日本オペレーションズ・リサーチ学会
2004年春季研究発表会

作物栽培環境・品質データからのデータマイニング手法

01013495 広島県立大学 *奥原 浩之 OKUHARA Koji,
広島県立大学 濱谷 久美子 HAMATANI Kumiko,
広島県立大学 田中 稔次朗 TANAKA Toshijiro,
01005194 大阪大学 石井博昭 ISHII Hiroaki

1．はじめに

近年では情報技術の発展により大規模かつ分散したデータをまとめて取り扱うことが実現されてきている．それゆえ，収集されたデータに目的とする分析と関連ない情報が混在する傾向も高くなる．このことは不適切なモデル作成によるあやまった分析結果を導く恐れが増すことを意味している．そのような場合でも客観的に内包されているルールを抽出できる理論としてラフ集合 [1] による分析がある．

本研究では観測データから順モデルを構成することで着目する二変数間の関係を明らかにする手法，ならびに逆モデルを構成することで着目する変数の望ましい値を得るため制御可能な変数の値の設定法，さらに着目する変数の条件付確率密度関数の推定法といった3つの手法が実現できるシステム [2] に適用可能なラフ集合によるデータマイニングを提案することで，ラフ集合によるルール抽出と確率ニューラルネットの融合を図ることを目的とする．

2．逆モデルを構成するシステム

いま，確率ベクトル $\mathbb{Z}_s$ に任意の定数行列 $\mathbb{B}^{\mathrm{T}} \in \Re^d \times \Re^d$，定数ベクトル $\mathbb{m} \in \Re^d$ によるアフィン変換を施した結果，

$$\mathbb{Z}'_s = \mathbb{B}^{\mathrm{T}}\mathbb{Z}_s + \mathbb{m}$$

となり，$\mathbb{Z}'_s$ は $[\mathbb{X}_s, \mathbb{Y}_s]^{\mathrm{T}} \in \Re^d$ で構成されるものとする．ここで，$\mathbb{X}_s \in \Re^n$，$\mathbb{Y}_s \in \Re^m$ であり，$d = n + m$ である．いま，確率ベクトル $\mathbb{X}_s$，$\mathbb{Y}_s$ の共分散行列を $\mathbb{C}^x \in \Re^n \times \Re^n$，$\mathbb{C}^y \in \Re^m \times \Re^m$，平均ベクトルを $\mathbb{m}^x \in \Re^n$，$\mathbb{m}^y \in \Re^m$ とし，$\mathbb{X}_s$ と $\mathbb{Y}_s$ の相関行列を $\mathbb{C}^{xy} \in \Re^n \times \Re^m$ とする．

このとき，第 k 入力ニューロンにおいて確率ベクトル $\mathbb{Z}'_s$ は平均ベクトル $\mathbb{m}'_k = \mathbb{B}^{\mathrm{T}}\mathbb{m}_k + \mathbb{m}$，共分散行列 $\Sigma'_k = \mathbb{B}^{\mathrm{T}}\Sigma_k\mathbb{B}$ をもつ確率密度関数

$$p_k(\mathbb{Z}'_s|\phi'_k) \quad = \quad N_d(\mathbb{Z}'_s, \phi'_k)$$

に従う．さらに，

$$p(\mathbb{Z}'_s|\mathrm{w}, \phi') \quad = \quad \sum_{k=1}^{K} p(k) N_d(\mathbb{Z}'_s, \phi'_k)$$

に従う．ここで，パラメータ ϕ'_k は集合 $\{\mathrm{m}'_k, \Sigma'_k\}$ を表し，ϕ' で集合 $\{\phi'_1, \phi'_2, \cdots, \phi'_K\}$ を表す．

システムの出力ベクトルを

$$\mathrm{E}[\mathbb{Y}_s|\mathbb{X}_s] = \int_{\Re^m} \mathbb{y} p(\mathbb{Y}|\mathbb{X}_s, \theta') d\mathbb{y}$$

で与えることとする．ここで，パラメータ θ' で w と ϕ' の集合を表す．システムにおける条件付き確率密度関数は

$$p(\mathbb{Y}|\mathbb{X}_s, \theta') = \sum_{k=1}^{K} \alpha(k) p_k(\mathbb{Y}|\mathbb{X}_s, \phi'_k)$$

である．ただし，

$$\alpha(k) = \frac{p(k) p_k(\mathbb{X}_s|\phi'_k)}{\sum_{k=1}^{K} p(k) p_k(\mathbb{X}_s|\phi'_k)}$$

であり，システムの出力ベクトルは

$$\begin{aligned} \mathrm{E}[\mathbb{Y}_s|\mathbb{X}_s] \quad &= \quad \sum_{k=1}^{K} \alpha(k) \int_{\Re^m} \mathbb{y} p_k(\mathbb{Y}|\mathbb{X}_s, \phi'_k) d\mathbb{y} \\ &= \quad \sum_{k=1}^{K} \alpha(k) \mathrm{E}_k[\mathbb{Y}_s|\mathbb{X}_s] \end{aligned}$$

となる．いま，共分散行列 $\mathbb{C}^x_k$ の逆行列を $\{\mathbb{C}^x_k\}^{-1}$ で表し，行列 $\mathbb{D}_k$ を $\mathbb{D}_k = \mathbb{C}^{yx}_k \{\mathbb{C}^x_k\}^{-1}$ で与えると，

$$\mathrm{E}_k[\mathbb{Y}_s|\mathbb{X}_s] = \mathbb{m}'^y_k + \mathbb{D}_k(\mathbb{x}_s - \mathbb{m}'^x_k)$$

であることから，

$$\mathrm{E}[\mathbb{Y}_s|\mathbb{X}_s] = \sum_{k=1}^{K} \alpha(k) \left\{\mathbb{m}'^y_k + \mathbb{D}_k(\mathbb{x}_s - \mathbb{m}'^x_k)\right\}$$

