関係データベースシステムを結合した関係型データマイニング法実装の改善

犬塚研究室　ネットワーク系
No. 17115128　日比野 仁志

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．関係型データマイニングをおこなう手法 SQL_MAPIX[1] は，データベース内の事例からそれが持つ性質を抽出してデータマイニングに用いる手法である．

現在までの SQL_MAPIX は，パターン導出のために背景知識テーブルを結合する操作が必要であり，この操作のメモリ効率がデータ量の限界を決めている．そこで，本研究では，背景知識が持つバイアス情報を利用して背景知識のテーブルを圧縮しテーブル数を減らすことで，マイニングをおこなうことができるように拡張したアルゴリズムを示す．

2　SQL_MAPIX

例えば，ある家族関係のデータベースにおいて事例 grandfather(koji) に関して，次の式が成り立つとする．

$$\mathbf{grandfather}(\mathbf{koji}) \leftarrow \mathbf{parent}(\mathbf{koji}, \mathbf{yozo}) \wedge \mathbf{parent}(\mathbf{yozo}, \mathbf{kyoichi}) \wedge \mathbf{male}(\mathbf{kyoichi}).$$

これは「koji が kyoichi という孫息子をもつ」という事実である．SQL_MAPIX の目的は，このようなパターンの頻度を調べ，頻出なものを枚挙していくことである．SQL_MAPIX では事例のテーブルと背景知識のテーブルを JOIN することによって，このようなパターンを求めている．

SQL_MAPIX の概要は以下のようになる．

1. 選択した事例を説明し得るリテラルを全て結合したものである関連リテラル集合を生成する
2. 性質を抽出し，性質を変数化してパターン化した性質アイテムを生成する
3. 閾値以上の頻度を持つアイテムセットを生成する

1,3 はデータベース側，2 は Prolog 側でおこなっている．

3　背景知識の圧縮

このアルゴリズムでは，背景知識のテーブルを圧縮することにより，関連リテラル集合を SQL_MAPIX と同じ形式で生成する．このアルゴリズムの概要を以下に示す．

1. 述語のバイアスに基づき, 引数の個数, モード, 型が同じ述語の関係表を一つの表にまとめる
2. まとめた表に対応する関係表の情報を拡張バイアスとして生成する
3. SQL_MAPIX で拡張バイアスを参照して関連リテラル集合を生成する

次のような背景知識のバイアスとテーブルがあるとする．

bias(female,[+],[human])，bias(male,[+],[human])

female
haruko
tatsumi

male
ichiro
jiro

まず，バイアスの第 2,3 引数が一致するテーブルを UNION する．このとき，元のテーブル名を表す列を加える．そして，このテーブルに対応する e_bias を作成する．

predicate	value
female	haruko
female	natsumi
male	ichiro
male	jiro

e_bias(female&male,[n,+],[n,human])

この e_bias と UNION したテーブルを使用して SQL_MAPIX で関連リテラル集合を生成すれば，同様の結果を得ることができる．

4　実験結果

英文の構造を表形式で表現したデータを使用して，事例数に対する性質アイテムの平均数を求めた．このデータには背景知識のテーブルが 53 個あったが，今回の提案アルゴリズムを適用した結果，3 個に圧縮することができた．図 1 により，事例数の増加に伴って生成された性質アイテム数も増加する．これにより，正常に性質アイテムが生成されていることが確認できた．

事例数	1	10	100	1000
性質アイテム数	16.8	139.4	856.3	4396.8

図 1: 事例数に対する性質アイテム数

5　まとめと今後の課題

このアルゴリズムによって，関連リテラル集合を生成し，性質アイテムを生成できたことが確認された．今後の課題としては，現段階のアルゴリズムでは関連リテラル集合のテーブルの列数が制限されているため，この制限を取り除く必要がある．

参考文献

[1] 牧野 敏行，犬塚 信博：“関係データベースシステムを結合した論理データマイニングの実装”，WiNF 2008, pp.123-128，2008.