信頼区間を用いた一対多関係の推定

菊地 真人 † 梅村 恭司 † 山本 英子 ‡ 吉田 光男 † 岡部 正幸 †
† 豊橋技術科学大学 ‡ 岐阜聖徳学園大学
† {kikuchi14@ss.cs, umemura@ss.cs, yoshida@cs, okabe@imu}.tut.ac.jp
‡ eiko@gifu.shotoku.ac.jp

1 はじめに

本稿では，データ集合から一対多関係を推定する問題を扱う．ここで述べる一対多関係とは，例えば，実世界における地名間で成り立つ関係である．一対多関係の「一」をある都道府県とすると,「多」はそこに属する複数の市郡となる．これまでにもデータ集合から一対多関係を推定する問題は研究されている [1, 2].

一対多関係は相関ルールで表現することもできる．データベースから相関ルールを発見する代表的な手法として，アプリオリアルゴリズムがある [3]．この手法では，支持率と信頼度という尺度が用いられる．支持率はデータベース全体に対してあるアイテム集合が含まれる割合である．信頼度は支持率によって表現できる条件付き確率であり，相関ルールの“面白さ”の推定値となる．アプリオリアルゴリズムは，しきい値として最適な最小支持率を設けることで相関ルールを精度よく推定できる．

一般的に，ユーザが最適な最小支持率の値を調整することになるが，データ集合によってその値は異なる．最小支持率を用いると，その値よりも小さい支持率を持つ相関ルールは推定の対象としないため，結果として，発見されるべき面白い相関ルールを推定できないという問題がある．そのため，一対多関係の強さの推定値として最小支持率を用いる相関ルールの信頼度を採用すると，支持率の小さい発見されるべき一対多関係を推定できないことが問題となる．

そこで，本稿では，最小支持率を調整する代わりに信頼度を区間推定し，その下限値を一対多関係の強さの推定値とすることを提案する．この信頼区間は事前分布が一様分布であると仮定して求め，事後分布の期待値の区間とした．

提案手法における一対多関係の推定性能を評価するために，地名（都道府県市郡）を対象とした実験において，最小支持率を調節する相関ルールの信頼度と一対多関係の推定性能を比較する．実験では，実世界の地名間にある一対多関係から，ある都道府県名とそこに属する市郡名からなる二種類の組を取り出して，それらの組に現れる地名を一つの要素とするデータ集合を用いる．このデータ集合を用いて，各要素から都道府県名とそこに属する市郡名からなる二種類の組を正しく分類する性能を測定する．この実験により，提案手法は，最小支持率を調節する相関ルールの信頼度よりも良い結果を示すことを報告する．

2 問題定義

あるデータ集合 $D = \{t_1, t_2, \cdots, t_n\}$ に存在するアイテム集合を $I = \{i_1, i_2, \cdots, i_m\}$ とする．データ集合を構成する各要素 $t_k (t_k \subseteq I)$ をトランザクションと呼び，長さ m のアイテム集合とは m 個のアイテムの組み合わせによって構成されたアイテム集合のことである．ここで，各アイテムを実世界に存在する地名（都道府県市郡名）と考えるとアイテムの集合 I の例は次のようになる．

$$I = \{\text{東京都}, \text{大阪府}, \text{神戸市}, \text{北海道}, \text{江別市}\}$$

一対多関係 $\langle x, y\rangle$ は一つのアイテムからなるアイテム集合 x と y の間で定義される関係である．一対多関係 $\langle x, y\rangle$ は次の二つの定義を満たす．

$$\forall a \subset I; \forall b \subset I; \forall c \subset I; \langle a, c\rangle \wedge \langle b, c\rangle \quad \rightarrow \quad a = b \ (1)$$

$$\exists a \subset I; \exists b \subset I; \exists c \subset I; \langle a, b\rangle \wedge \langle a, c\rangle \quad \rightarrow \quad b \neq c \ (2)$$

式 (1) は，関係の右のアイテム集合が等しい場合は左のアイテム集合も等しいという定義である．式 (2) は，この関係には一対一ではないアイテム集合が存在するという定義である．これらの定義を満たす関係の集合を R と定義し，正解集合と呼ぶ．一つのアイテムからなるアイテム集合をそれぞれ $S_c = \{i_c\}$，$S_p = \{i_p\}$ とすると，正解集合 R は次のように定義される．

$$R = \{\langle S_c, S_p\rangle \mid S_c, S_p \subset I\}$$

一対多関係の「一」を都道府県,「多」を市郡とするとき，正解集合 R の例は次のようになる．

$$R = \{\langle\{\text{北海道}\},\{\text{札幌市}\}\rangle, \langle\{\text{北海道}\},\{\text{釧路市}\}\rangle\}$$

