確率的知識の帰納学習

中藍 洋一郎　古関 義幸　田中 みどり

日本電気(株) C&Cシステム研究所

1　はじめに

近年、帰納的学習に関する研究が盛んに行われているが、その多くは分類規則を帰納的に学習する問題を対象としたものである。この問題は分類問題と呼ばれ、各データに与えられるいくつかの属性値をもとに、それぞれどのクラスに属するのかを決定する問題である。与えられた例題に基づいて、適切な分類規則を学習するために様々な方式、アルゴリズムが提案されている。その代表的なものとして、分類木の学習アルゴリズムID3等が知られている[3]。しかし、一般には、予めどのようなクラスが存在するのか判らないような問題や、決定的にクラスが決まらない確率的な問題も多い。

そこで、本論文では、そのような確率的に発生する事象を対象として、過去に観測された情報を基に帰納学習を行う問題について考える。

2　確率モデルの帰納学習

確率的に発生する事象のひとつの例として、機器の故障があげられる。ある装置を構成する各部品の「壊れやすさ」を考えると、一般には一様ではなく、ある傾向を持っている場合が多い。例えば、ある種類の部品は非常に壊れやすいとか、1年以上前に導入した部品に、新しいものよりも故障の頻度が極端に高いといった傾向を持つ場合も多い。しかし、一般に部品の種類や新しさの他にも、設置場所、温度その他、様々な属性が考えられるため、各事象の発生確率に影響を与える属性（あるいはその組み合せ）を帰納的に見つけていく必要がある。

表1　与えられる観測結果

事象	属性				頻度
	属性 a_1	属性 a_2	・・・	属性 a_n	(回)
x_1	v_{11}	v_{12}	・・・	v_{1n}	n_1
x_2	v_{21}	v_{22}	・・・	v_{2n}	n_2
⋮	⋮	⋮	⋱	⋮	⋮
x_m	v_{m1}	v_{m2}	・・・	v_{mn}	n_m

ここでは、表1に示すような形式で過去の発生事象に関する情報が与えられるものと考える。このような確率モデルの学習問題を解決するため、我々は推定木を用いて確率確率モデルを表現する方式を導入し、最も適切な確率モデルを選択するためにMDL基準を採用した[2, 6]。以下、簡単にその概要を述べる。

2.1　推定木

下図に示すように、推定木は●で示される分岐点と、○で示される葉とから構成される。

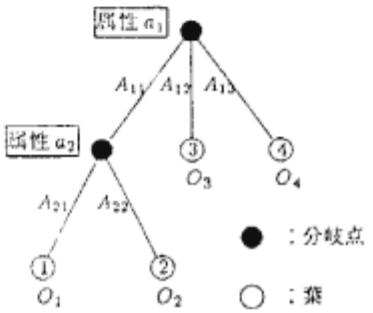

図1　推定木

各分岐点●にはそれぞれある属性 a_i が対応し、（図では属性 a_1, a_2）、そこから下へ出る枝にはそれぞれ属性 a_i の取り得る値の部分集合 $A_{i1}, A_{i2}, \ldots A_{il}$（$l$ はその分岐点から下へ出る枝の本数）が対応する。ここで、各 A_{ij} は、排他的かつ $Dom(a_i)$ を網羅するものとする。推定木の意味について、以下に例を用いて説明する。

例えば、ある装置を構成する部品の内のどれかが故障するという状況を考える。ここで、装置を構成する部品は古いものと新しいものが各々同数あるものと仮定する。もし、過去に古い部品が20回壊れたのに対し、新しい部品が1回しか壊れなかったとすると、「古い部品」と「新しい部品」の故障確率には差があると考えるのが自然である。そのような確率モデルを表現する推定木は図2のようなものになる。

一方、過去の故障回数がそれぞれ2回と1回の合計3回しかなかった場合、古い方が壊れやすいと結論付けるのは危険である。これは、本来は古いものも新しいものも故障確率に大差はなく、偶然にそのような観測結果が得られた可能性が十分あると考えられるためである。従って、このような場合には、図2のような推定木によって、古いものと新しいものの2つのグループに分けてしまうことは適当ではないと考えられる。このよう

Inductive Learning of Probabilistic Knowledge
Y. Nakakuki, Y. Koseki and M. Tanaka
C&C Systems Research Labs., NEC Corporation

