博士（工 学） 工 藤 嘉 晃

学 位 論 文 題 名

A Study on Appropriate Abstraction for Data Mining

（データマイニングにおける適切な抽象化に関する研究）

学位論文内容の要旨

近年、データベースからの知識発見(Knowledge Discovery in Databases; KDD)に関する研究が盛んに行われ、様々なKDDシステムが開発されている。KDDシステムは、基本的に複数のプロセスから構成されるが、主として、前処理とデータマイニングに大別できる。特に、データマイニングは、効率的に巨大なデータベースから有益な知識を発見するための中心的なプロセスであり、KDDに関する多くの研究はこの部分の研究に集中している。しかしながら、データベースにおけるデータ値が、非常に具体的、または詳細に記述されている場合、既存のマイニングアルゴリズムでは、マイニングの目的に無関係なデータの性質・側面までも結果的に利用され、それゆえに、ユーザーにとって理解が難しく、容易に分析できない複雑なルールが数多く生成される。また、このような状況では、ルール生成にかかる計算コストも増加する可能性がある。本論文では、これらの問題に対処するために、データ値の記述レベルをある程度抽象化することにより、目的に関連するデータの性質・側面のみに注目し、かつ、記述を簡略化するという意味でのデータ抽象が有効であるとの立場に立ち、その手法の開発を行なう。

第一章では序論として、上記のような研究背景、本研究の動機と概要について詳しく述べている。

第二章では、データ抽象の代表的な手法の一つとして属性指向アルゴリズムを取り上げ、その問題点を指摘している。このアルゴリズムは、単一継承の木構造に制限された概念階層に基づき、データベースの一般化を行なう。このとき、抽象レベルは各属性における属性値の数の上限により決定される。しかしながら、単一継承の階層を用いた抽象化では、データの一つの側面のみしか抽象化できず、マイニングの目的を達成するために保存すべき重要なデータの性質・側面を失う可能性がある。また、属性値数の上限だけで、目的に対し適切な抽象レベルを決定することが困難であることを示す。

第三章においては、前章で指摘した問題点を解決するための方法を与えている。具体的には、マイニングの目的を、「与えられたクラスにデータを正確に分類するルールの発見」と定め、その目的に照らして簡潔でかつ高精度の分類ルール、特に、決定木を生成するために適切な抽象化を選択する基準を提案する。決定木生成の際に、重要なデータの側面・性質はクラス分布である。抽象化後においてもクラス分布が保存されるならば、その抽象化は適切であるといえる。逆に、クラス分布が保存されない場合は、抽象化による情報損失が生じ、決定木の分類精度が低下する。そこで、本章では、クラス分布保存の基準を考案し、その基準に基づいて可能なデータ抽象の集合から適切なものを自動的に選択する手法を開発している。クラス分布が類似する属性値は、クラスを分類する能力が同じであることから、決定木の生成において区別する必要がない。したがって、「ほぼ類似したクラス分布をもつ属性値を一つのグループにまとめる抽象化を選択する」という第一の基準を与

える。この基準を満たす抽象化は、抽象化後の情報量の損失を最小限に抑えることができる。このことから、第二の基準「クラスに関する情報量をほぼ保存する抽象化を、適切な抽象化の候補とする」を提案する。この基準により、分布間距離が大きな分布でも、その生起確率が小さければ例外的分布として扱うことができる。これにより、情報量の損失を一定の範囲に制限し、かつ、より簡略化されたデータベースを構築することが可能となる。本章では、最後に、これらの情報量基準によって、適切な抽象化を選択できる Information Theoretical Abstraction (ITA) を提案する。また、ITA よる各属性への抽象化の適用は、他の属性と独立に行なうために、複数の属性値に相関性がある場合、一つの属性に対して最適なデータ抽象が、相関のある属性で条件付けたときに、必ずしも最適であるとは限らないという問題を指摘する。

第四章では、複数属性のもとでの最適な抽象化を求めるために、決定木の展開プロセスにおける各属性選択ステップに対して、最適な抽象化の適用を行なう逐次 ITA を提案し、この方式によって、データ抽象による性能劣化を限りなく抑制できることを示す。さらに、多くのマイニングアルゴリズムで問題となっている、デフォルトルールの大量生成を防ぎ、隠れたルールを発見することも試みる。ここでデフォルトルールとは、より多くのデータに対して高い確信度で成立する常識的なルールを指している。こうしたデフォルトルールの生成を抑制するために、高い支持度を持つルールは多数のインスタンスを持つ抽象データ(タプル)によりサポートされることが多い事実に着目し、この基準を満たす抽象データを除去し、デフォルトルール生成の原因を排除する手法を提案する。具体的には、属性指向アルゴリズムやその拡張である ITA が持つ vote 値を利用し、除去すべき抽象データを決定している。これにより、デフォルトルールを大幅に除去し、例外的なルールを発見することが可能となることを示している。

第五章では、ITA や逐次 ITA でデータ抽象の知識源として用いている階層情報に不具合がある場合、マイニングされたルールの品質に劣化が生じる問題を回避するために、階層を洗練する手法を与えている。ここでの洗練化とは、マイニングされるルールの品質向上に寄与できる、階層における中間概念を生成することを意味している。この目的のために、可能な中間概念を生成できる探索空間とその枝刈り手法を提案する。さらに、辞書に照らして無意味な候補を除去するために、候補となるデータ抽象を所与の上界と下界に基づき制限する探索法を与える。

