

概念形成における特徴次元への重み付けのモデル化

Modeling the feature dimension weighting in unsupervised category learning

西田 豊
Yutaka Nishida

大阪大学 大学院人間科学研究科
Graduate School of Human Sciences, Osaka University
nishida [at] bm.hus.osaka-u.ac.jp

Abstract

Selecting features is an important problem in category learning because of our limited cognitive resource. We introduce the feature dimension weighting and present an algorithm that calculates the weighting through the optimization process of unsupervised category learning. This model allows us to identify the important features for category learning.

Keywords — dimension weighting, unsupervised category learning, regularization method

1. 特徴次元の重み付けと選択

事例を分類し概念を形成することは、人の認知活動のうち最も基礎的で重要なもののうちの1つである。概念形成時にどの特徴を用いるかを選択することは重要な問題である。観測された特徴をすべて用いて概念を形成することは可能であるが、特徴数が多い場合は認知的負荷が高い。また、観測された事例によく当てはまる概念を形成したとしても、その概念が後から観測される事例に対してよい概念であるとは限らない。そのため、観測された全特徴を用いるのではなく、なるべく少数の特徴を用いて概念を形成していると考えるのが自然である。

より少ない特徴次元を用いて概念形成を行う場合、2つの方法が考えられる。1つは相関の高い特徴をまとめて、新たな重みつき合成特徴を創出する方法である。もう1つは概念形成に有用であると考えられる特徴のみを用いる方法である。

前者のような方法には、次元縮約を伴うクラスタリング法として提案されたREDKM (De Soete & Carroll, 1994) がある。概念形成のモデルとして提案されたわけではないが、このモデルを認知モデルとして考えることもできる。合成された特徴量を元に概念を形成するため少数の特徴次元を用いているといえるが、特徴次元を合成する過程があるため認知的負荷の軽減にはつながらないと考えられることもできる。

後者のような考え方には、重み付けパラメータ

が導入される。特徴が概念形成に有用であるかの判断には特徴次元への重み付けの大小によって解釈される。すなわち特徴次元への重み付けが大きいほど、その特徴は概念形成に有用であると考えられる。このような考えを基にした教師つき概念形成モデルにALCOVE (Kruschke, 1992) がある。

ALCOVEは教師つき学習のアルゴリズムであり、その派生モデルも多く研究されている。しかしながら、教師つき学習と同様に重要である、教師なし学習における概念形成における特徴次元への重み付けを考慮したモデルは考えられていない。そこで本研究では、特徴次元への重み付けパラメータを導入し、Miyamoto, Ichihashi, Honda (2008) によって提案されたエントロピー正則化法を用いて、概念形成の最適化の過程で重み付けパラメータも計算するアルゴリズムを提案する。

2. 特徴次元への重み付けのモデル化

データを $N \times P$ の行列 $X = \{x_{ij}\}$ とする。 N と P はそれぞれ事例数と特徴次元数である。概念数を $K (k = 1, \dots, K)$ で表す。

提案モデルの目的関数は

$$F(\mathbf{U}, \bar{\mathbf{X}}, \mathbf{W}) = \sum_{k=1}^K \sum_{i=1}^N \sum_{j=1}^P u_{ik} w_j (x_{ij} - \bar{x}_{kj})^2 + \lambda^{-1} \sum_{j=1}^P w_j \log w_j. \quad (1)$$

とあらわされる。

$\mathbf{U} = \{u_{ik}\}$ は事例 i が概念 k に含まれるか否かを示す $N \times K$ のメンバシップ行列である。事例 i が概念 k に含まれる場合は1、含まれない場合は0となる。 $\bar{\mathbf{X}} = \{\bar{x}_{kj}\}$ は概念 k のセントロイドを示す $K \times P$ の平均行列である。 $\mathbf{W} = \{w_j\}$ は次元 j への注意重みを対角要素に持つ重み付け行列である。ここで $\lambda > 0$ は w についての正則化パラメータである。

