概念学習システムにおける事例のモデル化についての検討

4M－1

増田 剛*　坂本 憲広**　牛島 和夫*

* 九州大学 大学院 システム情報科学研究科　** 九州大学 医学部附属病院 医療情報部

1. はじめに

概念学習は事例から一般的な概念記述を帰納する学習である。様々な情報がデータベースに蓄えられている今日、それらのデータの機械的な解析に、決定木を用いた概念学習の技術を応用できる場面が多くあると考えられる。それゆえ、決定木学習システムにおいて応用分野に則したより複雑なデータを扱うための拡張が試みられている。例えば文献[2]や[3]では、階層構造を持つ属性を含む事例からの学習手法について述べられているが、これらの手法は、ある種の属性だけを考慮しており、様々な構造を持つ事例を扱う一般的な方法というわけではない。

そこで本論文では、事例をオブジェクト指向技術を用いてモデル化することによって、複雑な構造を持つ事例から学習を行なう一般的な枠組を提案する。

2. 決定木学習システム

まず、一般的な決定木学習システム[1]を簡単に紹介し、既存システムの限界ついて述べる。

2.1 属性 - 値記述

一般的な決定木学習アルゴリズムが扱う事例は、属性 - 値記述で表現されている。つまり各事例は共通の属性集合を持ち、各属性値のとりうる値は離散値又は連続値である。また各事例は、自分が属するカテゴリを示すラベルを持つ。

2.2 学習アルゴリズム

決定木学習アルゴリズムは、適当な属性に基づく一つのテストを選択しながら、事例集合を繰り返し分割していく分割 - 統治法に基づくものである。

このアルゴリズムで最も重要な部分は事例集合を分割するためのテストの選択である。一般的なアルゴリズムでは事例の持つ全ての属性について、各属性に関するテストで分割したときの分割の良さをなんらかの指標に従って計算し、最も良い結果を与えるテストで事例集合を分割する。

2.3 属性 - 値記述の問題点

属性 - 値記述の問題点として、構造を持つ事例や時系列データを含む事例を適切に表現できないということが挙げられる。例として時系列データを含む医療データを属性 - 値記述で表現する場合を考える。図1は属性 - 値記述で表現された事例である。実際の診療記録では、「体温」や「血圧」は日時は異なっているが、関連する一つの項目である。しかし、属性 - 値記述による表現ではこのように別々の属性として表現しなければならない。これは実際の構造を反映しておらず、事例の適切な記述であるとは言えない。また、本研究で対象としている属性間の関係を考慮しない学習モデルにおいては、関連する項目を別々の属性とすることは、特に有用な知識の抽出を妨げることになりうる。実際のデータの構造から考えると、事例は図2のような形で表現されるべきであると考える。同様に、事例が木構造やグラフ構造を持つ場合も、その事例の適切な属性 - 値記述表現を得ることは難しい。

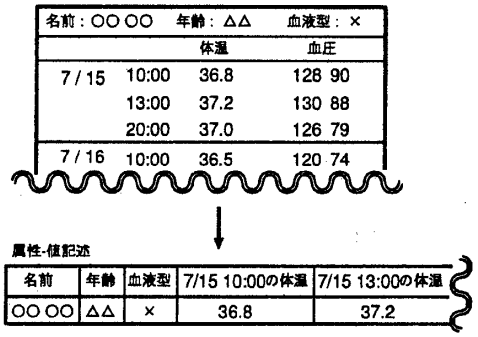

図 1: 属性 - 値記述の問題点

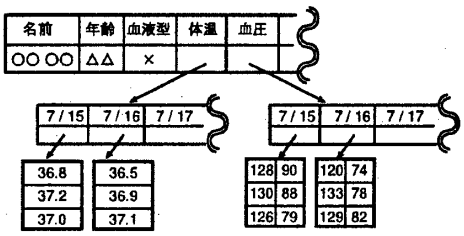

図 2: 事例の理想的な表現

3. 事例のモデル化

前節のような属性 - 値記述の問題の解決策として、本研究では事例が持つ属性として、離散値や連続値だけでなく様々な構造の属性を扱うことができるようにする。本節では以下の二点について述べる。

- 様々な構造を持つ属性を含む事例を一般的にどうモデル化すればよいか。
- そのような事例を決定木アルゴリズムでどう扱えばよいか。

3.1 事例オブジェクト

一つの事例は、属性 - 属性値の組の集合と一つのカテゴリを持つと考えることができるので、事例をモデル化したクラス **Case** を図3のように設計した。事例の属するカテゴリはクラス **Category** のインスタンスで、各属性はクラス **Attribute** のインスタンスとして表現される。ここで、**Attribute** は一般的な属性の振舞いを規定した抽象クラスで、図4中の四つの抽象メソッドを提供する。離散属性や連続属性のような具体的な属性は、**Attribute** を継承した各サブクラスとして表現される。これらは **Attribute** の持つ四つの抽象メソッドを具体的に

A Model of Cases in Concept Learning System by Object-Oriented Technology.
Gou Masuda*, Norihiro Sakamoto** and Kazuo Ushijima*
*Graduate School of Information Science and Electrical Engineering, Kyushu University.
**Department of Medical Informatics, Kyushu University Hospital.

