不完全フィードバック情報を用いた混合カテゴリ学習モデル
A mixture category learning model using incomplete feedback information

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 @ bm.hus.osaka-u.ac.jp

Abstract

In a real learning situation, it is rare that we can use complete feedback information. Therefore we should learn categories from labeled and unlabeled data. This problem is called semi-supervised learning. This article presents a simple extension of a mixture model to represent such a situation.

Keywords — category, semi-supervised learning, mixture distribution

1. Supervised or Unsupervised

認知心理学の分野において，カテゴリ学習は古くから研究されており，数理モデル化が進んでいる．これまでのカテゴリ学習モデルは，教師つき学習，つまり正答か誤答かが常にフィードバックされる学習が多かったが，現実世界に置ける学習ではそのような場面は限られると考えられる．

例えば，子どもが犬というカテゴリを学習する場面について考えてみよう．母親が4本足で，毛がフサフサしていて，尻尾があって，…，といった生き物を指さして「わんわん」と子どもに教える．この場合はフィードバックがなされている．しかし，子どもが目にする犬(もしくは他の生き物)すべてに対して母親はフィードバックを与えるわけではない．

この様に，現実の学習場面では，少数のフィードバックありの事例と，多数のフィードバックなしの事例から学習を成立させていると考えられる．

2. Prototype or Exemplar

また，カテゴリ学習のモデルとしてよく取り上げられるのは，prototype modelとexemplar modelである．prototype modelは観測した事例をカテゴリの代表値(平均値)と比較し，事例がカテゴリに属するか否かを判断するモデルであるのに対し，exemplar modelは観測した事例をカテゴリ内の全成員と比較し，事例がカテゴリに属するか否かを判断するモデルである．

数理的な観点から見れば，prototype modelはカテゴリに1つの分布を仮定し，その平均値と事例との距離を計算していることになる．またexemplar modelはカテゴリの成員ごとに分布を仮定し，その分布をすべて足し上げたものをカテゴリ全体の分布とし，全成員と事例との距離を計算している．

これまでの研究によってprototype model，exemplar modelいずれのモデルが優位かについて議論がなされてきた．しかし，これら2つのモデルは混合分布の導入により，1つのモデルが有するパラメータの両極端の場合の下位モデルであることが示されている(Rosseel, 2002)．

本研究では現実の学習場面に即したカテゴリ学習を表現するため，prototype model，exemplar model両方の表現を可能とする混合分布クラスタリングモデルを，フィードバックが得られているデータと得られていないデータの両方から学習するよう拡張した．

3. Model

3.1 learning phase

いま，C個のカテゴリ$\{c = 1, ..., C\}$があり，そのC個のカテゴリがさらにK個のサブカテゴリ$\{k = 1, ..., K\}$に分かれると仮定する．所属するサブカテゴリが既知の事例と未知の事例から，カテゴリを学習する問題を考える．N個の事例$\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_\ell, \mathbf{x}_{\ell+1}, ..., \mathbf{x}_N\}$が与えられており，はじめの$\ell$個の事例は所属するサブカテゴリがわかっていて，残りの$n-\ell$個の事例は所属するサブカテゴリがわからないとする．さらに，$\mathbf{x}_i$がサブカテゴリkに所属する場合$y_{ik} = 1$，所属しない場合$y_{ik} = 0$となるような所属度行列$Y = \{y_{ik}\}$を考える．カテゴリはサブカテゴリの成員である$\mathbf{x}$によって構成される．このとき，$\mathbf{x}$が生起する確率は，

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}). \quad (1)$$

によって表される．π_kは$\mathbf{x}$がK個のサブカテゴリのうちどのサブカテゴリから生じるかを示す確率(サブカテゴリの混合比率)，$f_k(\mathbf{x})$はサブカテゴリ

kの確率密度関数である．いま$f_k(\mathbf{x})$を正規分布であるとすると，$p(\mathbf{x})$はK個の正規分布を足し合わせたものとなり，正規混合分布となる．

$$p(\mathbf{x}, k; \pi_k, \mu_k, \Sigma_k) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}; \mu_k, \Sigma_k). \tag{2}$$

ここで$\phi(\mathbf{x}; \mu_k, \Sigma_k)$は平均$\mu$，分散共分散行列$\Sigma$を持つ正規分布で，$x$の次元を$d$としたとき

$$\phi(\mathbf{x}; \mu_k, \Sigma_k) = \left(\frac{1}{2\pi}\right)^{\frac{d}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}||\mathbf{x} - \mu_k||^2_{\Sigma_k^{-1}}\right\}. \tag{3}$$

