平成 16 年度　卒業論文概要

入学年度 平成 13 年	**学生番号** 13117751	**氏名** 元山 純一
卒業研究題目 事例から抽出した性質に基づく関係的知識発見に関する研究		和田・犬塚 **研究室**

1. はじめに

データマイニングとは，大量のデータから隠された知識や新しい規則を発見するプロセスである．

データマイニングの手法として注目されているものの一つに，**帰納論理プログラミング**(Inductive Logic Programming : ILP) がある．これは述語論理を使用することで豊かな表現力を持ち，また論理的な記述を自然におこなえることによって可読性の高い解析がおこなえると考えられている．これらのことから，データマイニングの有力な手法として考えられている．

本研究では，一般的なデータマイニングの手法である**相関ルール**の考え方を使用して，ILP の枠組みでデータマイニングをおこなう手法を提案する．

これは対象となるデータの性質を取り出すことで，データ間の規則性を網羅的に発見することを可能にする手法である．

2. 相関ルール

相関ルールとは，マーケットバスケット分析を目的として提起された方法論である．

マーケットで売られている個々の商品を**アイテム**，一人の顧客が購買したアイテムのリストを**トランザクション**という．また，全アイテム集合を I とし，その部分集合をアイテムセットという．ここでトランザクションの集合 D に対して，その中にアイテムに関する規則性を見つけることが目的である．ここで規則性は

$$\emptyset \neq X,\ Y \subset I \ \text{ かつ } \ X \cap Y = \emptyset$$

を満たす X, Y について

$$X \Rightarrow Y$$

の形式で表わされ，相関ルールという．

すべてのトランザクションに対して，ルール中のすべてのアイテムを持つトランザクションの割合を**支持度**(support) と呼ぶ．また，すべてのトランザクションに対して，アイテムセット X を持つトランザクションのうちアイテムセット Y も持つトランザクションの割合を**確信度**(confidence) と呼ぶ．相関ルール導出は，これらの閾値を与えることで閾値以上のルールを網羅的に導出することをおこなっている．

3. 提案手法

本論文で目的としている問題は，ILP の枠組みでルールを取り出したい対象である**事例**E とその事例に関する知識である**背景知識**$\mathcal{B}$ を使用して，事例を特徴付けるようなルールを枚挙することである．

述語論理の上でおこなうため，入力には事例または背景知識の述語の形式も述語の集合 $\mathcal{P}$ として与えられる．また事例は，対象として正しいものである正事例 E^+ と間違っているものである負事例 E^- が与えられている場合を考える．

これは以下のように定式化される．

ILP の枠組みでのルール導出

Given

$\mathcal{P}$	: 述語の集合
$\mathcal{B}$	: 背景知識 (事例についての既知の知識)
$E^+ \cup E^-$	: 事例の集合 (特徴を取り出す対象 q の集合)

Enumerate

ルール　$R \Rightarrow q$: 所定の条件を満たす

R は背景知識 $\mathcal{B}$ から生成された，対象 q の特徴である．

背景知識は関連する述語について，関連する事実を表わす基礎 (変数，論理記号を含まない) アトムの集合として与える．また述語の各引数は，入力引数と出力引数として指定されているものとする．

このとき入力引数のみの述語は，事例に関するある具体的な事実を表わす述語であり，**判定述語**と呼ぶ．また，入力引数と出力引数をもつ述語は，事実と事実の関係性を表わす述語であり，**経路述語**と呼ぶ．

ここで $\mathcal{B} \cup \{$ 事例 $e\}$ の各アトムをノードとし，2 つのアトム a, b に対し a のある出力引数と b のある入力引数が同じ項をもつときに $a \rightarrow b$ を辺としたグラフを考える．このグラフには，各判定述語のアトム a_0 に対し，e から a_0 へのすべてのパス上の経路述語のアトムの集合 $\{a_1, \cdots, a_n\}$ がある．これについて，$a_0\ a_1, \cdots, a_n \Rightarrow e$ の各項を (異なる項は異なる変数に) 置き換えたものを**性質**とする．

さらに，相関ルール導出の枠組みを利用するために，事例に関する性質を取り出して一つのアイテム (**性質アイテム**) とし，事例をその性質の集合，すなわちトランザクション (**事例トランザクション**) と考える．これらを与えられた事例と背景知識を使用して生成することで，ILP の枠組みにおいて相関ルール導出と同様にしてルールを網羅的に導出する手法である．また，これは相関ルール導出と同じように支持度と確信度を使用して導出する．

この提案手法は以下のように定式化される．

性質の抽出に基づくルール導出

Given

$I = \{i_1, i_2, \cdots, i_k\}$	: 性質アイテムの集合
D	: 事例トランザクションの集合
minsup	: 支持度の閾値
minconf	: 確信度の閾値

Enumerate

s.t. $R \Rightarrow e$, ただし，$R = i_1 \wedge i_2 \wedge \cdots \wedge i_r$

$R \Rightarrow e$ の支持度が minsup 以上

$R \Rightarrow e$ の確信度が minconf 以上

また事例の性質を取り出すときは，正事例のみから性質を取り出すことで，事例の特徴として必要のない性質を取り出さないようにする．

4. アルゴリズムの概要

アルゴリズムの流れを簡略化したものを以下に示す．

(1) 正事例の集合からいくつか事例を取り出し，背景知識を使ってその事例についてのすべての性質を取り出す．

(2) (1) で取り出した性質の集合から性質を一つずつ取り出し，それぞれを一つのアイテムとする．

(3) 事例の性質と相関ルール導出の枠組みを使って事例の特徴をルールとして取り出す．

5. 実験

実験として,「貨物列車の分類問題 (East-West Challenge)」をデータとして使用し，性質を取り出すために利用する事例の個数と支持度，確信度の閾値を変化させながら，以下の三つを記録した．

- 支持度，確信度の計算をおこなうアイテムセットの数
- 閾値以上の支持度をもつアイテムセットの数
- 閾値以上の確信度をもつアイテムセットの数

考察として，以下のことがあげられる．

- 利用する事例の数を増やすことで導出されるルールは増加する．
- 性質の抽出に利用する事例は，負事例からよりも正事例から取り出したほうが導出できるルールは多い．

6. まとめ

事例の性質に注目することで，強力なアプローチである ILP の枠組みでデータマイニングの一般的な手法である相関ルールの導出を用いて，述語論理で表わされた可読性の高いルールの導出をおこなった．

今後の課題として，より複雑なルールの導出と実データへ適用した検証が必要である．