D-4-6 大規模データベースにおける相関ルール抽出のための属性値離散化法

Discrete Indication of Continuous Value of Attributes for Generation Methods of Association Rules on Very Large Database

大島 敬志　熊谷 貴禎　小林 学　平澤 茂一
Keishi Ohshima　Kiyoshi Kumagai　Manabu Kobayashi　Shigeichi Hirasawa
早稲田大学理工学部経営システム工学科
Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University

1 はじめに 大量データ中に埋もれた知識を発掘するデータマイニング手法のひとつに相関ルール抽出技法がある．その中でもよく知られたアルゴリズムに，2値属性を扱うアプリオリ[1]があり，これを多値属性にも適用できるように拡張したアプリオリも提案されている[2]．一方，連続値属性を含んだデータを扱う場合，連続値を離散化し他値属性に変換してアプリオリを適用する研究がある．本研究ではクラス情報を必要とせずに，複数アイテム間の共起情報を考慮しつつ，各連続値属性を離散化することを考える．そのために連続値属性のクラスタリングを行い，離散化する手法を提案する．また提案した手法をセンサスデータに適用し，その有効性を示す．

2 相関ルール抽出の連続値属性への適用と問題点 本研究では相関ルールの抽出アルゴリズムとして，多値属性を対象としたアプリオリ[2]を用いる．このアプリオリは多値属性のみのアイテム間の共起情報を解析する手法であるため，連続値属性を含むデータに直接適用することは困難である．そこで連続値属性を離散化することでアプリオリを適用しようという研究が行われている．従来では連続値属性は等区間に離散化したり，情報エントロピーを用いて離散化している[4]．しかし殆どは他の属性の影響を考慮せずに各属性を独立に離散化を行っている．また鷲尾らは連続値属性を全て独立した正規分布と仮定し，情報量基準を用いることで複数属性値間の依存性に基づいて離散化を行っている[5]．しかし連続値属性の中でも年齢など正規分布に近似できないものも多いため，適用に留意する必要がある．

このように連続値属性を含むデータには，複数アイテム間の共起情報を十分に保存しつつ，各属性アイテム間を離散化する手法が求められている．次節ではこの問題点を踏まえ，連続値属性に対しクラスタリングを行うことで離散化する手法を提案する．

3 提案手法 各連続値属性を$A_1, A_2, \cdots, A_n$とし，これらについてのクラスタリングを行い，共起情報を保持できるような区切り点をみつける．ここではクラスタリングの手法として凝集法を用いることとする．凝集法では距離が一定の値よりも大きくなった時にクラスタリングを終了するので，外れ値のデータなどはそのデータひとつだけのクラスタとして処理されることが少なくない．アプリオリでは，最小支持度を満たさないルールは削除されるので，これをクラスタについても同様に考え，最小支持度を満たさないクラスタについては削除する．

またアプリオリを適用するためには，属性平面上の各属性軸に垂直な離散化を行わなければならない．そこで図1のように各クラスタ内のデータの分布を属性軸からみたときに，正規分布に近似して考える．そこで属性値の離散化によりデータが誤って分類される個数が最小になるような交点が区切る点であると考え，(1)式を満たすαを求める．

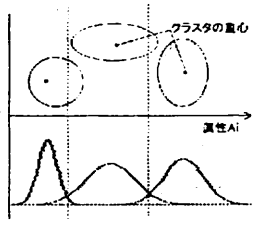

図1 区切り点の決め方

$$n_1 Q((\alpha-\mu_1)/\sigma_1) = n_2 Q(-(\alpha-\mu_2)/\sigma_2) \quad (1)$$

$$Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) dy \quad (2)$$

ただし，隣り合うカテゴリに含まれるデータのそれぞれの平均をμ_1, μ_2，標準偏差をσ_1, σ_2，対象とするクラスタに含まれているデータの個数をn_1, n_2とする．このとき誤りの割合の多い区切り点は区切るべきではないと考え，(3)式の条件を満たすαを区切り点とする．ただしθはユーザが与える閾値である．

$$\max(Q((\alpha-\mu_1)/\sigma_1), Q(-(\alpha-\mu_2)/\sigma_2)) \le \theta \quad (3)$$

