

ファジィ集合を用いた概念形成モデルの提案 A Proposal of Conceptual Clustering Model Using Fuzzy Sets

林 勲*, 津田 エミ子*, 小澤 順**

I.Hayashi, E.Tsuda, J.Ozawa

* 阪南大学 経営情報学部

** 松下電器産業(株) 中央研究所

* Dept. of Management and Computer Sciences, Hannan University

** Central Research Laboratories, Matsushita Electric Industrial Co., Ltd.

Abstract. A new conceptual clustering model is discussed here to construct conceptual trees using membership values. In this paper, a new similarity measure is first defined to scale the relationships between membership values in a class or classes. Next, we discuss how to assign the data to any class of the conceptual tree. we define here the four kinds of operators to do it, which are placing operator, creating operator, merging operator and splitting operator. Since the new model is formulated using fuzzy sets, the model is called here the fuzzy conceptual clustering model. The fuzzy conceptual clustering algorithm is formulated and applied to a simple numerical example to verifying its effectiveness.

1. はじめに

ある対象物を分類する場合、人間は分類作業の過程で典型的な概念を暗黙的に形成し、対象物との類似性を計測しているといわれている¹⁾。この概念形成モデルをプロトタイプモデル^{2),3)}といい、認知心理学の観点から多くの臨床データや実験により妥当性が証明されている。S.K.Reed³⁾は顔図形を2つのカテゴリーに分類する実験で、被験者がプロトタイプを含む新たな顔図形を高正解率で分類できることを証明した。また、J.R.Anderson⁴⁾は、人間の分類作業や再認テストから知識をルールとして獲得し、知識ネットワークを形成している。一方、機械学習において、分類木を構築する概念クラスタリング手法⁵⁾が提案されている。その一手法として、COBWEB^{6),7)}は有用である。対象データが与えられた場合、データ間やクラス間の類似度を計算し、分類指標が最大となる分類木のクラスにデータを割り当てる。この手順を繰り返し、分類木の構造を変更し、結果の分類木を得る。

本論文では、ファジィ集合を用いて人間の主観量と属性の重みを取り扱う新たな概念クラスタリング手法⁸⁾を提案する。このモデルをファジィ概念形成モデルと呼ぶ。データが与えられた場合、メンバーシップ値からデータ間やクラス間の類似指標を計算し、分類木を構成する。ただし、クラス内類似指標とクラス間類似指標に加えて、クラス間の占有率を新たに定義した分類指標を提案する。COBWEBでは、クラス間の占有率はクラス間類似指標と等しくなる。メンバーシップ値の場合には、クラス間類似指標が等しい場合でも、クラス間に占有率の差が生じるので、この違いを分類指標に考慮した。分類木はCOBWEBと同様な4つの変更操作により構成される。ただし、従来の4操作に修正を加えた新たな操作手順を提案

する。データは分類木の中で分類指標が最小となるクラスに割り当てられる。ここでは、本手法とCOBWEBの拡張手法を数値例に適用し、本手法の有効性を検証する。

2. 分類指標

いま、 n 個の属性 $x_j, j=1,2,\dots,n$ に対して、 N 個のデータ $x_{ij}, i=1,2,\dots,N$ が与えられているとする。データ x_{ij} を分類木の L 個のクラス $C_k, k=1,2,\dots,L$ に分類することを考える。ただし、属性 x_j には、 $m(j)$ 個のファジィ集合 $F_{js}, s=1,2,\dots,m(j)$ が設定されており、各クラス C_k にはデータ $x_{ij}, i=1,2,\dots,N(k)$ が含まれているとする。クラス C_k の分類指標 I を次のように定義する。

$$I = \sum_{j=1}^n \mu_j \sum_{k=1}^L \sum_{s=1}^{m(j)} (w_1 I_1 + w_2 I_2 + w_3 I_3) \quad (1)$$