前節で述べたように，一対多関係は相関ルールを用いて表すことができる．相関ルールは，$X \Rightarrow Y$ と表され，矢印の左側にある X を条件部，右側にある Y を帰結部と呼ぶ．これは，データ集合のあるトランザクションに X というアイテム集合が含まれていれば，Y というアイテム集合も含まれている可能性が高いということを意味する．なお，$X \cap Y = \emptyset$ である．関係 $\langle S_c, S_p \rangle$ は相関ルールを用いて $S_c \Rightarrow S_p$ と表すことができる．

本稿では，次の手順で正解集合 R を推定し，正解判定を行う．まず，各トランザクションに含まれる二つのアイテムからなるすべての組を相関ルールとして求め，頻度情報をもとに各相関ルールの強さの推定値を計算する．そして，その値が高いほどアイテム間の関係性が高いと推定する．最後に，各トランザクションに含まれる二つのアイテムからなる組が持つ関係 $\langle S_c, S_p \rangle$ が R に含まれるかどうかで一対多関係の正解判定を行う．

3 類似尺度

本稿では，類似尺度を用いて一対多関係の強さを推定する．この類似尺度は，一対多関係となるアイテムがトランザクションに同時に現れる頻度をもとに計算される．前述したように，一対多関係は相関ルールで表現できる．そこで，データベースから相関ルールを発見する手法であるアプリオリアルゴリズムにおいて，相関ルールの“面白さ”の推定値として用いられる信頼度を比較対象とすることにした．提案手法では信頼度を区間推定し，その下限値を類似尺度として用いる．以降で，それぞれの類似尺度について示す．

3.1 信頼度

データベースから相関ルールを発見する代表的な手法として，アプリオリアルゴリズムが知られている [3]．アプリオリアルゴリズムでは，支持率（Support）を用いて信頼度（Confidence）を計算し，相関ルールの“面白さ”を評価する．信頼度は式 (3) のように定義され，アイテム集合 X を含むトランザクションに対するアイテム集合 Y を含むトランザクションの割合を表し，相関ルールの面白さの推定値となる．言い換えれば，信頼度はトランザクションにアイテム集合 X が含まれるときにアイテム集合 Y が含まれる条件付き確率 $p(Y|X)$ の最尤推定値である．

$$Confidence(X, Y) = \frac{Support(X \cup Y)}{Support(X)} \tag{3}$$

支持率 $Support(X)$ はデータベースの全トランザクションに対するアイテム集合 X を含むトランザクションの割合，支持率 $Support(X \cup Y)$ は全トランザクションに対するアイテム集合 X とアイテム集合 Y をともに含むトランザクションの割合を表す．

アプリオリアルゴリズムは式 (4) を満たす相関ルールの信頼度を計算する．

$$Support(X \cup Y) \geq minsup \tag{4}$$

最小支持率（minsup）をしきい値として設け，それ以上の支持率 $Support(X \cup Y)$ を持つ相関ルールの信頼度を求める．一般的に，この最小支持率はユーザが指定し，データ集合によって異なる．最小支持率を設けることによって，統計的に不安定な相関ルールを無視して信頼度を計算できるが，支持率の小さい発見されるべき面白い相関ルールを推定することができなくなるという問題がある．

3.2 提案手法

信頼度は条件付き確率 $p(Y|X)$ の最尤推定値と解釈できる．一般的には $p(Y|X)$ を推定するために最尤推定値や期待値が用いられる．本稿では，間違った一対多関係をできる限り拾いたくないと考え，適合率を保証して一対多関係を推定する．そこで，$p(Y|X)$ について信頼区間を求め，その下限値を一対多関係の強さの推定値とすることを提案する．

一対多関係の判定対象となる関係における $p(Y|X)$ の真の値を θ とする．θ の事前情報がないため，事前分布 $\pi(\theta)$ は一様分布と仮定する．θ の事後分布を求め，θ の期待値 $\overline{\theta}$ に対する片側 $100(1-\alpha)\%$ 信頼区間を求める．片側 $100(1-\alpha)\%$ は信頼係数と呼ばれ，信頼区間を求める際のパラメータとなる．本稿では，この信頼区間の下限値を一対多関係の強さの推定値とする．信頼区間の計算については，近似公式の誤差が影響すると考え，直接に数値積分で求める．