この目的関数に課される制約は2つある。1つめ

は所属度を示すパラメータ u に関するもので,

$$\sum_{k=1}^K u_{ik} = 1, u_{ik} \in \{0, 1\} \quad (2)$$

である。(2)は $\mathbf{U}$ の行和は1であり、各値は0か1であることを示している。すなわち、事例はいずれかの概念に所属し、かつ1つの概念のみに所属するということである。

2つ目は重み付けを示すパラメータ w に関するもので,

$$\sum_{j=1}^P w_j = 1, 0 \leq w_j \leq 1. \quad (3)$$

である。(3)は $\mathbf{W}$ の対角和が1であり、各値は0から1の値をとることを示している。すなわち、特徴次元への重み付けは全体で1であり、各特徴次元へ配分するということである。

最適なパラメータ $\mathbf{U}$, $\bar{\mathbf{X}}$, $\mathbf{W}$ を求めることが目的となる。

3. アルゴリズム

目的関数 (1) の値を最小にするパラメータ $\mathbf{U}$, $\bar{\mathbf{X}}$, $\mathbf{W}$ を求めるために、次の3ステップを反復させるアルゴリズムを用いる。

- **U step:** $\bar{\mathbf{X}}$, $\mathbf{W}$ を固定し、 $F(\mathbf{U}, \bar{\mathbf{X}}, \mathbf{W})$ を最小にする $\mathbf{U}$ を求める。
- **$\bar{\mathbf{X}}$ step:** $\mathbf{U}$, $\mathbf{W}$ を固定し、 $F(\mathbf{U}, \bar{\mathbf{X}}, \mathbf{W})$ を最小にする $\bar{\mathbf{X}}$ を求める。
- **W step:** $\mathbf{U}$, $\bar{\mathbf{X}}$ を固定し、 $F(\mathbf{U}, \bar{\mathbf{X}}, \mathbf{W})$ を最小にする $\mathbf{W}$ を求める。

それぞれのステップでは以下のように最適解が求まる。

$$u_{ik} = \begin{cases} 1, & \text{if } \sum_{j=1}^P w_j (x_{ij} - \bar{x}_{kj})^2 \leq \sum_{j=1}^P w_j (x_{ij} - \bar{x}_{lj})^2 \\ 0, & \text{other.} \end{cases} \quad (4)$$

$$\bar{x}_{kj} = \frac{\sum_{i=1}^N u_{ik} x_{ij}}{\sum_{i=1}^N u_{ik}} \quad (5)$$

$$w_j = \frac{\exp \left\{ -\lambda \sum_k^K \sum_i^N u_{ik} (x_{ij} - \bar{x}_{kj})^2 \right\}}{\sum_{j=1}^P \exp \left\{ -\lambda \sum_k^K \sum_i^N u_{ik} (x_{ij} - \bar{x}_{kj})^2 \right\}} \quad (6)$$

4. 数値例

本節では、簡単な数値例を示して、提案モデルの有効性を検討する。使用するデータセットは豊田 (2008) の動物の特徴データである。表(2)に示す30種の動物について、表(1)に示す13個の特徴が、ダミー変数(特徴を有していれば1, 有していなければ0のバイナリデータ)によって表現されている。

動物特徴データに概念数 K を2~5に変化させて提案手法を適用した。正則化パラメータはすべての条件で $\lambda = 1$ とした。50個の初期値を用いて、最も目的関数の値が小さくなった解を採用した。特徴次元への重み付けパラメータを表(1)に、所属を示すメンバシップパラメータを表(2)に示す。

まず概念数2の結果から解釈する。表(2)から30種の動物が「ほ乳類」と「鳥類」にカテゴリ化されている様子が見えてくる。表(1)からは「4本足」と「羽」の次元への重み付けが大きいことがわかる。以上からほ乳類と鳥類に分類するために用いられる特徴次元は「4本足」と「羽」であることが読み取れる。

