Knodiasにおけるデータの性質に着目した相関ルール抽出の効率化

2W－7

山崎 高日子，三石 彰純，小幡 康，田中 秀俊，白石 將†
†三菱電機（株）情報技術総合研究所

1 はじめに

相関ルールを構成するアイテム間には、多くの場合既知の関係が存在する。

そこで、我々は、データマイニングシステムKnodiasにおいて、アイテム間に存在する既知の因果関係と排反関係を活用して、無意味なルール抽出や無駄な探索を行なわない機能をもつエンジンを開発した。

すなわち、①アイテムごとに相関ルールの条件部あるいは結論部への出現可否を指定（デフォールトは条件部／結論部共に出現可能）し、指定されていない組み合わせの相関ルールは探索を行なわない。②排反関係にあるアイテム（例えば、”男”と”女”）の組み合わせは探索対象から除外する。

この機能により、無意味な出力の抑制の効果が得られる。本稿では、その手法と出力ルール数の測定結果について報告する。

2 条件部、結論部出現指定の手法

正の相関ルール抽出と負の相関ルール抽出の場合とで、実現手法が異なるので、以下それぞれに分けて説明する。

2．1 正の相関ルール抽出における手法

アイテムの組を昇順に並べたハッシュ木を図1に示す。ここで、L_kはrootからの階層距離がKであるノードの集合であり、l_kはその要素であって長さKのアイテムセットを示す。かかるハッシュ木に対し条件部、結論部出力指定情報としてノードに付加することとする。すなわち、L_1ノードについては条件部（左辺）指定がされているものであればL（LHS(Left Hand Side bit) ON)、結論部（右辺）であればR（RHS ON)なる情報を付加する。そして、L_k(k>1)ノードではLHSの情報のみを付加する。なぜなら、Knodiasではルールの結論部に複数のアイテムがくることはないからである。L_k(k > 1)においてノードのLHSをONにする規則を図1中に示す。この時、そのアイテムセットが条件部になりうることが示される。

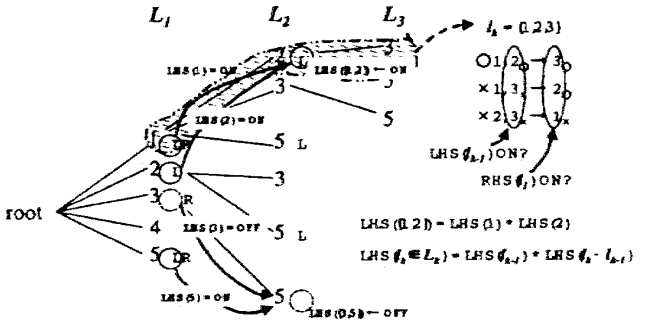

図1 正の相関ルール抽出における手法

これにより、長さKのアイテムセットl_kを作成後、サブセットl_{k-1} ,l_1により$l_{k-1} \rightarrow l_1$なるルールを作成する段階で、l_1のRHSがONであることの他にはl_kのLHSがONになっていることを確認するだけで足りることになる。

2．2 負の相関ルール抽出における手法

この場合、図2のように、○ノードと□ノードの2種類にノードを分類する。ここで○ノードはLHSがONであるノードであり、長さKのルールl_kの条件部になりうるサブアイテムセットl_{k-1}である。□ノードはl_kになりうるアイテムセットである。□ノードの生成規則は、まず、LHSがONであり、かつ最小支持度を満たすノードを○ノードとして残しておき、次にl_{k-1}の○ノードにRHS(l_1)=ONとなるl_1をつないでl_kにすることによる。

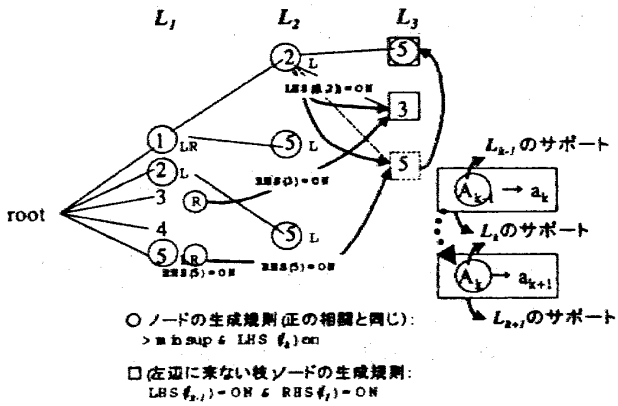

図2 負の相関ルール抽出における手法

Mining Algorithm for Extracting Rules with Constraints.
Takahiko YAMAZAKI, Akitoshi MITSUISHI, Yasushi OBATA
† Mitsubishi Electric Corporation

