博 士（情報科学） 設 樂 洋 爾

学 位 論 文 題 名

Descriptive Decision Rule Mining for Revealing Multiple Aspects of Databases

（データベースの多面的な解釈を与える
記述的決定規則ルールマイニングに関する研究）

学位論文内容の要旨

データベースから知識を発見する技術として、データマイニング、あるいはKDD (Knowledge Discovery in Database) と呼ばれる技術が注目を集めている。とりわけ、データベースに潜在するパターンをIf-then型のルールとして取り出すルール抽出手法は、分析者が解釈しやすい知識表現を与え得るため、知識発見において重要な役割を果たす。

本研究では、ルール結論部の属性が固定されている場合、すなわち決定規則ルールの抽出を扱う。ある特定の決定属性 (クラス属性) が複数の説明属性によってどのように説明されるかをルール「A ならば B」の形で表現する。決定規則ルールを得る手法はこれまでも広く研究されており、様々な手法が提案されている。しかし、従来の決定規則ルール抽出においては、知識発見に貢献するルールを求めるよりも、未知のインスタンスに対して精度よい予測を行うルールの抽出に主眼が置かれていた。従って、従来の方式ではルール集合の予測あるいは識別性能が高くなるような基準が設定され、その基準を最大化するルールが探索されていた。しかし、探索はヒューリスティクスに基づいて行われるため、ルール抽出の手続きをアルゴリズムの形で定義できても、それによって得られるルール集合の性質については陽に定義を与えることができないといった状況であった。

従来の多くの決定規則ルール抽出手法はSeparate-and-Conquer型あるいはDivide-and-Conquer型に分類される。Separate-and-Conquer型では、まずルールを一つ求め、そのルールによって被覆されるインスタンスをデータから削除し、そこから次のルールを生成する。一方、Divide-and-Conquer型では、インスタンスの空間をクラスの分布がおよそ均一になるまで、再帰的に分割を繰り返す。決定木に基づくルール抽出手法はこれにあたる。これらの方式に共通する問題点として、"splintering problem"(分化問題) が指摘されている。これは、ルールの生成が進むにつれて使用できるインスタンスの数が減少していくため、はずれ値であるようなインスタンスを無視できなくなり、またノイズに敏感になり過ぎるという指摘である。また、ルールはルール集合を探索する過程で順次追加されていくため、得られるルールがその探索のコンテキストに依存するという問題もある。これは、データベースの解釈を行うためのルールを抽出するという観点からは望ましい性質ではない。

ルール抽出手法の適用可能性にも注意が必要である。ルール抽出の対象となるデータベースには、多くの場合、数値属性や順序属性、カテゴリカル属性が混在する。実問題への適用を考慮すると、これら多種の属性を統一的に扱えることが望ましい。また、多くの値をとるカテゴリカル属性

がデータベースに含まれる場合に、ルール前提部で属性値を一つしか指定できない手法では、ルールが細分化し過ぎるという問題もある。このような場合には、属性値を適切に纏め上げ、ルール前提部で属性値集合を指定することで、一般性を保つことができると考えられる。加えて、データベースにはしばしば欠損値が含まれる。よって、欠損が含まれる場合でも適切なルールの抽出が行えることが望ましい。

本研究では、これらの問題意識のもと、決定規則ルール抽出手法を以下のように構成した。まず、抽出すべきルール集合を、部分クラス法 (Kudo *et al.*, 1993) の枠組みに基づいて「無矛盾性」と「極大性」の観点から与えられた同一決定属性を持つインスタンス集合の部分集合族 (部分クラス族) として定義する。カテゴリカル属性、順序属性、数値属性を同時に扱えるようすべての属性値をべき集合の枠組みで扱う。次に、部分クラス族の (被覆が大きいなど) 性質の良い部分集合を確率アルゴリズムによって計算コストを抑えて求める。ルールの抽出は無矛盾性を維持しながら、可能な限りルールを一般化していくことで行う。このとき、各々のルールは既に得られているルール集合には依存せず抽出される。その結果、ルールはデータを多面的に被覆し、多様な属性で説明するものとなる。計算機実験により提案ルール抽出法にて実データから解釈のしやすいルールの抽出が行えることを確認した。

さらに幅広い問題に適用可能にするため、属性がアイテム集合で表現されるトランザクション型のデータベースからのルール抽出に関しても検討を行った。例えば、表型のデータベースや、テキストデータもアイテム集合へと変換することができる。属性値としてアイテム集合が与えられる場合でも、アイテム集合から無矛盾性を維持してアイテム集合の一般化を行うことで、表型データベースの場合と同様の枠組みでルールの抽出が行えることを確認した。また、提案ルール抽出手法を Associative Classifier の枠組みで捉えると、従来の Class Association Rule を用いる方式では抽出が難しい、低サポート高コンフィデンスのルールを抽出する手法とみることができる。実際に識別子を構成し、計算機実験により良好な識別性能が得られることを確認した。

提案ルール抽出手法の特長を以下にまとめる。

・未知インスタンスを予測するためのルール抽出に主眼を置くのではなく、データの解釈を行うためのルールを抽出することができる。
・抽出されるルールの性質がアルゴリズムの基準から陽に定義される。
・既に得られているルール集合に依存してルールを決定するのではなく、ルール一つ一つを独立して抽出するため、多様なルールを抽出できる。
・数値属性、順序属性、カテゴリカル属性が混在しているデータを統一的に扱える。
・カテゴリカル属性の属性値をグルーピングしたルールが発見される。
・データベースに欠損値が含まれる場合でも適切に扱える。

