1N-1　不確実な知識を用いた推論システムの構築とその評価

鈴木 佐俊† 鈴木 誠† 平澤 茂一†
† 早稲田大学大学院理工学研究科

1 はじめに

昨今，大規模なデータベースを分析し，マーケティング等に応用する手法が注目されている．これらのデータベースに蓄積したデータには通常不確実性が含まれる．そのため，不確実性を取り扱う推論手法の一つとして，古くからBayesian Network(BN)[1]が注目されてきた．

佐藤らはBNをもとに帰納推論と演繹推論に大別し，不確実性を含む推論の枠組みを提案した[2]．また，鈴木誠らは演繹推論に焦点をあて，取り扱う不確実性を2つに分類したモデル化を行い，従来BNが扱ってきたモデルクラスよりも広いモデルクラスを扱うことのできる演繹推論手法を提案した[3]．

本研究は佐藤らの提案した基礎理論を下に鈴木誠らの演繹推論手法を用いた推論システムを構築する．そして実際のデータによるシミュレーションを行い，システムの有効性を示す．

2 従来研究

確率世界論理に基づく佐藤らの推論方式の枠組みではBNを拡張して，ある事象同士の確率的関係を2値属性の多元分割表形式で表現する．

帰納推論法は分割表をもとに属性間の確率的関係を推論する．関係モデルを選択し，そのパラメータ推定を行う．この関係モデルから導き出された各分割表セルの推測値が演繹推論の事前知識(d確信度)として利用される．

演繹推論とはいくつかの基本式が与えられた下で，残りの基本式の条件付確率を求める問題であり，確率変数の推定を行うプロセスである．

3 不確実性を含む推論システムの概要

本研究で構築したシステムは，佐藤らの枠組み[2]を下に鈴木誠らによる演繹推論手法[3]を用いた．データベース中の各属性を基本式と呼ぶ．基本式数がk個の場合，観測される個体は2^k個のセルのいずれかに分けられる．この2^k個のセルの個体頻度c_iを格納した分類表が分割表である．この分割表を帰納推論部に入力する．

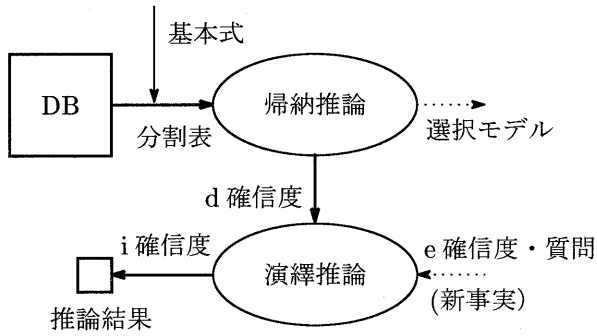

図 1: 構築システムの概要

不完全データ，欠損データは無いと仮定する．帰納推論における候補モデルは対数線形モデル(LLM)を使用した．LLMにはある交互作用項がない場合にそれを含むより高次の交互作用項は存在しないという階層クラスの仮定がある．3属性(i,j,k)のLLMは式(1)で表される．

$$\begin{aligned}\log p_{ijk} &= \mu+\theta_i+\theta_j+\theta_k+\theta_{ij}+\theta_{ik}+\theta_{jk}+\theta_{ijk}\\ &= \log p_{000}+\log\frac{p_{100}}{p_{000}}+\log\frac{p_{010}}{p_{000}}+\log\frac{p_{001}}{p_{000}}\\ &\quad+\log\frac{p_{000}p_{110}}{p_{100}p_{010}}+\log\frac{p_{000}p_{101}}{p_{001}p_{100}}\\ &\quad+\log\frac{p_{000}p_{011}}{p_{001}p_{010}}+\log\frac{p_{100}p_{010}p_{001}p_{111}}{p_{000}p_{110}p_{101}p_{011}}\end{aligned}\quad(1)$$

帰納推論部からは選択されたLLMの1つとd確信度が出力される．LLMを仮定するとき，分割表内の値0のセルの取り扱いが問題となるが，各0セルに対して共通の微小量ϵを与える手法を用いた[4]．このシステムでは候補モデルの中から1つのモデルを選択する方法としてAICによるモデル選択を行った．

演繹推論部ではドメイン全体の不確実性に対応するd確信度と，観測した新事実としてe確信度が入力として与えられ，質問となる基本式のi確信度が推論結果として出力される（図1）．

4 実験

本研究では演繹推論に入力するd確信度として(1)すべての候補モデルについてAIC最小のものをモデルとして選択する方法と，(2)分割表の各セルの値c_iをデータ数nで割った値$\bar{p}_i = c_i/n, (\bar{P} \ni \bar{p}_0,\ldots,\bar{p}_i)$

A Reasoning System Using Uncertain Knowledge and its Evaluation

† Satoshi Suzuki, Makoto Suzuki, Shigeichi Hirasawa
Graduate School of Science and Engineering, Waseda University (†)
3-4-1 Ohkubo,Shinjuku,Tokyo 169-8555,Japan