ただし，$||\mathbf{x}||^2_G = \mathbf{x}'G\mathbf{x}$である．

本研究では，EMアルゴリズムを用いてパラメータを推定する．EMアルゴリズムによる正規混合分布のパラメータμ_k, Σ_kは次式によって更新される(McLachlan & Basford, 1987).

$$\mu_k^{(t+1)} = \frac{1}{N\pi_k^{(t+1)}} \sum_{i=1}^{N} q^{(t)}(k|\mathbf{x}_i)\mathbf{x}_i. \tag{4}$$

$$\mathbf{\Sigma}_k^{(t+1)} = \frac{1}{N\pi_k^{(t+1)}} \sum_{i=1}^{N} q^{(t)}(k|\mathbf{x}_i)\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'. \tag{5}$$

ただし$p_k^{(t+1)}$，$q^{(t)}(k|\mathbf{x}_i)$，$\tilde{\mathbf{x}}_i$は

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} q^{(t)}(k|\mathbf{x}_i). \tag{6}$$

$$q^{(t)}(k|\mathbf{x}_i) = \begin{cases} y_{ik}, & \text{if } i = 1, ..., \ell. \\ \dfrac{p(\mathbf{x}_i, k; \pi_k, \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} p(\mathbf{x}_i, k'; \pi_{k'}, \mu_{k'}, \Sigma_{k'})}, & \text{if } i = \ell + 1, ..., N. \end{cases} \tag{7}$$

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mu_k^{(t+1)}. \tag{8}$$

3.2 discrimination phase

学習を終えた後，新規事例がどのカテゴリに属するか否かを判断するプロセスである．まず，新規事例と各サブカテゴリとの類似性が計算される．類似性は以下のように定義される．

$$S_k = \exp\left\{-(\mathbf{x} - \mu_k)'\Sigma_k^{-1}(\mathbf{x} - \mu_k)\right\}. \tag{9}$$

μ_k,Σ_kは，learning phaseで学習された，各サブカテゴリの平均と共分散である．次に，類似性から各カテゴリへの分類確率を計算する．

$$P(R_{\mathcal{A}}|x) = \frac{\sum_{k\in\mathcal{A}} \pi_k S_k}{\sum_{\mathcal{C}} \sum_{j\in\mathcal{C}} \pi_j S_j}. \tag{10}$$

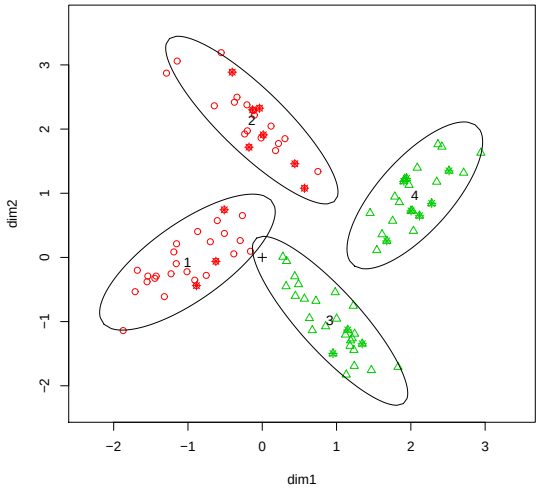

図 1 人工データと推定されたμとΣのプロット

4. Numerical Experiment

以上のモデルの動作を人工的に発生させたデータを用いて確認する．カテゴリ数$C = 2$,サブカテゴリ数$K = 4$の場合を考える．教師つきデータを20個，教師なしデータを80個，2変量正規分布から発生させた．

人工データの布置を図1に示す．＊はカテゴリラベルが既知のデータである．1 から4 の数字はlearning phase で推定された各サブカテゴリのμ_kの座標を，楕円は推定されたΣ_kの95%領域を示している．いま新規事例として＋(0, 0)を観察したとする．discrimination phaseで計算された，＋を○のカテゴリへ分類する確率は0.283，△のカテゴリへ分類する確率は0.717 となった．

5. Discussion

数値実験により，混合分布の各要素分布の平均値と分散共分散行列が正しく推定されることが示された．また，分類確率の計算もうまく言っているといえるだろう．

今後は，実験により集められデータに対して，このモデルがどの程度当てはまるのかを検討する必要がある．また今回はバッチ型のEMアルゴリズムを用いたが，人の認知活動をモデル化する上では，オンラインEMアルゴリズムを用いる必要があると考えられる．

Reference

McLachlan, G. J., & Basford, K. E. (1987). *Mixture models: Inference and applications to clustering.* New York: Mercel Dekker.

Rosseel, Y. (2002). Mixture model of categorization. *Journal of Mathematical Psychology*, **46**, 178-210.