* 487-499.

[Ohno & Yamaguchi, 2003] Ohno, M and Yamaguchi, K. (2003). Variable Selection using the Apriori Algorithm for Discriminant Analysis. *The proceedings of the 4th ARS Conference of the IASC.* 23-237.

[Goodman, 1971] Goodman, L. A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics.* Vol.13. pp 91-61.

[大野，垂水, 2003] 大野学, 垂水共之. (2003). 判別分析におけるアプリオリアルゴリズムを利用した変数選択手法の改良. 日本計算機統計学会第 17 回シンポジウム論文集. 95-98.

[大野，垂水, 2004] 大野学, 垂水共之. (2004). 特定のアイテムを含んだアソシエーションルール

表 5: Model3 の数値実験結果

Model 3 : $X_1X_2X_3X_4$						
minsup	1st. Acc	Overall. Acc.	NS. Ave.	NS. Var.	DR. Ave.	DR Var.
200	0.00	0.00	19.0	0.00	0.00	0.00
175	0.00	0.00	18.73	0.34	0.04	5.00×10^{-3}
150	0.00	0.00	20.22	4.88	0.60	2.00×10^{-2}
125	0.00	0.00	21.89	5.29	0.87	2.55×10^{-4}
100	0.00	0.00	22.1	6.31	0.88	1.76×10^{-4}
75	0.07	0.07	20.55	3.74	0.95	1.32×10^{-4}
50	0.97	0.97	19.37	6.37	0.98	4.76×10^{-7}
25	0.88	0.88	18.56	6.02	0.99	2.78×10^{-7}

の抽出. 日本計算機統計学会第 18 回シンポジウム論文集. 95-98.

[坂元, 1985] 坂元慶行. (1985). カテゴリカルデータのモデル分析. 共立出版.

[福田, 森下, 1995] 福田剛志, 森下真一. (1995). 相関ルールの可視化について. 電子情報通信学会技術研究報告. **95**-81, 41-48.