2007－MPS－63 (9)
2007／3／3

重要度ランク付きパターン識別問題に対するファジィシステムの構築

横田泰之　中島智晴　石渕久生

本論文では，パターン識別問題における重要度ランクを考慮したファジィパターン識別器の構築手法を提案する．癌の医療診断のように，正答率が多少犠牲になっても損失をできるだけ抑えることが必要なパターン識別問題が実世界には数多く存在する．このような場合に識別の重要度を考慮したパターン識別器の構築が必要となる．本論文では，識別の重要度がパターンごとにランクの形式で与えられているという仮定のもとでのファジィパターン識別器を構築する手法を提案する．提案手法により，重要度ランクを考慮したファジィIf-Thenルールが学習用パターンから生成される．数値実験から，提案手法を用いると，パターンの重要度ランクを考慮したファジィパターン識別器が構築できることを示す．

Construction of Fuzzy Systems for Pattern Classification Problems with Importance Ranks

YASUYUKI YOKOTA, TOMOHARU NAKASHIMA and HISAO ISHIBUCHI

In this paper we propose a method for constructing fuzzy classificationsystems that considers importance ranks of classification. As in medicaldiagnosis problems, there are many real-world pattern classificationproblems where minimizing misclassification cost is necessary even ifclassification accuracy is degraded. In this case classification systemsthat consider classification importance are useful. It is assumed inthis paper that the classification importance of a pattern is given as arank. Fuzzy if-then rules are generated from the given training patternsunder this assumption. Experimental results show that constructed fuzzyclassification systems by the proposed method can successfully considerthe classification importance.

1. はじめに

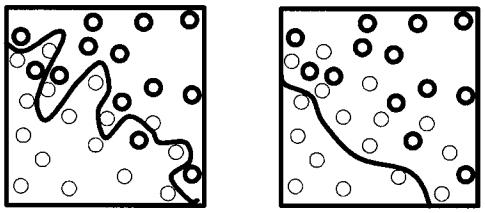
ファジィルールに基づくファジィシステムは，ニューラルネットワークと同様に高い近似能力を持つ非線形システムであることが知られている[1]．ファジィシステムは，制御や関数近似，パターン識別といった様々な分野で利用されている[2]．上記の研究の他にも，ファジィシステムの性能を落とすことなく解釈可能性を最大化する研究も近年よく行われている[3]．

一般に，パターン識別問題に対してファジィシステムやニューラルネットワークといった学習システムを適用して識別器の設計を行う研究では，学習用データに対する正答率の最大化のみを目的としている場合がほとんどである．しかし，実世界の問題では，正答率の最大化のみでは適切とはいえない場合が数多く存在する．例えば，癌の医療診断を例に挙げる．この例における誤診断として，(a) 陽性（癌である）の患者を陰性（癌ではない）とする誤診断と，(b) 陰性の患者を陽性とする誤診断の２種類が考えられる．どちらの誤診断もできるだけ起こらないようにするべきではあるが，(a) と (b) を比較すると，(a) は (b) に比べて許容できない誤診断である．これは，陽性の患者に対する診断の重要度が，陰性の患者に対する診断の重要度より高いことが理由と考えられる．誤識別の種類により異なる損失が発生する問題例として，他にもダイレクトマーケティングや侵入発見，詐欺発見が考えられる．

本論文では，このような問題を考慮し，パターン $\mathbf{x}_p$, $p = 1, 2, ..., m$ に対して，重要度ランク $Rank_p$ が与えられている状況を仮定する．ここで，パターン $\mathbf{x}_p$ は，$Rank_p$ が１の場合に最も重要なパターンであり，重要度ランク $Rank_p$ が大きくなるにつれて重要ではなくなる．よって，ランク１のパターンはランク２のパターンより，ランク２のパターンはランク３のパターンより優先的に識別するような識別器が必要不可欠となる．

そこで，本論文では，定量的ではない重要度を考慮する状況に対して適用可能である，重要度ランクを考慮したファジィパターン識別器の構築手法を提案する．図１に，本論文で構築するパターン識別器と従来のパ

