関係データマイニングにおける頻出飽和パターン発見の分散化

谷本　翔一 † 世木　博久 †
† 名古屋工業大学 情報工学科

1　はじめに

関係データマイニング (multi-relational data mining: MRDM)[1] においては飽和パターンのマイニングアルゴリズムはいくつか提案されているが ([2], [5] など), その分散化についてはあまり研究されていない. しかし, 効率化やプライバシーの観点から, 各所に蓄積された分散データベースを対象としてマイニングを行うことは重要である.

アイテム集合を対象にした分散データベースからの飽和パターンの計算方法として Lucchese らの方法 [3] がある. 本研究では, それを MRDM に適用した手法を提案し, その正当性を示す. また具体的なデータベースを対象にして提案手法が正しく飽和パターンをマイニングすることを確認する.

2　準備: 関係データマイニングと飽和パターン

従来のアイテム集合に対するデータマイニングが 1 つのトランザクション・データベースを扱うのに対し, 関係データマイニング (MRDM) では, 複数の関係表で構成される関係データベースから, 述語論理式で表現されたパターン (連言) を発見する. 関係データベースを用いることで, 対象の持つ様々な属性や対象間の関係を表現できる高い記述能力を持つ.

例 1 図 *1* のデータベース DB_{ex} は *5* つの関係表 *customer, buys, parent, male, female (*以降, 関係名をその先頭文字で略記する*)* を持ち, それぞれは顧客関係, 顧客の購買関係, 親子関係, 性別情報を表す. マイニングされるパターンは連言形式であり, 例えば連言 C = *c(X),b(X,Y),p(X,Z),m(Z)* である. ここで, *c(X)* は関係 *customer* に対応する述語で「X が顧客である」ことを表す. 連言 C は「顧客 X は商品 *(*アイテム*)* Y を買い, X は子供 Z を持ち, Z は男性である」ということを表す.

MRDM においては, 通常ユーザが注目したいパターンを指定するために注目述語 (キー) の概念を導入する. 例えば例 1 の場合で顧客にユーザの関心がある場合には述語 customer をキーとする.

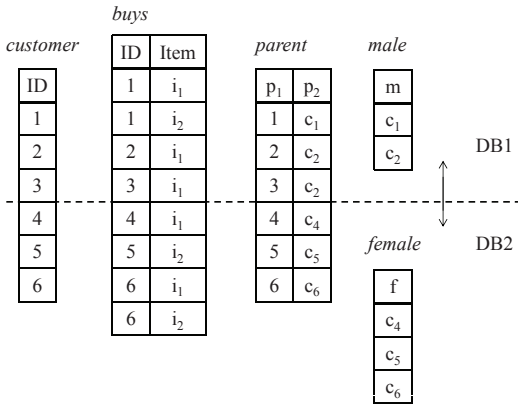

図 1: MRDM の対象とするデータベースの例 DB_{ex}

キーの概念を用いて連言 C がデータベース (DB) に出現する頻度 (サポート) を次のように定義する. 連言 C に含まれる変数の集合を Var(C), 変数への代入を θ と書く. $C\theta$ が DB に 出現する とは, $DB \models C\theta$ が成り立つことを言う. 代入 θ のキー変数 (キーに出現する変数) に対する制限を θ_{key} と書く. この時, 連言 C の**サポート** σ は $\sigma = |\{\theta_{key} | DB \models C\theta\}|$ で定義され, σ がユーザの設定した最小サポート min_sup 以上である場合, C は頻出であると言う.

例 2 例 *1* の DB_{ex} で, 連言 C の *3* つの変数 *X,Y,Z* への代入 θ は, $\{(1, i_1, c_1), (1, i_2, c_1), (2, i_1, c_2), (3, i_1, c_2)\}$ の *4* つである. 従って θ_{key} を考えると連言 C のサポートは *3* である.

MRDM では通常, マイニングの対象とする連言に出現する各変数はキー変数と”関連している”という条件を課す. キーを $key(X)$ (X はキーに現れる変数), l をリテラルとする時, l が $key(X)$ に 関連している とは ($key(X) \sim l$ と書く), (i) $X \in Var(l)$, あるいは (ii) ある l_1 が存在し, $key(X) \sim l_1$ かつ $Var(l_1) \cap Var(l) \neq \emptyset$ を満たすことを言う. 連言に現れるリテラルがすべてキーに関連していることを**キー関連**であると言う.

データマイニングでは頻出パターンをすべて求める代わりに, 代表となるようなパターンだけを効率的に求める方法が研究されている. 関数 f を「変数への代入の集合 S を受け取り, S を満たす最大の連言を返す」関数と定義し, 関数 g を「連言 C を受け取り, C が満たす変数への代入の集合を返す」関数と定義する. こ

Distributed Mining of Frequent Closed Patterns in Multi-Relational Data

† Shoich Tanimoto (cht15100@stn.nitech.ac.jp)
† Hirohisa Seki (seki@nitech.ac.jp)
Dept. of Computer Science, Nagoya Inst. of Technology (†)
Showa-ku, Nagoya, 466-8555 Japan

のとき $f \circ g(C) = C$ となる連言を**飽和パターン (飽和連言)** と呼び, これを代表のパターンとして扱う.