以上のように区切り点を求め，各連続値属性の離散化を行う．

4 提案手法の評価及び考察 まず人工データにより提案アルゴリズムが意図通り動作していることを確認した．次にシミュレーションには米国のセンサスデータ[3]を用いた．このデータの連続値属性に対しクラスタリングしたところ，各クラスタを属性軸から見たときの歪度と尖度は大部分が−0．5〜0．5，2．0〜4．0の値をとった．よって正規分布に近似しても問題ないと考えられる．以下では連続値属性を等間隔で区切ったものを従来手法とし，これとの比較を行う．このとき区切り数は提案手法と同数とする．

I) 相互情報量による評価

各クラスタを$x_1, x_2, \cdots, x_m$とし，属性A_i（$i=1\cdots n$）の離散化された区間の集合を$\{y_{i,1}, y_{i,2}\cdots, y_{i,t_i}\}$とする．ただし$t_i$は属性$A_i$の区間の数とする．このとき離散化された区間を知ったもとでのクラスタに関する相互情報量は

$$I(X;Y) = \sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{t_i} P(x_k, y_{i,j}) \log_2 \frac{P(x_k, y_{i,j})}{P(x_k)P(y_{i,j})} \quad (4)$$

で求められる．センサスデータの結果は提案手法が1．93，従来手法が1．57となり提案手法の方が大きくなった．これは提案手法が各クラスタを考慮して離散化している効果のためと考えられる．

II) 抽出されたルールの具体例による評価

支持度10%，確信度50%で抽出したルールを比較した．表1は抽出されたルールに含まれる離散化後の属性値の具体例を，表2は実際に抽出されたルールの具体例を示している．

表1：抽出されたルールの具体例①

従来手法	{労働時間25〜50時間}
提案手法	{労働時間31〜39時間}，{労働時間1〜30時間}，{労働時間40〜42時間}，{労働時間43時間以上}

表2：抽出されたルールの具体例②

従来手法	{年齢17〜33才，自営業→教育年数6〜11年} (支持度20%，確信度71%)
提案手法	{年齢17〜26才，自営業→教育年数7〜11年} (支持度12%，確信度78%)

表1は従来手法では等間隔で区切っているために{労働時間25〜50時間}にトランザクションが集中していること，提案手法では最小支持度を満たすクラスタを考慮して区切り点を求めたためアイテムが残りやすくなったことなどが考えられる．また表2は提案手法の方が年齢，教育年数の属性において短い区間でルールが抽出され，しかも確信度が高い値になっている．これは複数アイテム間の共起情報を考え離散化したため，適切な区域のルールが抽出され確信度の高いルールが抽出されたと考えられる．

III) 平均確信度による評価

平均確信度は抽出されたルールの信頼性を評価する1つの基準と考えられる．両手法で抽出ルール数を10000にしたときの平均確信度の値をみると従来手法が73．1%，提案手法が75．1%となり提案手法のほうが2．0ポイント高くなっていた．これはII) の抽出されたルールの具体例の②で示したような確信度の高いルールが数多く抽出されたためだと考えられる．

5 むすび 従来は連続値属性について他の属性の影響を考慮せずに離散化を行っていたが，今回の提案手法により，複数アイテム間の共起情報を失わずに連続値を離散化することが可能となった．

<参考文献>

[1] R,Agraval R,Srikant, "Fast Algorithms for Mining Association Rules", 1994, Proceeding of the 20th VLDBConference PP.487-499
[2] 寺邊正大ら，"相関ルールにもとづく属性生成手法" 2000.1，人工知能学会誌 Vol.15 No.1 PP187-197
[3] http://www.ics.uci.edu/pub/machine-learning-databases
[4] 松本一則ら，"実時間網管理への定性的診断知識の適用手法"，1993，信学技法 AI93-37 PP9-14
[5] 鷲尾隆，元田浩，"構造データ及び数値データに対する相関ルールマイニングの拡張"，2000.9，人工知能学会誌 Vol.15 No.5 PP.759-767