に、与えられたデータに対してどのような推定木を学習すれば良いかを判断するための基準について次に述べる。

図２　推定木の例

2.2　MDL基準によるモデル選択

推定問題を解決するために我々は前節において、「推定木」による確率モデルの表現方法を採用した。そこで、与えられたデータに対して考えられる推定木の中で、「将来起こる事象の予測に最適な」推定木を選択するための基準として Minimum Description Length criterion, MDL基準を用いる。この基準は、Rissanen[4, 5] によって提案されたものである。彼の主張は、確率的に生起する事象について、「過去における観測データを基に将来起こる事象を最も適切に推定できる確率モデルは、その記述長が最も短いモデルである」というものである。ここで、確率モデルの記述長は、

1. そのモデル自身の記述長と、
2. そのモデルを用いた場合に、与えられたデータを記述するのに必要な記述長

の和として計算される（単位は共に bit）。この考え方によれば、推定木の記述長は、

1. 木の記述自体に必要な情報量と
2. その木の表わすモデルの、与えられたデータに対する対数尤度

の和となる [2, 6]。このMDL基準を用いることで、生起確率の適切な推定に必要な属性（またはその組み合せ）を見つけることができる。従って、各事象の起こる確率の適切な推定を行うことが可能となる。

3　問題解決への応用

2章で述べたような推定木とMDL基準を用いた経験的知識（確率モデル）の学習機能を問題解決に応用することで、「経験を積むに従って効率の良い問題解決が可能となる適応型の問題解決システム」を構築することが可能となる。我々は、故障診断問題を対象として、上記学習方式の有効性を確認した [1, 6]。モデルベースの診断においては、考え得る全ての故障原因の数が非常に多くなるため、効率よく診断を行うためには、各々の原因の起こりやすさ（生起確率）を考慮した診断戦略が必要となる。全く経験のない状況では、各部品の故障確率は一様と判断されるされるために、あまり効率の良い診断を行うことができない。しかし、経験を積むに従って、故障発生の確率分布を学習することで、効率のよい診断を行うことが可能となる。

この他にも、多くの問題に対して確率モデルの学習機能が有効であるものと考えられる。例えば、EBLやchunkingによる学習を行って得られた知識は常に有効である訳ではなく、実際にいくつかの問題に対して適用し経験を積むことによって、初めてどのような場合に有効であるのかが判る場合が多い。従来は単純に、得られた知識がどのくらいの確率で有効であるかを調べ、その確率が高いものを利用し、低いものは利用しない等の戦略がとられていた。しかし、一般には役に立つ確率の低い知識でも、ある状況では非常に有効である可能性があり、また全くその逆の場合も考えられる。従って、より効率よく問題解決を行うためには、そのような状況に応じた問題解決戦略が必要となる。従って、このような分野においても、提案する学習方式は非常に有効であると考えられる。

4　おわりに

確率的に発生する事象を対象として、過去の観測データに基づいて適切な確率モデルを学習する方式について述べ、効率的な問題解決への応用について論じた。今後は、高速な学習アルゴリズムの開発を進めていく予定である。

謝辞

本研究は、第５世代コンピュータプロジェクトの一環として行われたものである。日頃御世話になっている(財)新世代コンピュータ技術開発機構 新山室長に感謝いたします。

参考文献

[1] Koseki, Y., Nakakuki, Y., and Tanaka, M., "An adaptive model-Based diagnostic system," *Proc. PRICAI'90*, Vol. 1, pp. 104–109, 1990.

[2] Nakakuki, Y., Koseki, Y., and Tanaka, M., "Inductive learning in probabilistic domain," *Proc. AAAI-90*, Vol. 2, pp. 809–814, 1990.

[3] Quinlan, J. R., "Induction of decision trees," *Machine Learning*, Vol. 1 (1), pp. 81–106, 1986.

[4] Rissanen, J., "Modeling by shortest data description," *Automatica*, Vol. 14, pp. 465–471, 1978.

[5] Rissanen, J., "A universal prior for integers and estimation by minimum description length," *Ann. of Statist.*, Vol. 11, pp. 416–431, 1983.

[6] 中島洋一郎、古関毅幸、田中みどり「確率モデルの学習方式と診断への応用」 情報処理学会研究報告（１月発表予定） 1990.