$$0 \leq w_1, w_2, w_3 \leq 1, w_1 + w_2 + w_3 = 1$$

$$I_1 = \sum_{i=1}^{N(k)} |\mu_{F_{is}^k}(x_{ij}) - \mu_{F_{is}^k}| \log(2 / (\mu_{F_{is}^k}(x_{ij}) + \mu_{F_{is}^k})) / 2 \quad (2)$$

$$I_2 = \sum_{k=1}^{L C_2} (|\mu_{F_{is}^k} - \mu_{F_{it}^k}| + 1) \log(2 / (|\mu_{F_{is}^k} - \mu_{F_{it}^k}| + 1)) / 2 \quad (3)$$

$$i=1,2,\dots,m(j), i \neq s$$

$$I_3 = \sum_{k=1}^{L N(k)} (\sum_{i=1}^{N(k)} \mu_{F_{is}^k}(x_{ij}) \log(\sum_{k=1}^{L N(k)} \mu_{F_{is}^k}(x_{ij}) / \sum_{i=1}^{N(k)} \mu_{F_{is}^k}(x_{ij}))) / \sum_{k=1}^{L N(k)} \mu_{F_{is}^k}(x_{ij}) \quad (4)$$

$$\mu_{F_{is}^k} = \sum_{i=1}^{N(k)} \mu_{F_{is}^k}(x_{ij}) / N(k) \quad (5)$$

ただし、 μ_j は属性 x_j に対する重みであり、 w_1, w_2, w_3 はそれぞれ、 I_1, I_2, I_3 に対する重みである。

I_1 はデータ x_{ij} のメンバーシップ値 $\mu_{F_{ik}(x_{ij})}$ に対するクラス C_k のクラス内類似指標を表している。 $\mu_{F_{ik}(x_{ij})}$ のクラス C_k での平均値からの偏差を計算し、情報エントロピーを計算する。平均値のデータはプロトタイプを形成する¹⁾。 I_1 の値は小さいほどよい。次に、 I_2 はクラス C_k のクラス間の類似指標を表している。クラス間での $\mu_{F_{ik}(x_{ij})}$ の平均値の差に対する情報エントロピーを計算する。 I_2 の値は小さいほどよい。また、 I_3 はクラス C_k の $\mu_{F_{ik}(x_{ij})}$ に対するクラス間の比率を表している。ここでは、この指標をクラス比類似指標と呼ぶ。クラス C_k のメンバーシップ値 $\mu_{F_{ik}(x_{ij})}$ に対する占有率から情報エントロピーを計算する。 I_3 の値は小さいほどよい。

分類指標 I は、 I_1, I_2, I_3 の重みつき平均である。したがって、 I の値が小さいほど、その分類木の構成が正しいことを表す。COBWEBでは、クラス内類似度とクラス間類似度から分類指標が定義されている。しかし、メンバーシップ値を取り扱う場合、クラス比類似指標を考慮する必要がある。

類似指標 I_1, I_2, I_3 を用いて、与えられたデータによって生じるクラス内の矛盾の度合いを次のように定義する。

$$C = 2I_1 / (I_2 + I_3) \quad (6)$$

クラス内類似指標 I_1 の値が I_2 や I_3 の値よりも大きい場合には、クラス内部の非類似度がクラス間の非類似度よりも大きくなるので、分類木のクラスの構成は矛盾した状態といえる。この場合、 $C > 1$ となる。一方、 I_1 の値が小さい場合には、それぞれのクラスでの類似度は高いといえるので、矛盾の度合いは小さい。この場合、 $C < 1$ となる。

3. 分類木の構築操作とアルゴリズム

与えられたデータに対して、次の4操作の中で、分類指標が最小となる分類木のクラスにデータを追加させる。

- (操作1) 既存クラスへの付加
- (操作2) 新たなクラスの生成
- (操作3) 既存クラスの統合
- (操作4) 既存クラスの解体

操作1の既存クラス付加とは、分類木の既存クラスにデータを追加する操作をいう。図1に事例を示す。クラス C_p の分類木にデータ c を与えた場合、既存クラス C_p へ付加される。操作2の新規クラス作成とは、新規クラスにデー