信頼度と支持率はトレードオフであり，信頼度が同じでも支持率の値によって相関ルールの価値は異なるという考え方 [4] を本研究は直接的に実現する．支持率が小さければ相関ルールの価値も相対的に低くなることが望ましい．信頼区間の下限値を一対多関係の強さの推定値とすることによって，低頻度で高い信頼度を持つ一対多関係の強さは弱くなる．

Algorithm 1 データ集合の生成アルゴリズム

```
D* := φ; k := 0;
while k < 1000 do
  j := 0; t_k := φ;
  while j < 2 do
    R からランダムに ⟨S_c, S_p⟩ を取り出す;
    t_k := t_k ∪ S_c ∪ S_p;
    j := j + 1
  end while
  D* := D* ∪ {t_k};
  k := k + 1
end while
```

表 1: 人工のデータ集合に関する情報

トランザクション数	1000
候補となる組の種類	4469
候補となる組の出現数	5934
正解集合に含まれる組の種類	975
正解集合に含まれる組の出現数	2000

4 実験準備

実験では，文献 [1, 2] と同様に地名（都道府県市郡）を一対多関係の推定対象として用いる．地名を用いる理由としては，実世界において地名間には一対多関係が成り立っていること，地名間の一対多関係は定まっているため，正解判定が容易であることが挙げられる．実験では，地名間にある一対多関係から二種類の組を取り出して，それらの組に現れる地名からなるトランザクションを要素とするデータ集合を用いる．このデータ集合を用いた判定結果をランクと再現率のグラフをもとに評価する．

4.1 実験に用いるデータ集合

実世界のデータ集合では，アイテム集合に対する出現頻度の偏りが観測される．この偏りを考慮せずに，類似尺度における一対多関係の推定性能を評価するため，一対多関係の推定対象として人工的に生成したデータ集合を用いて実験した．用いたデータ集合は文献 [1] と同様の方法で正解集合 R から生成した．この生成アルゴリズムをアルゴリズム 1 に示す．

生成されたデータ集合に含まれるトランザクションは，例えば，$t_k = \{$ 北海道, 札幌市, 愛知県, 名古屋市 $\}$ などが考えられる．これは正解集合 R から $\langle\{$ 北海道 $\}, \{$ 札幌市 $\}\rangle$ と $\langle\{$ 愛知県 $\}, \{$ 名古屋市 $\}\rangle$ という二種類の組の関係を抽出して，これらに現れる 4 つのアイテムからなるトランザクションである．データ集合のトランザクション数を無限に増やすことはできないため，このようなトランザクションを 1000 個持つデータ集合を生成した．

実験では，トランザクションごとに含まれる二つの地名からなる組をすべて求め，類似尺度の値を計算する．そして，その値によって組を降順に並べ，ランク付けをする．類似尺度の値が高いほど組の関係性が強いと判断する．生成したデータ集合を用いて，トランザクションごとに組み合わされた二種類の組の関係を類似尺度の値によって正しく分離することを試みる．人工のデータ集合に関する情報を表 1 に示す．正解集合 R に含まれる全正解数は 1215 であるが，データ集合に含まれない正解があることによって，実態に合った評価ができると考える．

4.2 評価方法

実験では，ランクが上位 4000 件の関係について正誤を確認し，横軸をランク，縦軸を再現率として上位からランク—再現率曲線を描く．再現率の定義を次に示す．

$$\text{再現率} = \frac{\text{あるランクまでの正解数}}{\text{データ集合に含まれる正解数}}$$

ランク—再現率曲線を用いて，類似尺度の値の上位に着目し，類似尺度の一対多関係を推定する性能を評価する．

4.3 類似尺度のパラメータ設定

提案手法は信頼係数，信頼度は最小支持率というパラメータを持つ．そのため，実験を行うにあたってパラメータを設定する必要がある．実験では，信頼係数を変化させて上位での適合率が高く，下位の再現率を保持できるような値を探した．その結果，片側 99% ($\alpha = 0.01$) が最適な信頼係数となった．信頼度については，実験で最小支持率を変化させてその振る舞いを観察する．

5 実験結果

類似尺度の値が上位 4000 位までのランク—再現率曲線を図 1 に示す．グラフ上の点と原点を結んだ線の傾きが適合率に比例する．$|D|$ はデータ集合に含まれるトランザクション数である．最小支持率が $1/|D|$ のときは，データ集合に 1 回以上含まれる組を関係推定の対象とするため，最小支持率を設けないことを意味する．