このようにノードを○と□で区別する理由は、○ノードは条件部になりうるので最小支持度による枝刈りが必要なのに対し、□ノードのみの場合はもはやかかる枝刈りの必要がないからである。

3 排反関係削除の手法

アイテムセット生成にあたっては、図3のように、一般に元データとなるＲＤＢから、２値データを要素とするレシート形式に変換する。この変換によって、アイテム「性別男」、「性別女」、「検査+」、「検査-」が異なるアイテムとして同一のレベルに並んでしまうが、実はもともとの属性が共通の属性、すなわち、例えば、「性別男」と「性別女」の組み合わせは排反であるので、これを含むアイテムセットは意味がない。そこで、アイテムセット生成時にかかる排反関係リスト中の組み合わせをサブアイテムセットに含むものは排除する。

かかる手法は、正の相関ルール抽出の場合は、もともと排反の組み合わせの支持度が０であるのでいずれにせよ、候補集合からアイテムセット生成時点で枝刈りされるが、負の相関ルール抽出の場合には支持度が低いアイテムセットがノードとして残るため、

男 －＼→女

のごとき周知のルールの出力を抑止するのに有効である。

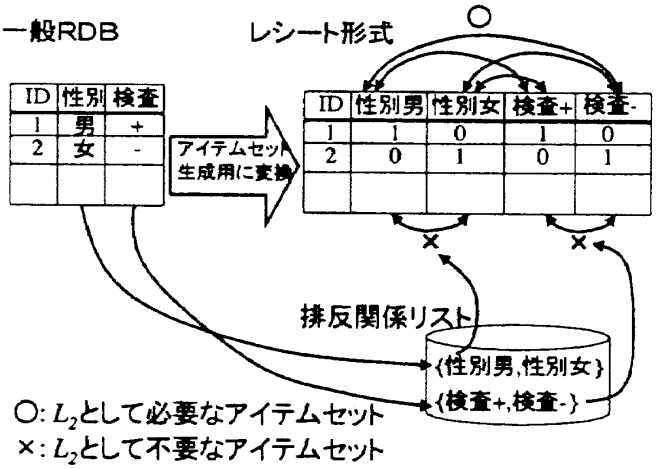

図3 排反関係出力抑止の手法

4 性能測定

4．1 測定方法

レコード数 5770 件、属性数 337 の健康診断 RDB を元データとして、上記出力抑止手法を使った場合とそうでない場合についてそれぞれ測定した。

パラメータとしては、ルールの最大長を３として、χ^2値を変化させた。

条件部結論部指定については、性別、年齢、問診などの属性について条件部指定に、検査結果などを結論部指定にして行なった。

4．2 結果

図４に負の相関ルール抽出についての結果を示す。出力抑止の効果が大幅に出ていることがわかる。出力ルールが多いと人手による検証が大変であるので、二つの出力手法を適切に使いながら、有効ルールの発見を効率化するべきである。

χ^2値	100	200	300	400
出力抑止なし	12715	2761	952	96
排反抑止	1767	160	24	6
排反抑止+条件部結論部出現指定	38	8	3	0

図4 出力ルール数 測定結果

なお、正の相関ルール抽出についても条件部結論部指定について同様な効果を測定できた。

5 おわりに

今後は出力ルール数の削減だけでなく、それに伴う使用メモリ量の削減、実行時間の短縮化向上のためにアルゴリズムを改良をしていくことが課題である。

参考文献

[1]Agrawal, R., Srikant, R. : "Fast Algorithm for Mining Association Rules" , Proc. VLDB '94.

[2]三石, 他: Knodiasにおけるデータマイニング方式., 第56回情処全国大会1998.

[3]Srikant, R., Agrawal, R. : "Mining Association Rules with Item Constraints" , Proc. KDD '97.