概念数が3になると先ほどの「ほ乳類」がさらに「蹄」を持つ種と持たない種に細分化されている。重み付けパラメータを見ると「4本足」、「羽」への重み付けが減少し、「蹄」への重み付けが増加している。

概念数が4になると先ほどの「鳥類」が「飛ぶ」ことを好む鳥とそうでない鳥に細分化された。重み付けパラメータを見ると「飛ぶ」への重み付けが増加していることがわかる。

最後に概念数が5の場合は「飛ぶ」ことを好む鳥のうち「泳ぐ」ことも好む鳥に細分化された。この場合も「泳ぐ」への重み付けがわずかではあるが増加している。

概念数を2~5まで通してみると、「体のサイズ(小, 中, 大)」、「鬣」、「狩猟」、「走る」への重み付けはほとんどなされなかった。従ってこれらの特徴次元は「ほ乳類」か「鳥類」かといった分類にはあまり必要の無い情報であることがわかる。

逆に2つの概念に分類されるときに大きく重みづけられた「4本足」と「羽」は概念数を変化させても大きく重みづけられていたことから、「ほ乳類」か「鳥類」かの分類に最も重要な特徴次元であることがわかる。

以上の結果から、概念を形成する上で重要な特徴次元へ重み付けられていることが分かる。すなわち、本モデルによって得られた重み付けパラメータは認知モデルとして解釈可能であることが示された。

5. ファジィメンバシップへの拡張

2節で提案したモデルは事例がある概念に所属するか否か、すなわちメンバシップがバイナリで表されるものであった。しかしながら、我々が持つ概念構造はこのような決定論的なものではなく、確率的に変動するとも考えられる。例えば、プロトタイプから離れた事例について評定する場合、同一人物であっても文脈によって概念に含まれるか否かの判断が異なる場合がある (McClosky & Glucksberg, 1978)。

そこで本節ではメンバシップパラメータが $[0, 1]$ の実数値を取るように拡張する。Miyamoto, Ichihashi, Honda (2008) のエントロピー正則化法をメンバシップパラメータ u についても適用することによって達成する。したがって、ファジィモデルの目的関数は

$$G(\mathbf{U}, \bar{\mathbf{X}}, \mathbf{W}) = \sum_{k=1}^K \sum_{i=1}^N \sum_{j=1}^P u_{ik} w_j (x_{ij} - \bar{x}_{kj})^2 + \phi^{-1} \sum_{k=1}^K u_{ik} \log u_{ik} + \lambda^{-1} \sum_{j=1}^P w_j \log w_j. \quad (7)$$

とあらわせる。ここで $\phi > 0$ は u についての正則化パラメータである。メンバシップパラメータ u に関する制約が (2) から次のように変更される。

$$\sum_{k=1}^K u_{ik} = 1, 0 \leq u_{ik} \leq 1. \quad (8)$$

重み付けを示すパラメータ w に関する制約は(3)と同様である。

アルゴリズムは (7) を最小にする $\mathbf{U}$, $\bar{\mathbf{X}}$, $\mathbf{W}$ を求める3ステップを反復させるアルゴリズムを用いる。 $\bar{\mathbf{X}}$, $\mathbf{W}$ のステップは(5), (6)と同様であるが、 $\mathbf{U}$ を求めるステップではファジィ化にともない以下のように最適解が求まる。

$$u_{ik} = \frac{\exp \left\{ -\phi \sum_j^P w_j (x_{ij} - \bar{x}_{kj})^2 \right\}}{\sum_{k=1}^K \exp \left\{ -\phi \sum_j^P w_j (x_{ij} - \bar{x}_{kj})^2 \right\}} \quad (9)$$

6. まとめと展望

本研究では人は概念を形成する際、観測されるすべての特徴次元を等しく用いているわけではなく、概念を形成するために有用である特徴に重み付けを行いより効率的に概念形成を行っているという仮定のもと、特徴次元への重み付けを概念