タを追加する操作をいう。図2に事例を示す。与えられたデータ c は既存データ b とともに新規クラス C_q に追加される。操作3のクラス統合とは、分類指標の小さい2つのクラスを統合し、与えられたデータを追加する操作をいう。図3に事例を示す。分類指標の小さいクラス C_q とクラス C_r とがクラス C_s に統合している。操作4のクラス解体とは、分類指標が最小なクラスを解体する操作をいう。図4に事例を示す。データ a, b が所属するクラス C_q が解体され、データ a, b は上位クラス C_p に所属する。

分類木は次のアルゴリズムにしたがって構築される。

- 手順1) データを逐次的に選択する。
- 手順2) 分類木のクラスの中で、4つの操作から適応できる操作を選択する。
- 手順3) 分類指標 I を計算し、最小となる操作を選択して、データをそのクラスに追加する。ただし、分類指標の値が同値の場合には、適用する順で最後の操作を採用する。
- 手順4) すべてのデータが分類木に分類された時点で、アルゴリズムを終了する。

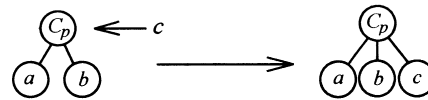


図1 既存クラスへの付加の事例

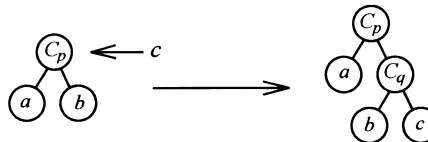


図2 新規クラスの作成の事例

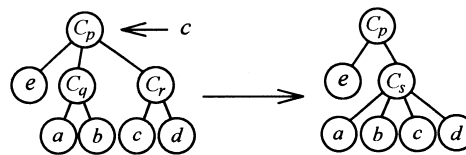


図3 既存クラスの統合の事例

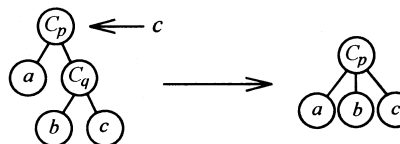


図4 既存クラスの解体の事例

4. 数値例

ファジィ概念形成モデルの有用性を検討するために、アルコール飲料の分類の事例を取り上げる。いま、表1のような、原料、製造法、アルコール度数、発酵・貯蔵期間、価格の5つの属性に対する各種のアルコール飲料のデータが与えられているとする。各属性に対して、ファジィ集合を設定する。図5に発酵貯蔵期間のファジィ集合を示す。全体では、16種類のファジィ集合を設定した。

表1 アルコール飲料のデータ

種類	原料	製造法	度数	発酵貯蔵期間	価格(円)
ウイスキー	麦, 穀物	蒸留法	40	10年	2,000
ワイン	ブドウ	発酵法	10	2年	3,500
ビール	麦, 米	発酵法	5	1カ月	600
ブランデー	ブドウ, 果実, 穀物	蒸留法	60	6年	5,000
清酒	米, 麦, 穀物	発酵法	15	2カ月	1,500

