記述長最小基準を用いた自然現象予知のための確率的ルール学習

6C−3

白土　保
郵政省通信総合研究所

1．まえがき

筆者は自然現象予知のための確率的ルールを学習する手法について，太陽フレア予知の分野において検討してきた[1]．しかしこの手法においては，得られた知識の良し悪しの客観的な評価が欠けていた．今回は，筆者のいままでの手法における確率的ルール生成プロセスの最終段階に記述長最小基準を適用することにより，客観的な確率的ルールの学習を試みた．仮想的な確率モデルを用いて手法の妥当性の確認をしたところ，統計的に許容される誤差の範囲内で適切な確率的ルールが学習されることが分かった．

2．自然現象予知のための確率的ルールの学習

ここでいう自然現象予知とは，ある物理対象のある時点での状態を知った時，ある自然現象が近い将来発生する確率を求めることである．図1に確率的ルールの形式を示す．

IF　$\mathbf{a}=\mathbf{a}_1$
　or　・・・
　or　$\mathbf{a}=\mathbf{a}_n$
THEN　Probability$=\hat{p}$

図1　確率的ルールの形式

図中$\mathbf{a}$，$\mathbf{a}_1$，・・・，$\mathbf{a}_n$を属性ベクトルと呼ぶ．属性ベクトルは物理対象の状態を表し，a_iの組$\{a_1, \cdots, a_m\}$の形式をとる．それぞれのa_iにはA_iあるいはA_iの範囲を示す表現が入る．A_iの範囲を示す表現については後述する．ここではA_iについてのみ説明する．A_iは，i番目の属性の属性値（連続量）の値域を一定の幅でいくつかの区間に分割し，ある区間に落ちた属性値をその区間を代表する標識で表したものである（区間の中央値を標識の名前にする）．属性ベクトルは，予知をするために十分な属性の情報をすべて含んでいるとは限らない．つまり属性ベクトルは，限られた属性空間の上に物理状態を射影したものである．確率的ルールは，ある時点の物理対象の状態を表した属性ベクトル$\mathbf{a}$がルールの条件部の記述に含まれる$\mathbf{a}_1$，・・・，$\mathbf{a}_n$のいづれかに一致するとき，ある自然現象が近い将来生起する確率がそのルールの結論部の値$\hat{p}$で与えられることを意味する．予知の期間（どの程度先を予知するか）は具体的な問題ごとに異なり，以下で述べる履歴データが与えられた時点で決められる．図2に確率的ルール学習プロセスのおおまかな流れを示す．

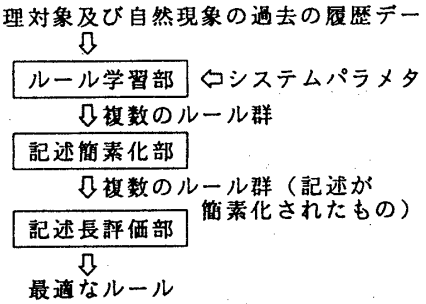

図2　ルール学習プロセス

物理対象及び自然現象の過去の履歴データは，属性ベクトルとラベルで表される．ラベルは属性ベクトルが表す状態に対して自然現象が生起した時に1，生起しなかった時に0の値をとる．この学習データがルール学習部に与えられると，複数の確率的ルール群が生成される．この際，システムパラメタを変化させることによりいろいろな複雑さをもった複数のルール群を生成する．記述簡素化部では，各ルールの条件部の記述を，そのルールは含み，他のルールは含まないような，できるだけ簡単な記述に書き換える．具体的には，Aの範囲を意味する“Ai〜Aj”や任意のAを意味する“＊”の表現を含む記述に変換する．ルール学習部及び記述簡素化部の詳細については文献(1)で詳しく報告したのでここでは省略する．記述長評価部では，ルール群の記述長を評価し，記述長最小のルールを最適なルールとして選択する．記述長Lは次式で与えられる．

Acquisition of stochastic rules for physical phenomenon prediction by MDL criterion