ターン識別器の違いをイメージで示す．



(a) 従来手法　　(b) 提案手法

図 1　提案手法のイメージ（太線の○：ランク 1 のパターン，○：ランク 2 のパターン）．

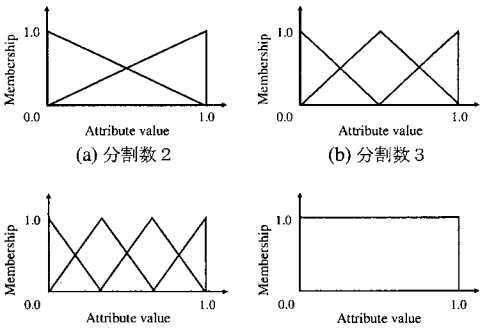


図 2　三角型ファジィ集合と *don't care*

図 1(a) は，従来手法により構築されたパターン識別器の識別境界を示している．従来手法では，パターンが平等に扱われているために，誤識別の数がどちらの重要度ランクのパターンに対してもほぼ同数存在する．一方，図 1(b) では，ランク 1 のパターンを誤識別しないようなパターン識別器が構築されているため，ランク 2 のパターンには誤識別があるが，ランク 1 には誤識別は無い．

正答率最大化を目的とする場合，パターン識別器の性能評価指標としては全体の正答率のみでよいが，本論文では重要度ランクが与えられている状況を考えているため，パターン識別器の性能評価指標として，全てのパターンを含めたランクごとの正答率 CR_q を評価関数として用いることにする．

$$f_q(S) = CR_q, \quad q = 1, 2, ..., Q \tag{1}$$

ここで，S は識別システム，q は重要度ランク，Q は全てのクラスを含めた最大ランク数を示す．また，全体の正答率 CR_{all} も用いる．

本研究では，石渕ら[4] が提案しているファジィ If-Then ルールの生成手法を改良し，重要度ランクを考慮したコストセンシティブなファジィパターン識別器を構築する手法を提案する．

2. ファジィパターン識別器

まず，本章では，従来手法である[4] の手法について，ファジィ If-Then ルール，メンバシップ関数，結論部クラスと確信度の生成手順について述べる．

n 次元 M クラスパターン識別問題として，m 個の学習用パターン $\mathbf{x}_p = (x_{p1}, x_{p2}, ..., x_{pn})$，$p = 1, 2, ..., m$ が与えられていると仮定する．ここで，x_{pi} は p 番目の学習用パターン $\mathbf{x}_p$ における i 番目の属性値である．また，一般性を失うことなく，パターン入力空間は n 次元超立方体 $[0, 1]^n$ であるとする．つまり，全ての属性値が区間 $[0, 1]$ 内の実数値であると仮定する．

本論文のファジィパターン識別器では，以下の形式のファジィ If-Then ルールを用いる．

$$R_j: \text{ If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{and } x_n \text{ is } A_{jn} \text{ then Class } C_j \text{ with } CF_j \tag{2}$$

ここで，R_j は j 番目のファジィルールを示すルールラベル，A_{ji} は i 番目の属性に対する条件部ファジィ集合，C_j は結論部クラス，CF_j はルールの確信度である．パターン空間をファジィ分割するメンバシップ関数として，三角型ファジィ集合と *don't care* を用いることにする（図 2）．図 2 には，それぞれ属性軸を 2 分割，3 分割，4 分割する三角型メンバーシップ関数と，メンバシップ値が常に 1.0 である *don't care* が示されている．

i 番目の属性軸を L 分割する場合の r 番目のファジィ集合 A_r^L 三角型メンバシップ関数は以下のように定式化することができる（$i = 1, 2, \ldots, n$, $r = 1, 2, \ldots, L$）．

$$\mu_r^L(x) = \max\{1 - \frac{|\, x - a_r^L \,|}{b}, 0\} \tag{3}$$

ここで，a_{ir} と b は以下で定義される．

$$a_r^L = \frac{(r-1)}{(L-1)} \tag{4}$$

$$b = \frac{1}{(L-1)} \tag{5}$$

2.1 結論部クラスと確信度

文献 4) で提案されているファジィ If-Then ルールの同定手法では，ファジィ If-Then ルールの条件部をあらかじめ決めておけば，学習用パターンから自動的に結論部クラスと確信度が決定される．以下に，ファジィ If-Then ルール R_j の結論部クラス C_j と確信度 CF_j の生成手順を示す．

【ファジィパターン識別器の生成】

Step 1: ファジィIf-Thenルール R_j の条件部 $A_{j1}, A_{j2}, \ldots, A_{jn}$ に対して，各クラスの $\beta_h(R_j)$, $h = 1, 2, \ldots, M$ を計算する．

$$\beta_h(R_j) = \sum_{\mathbf{x}_p \in \text{Class } h} \{\mu_{j1}^L(x_{p1}) \cdot \ldots \cdot \mu_{jn}^L(x_{pn})\} \quad (6)$$