3 分散データベースからのマイニング

飽和パターンの分散マイニングの目的は, 部分データベースに対して独立してマイニングを行い, その結果を基に部分データベースによって構成される全体データベースから得られる結果を導出して, それが全体データベースを直接にマイニングした結果と同一となることを示すことである. 本研究ではキーアトムを含む飽和パターンを計算するために必要となるデータはその部分データベースに存在すると仮定する. 例えば, 図1の DB_{ex} を全体データベースとした場合, DB1 と DB2 の2つに分割した部分データベースはこの条件を満たしていることが分かる.

以下では簡単のため, マイニングしたい全体データベースを2分割とし, 最小サポート $min\_sup = 1$ の場合について述べる. 全体データベース DB とし, その飽和連言の集合を $\mathcal{C}$ とする. また, DB の2分割である部分データベースを DB_1, DB_2 とする. 部分データベースの飽和連言の集合を $\mathcal{C}_1$, $\mathcal{C}_2$ とする. この時, $\mathcal{C}$ は $\mathcal{C}_1$ と $\mathcal{C}_2$ から次のようなマージ関数 $\oplus$ を用いて構成される.

命題 1 *$DB = DB_1 \cup DB_2$, $\mathcal{C}$, $\mathcal{C}_i$ を DB, DB_i の飽和連言の集合とする $(i = 1, 2)$. この時, 以下が成り立つ.*

$$
\begin{aligned}
\mathcal{C} &= \mathcal{C}_1 \oplus \mathcal{C}_2 \\
&= (\mathcal{C}_1 \cup \mathcal{C}_2) \cup \\
&\quad \{C_1 \cap C_2 \mid (C_1, C_2) \in (\mathcal{C}_1 \times \mathcal{C}_2), \\
&\quad Var(C_1) = Var(C_2), \\
&\quad C_1 \cap C_2 \text{はキーと関連している.}\}
\end{aligned}
$$

ここで $C_1 \cap C_2$ は, C_1 と C_2 に共通に現れるすべてのリテラルを l_i $(i = 1, \ldots, k, k \geq 1)$ とした時, その連言 $\wedge_{i=1}^{k} l_i$ を表す. マージ関数 $\oplus$ は $\mathcal{C}_1$ と $\mathcal{C}_2$ の和集合演算 $(\mathcal{C}_1 \cup \mathcal{C}_2)$ と, $\mathcal{C}_1$ と $\mathcal{C}_2$ に現れるすべての連言 C_1 と C_2 の共通部分演算 $(C_1 \cap C_2)$ の2つからなる. その際, $C_1 \cap C_2$ がキーと関連している条件を考慮する点が MRDM では必要になる.

命題1は一般の場合, つまり部分データベースの数が $N(\geq 2)$ の場合にも同様に成り立つ.

4 実験

突然変異データベース[†]を用い, $min\_sup = 1$, 連言 C に含まれる変数の数 $|Var(C)| \leq 4$ とし, 飽和連言を求めるアルゴリズムとして ffLCM [5] を利用し, データベースの分割数 N を 2,4,8 の3通りの方法で実験を行った.

表 1: データベースの分割数 N に対する飽和連言の数: n_2 には和集合演算から得られる飽和連言は含まない.

	N=1	N=2	N=4	N=8
和集合で得られる飽和連言数 n_1	2346	2346	2291	2223
共通部分演算で得られる飽和連言数 n_2	-	0	55	123
$n_1 + n_2$	2346	2346	2346	2346

実験では, 共通部分演算で得られる飽和連言数の, 全体データベースから得られる飽和連言数に対する割合は $N = 8$ の時 5% であった. これは, 部分データベースからのマイニング結果の単純な和集合だけでは正確な解が得られず, マージ演算が必要であることを示している.

5 まとめ

本研究では, Lucchese らのアイテム集合を対象にした分散データベースの飽和パターン計算方法を MRDM に適用する手法を考察した. そして, パターンにキーの概念を導入する場合にはキー関連であるという条件を考慮に入れることで, アイテム集合の場合と同様に飽和パターンが計算できることを示した. この論文では最小サポート $min\_sup = 1$ の場合を扱ったが, $min\_sup \geq 2$ の場合も, [3] と同様に通常のアイテム集合マイニングで使われる Partition [4] の方法を用いてマイニングすることが考えられる. 今後の課題として, 本提案手法の効率的な実装が挙げられる.

参考文献

[1] Dzeroski, S., Lavrač, N., editors. Relational Data Mining. Springer-Verlag, Inc. 2001.

[2] Garriga,G. C., Khardon, R., and De Raedt, L.: On Mining Closed Sets in Multi-Relational Data. IJCAI 2007, pp.804-809 (2007)

[3] Lucchese, C., and Orlando, S.: Distributed Mining of Frequent Closed Itemsets: Some Preliminary Results. International Workshop on High Performance and Distributed Mining (2005).

[4] Savasere, A., Omiecinski, E., and Navathe, B. S.: An Efficient Algorithm for Mining Association Rules in Large Databases. In VLDB'95, pp. 432-444 (1995)

[5] Seki, H., Honda, Y., and Nagano, S.: On Enumerating Frequent Closed Patterns with Key in Mutirelational Data. LNAI 6332, pp. 72-86 (2010)

[†] http://www.comlab.ox.ac.uk/activities/machinelearning/mutagenesis.html