本研究における成果は、決定規則ルール抽出においてデータベースの特徴を表現する記述的なルールを求める手法を提案したことにある。提案手法は数値属性、順序属性、カテゴリカル属性の混在するデータベースに対応し、またカテゴリカル属性については属性値のグルーピングを自動で行うことで解釈しやすいルールを効率良く抽出することができる。計算機実験により、提案法が実データに対しても有効に働くことを確認した。

学位論文審査の要旨

主　査　教　授　工　藤　峰　一
副　査　教　授　宮　腰　政　明
副　査　教　授　原　口　　誠
副　査　准教授　中　村　篤　祥

学位論文題名

Descriptive Decision Rule Mining for Revealing Multiple Aspects of Databases

(データベースの多面的な解釈を与える記述的決定規則ルールマイニングに関する研究)

知識発見あるいはデータマイニングという分野が注目を集めるようになってから早10年以上が過ぎている. この間, 産業応用を含め, 理論的検討も数多くなされてきている. 分野の動機と目的から, 方法論の評価が「実際の知識発見に役に立つか」という主観的なものにならざるを得ないという難しさがある一方で, 膨大なデータから隠れた規則や知識を効率良く探り当てるという挑戦的課題は多くの研究者を惹きつけてきた.

本研究では, これまでの各種方法論の多くが高速アルゴリズムを開発するということに偏重していたことを憂いて原点に立ち戻り, 知識発見に有効なアルゴリズムの開発を行ったものである. 従って, 利用者であり, かつ, 評価者である得る「人」を強く意識した研究となっている. 特に, 決定規則という形での知識発見を扱っている. 問題意識は,1) どんなデータに関しても容易に適用可能なものであるか,2) 抽出した規則が人にとって解釈しやすいものであるか,3) 人が一覧しやすい規則の提示を行えるか, として具体化されている. 実際, これまでの多くの方法は, トランザクション (1データ単位) が離散属性, あるいは, 連続属性のみで表現されている場合にのみ適用可能なものが殆どであり, 順序を持たないカテゴリカル属性が含まれる場合, 特殊な前処理を必要としていた. また, アンケート調査などでは回答者は全ての質問に答えるとは限らず, 結果として欠損値が発生する. このような欠損値を含むデータに関してそのままの形で適用可能な方法は限られている. また, 解釈のしやすさよりも, 予測として精度の高い規則を求める方式が多く, 解釈のしやすさがおざなりになっていたことは否めない. 加えて, 各種アルゴリズムは非常に多くの規則を生成することが多く, 実際にそれらをユーザが吟味することは現実的ではなかった.

本研究では, これらの反省から,1)〜3) を総合的に評価しつつ新たな提案を行ったものである. 基本方針を,「同一決定属性を有するトランザクション集合を極大基準で抽出する」としてアルゴリズム設計を行っている. 極大性は明らかに簡潔な規則を導く. また, 重複を積極的に許すことは, 類似事象を多視点から解釈することを許すという利点がある. さらに, この方法では, 計算量の問題から

ないがしろにされることが多かった「少数事例にしか当てはまらないものの貴重な知識」を発見することも可能にしている. 客観評価指標の一つである (サポート, コンフィデンス) 対の言葉で言い換えれば, これまで抽出が困難であった低サポート高コンフィデンスなルールがこの方法では抽出可能となっている. 実際のところ, 高サポートの規則は, ユーザが既に知っている「当たり前の規則」になることが多く, 適用範囲は狭くとも信頼性の高いルールが新たな発見を生むことも少なくない.

これまでの方法では, 前処理によりカテゴリカル属性や連続属性を離散化することを多く行っていたものの, これでは合理性を欠く順序が入るきらいがあった. 提案方法では, すべての属性値を部分集合の族において統一的に定式化することでこの問題を回避している. さらに, 欠損値の扱いにおいても, 所与の問題知識に応じて, 欠損値を「すべての値となり得る可能性がある場合」,「どの値でもない場合」に分けて自然な形で取り扱う方法を提案している. これにより, 他のトランザクションにおける代表値で置き換えるよりもデータを適切に扱える可能性は高い. 提案方法ではこれらの統一的方法論により, 殆どすべの種類のデータを同一手法で扱うことが可能となり, 方法論の適用範囲を格段に広げることに成功している.

さらに, 利用者にとっての利便性向上を目指して, 抽出した規則の評価に記述の簡潔さを取り入れる評価式を開発するとともに, 規則集合全体の評価を ROC 曲線における面積 (AUC) で行っている. 結果として, より簡潔かつ評価の高い規則を適切な表示順において提示することができるようになっている. 加えて, トランザクションがアイテムセットである場合に, これまでの方法論と比べてより高い推定精度を達成している.

本論文による成果は以下にまとめられる.

1. 質量属性が混在する, あるいは, 多数の欠損値が存在する多様なデータセットを同一の方法で統一的に, かつ, 合理的に扱うことを可能とした.
2. 提案方法により, 従来と比べてより多彩な規則を抽出できるようになるとともに, 従来軽視されていた, 適用範囲は狭いものの重要な知識発見に繋がり得る規則を抽出できるようになった.
3. 利用者の負担を考慮して, 解釈しやすい規則を高順位で表示する方法を考案するとともに, そのために規則集合を適切に評価する方法を提案した.

これらはいずれも, 従来の方法論による制約を一定程度減ずるものである.

これを要するに, 著者は, 知識発見, 特に, 膨大なデータから決定規則を抽出する問題において, 既存方法の適用範囲を大きく広げるとともに, 真の知識発見へと繋がる評価方式を検討し, その有効性を具体的に示した. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.