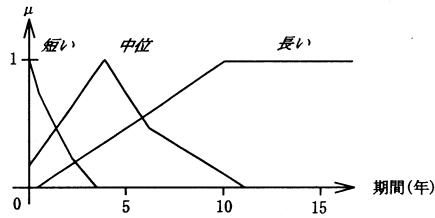


図5 発酵貯蔵期間のファジィ集合

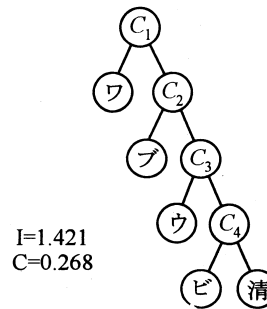


図7 結果の分類木

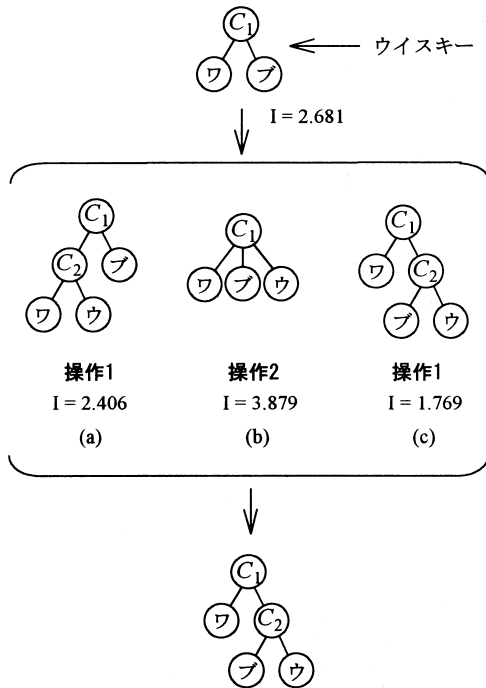


図6 分類木の構築過程

実験として、ワイン→ブランデー→ウイスキー→ビール→清酒の順にデータを与えて分類木を構成した。分類木の構成過程を図6に示す。ワインとブランデーのクラス C_1 から構成されている分類木に、ウイスキーのデータを加える。ここでは、操作1と操作2が適用できるので、候補として、(a), (b), (c)の分類木が考えられる。分類指標を計算し、 $I(a)=2.406$, $I(b)=3.879$, $I(c)=1.769$ より、(c)の分類木を採用する。これらの手順を繰り返し、最終的に図7の分類木を得た。同じクラス C_4 に所属するビールと清酒は、表1より属性値が類似していることがわかるので、結果の分類木は納得ができる。また、ウイスキーとブランデーとは製造法やアルコール度数が類似しており、この点も分類木に反映されている。

次に、本手法の有効性を検討するために、本手法とCOBWEBの拡張手法との比較をおこなった。COBWEBの拡張手法の分類指標を次のように定義した。

$$I = 1 - \sum_{j=1}^n \mu_j \left(\frac{1}{\prod_{s=1}^{m(j)} \mu(x_{ij}|C_k)} + \frac{1}{\prod_{k=1}^L \mu(C_k|x_{ij})} \right) / 2 \quad (7)$$

$$\mu(x_{ij}|C_k) = 1 - \sum_{i=1}^{N(k)} |\mu_{F_{is}^k}(x_{ij}) - \mu_{F_{is}^k}| / N(k) \quad (8)$$

$$\mu(C_k|x_{ij}) = |\mu_{F_{is}^k} - \sum_{k=1}^L \mu_{F_{is}^k}| / L \quad (9)$$

ここで、 $\mu(x_{ij}|C_k)$ はクラス C_k のクラス内類似指標であり、 $\mu(C_k|x_{ij})$ はクラス間類似指標である。なお、 $a \dot{+} b = a + b - ab$ である。分類指標 I の値が小さいほど、分類木の構成が正しいことを表す。

比較実験では、30種類のデータを異なる順で分類木に与え、本手法と拡張手法によって得られた分類木の分類指標と矛盾量を比較した。結果を表2に示す。本手法によって得られた分類木の分類指標 I_{AA} に対して、拡張手法の(7)式を計算した値を I_{AB} としている。一方、拡張手法によって得られた分類木の分類指標 I_{BB} に対して、本手法の(1)式を計算した値を I_{BA} としている。 $I_{AA} < I_{AB}$ かつ $I_{BA} < I_{BB}$ の場合、本手法が優位であることを示している。 $I_{AB} < I_{AA}$ かつ $I_{BB} < I_{BA}$ の場合には、拡張手法が優位であることを示している。一方、 $I_{AA} < I_{AB}$ かつ $I_{BB} < I_{BA}$ の場合には、各手法が優位であることを主張しているの、比較不能となる。 $I_{AB} < I_{AA}$ かつ $I_{BA} < I_{BB}$ の場合には、両手法以外に優れた手法が存在する。本手法が優位である場合が全体の40%を占め、両手法の分類指標が同値である場合を含めると、63%を占める。また、拡張手法が優位である場合や両手法以外に優れた手法が存在する場合は0%であり、この点からも本手法の優位性が示せる。さらに、両手法の矛盾量を比較すると、どの場合でも、 I_{AA} の矛盾量が I_{AB} の矛盾量よりも値が小さく、本手法の分類木が矛盾する程度が低いことがわかる。