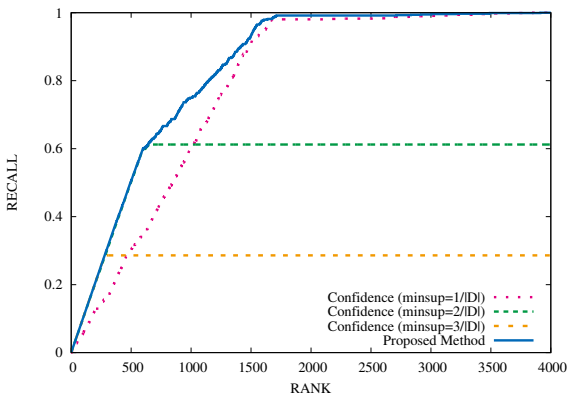

図 1: 人工のデータ集合におけるランク—再現率曲線

最小支持率を $1/|D|$ として，データ集合に含まれるすべての組について信頼度を計算すると，頻度が低いにもかかわらず，信頼度が高い組（不正解となる組）が推定値のランク上位となる傾向がある．この傾向は上位の適合率を低下させる原因になる．そこで，最小支持率を $2/|D|$ とすると，出現頻度が 2 回以上の組が上位となり，そのような不正解の組が取り除かれるため，上位における一対多関係の適合率は向上したと考える．しかしながら，データ集合において出現頻度が 1 となる組は一律に推定対象から取り除かれ，このときに正解の組も共に取り除かれてしまう．このため，下位の再現率が低下したと考える．最小支持率を $2/|D|$ から $3/|D|$ に上げると上位の適合率は向上せず，下位の再現率がさらに低下した．これは最小支持率を上げすぎたため，多くの正解と判断される組が取り除かれてしまうことが原因と考える．このことから，最小支持率はこれ以上変化させても意味がなく，$2/|D|$ を最適な最小支持率とすることが妥当と考える．

提案手法は，ランク上位では，最適な最小支持率を設けた信頼度とほぼ等しい適合率となる．このことから，提案手法は最適な最小支持率の信頼度と同様に，不正解の組を取り除いていると考える．下位では，最適な最小支持率を設けた信頼度よりも高い再現率となる．これは頻度の低い組を正しく扱っていると考える．

以上のことから，上位での適合率が高く，下位の再現率を保持できる信頼係数を設けた提案手法は，同様に最小支持率を設けた信頼度よりも上位における一対多関係の推定性能が高くなる可能性を示唆した．ここでいう推定性能はランク上位の組のうち，正解集合に含まれる組をどれだけ推定できたかという指標である．すなわち，それぞれの持つパラメータである最小支持率の最適値，信頼係数の最適値を比較すると提案手法の方が優れている．一般に頻度の低いデータは無視されることが多いが，提案手法はそれを利用する方法の一つとなっている．

6 おわりに

本稿では，データ集合から一対多関係を推定する問題を扱い，最小支持率を調節する信頼度に変わる手法を提案した．提案手法では，最小支持率を調節する代わりに信頼度を区間推定し，その下限値を一対多関係の強さの推定値とした．この信頼区間は事前分布が一様分布であると仮定して求め，事後分布の期待値の区間とした．

提案手法の一対多関係を推定する性能を評価するために，地名（都道府県市郡）を対象とした実験において，最小支持率を調整する信頼度と性能比較を行った．実験により，提案手法は最適な信頼係数を設定すれば，最適な最小支持率を設けた信頼度よりも一対多関係の推定性能が高くなる可能性を示唆した．

今後は，提案手法のパラメータである信頼係数を自動調整する手法の検討，提案手法の高速化に取り組む．

謝辞

本研究は，平成 27 年度岐阜聖徳学園大学研究助成金を受けた．

参考文献

[1] 山本 英子, 梅村 恭司. コーパス中の一対多関係を推定する問題における類似尺度. 自然言語処理, Vol. 9, No. 2, pp. 45–75, 2002.

[2] 岡部 正幸, 梅村 恭司. 頻度差が著しい場合における一対多関係を推定する類似尺度. 情報学シンポジウム講演論文集, Vol. 2005, pp. 129–136, 2005.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In *Proceedings of the International Conference on Very Large Databases*, pp. 487–499, 1994.

[4] Tobias Scheffer. Finding association rules that trade support optimally against confidence. In *Principles of Data Mining and Knowledge Discovery*, pp. 424–435. 2001.