学習の最適化の過程で計算するモデルとそのアルゴリズムを提案した。提案モデルの有用性は動物データを通して検証された。その結果、概念を形成するに当たって、有用であると考えられる特徴次元へ大きな重み付けがされており、認知モデルとして解釈可能であることが示された。

また、人の概念構造はファジィであるとの考えから、 $[0, 1]$ のメンバシップパラメータを得られるように拡張を行った。しかしながら、人の概念形成においてメンバシップがバイナリかファジィのどちらで表現されるかは行動実験とモデルフィッティングによって検証されるべき事項である。

本モデルの問題点として、あらかじめ概念数 K を設定しなければならないことがあげられる。人の認知処理を考えた場合、 K をいくつか設定し、モデル比較を行うといったメタ認知的な処理が行われていると考えられる。

メタ的な処理として現在考えられる方法には次のようなものがある。1. K を変化させ得られた目的関数の値の減少幅が少なくなるところの K を採用する方法、2. 概念間の平方和と概念内の平方和の比が最も大きくなる K を採用する方法 (e.g., Calinski & Harabasz, 1974), 3. ギャップ統計量 (Tibshirani, Walther & Hastie, 2001)を用いる方法などが考えられる。しかしながら、どのような処理が心理学的に妥当であるかを検証するためには、行動実験の結果との比較が必要である。

文献

- Calinski, R. B. & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3**, 1-27.
- De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand & B. Burtschy (Eds.), *New approaches in classification and data analysis*, Berlin, Springer-Verlag.
- Kruschke, J. K., (1992). ALCOVE: An Exemplar-Based Connectionist Model of Category Learning, *Psychological Review*, **99**, 22-44.
- McClosky, M. E., & Glucksberg, S. (1978). Natural categories: well defined or fuzzy sets? *Memory & Cognition*, **6**, 462-472.
- Miyamoto, S., Ichihashi, H., & Honda, K. (2008). *Algorithms for fuzzy clustering: methods in c-means clustering with applications*. Berlin: Springer-Verlag.
- Tibshirani, R., Walther, G., & Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistics *Journal of the Royal Statistical Society Series B*, **63**(2), 411-423.
- 豊田秀樹 (編著) (2008). データマイニング入門. 東京図書.

表 1 概念数別重み付けパラメータ

概念数	小	中	大	2本足	4本足	毛	蹄	鬣	羽	狩猟	走る	飛ぶ	泳ぐ
II	.01	.00	.00	.12	.32	.12	.00	.01	.32	.00	.00	.04	.04
III	.01	.00	.00	.10	.24	.10	.24	.01	.24	.00	.00	.03	.03
IV	.02	.00	.00	.08	.20	.08	.20	.01	.20	.01	.00	.20	.02
V	.02	.00	.00	.07	.18	.08	.18	.01	.18	.01	.00	.18	.09

表 2 概念数別メンバシップパラメータ

	II		III			IV				V				
	A	B	A	B	C	A	B	C	D	A	B	C	D	E
イヌ	1		1			1				1				
オオカミ	1		1			1				1				
キツネ	1		1			1				1				
クマ	1		1			1				1				
サル	1		1			1				1				
タヌキ	1		1			1				1				
チーター	1		1			1				1				
トラ	1		1			1				1				
ネコ	1		1			1				1				
ハイエナ	1		1			1				1				
ヒョウ	1		1			1				1				
ライオン	1		1			1				1				
イノシシ	1			1			1				1			
ウシ	1			1			1				1			
ウマ	1			1			1				1			
シカ	1			1			1				1			
シマウマ	1			1			1				1			
ゾウ	1			1			1				1			
ブタ	1			1			1				1			
ペンギン		1			1			1				1		
ダチョウ		1			1			1				1		
ニワトリ		1			1			1				1		
カナリヤ		1			1				1				1	
スズメ		1			1				1				1	
タカ		1			1				1				1	
フクロウ		1			1				1				1	
ワシ		1			1				1				1	
ハト		1			1				1				1	
アヒル		1			1				1					1
ガチョウ		1			1				1					1