第六章では、提案した各手法の有用性を検証するために行なった実験の結果とその考察を示している。まず、ITA により一般化されたデータベースから抽象的な決定木を生成する実験を行ない、元の決定木と比較して、分類精度の劣化をできるだけ抑え、かつ、決定木のノード数を大幅に減少できることを実証する。次に、決定木での分類問題以外に、パターン認識の分野などでよく用いられる SVM(Support Vector Machine)やブースティング学習器として著名な AdaBoost に対して、本手法の有用性を確かめる予備実験を行ない、決定木以外の分類器に対するデータ抽象の効用について述べている。 さらに、逐次 ITA による抽象決定木を考察し、ITA を用いた場合よりも、さらに学習データに殆ど依存しない抽象決定木を生成できることを示す。このことから、逐次 ITA は ITA と同様に、過学習を回避することができ、ITA の場合よりもノード数は多いが、より精度の良い決定木を生成できるとの結論を得る。また、vote 値に基づく基準により、第五章で述べた多数のデフォルトルールを除去し、少数のデータ群でしか成立しない例外的なルールを発見できることも実証する。最後に、データ抽象の探索手法に基づく予備実験を行ない、大幅にその候補の数を減らせることを示している。

第七章では、本論文の総括を与え、残された研究課題について述べている。

学位論文審査の要旨

主　査　教　授　原　口　　誠
副　査　教　授　北　島　秀　夫
副　査　教　授　田　中　　譲

学位論文題名

A Study on Appropriate Abstraction for Data Mining

(データマイニングにおける適切な抽象化に関する研究)

情報通信ネットワークを介した大量データの収集・蓄積とその高速な計算機処理が可能になったことから、データベースからの知識発見(KDD)の研究が活発に行われるようになった。KDDが行う主要な処理は、所与の大規模データからユーザにとって有意味だと判断される知識、とりわけ、ある一定の精度を持つルール形式の知識を効率的に発見・検出することだと言われている。しかしながら、データの記述レベルがあまりにも詳細かつ具体的な場合、獲得されるルールの可読性は低く、また、精度条件を満たすルールの総数は一般には数千個以上となることが多く、その妥当性をチェックしなければならないユーザーの負担は決して無視できるものではない。本論文ではこうした問題に対処するために、データ記述の抽象化により、出力ルールの可読性の向上とルール数の減少を同時に達成することを試みている。その際、マイニングの目的に照らして必要な情報の捨象を防ぐために、適切な抽象化に基づくデータマイニング手法を、決定木の形で表現される分類ルール発見問題に対して新たに導入し、高い可読性と精度落ちを最小化するより小さなサイズを持つ決定木構築が可能であることを実験的に示している。さらに、そうした抽象化が有効となるための一つの十分条件を明らかにしている。

第一章では序論として、本研究の動機と概要を述べ、本論文の位置付けを行っている。

第二章では、データベースの汎化に関する代表的な手法である属性指向アルゴリズムについて考察し、複数の可能なデータ抽象を処理できることが、汎化されたデータベースから高精度の有用なルールを獲得するために必要になると指摘している。

第三章では、データ抽象の適切さの基準とこれに基づくデータ抽象化手法を提案し、さらに、データ抽象が出力ルールのサイズ減少に寄与するための十分条件を考察している。先ず、事後分布情報に基づくデータマイニングにおいて重要となるクラス分布を保存するデータ抽象は、抽象化前後での情報量損失を最小化する性質を持つことから、相互情報量の損失をできるだけ抑える抽象化を選択する基準を提案している。さらに、こうした抽象化が、出力ルール数の減少に寄与することをより一般的に示すために、事後クラス分布に対するクラスター概念を定義し、一つの抽象事後分布に対応する抽象化前の事後分布がクラスターを形成できる場合に、出力ルール数の減少を実現できることを示している。こうしたデー

タ抽象の選択基準とその性質に基づいてマイニングを行うＩＴＡ (Information-theoretic Abstraction) システムを設計している。

第四章では、前章で導入したＩＴＡの問題点を分析し、さらなる洗練化を行っている。先ず、ＩＴＡでは、属性毎に独立して最適なデータ抽象を定めているが、属性間に相関性がある場合には、属性毎の抽象化による誤差が累積し、よって、出力ルールの精度に影響を与えると指摘している。この問題に対処するために、属性選択の各段階で、他の属性に対して選択された抽象化に依存して最適な抽象化を決定する逐次ＩＴＡ (Iterative ITA) システムを提案・設計している。次に、高精度で有用だが比較的低い支持率のために検出されにくいルールを発見する問題に対しても、データ抽象は有効であることを示している。すなわち、抽象化によって得られる抽象データのうち、ある一定の投票比を持つものを排除した部分データベースに限定することにより、比較的小規模の集団に対してもＩＴＡによって高精度のルールを検出できる手法を提案している。

第五章では、ＩＴＡや逐次ＩＴＡがその知識源として利用している電子化辞書が不具合を持つ場合、ルールの品質が劣化する問題を指摘し、中間概念の自動生成により不具合を除去する手法を与えている。

第六章では、第３章から第５章において提案した各手法に対する実験結果を述べている。すなわち、精度の劣化を最小限に抑えながら、同時に出力ルール数の大幅な減少を実現できることを実験的に確認している。特に、逐次ＩＴＡはＩＴＡ同様に、出力ルール数の大幅な減少を実現し、また、精度落ちに関しても最も優れた特性を持つことを実証している。さらに、投票比による部分データベースへの制限によって、制限をおかない場合は抽出されない高精度のルールを検出できることを確認している。

第七章では、本論文の総括を与え、残された研究課題について述べている。

これを要するに、著者はデータマイニングにおけるデータ抽象の効用に関する新知見を得たものであり、大規模データベースからの知識発見に対して工学上貢献するところ大なるものがある。よって著者は、北海道大学博士（工学）の学位を授与される資格あるものと認める。