をd確信度として使用する方法，で実験を行う．

4.1 実験方法

仮想データによる実験 k個の属性を持ち，r項($r \leq 2^k$)の任意のモデルを真のモデルM^*とし，それをもとに1万件の仮想データD'を作成する（図2）．M^*からは真のd確信度d^*が計算できる．

D'はM^*と乱数を用いて作成され，正規分布に従う誤差を含む．D'の$\bar{P}$はd'である．またD'から帰納推論を行い，推定したモデル$\hat{M}$から求めたd確信度を$\hat{d}$とする．$d^*, d', \hat{d}$をもとに演繹推論を行った結果のi結合確信度を$i^*, i', \hat{i}$とする．

d'，$\hat{d}$とd^*の差はそれぞれの誤差と言える．仮想データを用いた実験では$d', \hat{d}$それぞれに含まれる二乗誤差を求める．

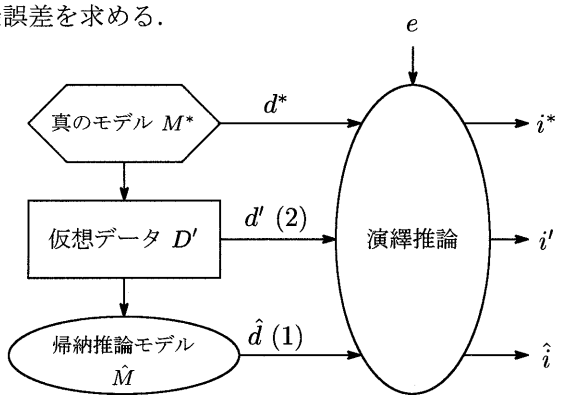

図2: 仮想データによる実験

実データによる実験 構築したシステムが実際のデータに対してどの程度対応できるのかを評価する．以下に示す2種類の実データ（D_1, D_2）について帰納推論，演繹推論を行った．

D_1：車に関するアンケート[6]

D_2：甲状腺疾患の患者に関するデータ[6]

4.2 結果

$\hat{d}$に含まれる誤差は帰納推論の際に真のモデルを選択することができれば小さくなると考えられるが，実際には$\hat{d}$，d'ともに同じ傾向を示し，$k = 4$のとき0.004付近で横ばいであった．M^*がLLMの階層クラスに含まれている場合には，$\hat{d}$の二乗誤差の分散は帰納推論に使用するデータ数が増えるに従い，減少していく傾向が見られた．

実データによる実験では，表1の結果が得られた．

表1: 実データ試験による誤答率

	k	N	(1)AIC	(2)$\bar{P}$
D_1	5	863	0.265357	0.265357
D_1	6	863	-	0.237543
D_1	7	863	-	0.165701
D_2	4	4433	0.037672	0.066772
D_2	5	4433	0.066772	0.066998

5 考察とまとめ

M^*がLLMの階層クラスに含まれていない場合，AICを用いてM^*を選択することはできないが，平均二乗誤差は真のモデルがLLMの階層クラスに含まれている場合よりも小さくなっていることがあった．これは単なる平均二乗誤差が評価基準として適切でないことを示している．本来ならば，(1)M^*を選択したか否か(2)M^*を選択した場合は二乗誤差，という2段階で損失を考えるべきであるが，今回は帰納推論をしない場合との比較を考えているため，上記のような損失は考慮しなかった．

また$d', \hat{d}$に含まれる誤差にほとんど差がなかった．帰納推論の結果として，交互作用項が削減されたモデルが選択されるということは知識圧縮が行われているということである．d確信度として$\bar{P}$を使用しても出力結果に有意な差が見られないことから圧縮によるひずみがほとんどなく推論できていると言える．

実データによる実験では，用意できたデータについてはAICによるモデル選択を行った方がd確信度として$\bar{P}$を使用するよりも誤答率が低くなる傾向があった．ただし，その差は微小であり，属性数kが増加した場合には計算量削減のために$\bar{P}$を使用することも十分可能であると考えられる．

今後は，より多くの属性，多値属性が扱えるようにシステムを改良する予定である．

参考文献

[1] Pearl,J.: Fusion, Propagation, and Structuring in Belief Networks, *Artificial Intelligence*, Vol.29, pp.241-228(1986).

[2] 佐藤淳一，松嶋敏泰，平澤茂一: 不確実な知識の表現法と推論法について，第18回情報理論とその応用シンポジウム予稿集 pp. 629-632 (1995).

[3] 鈴木 誠，松嶋敏泰，平澤茂一: 不確実な知識を用いた推論のモデル化と推論法について，情報処理学会論文誌，Vol. 41, No. 1, pp. 1-11 (2000).

[4] 松田紀之: 質的情報の多変量解析，朝倉書店 (1988).

[5] 廣津千尋: 離散データ解析，教育出版 (1982).

[6] http://www.ics.uci.edu/pub/machine-learning-databases/