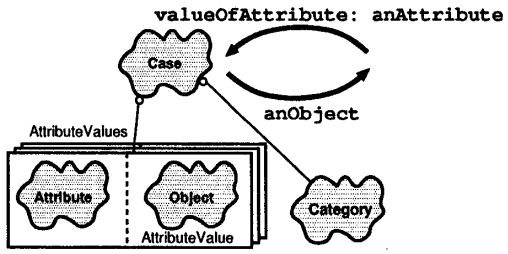

図 3: 事例オブジェクト

記述している。事例が持っている属性集合は、実際にはこのようなサブクラスのインスタンスである。また、各属性値には任意のオブジェクトをとることができ、事例の属性値は、valueOfAttribute: メソッドで参照することができる。抽象クラスを用いることで、様々な構造の属性をアルゴリズムの中で統一的に扱うことが可能となる。

3.2 学習アルゴリズム

このようにオブジェクトとして表現された事例を、決定木学習アルゴリズムの中でどのように扱うかについて述べる。第2.2節のアルゴリズム中で変更する部分はテスト選択の部分である。事例の取りうる属性として任意の構造の属性を許すことは、属性に対するテストも各属性ごとに異なる形式のテストを可能にする。例えば、時系列データを扱うような属性には、その平均値とある閾値との大小を比較するテストを扱うことができる。また、階層構造を持つ属性では、ある値がある階層の中に含まれるかどうかを調べるテストを扱うことができる。

そこで、決定木アルゴリズムでこのような属性に関する任意のテストを扱えるように、テストを抽象化したクラス Test を導入する。さらに、ある属性に対するテストは関連する複数のテストの集合と考えられるので、Test の集合を表すクラス Tests も導入する。例えば、年齢 ≤ 35 と 年齢 > 35 という属性年齢に関する二つのテストは、それぞれ Test のインスタンスとして扱われる。これらはさらに Tests の1インスタンスとして扱われる。各属性は自分に適用可能な Tests の集合を makeTestsCollection メソッドによって生成することができる。これにより、これまでの離散属性や連続属性に対しても単に値の比較をするテストだけではなく、属性値にファジー関数を適用するテストなど、これまで扱えなかったテストが可能となる。

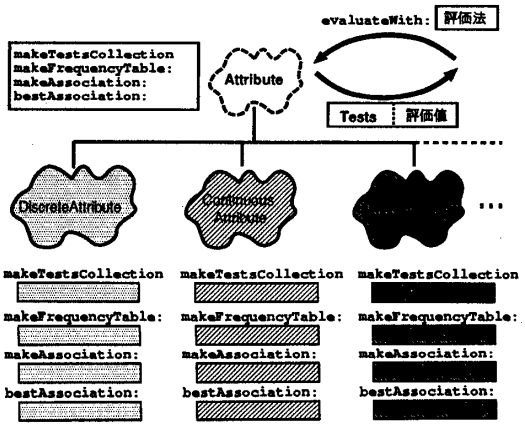

図 4: クラス Attribute の振る舞い

決定木アルゴリズムにおいて、各属性は evaluateWith: というメッセージとあらかじめ決められた評価法 (TestSelector のインスタンス) を受け取ると、自分自身に関する全てのテストの中で評価値を最良にするテストとその最良評価値の組を返す。evaluateWith: メソッドの中では、Attribute で定義されている四つの抽象メソッド (表1) を順に起動する。これらのメソッドは抽象的なインターフェースのみを記述しており、具体的なメソッドの中身は、離散属性や連続属性といった Attribute の各サブクラスごとにそれぞれ記述されている。つまり、これらのメソッドを記述するだけで決定木学習システムの中で種々の属性を扱うことが可能になる (図4)。

表 1: クラス Attribute の抽象メソッド

メソッド	振る舞い
makeTestsCollection	自分に対して適用可能な Tests の集合を作成
makeFrequencyTable:	ある一つの Tests で事例集合を分割する場合の事例のカテゴリの分布表を作成
makeAssociation:	一つの Tests を評価し、Tests と評価値の組を生成
bestAssociation	Tests と評価値の複数の組の中で最良のものを選択

4. 議論

前節で述べた事例オブジェクトを用いると、扱いたい属性を Attribute のサブクラスとして定義することで、任意の属性を含む事例に対して決定木学習アルゴリズムを適用できる。また、属性に対して適用可能なテストを各属性ごとに記述できるので、これまで扱えなかった種類のテスト (3.2節参照) を使うことが可能となる。これは決定木の表現力を高めることにつながる。

しかし本モデル化は、各事例では共通の属性集合を用いねばならないという制約を解決できていない。この問題を解決するためには、決定木のノードにおいて、属性ではなく事例自身がテストを受けて分類されていくというモデルにする必要があると考える。また決定木学習アルゴリズムにおいては、アルゴリズムの抽象度を高めるために、アルゴリズムの効率よりも構造について着目しているので、通常の決定木学習アルゴリズムと比較して効率が悪いことが予想される。

5. まとめ

本論文で提案した枠組を用いることによって、属性-値記述で表現された事例だけでなく、リストや木構造、グラフなどからの概念の抽出に決定木学習システムを応用できると考える。今後は、実際の応用分野のデータに対して提案方式を適用し、評価を行なう必要がある。

参考文献

[1] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[2] 田中英輝: 構造化属性を許す決定木アルゴリズム (1), 1996年度人工知能学会全国大会予稿集, 1996.
[3] 中島誠, 葉鈴如, 伊藤哲郎: 決定木による階層属性を用いた概念の帰納学習, 人工知能学会誌, Vol.10, No. 1, 1995.