6P－1

# 帰納論理プログラミングにおける学習過程からの相関ルールの抽出とその利用

岩男 康平　大原 剛三　馬場口 登　北橋 忠宏
大阪大学 産業科学研究所

## 1 はじめに

近年，帰納論理プログラミング(Inductive Logic Programming: ILP)は一階述語論理に基づいた強力な表現力やルールの再利用性などの特徴から，データベースからの知識発見もしくはデータマイニングと呼ばれる分野における知識発見手法の一つとして注目されている[1]. ILPによるデータベースからの知識獲得では，未知事例をある概念に分類するための分類ルールの獲得が主な目的となるのに対し，一般的なデータマイニング手法はデータベース中に存在するデータ間の規則性を相関ルールとして抽出することを目的としている. しかしながら，両者は獲得対象が異なるものの属性の組合せを探索するという点では共通している.

このような背景から本研究では，ILPシステムにおいて本来の目的である仮説の獲得と同時に仮説に関連する相関ルールを抽出する手法を提案する.

## 2 提案手法の概要

### 2.1 システム概要

提案システムでは，ILPシステムの仮説探索過程から取得した候補仮説，及びその被覆計算結果を利用することで，仮説生成に利用した属性間の相関ルールを抽出する. 提案手法の概要を図1に示す. なお，本研究ではトップダウン型のILPシステムを対象とする.

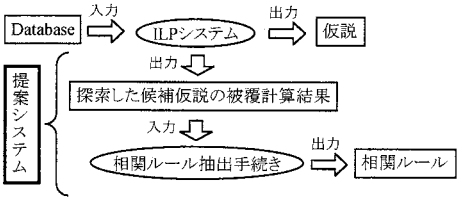


図1：提案手法の概要図

### 2.2 相関ルールの形式と評価方法

本研究における相関ルールの形式は，ILPシステムの出力である仮説と同様のホーン節形式とする. ホーン節形式のルール例を以下に示す.

$$A \leftarrow B_1, B_2, \cdots, B_m. \quad (1 \leqq m) \quad (1)$$

ここで，$A, B_1, B_2, \cdots, B_m$は仮説探索空間に現れるアトムとする. このルール(1)は，「アトムの連言$B_1 \wedge B_2 \wedge \cdots \wedge B_m$を満たす事例はアトム$A$も満たす」と解釈する. 以下，$A$を結論部，$B_1, B_2, \cdots, B_m$を条件部と呼ぶ.

相関ルールの評価基準としては，支持度(support)と確信度(confidence)を用いる. 支持度とはルールの条件部と結論部を同時に満たす事例の全事例に対する割合であり，ルール(1)の支持度を$support(A, B_1, B_2, \cdots, B_m)$により表す. 確信度とはルールの条件部を満たす事例が結論部も満たす割合であり，ルール(1)の確信度は$support(A, B_1, B_2, \cdots, B_m) / support(B_1, B_2, \cdots, B_m)$によって定義する. これらに閾値として最小支持度と最小確信度を設定しそれらを満たすルールのみを有益であると評価し，抽出対象とする.

## 3 仮説探索過程からの相関ルール抽出

### 3.1 候補仮説と相関ルールの関係

トップダウン型のILPシステムは，洗練化と被覆計算という2つの操作を繰り返しながら仮説空間を恒偽からより特殊な方向へと探索する. 仮説空間における最も特殊な仮説は最弱仮説(Most Specific Clause)と呼ばれ，仮説空間を構成する候補仮説の条件部は最弱仮説の条件部に表れるアトムの組合せからなる. 被覆計算とは候補仮説を評価するために，候補仮説が導く正例数$P$と負例数$N$を計算するものである. この$P$と$N$の合計値を全事例数で割った値は候補仮説の条件部の支持度に等しく，ILPシステムの目標概念を$L_0$とし候補仮説$L_0 \leftarrow A, B_1, B_2, \cdots, B_m$の満たす正・負例数を各々$P_1, N_1$，候補仮説$L_0 \leftarrow B_1, B_2, \cdots, B_m$の満たす正・負例数を各々$P_2, N_2$で表すと，2.2節のルール(1)の支持度と確信度は，次の計算式で求められる.

$$\text{支持度} \quad support(A, B_1, B_2, \cdots, B_m) = \frac{P_1 + N_1}{\text{全事例数}}$$

$$\text{確信度} \quad \frac{support(A, B_1, B_2, \cdots, B_m)}{support(B_1, B_2, \cdots, B_m)} = \frac{P_1 + N_1}{P_2 + N_2}$$

この計算式から，仮説探索過程で探索された候補仮説の条件部の支持度をすべて保持しておけば，後処理により相関ルールが抽出できるということが分かる.