Tamotsu SHIRADO
Communications Research Laboratory, MPT

$$L = L1 + L2 + L3 \qquad (1)$$

$$L1 = \sum_{i=1}^{s} \hat{N}i \times H(\hat{p}i) \qquad (2)$$

$H(p)$：pの平均情報量

$$L2 = 1/2 \times \sum_{i=1}^{s} \log_2(\hat{N}i) \qquad (3)$$

ここでＬ１はルール群と学習データの適合度を示す記述長，Ｌ２は確率パラメタの記述長、Ｌ３はルール群自体の記述長である．式中$\hat{N}i$はｉ番目のルールを作るに用いられた学習データの数である．ここで，各$\hat{N}i$は$\hat{N}i = \hat{N}i^{+} + \hat{N}i^{-}$とする．ただし，$\hat{N}i^{+}$及び$\hat{N}i^{-}$はそのルールを作るのに用いられた学習データのうちラベル＝１のデータの数及びラベル＝０のデータの数である．従って結論部の生起確率$\hat{p}i$は，$\hat{p}i = \hat{N}i^{+} / \hat{N}i$で与えられる．ｓはルール群に含まれるルールの数である．

Ｌ３は，ルールの条件部に含まれるすべての属性ベクトルの形式を調べ，それらの形式で表現しうる条件部記述の場合の数を計算して，その数のlogをとったものである．計算法の理論的な根拠は文献(2)，(3)を参照されたい．

３．仮想的なデータによるテスト

本手法の妥当性を確認するために，仮想的な確率モデルを仮定して動作テストを行った．確率モデルは，３個の属性及び各属性の属性値を線形に結合した関数を設定し，それに基づいて各属性ベクトルに対する自然現象の真の生起確率を割り当てることによって構成した．この確率モデル及び一様乱数を用いて学習データを1000個発生させた．この学習データに対して本手法を適用することにより，本来の確率モデルがどの程度の精度で復元されたかを評価した．すなわち，各ルールの結論部の値と，条件部に含まれる属性ベクトルそれぞれに対する真のモデルとの間で誤差平方和を計算し，ルール群全体についての総和を計算した．これを真のモデルとの間の誤差と呼ぶことにする．図３に処理結果を示す．

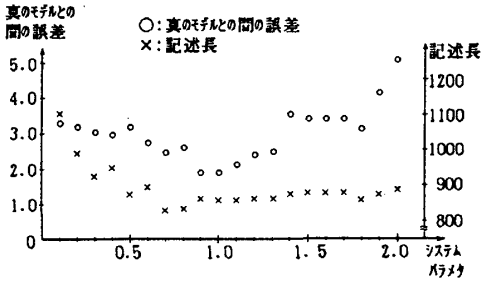

図３　テストデータ処理結果

図３をみると，真のモデルとの間の誤差および記述長の両方ともシステムパラメタに対してほぼＵ字型の変化をしていることが分かる．さらに，システムパラメタが0.7～1.1付近のときにどちらも極小になっていることが分かる．また式(4)により二項分布の分散による推定精度の評価をおこなったところ，分散値が2.86となり，システムパラメタが0.7～1.1付近の時に生成された複数のルール群と真のモデルとの間の誤差がいづれもこの範囲内に収まっていることが分かった．式中ｐｉは属性ベクトルｉに対する真の生起確率，Niは学習データ中の属性ベクトルｉの出現頻度を表す．ｔは異なった属性ベクトルの総数を示す．

$$\sum_{i=1}^{t} p\,i(1-p\,i)/Ni \qquad (4)$$

以上をまとめるとこのテスト結果に関しては，本手法によって生成された確率的ルール群のうち記述長最小（あるいはその近く）を与える確率的ルールが統計的に妥当な精度で真のモデルを復元していることが確認された．

４．むすび

記述長最小基準を自然現象の予知のための確率的ルールの学習に適用した手法を提案し，仮想モデルを用いたテストを行なうことにより手法の有効性を確認した．今回はひとつの実験結果に関してのみ報告したが，今後はデータ点数や確率モデルなどを変えテストを重ねることより，本手法がどの程度一般性を持ち，どの程度有効であるかを検証して行く．また実際の問題への適用も検討していく．

謝辞

日頃ご指導頂く電気通信大学伊藤秀一助教授，ならびに記述長最小基準の適用に関しご示唆頂いた日本電気大野和彦氏に深謝致します．また本研究の機会を与えていただいた塩見支所長，柳田室長，小川センター長，猪木室長に感謝致します．

参考文献

(1)白土　保：多変量クラスタ分析手法に基づく知識獲得機能を備えた太陽フレア予報システムの開発－ﾌﾟﾛﾄﾀｲﾌﾟ－，通信総研季報，Vol.38，No.1，pp.1-14(1992).

(2)山西健司，韓太舜：ＭＤＬ入門:情報理論の立場から，人工知能学会誌，Vol.7，No.3，pp45-52(1992).

(3)山西健司：ＭＤＬ入門:計算論的学習理論の立場から，人工知能学会誌，Vol.7，No.3，pp53-60(1992).