ここで，$\mu_{ji}(\cdot)$ はファジィ集合 A_{ji} のメンバシップ関数である．

Step 2: $\beta_h(R_j)$ が最大であるクラス $\hat{h}$ をファジィIf-Then ルール R_j の結論部クラスとする．

$$\beta_{\hat{h}}(R_j) = \max\{\beta_1(R_j), \ldots, \beta_M(R_j)\} \quad (7)$$

複数のクラスに対応する $\beta_h(R_j)$ が最大値をとる場合は，ルール R_j の結論部クラス C_j を決定することはできない．この場合，結論部クラスを $C_j = \phi$ と定める．一方，最大値をとるクラスが一意に決まる場合は，ファジィIf-Then ルールの結論部クラスを $\hat{h}$ と決定する．

Step 3: Step 2 で結論部クラス C_j が決定された場合，確信度 CF_j を次のように定める．

$$CF_j = \frac{\beta_{\hat{h}}(R_j) - \bar{\beta}}{\sum_{h=1}^{M} \beta_h(R_j)} \quad (8)$$

ここで，

$$\bar{\beta} = \frac{1}{M-1} \sum_{\substack{h=1 \\ h \neq C_j}}^{M} \beta_h(R_j) \quad (9)$$

2.2 未知パターンに対するファジィ推論

ファジィパターン識別器 S による未知パターン $\mathbf{x}$ の識別手順を以下に示す．

【未知パターン $\mathbf{x}$ のファジィ推論】

Step 1: 各クラスの α_h, $h = 1, 2, ..., M$ を計算する．

$$\alpha_h(\mathbf{x}) = \max\{\mu_{j1}^L(x_1) \cdot \ldots \cdot \mu_{jn}^L(x_n) \cdot CF_j | \ C_j = h\} \quad (10)$$

Step 2: $\alpha_h(\mathbf{x})$ の中で最大値を取るクラス $\hat{h}$ として未知パターン $\mathbf{x}$ を識別する．

$$\alpha_{\hat{h}}(\mathbf{x}) = \max\{\alpha_1(\mathbf{x}), \ldots, \alpha_M(\mathbf{x})\} \quad (11)$$

複数の $\alpha_h(\mathbf{x})$ が最大となる場合，未知パターン $\mathbf{x}$ の識別を拒否し，識別不能のパターンとする．また，未知パターン $\mathbf{x}$ に適合するルールが存在しない場合も，未知パターン $\mathbf{x}$ を識別不能とする．そうでない場合（すなわち，$\hat{h}$ が唯一存在する場合），未知パターン $\mathbf{x}$ をクラス $\hat{h}$ と識別する．

3. ランクを考慮したファジィパターン識別器

以下の手順により重要度ランクを考慮したファジィIf-Then ルールを生成する．

【重要度ランクを考慮したファジィパターン識別器の生成】

Step 1: ルール R_j が含む空間に存在するクラス h, 重要度ランク $Rank_p = q$ のパターン数 n_j^{hq}, $q = 1, 2, ..., Q$, $h = 1, 2, ..., M$ を計算する．ここで，ルール R_j が含む空間とは，ルール R_j に対応するメンバシップ関数の積 $\mu(\cdot)$ が $\mu(\cdot) > 0$ となる空間を示す．

Step 2: クラス h のパターンが存在する最大重要度ランク $\hat{q}^h$ を決定する．

$$\hat{q}^h = \min\{\ q \mid n_j^{hq} > 0,\ q = 1, 2, ..., Q\ \} \quad (12)$$

Step 3: 各クラスの $\hat{q}^h$ を比較して最大重要度のランク $q_{\min}$ を決定する．

$$q_{\min} = \min\{\hat{q}^1, \hat{q}^2, ..., \hat{q}^M\} \quad (13)$$

Step 4: 各クラスのパターン数 $n_j^{hq_{\min}}$ を比較して，最大であるクラス $\hat{h}$ をファジィIf-Then ルール R_j の結論部クラスとする．

$$n_j^{\hat{h}} = \max\{n_j^{1q_{\min}}, n_j^{2q_{\min}}, ..., n_j^{Mq_{\min}}\} \quad (14)$$

複数のクラスに対応する $n_j^{\hat{h}}$ が最大値をとる場合は，ルール R_j の結論部クラス C_j を決定することはできない．この場合，結論部クラスを $C_j = \phi$ と定める．一方，最大値をとるクラスが一意に決まる場合は，ファジィIf-Then ルールの結論部クラスを $\hat{h}$ と決定する．

上記の手続きでは，重要度ランクを考慮して n_j^{hq} を用いることにより，学習用パターンの重要度ランクはファジィIf-Then ルールの結論部クラス生成に影響している．