---

Discovery of Association Rules in Concept Learning with Inductive Logic Programming
Kouhei IWAO, Kouzou OHARA, Noboru BABAGUCHI, Tadahiro KITAHASHI
I. S. I. R., Osaka University

### 3.2 候補仮説の識別

仮説探索において探索される候補仮説の数は膨大であり支持度と共に保持する必要のある条件部のアトム構成情報をコンパクトに表すことが重要となる．そこで，候補仮説の条件部が最弱仮説の条件部に現れるアトムの組合せで構成されることに着目し，最弱仮説の条件部の各アトムにビットを割り当て，候補仮説の条件部のアトム構成情報をビット列で表すことを考える．このビット列は各候補仮説に対して一意に定まることから，当該ビット列を10進数に変換した値を候補仮説固有の識別値として用いることで，識別値から候補仮説を一意に再構成することが可能となる．例えば，最弱仮説が$L_0 \leftarrow L_1, L_2, \cdots, L_8$である時に，候補仮説$L_0 \leftarrow L_1, L_3, L_8$の条件部のアトム構成情報はビット列10000101で表され，その識別値は133となる．

### 3.3 相関ルール抽出手続き

仮説探索過程から出力された候補仮説の条件部の識別値と支持度を用いて相関ルールを抽出する手続きを以下に示す．

1. 識別値をkey値としたハッシュ表に支持度を登録する．
2. 評価可能な相関ルールをすべて抽出するまで以下を繰り返す．
   2-1. 最弱仮説の条件部における各アトムを組合わせて相関ルールを構成する．
   2-2. ルールの支持度と確信度を識別値でハッシュ表を検索することによって求める．
   2-3. 最小支持度と最小確信度を満たせば抽出する．

尚，操作2-1における相関ルールの構成はデータマイニングにおける代表的な相関ルール抽出アルゴリズムであるApriori[2]に従うものとする．

## 4 評価実験

提案手法を筆者らが提案したILPシステムG-REX[3]上に実装し，実験対象としてUCIにて公開されているmushroomとvoteの2つのデータベースを用いて相関ルール抽出実験を行った．各データベースの詳細を表1に示す．mushroomに関しては正・負事例各2,000個をランダムに抽出し，その半分を学習セット，残りをテストセットとした．voteに関しては事例を正負とも半分に分け一方を学習セットとし，残りをテストセットとした．なお，実験はSUNのワークステーションUltra1(CPU：UltraSPARC 143MHz, Memory：192MB)上で行った．

評価実験では提案手法の有効性を調べるために提案手法とAprioriによる相関ルール抽出数とその獲得時間を比較した．なお，最小確信度は0.9に固定し，最小支持度の増減による変化を測定した．各々の相関ルールの抽出数と獲得時間を表2に示す．また，獲得した相関ルールの有用性を調べるため仮説と相関ルールによる欠損値を持つ未知事例の分類実験を行った．結果を図2に示す．各実験結果は10回試行の平均値である．

表1：実験対象となるDB

| 実験データ | 目標概念 | 正例数 | 負例数 | 属性 | 属性値 |
|---|---|---|---|---|---|
| mushroom | 毒キノコ | 3,916 | 4,208 | 22 | 2~12 |
| vote | 共和党員 | 168 | 267 | 16 | 2 |

表２：相関ルールの抽出数と獲得時間（秒）
（最小確信度 0.9）

| DB | 最小支持度 | 0.3 | | 0.25 | | 0.2 | |
|---|---|---|---|---|---|---|---|
| | | 数 | 時間 | 数 | 時間 | 数 | 時間 |
| mush room | 提案手法 | 438 | 345.3 | 546 | 354.7 | 783 | 359.7 |
| | Apriori | 552 | 256.4 | 832 | 483.0 | 2021 | 2581 |
| vote | 提案手法 | 413 | 44.9 | 1496 | 47.9 | 5770 | 52.9 |
| | Apriori | 857 | 11.8 | 4758 | 55.0 | 17637 | 224.3 |

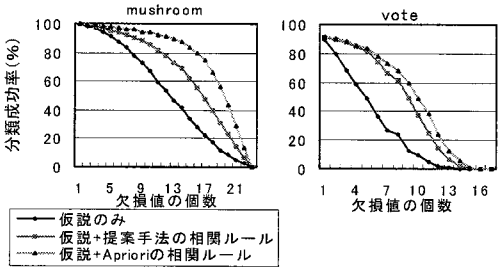


図２：欠損値を持つ未知事例の分類成功率
（最小支持度 0.2　最小確信度 0.9）

表2の結果において，Aprioriと提案手法は双方とも最小支持度の低下に伴いルール抽出数及び獲得時間を増加させている．指数的な伸びを示すAprioriに対し，探索空間がILPの仮説空間に限定されている提案手法は伸びが小さいという特徴を示している．また，提案手法のルール抽出数はAprioriよりも少ないが，図2の結果から相関ルールの欠損値の補完率にはルール数ほどの差がないと言え，抽出した相関ルールの仮説との関連性は提案手法の方が高いと考えられる．一方，ILPシステム自体にかかる実行時間はmushroomでは344.2秒，voteでは42.4秒であり，表2との比較から提案手法のILPシステムに対する負荷は小さいと言える．また，各実験結果は提案手法が事例数の異なる2つのデータベースに対しても同様に有効であることを示している．

## 5 まとめ

ILPの学習過程から相関ルールを抽出する手法を提案した．今後の課題として，有用なルールのより良い抽出法を検討するために欠損値の補完率が高い個々のルールの共通点を調べる予定である．

## 参考文献

[1] 特集：大規模データベースからの知識獲得，人工知能学会誌，Vol. 12No. 4, pp. 496-545(1997).
[2] Agrawal, R., Srikant, R. :Fast Algorithms for Mining Association Rules, Proc. of VLDB, pp. 487-499(1994).
[3] 高，大原，馬場口，北橋：例外関係に着目した不完全知識の獲得システムの実装と実験的評価，人工知能学会研究報告（1998).