となる [3]．

また，第 k 入力ニューロンにおける条件付き確率密度関数 $p_k(\mathbb{Y}|\mathbb{X}_s, \phi'_k)$ は平均ベクトル $\mathrm{E}_k[\mathbb{Y}_s|\mathbb{X}_s] \in \Re^m$，共分散行列 $\mathbb{D}'_k \in \Re^m \times \Re^m$

$$\mathbb{D}'_k = \mathbb{C}^y_k - \mathbb{D}_k \mathbb{C}^x_k \mathbb{D}^{\mathrm{T}}_k$$

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．

をもつ m 次元正規分布に従うことより，システムにおける条件付き確率密度関数 $p(\mathbf{Y}|\mathbf{X}_s, \theta')$ を求めることもできる．各ニューロンの平均ベクトルと共分散行列の学習はEMアルゴリズムにより推定することができる [4].

3．平均・分散を利用したルール抽出

いま，L 個のサンプルについて N 個の条件属性と M 個の決定属性からなる決定表を考える．確率ニューラルネットによる各ニューロンの平均ベクトルと共分散行列を利用してラフ集合によるルール抽出を行うことを考える．そこで，m 番目の決定属性を R 個のクラス

$$C_m^s \cap C_m^t = \phi, (s \neq t), \quad C_m^R \succ \cdots \succ C_m^r \succ C_m^1$$

に分類する．このとき，$x \in U$ が与えられると，少なくともクラス C_m^r に属している U の要素の集合である上側累積集合 $C_m^{\geq r}$ と，たかだかクラス C_m^r に属している U の要素の集合である下側累積集合 $C_m^{\leq r}$ が，

$$C_m^{\geq r} = \bigcup_{s \geq r} C_m^s, \quad C_m^{\leq r} = \bigcup_{s \leq r} C_m^s$$

で定義できる．

すべての基準の集合を W とするとき，$V \subseteq W$ を考える．任意の $v \in V$ について，$xO_m^v y$ が成立するとき，x は V において y を支配するといい $xD_m^V y$ で表し，区間値 $S[s_1, s_2]$, $T[t_1, t_2]$ が与えられたとき，属性のカテゴリ間の順序付けは

$$S[s_1, s_2] \succeq T[t_1, t_2] \leftrightarrow s_1 \geq t_1, \ s_2 \geq t_2$$

で行うものとする．これを $xD_m'^V y$ で表し，以下のように定義する [5].

$$xD_m'^V y \leftrightarrow g(x, n) \succeq g(y, n), \quad (\forall v \in V)$$

m 番目の決定属性において，$x \in U$ が与えられると，V において x を支配する U の要素の集合 $D_m^{+V}(x)$ と，V において x に支配される U の要素の集合 $D_m^{-V}(x)$ が

$$D_m^{+V}(x) = \{y \in U | y D_m^V x\}, \ D_m^{-V}(x) = \{y \in U | x D_m^V y\}$$

で定義できる．

支配集合 $D_m^{+V}(x)$ による累積集合 $C_m^{\geq r}$ の下近似集合 $V_*(C_m^{\geq r})$ と上近似集合 $V^*(C_m^{\geq r})$ は

$$V_*(C_m^{\geq r}) = \{x \in U | D_m^{+V}(x) \subseteq C_m^{\geq r}\},$$
$$V^*(C_m^{\geq r}) = \bigcup_{x \in C_m^{\geq r}} D_m^{+V}(x)$$

で定義できる．$V_*(C_m^{\geq r})$ に関する下近似集合から得られたルールと $V_*(C_m^{\leq r})$ に関する下近似集合から得られたルールの条件部において，すべての属性においてカテゴリが同じとなる集合から，確実にある $x^* \in U$ がクラス r に属するという次の if-then ルールが得られる．

IF $g(x^*, n_1) = g(x, n_1)$ *and* $g(x^*, n_2) = g(x, n_2)$
$\cdots$ *and* $g(x^*, n_N) = g(x, n_N)$,
THEN $x^* \in C_m^r$.

この抽出されたルールを満たすデータは確実にクラス r に属していることとなる．

5．おわりに

提案したシステムでは観測データを独立なデータに変換する前処理を施し，その後で中間処理として確率ニューラルネットの学習を利用する．元の観測データの各変数間に存在した相関は，前処理時に得られた情報を用いた後処理において考慮される．提案手法では中間処理において変数間に独立性が成立しているため，逆モデルの構成法が容易に実現でき，さらに，確率ニューラルネットによる学習時にパラメータの振動を抑える．そのうえで，確率ニューラルネットで得られた平均と分散の情報から，ラフ集合によるルール抽出を行うことが可能となる．

参考文献

[1] 日本ファジィ学会編, “ファジィとソフトコンピューティングハンドブック,” 共立出版株式会社, 2000.

[2] 福水健次，渡辺澄夫, “統計的推論を実現するニューラルネットワークとそのパターン認識への応用” 信学技報, NC92-36, pp. 83-90, 1992.

[3] 奥原浩之，石井博昭，内田誠, “ニューラルシステムを用いたデータマイニングによる意思決定支援,” 電子情報通信学会論文誌, Vol. J86-DII, No. 4, pp. 535-542, 2003.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm,” *J. Royal Statist. Soc. Ser. B*, **39**, pp. 1-38, 1977.

[5] 杉原一臣，石井博昭，田中英夫, “ラフ集合による新しいコンジョイント分析の提案,” 日本知能情報ファジィ学会誌, **15**, No. 4, pp. 421-428, 2003.

© 日本オペレーションズ・リサーチ学会．無断複写・複製・転載を禁ず．