確信度は，式 (6) で求めた β_h の比を用いて，以下のように求められる．

$$CF_j = \frac{\beta_{\hat{h}}(R_j)}{\sum_{h=1}^{M} \beta_h(R_j)} \quad (15)$$

ここで，式 (8) と比較すると，$\bar{\beta}$ を用いていない点が異なっている．

4. 数値実験

4.1 ランク付け

本論文では実験的に，パターンの重要度ランクはク

表 1　数値実験で用いた 8 種類のパターン集合

パターン集合	クラス数	パターン数	属性数	
			離散	連続
Appendix	2	106	7	0
Balance scale	3	625	4	0
Breast W	2	683	9	0
CMC	3	1473	7	2
Glass	7	214	0	9
Haberman	2	306	0	3
Hayes Roth	3	132	4	0
Iris	3	150	0	4

表 2　学習用パターンに対する評価実験におけるランク 1 の正答率 CR_1 の提案手法と従来手法を比較した実験数

CR_1 が従来以上の実験数	CR_1 が従来未満の実験数
104	0

ラスごとに与えた．

具体的には，ラスごとの場合，クラス h に属するパターンの重要度ランク $Rank_p$ を h と決定した．これをランク付け手法 1 とした．

4.2 実験結果

数値実験では，UCI Machine Learning Repository から 8 つのパターン集合（表 1）を用いて，提案手法と従来手法の学習用パターンに対する性能評価および評価用パターンに対する性能評価を行った．

評価用パターンに対する性能評価では，ten-fold cross-validation (10-CV) を 10 回繰り返してファジィパターン識別器の性能を調査した．

まず，学習用パターンに対する性能を調べた．各属性の分割数 L を $L = 3, 4, \ldots, 15$ と設定して数値実験を行った．

全ての実験結果を集計したものを表 2 に示す．この表は，従来手法の正答率 CR_1 と提案手法の正答率 CR_1 を比較し，提案手法の正答率 CR_1 が上がった実験数と下がった実験数を示している．

この実験結果より，提案手法を用いることで，重要度ランクの大きなパターンを正しく識別する識別器が構築されていることがわかった．

次に，評価用パターンに対する性能評価（10-CV）を行った．実験結果をまた，全ての数値実験を集計したものを表 3 に示す．表 2 と同様に，10-CV における全実験設定の組み合わせの実験結果数 104 個（パターン集合数 8 個，分割数 13 種類，ランク付け手法数 1 個）から，従来手法の正答率 CR_1 と提案手法の正答率 CR_1 を比較し，提案手法の正答率 CR_1 が上がった実験数と下がった実験数を示している．

表 3　10-CV におけるランク 1 の正答率 CR_1 の提案手法と従来手法を比較した実験数

CR_1 が従来以上の実験数	CR_1 が従来未満の実験数
104	0

この実験結果より，提案手法を用いることで，未知パターンに対する識別においても，重要度ランクの大きなパターンを正しく識別する識別器が構築されていることがわかった．

5. おわりに

本論文では，パターンの重要度ランクが与えられている状況でのコストセンシティブなファジィパターン識別器の構築手法を提案した．提案手法は，石渕ら[4]が提案したファジィ If-Then ルールの同定手法を改良したものであり，学習用パターンの重要度がルール生成の過程で考慮されている．

数値実験では，学習用パターンと評価用パターンに対する性能評価（10-CV）を様々な設定条件下で行った．数値実験より，ファジィパターン識別器を構成するファジィ If-Then ルールの生成において重要度を考慮することにより，従来手法と比較して重要度ランクの高いパターン集合の正答率を上げることができた．今後は，パターンごとにランクが与えられている場合の性能評価，理論的な性能評価，学習アルゴリズムによる正答率の向上，またコストセンシティブ識別器を構築してアンサンブル識別器を構築し，他のアンサンブル手法との比較を行うことを考えている．

参考文献

1) Kosko, B.: Fuzzy Systems as Universal Approximators, *Proc. of IEEE International Conference on Fuzzy Systems*, pp. 1153–1162 (1992).
2) Ishibuchi, H., Nozaki, K. and Tanaka, H.: Distributed Representation of Fuzzy Rules and Its Application to Pattern Classification, *Fuzzy Sets and Systems*, Vol. 52, No. 1, pp. 21–32 (1992).
3) Casillas, J., Cordon, O., Herrera, F. and Magdalena, L.: *Trade-off between Accuracy and Interpretability in Fuzzy Rule-based Modelling* (2002).
4) 石渕 久生，野崎 賢，田中英夫：分散型ファジィルールを用いたパターン認識，システム制御情報学会論文誌，Vol.4, No.12, pp.517–526 (1991).