5. おわりに

本論文では、ファジィ集合を用いた概念形成モデルを提案した。しかし、アルゴリズムでは、分類木の階層の差異が分類指標に考慮されていないので、正しい分類木を構成できない場合がある。今後、分類指標を再検討する必要がある。なお、本研究の一部は、1997年度阪南大学産業経済研究所共同研究「ソフトコンピューティングとその経済動学への応用(2)」の成果報告である。

表2 手法の比較の結果

	COBWEBの拡張手法		提案手法				回数	占有率 (%)
	I_{BB}	I_{BA}	I_{AA}	矛盾量	I_{AB}	矛盾量		
$I_{BB} > I_{BA}, I_{AA} < I_{AB}$	0.437	0.418	1.497	0.295	1.995	0.749	12	40
$I_{BB} < I_{BA}, I_{AA} < I_{AB}$	0.440	0.452	1.622	0.397	1.971	0.729	11	37
$I_{BB} = I_{BA}, I_{AA} = I_{AB}$	0.421	0.421	0.451	0.306	0.451	0.306	7	23
$I_{BB} < I_{BA}, I_{AA} > I_{AB}$	—	—	—	—	—	—	0	0
$I_{BB} > I_{BA}, I_{AA} > I_{AB}$	—	—	—	—	—	—	0	0
平均値	0.438	0.418	1.585	0.331	1.859	0.639		

参考文献

- 1) 市川伸一, 伊東裕司, 渡邊正孝, 酒井邦嘉, 安西祐一郎: “記憶と学習”, 認知科学5, 岩波書店 (1994)
- 2) M.I.Posner, S.W.Keele: “On the Genesis of Abstract Idea,” Journal of Experimental Psychology, Vol.77, pp.353-363 (1968)
- 3) S.K.Reed: “Pattern Recognition and Categorization,” Cognitive Psychology, Vol.3, pp.382-407 (1972)
- 4) J.R.Anderson, P.J.Kline, C.M.Jr.Beaesley: “General Learning Theory and its Application to Schema Abstraction,” The Psychology of Learning and Motivation, Vol.13, Academic Press, pp.277-318 (1979)
- 5) R.S.Michalski, Y.Kodratoff: “Research in Machine Learning: Recent Progress, Classification of Methods, and Future Directions,” Machine Learning III, pp.3-30, Morgan Kaufmann (1990)
- 6) D.H.Fisher: “Knowledge Acquisition via Incremental Conceptual Clustering,” Doctoral Dissertation, Department of Information and Computer Science, University of California, Irvine (1987)
- 7) 樺木哲夫, 岩井壯介, 片井修: “概念クラスタリング手法による決定モデルの動的学習とリアクティブルールの獲得”, 第18回知能システムシンポジウム, 第19回システムシンポジウム合同シンポジウム, pp.199-203 (1993)
- 8) 林 勲, 津田エミ子, 横井真也, 市川隆男: “ファジィ集合による概念形成モデルの一提案”, 第41回システム制御情報学会研究発表講演会 (1997)

【問い合わせ先】

580 大阪府松原市天美東5-4-33
 阪南大学 経営情報学部
 林 勲
 phone. 0723-32-1224 (ext.8412)
 fax. 0723-36-2633
 Internet ihaya@hannan-u.ac.jp
 URL http://www.hannan-u.ac.jp/~ihaya