既存の商品分類の階層構造の学習による商品の自動分類

プロフィットエンジニアリング研究研究

5215F016-7　土肥 怜生
指導教員　　大野 高裕

Automatic Hierarchical Classification of Products via Stick Breaking Process

DOI Satoki

1. はじめに

現実に存在する分類表や分類体系などの分類構造は, 次のような特徴を持つことが多い. 一つは, 「構造の階層性」である. 例えば, 野菜というカテゴリがあった時, その中にトマトといったようなカテゴリが存在している状況である. 二つ目に「曖昧なカテゴリの存在」がある. 現実に存在する分類構造には, 明確な区切りを設けず「その他」のようなカテゴリとして構造に柔軟さをもたせている場合がある. 三つ目に「新しいカテゴリへの許容性」がある. 新たな対象物が与えられた時, 現実には既存のカテゴリでは不適合の際, 新たなカテゴリを設ける必要が生じる. 上記に挙げた特徴が, 現実の分類構造に存在する一方で, こうした特徴に上手に適合した手法が確立されていない現状がある.

そこで本研究では, データに応じてモデル自体の複雑さも自動的に学習することが可能な統計モデルと言われている [1] ノンパラメトリックベイズ法の立場から, 現実に存在する分類構造の特徴を上手く捉えた手法の確立を目的する. ノンパラメトリックベイズ法の中でも, ディリクレ過程を対象の構造の広さと深さの二種の方向に適用し, 無限の広さと深さを持つ木構造を生成する木構造 Stick-breaing 過程 [2] をベースにモデリングを行う. 本研究では, 分類構造の事例として小売関連企業の持つ「商品分類」を対象として分類に取り組む. 本研究の取り組みにおいて, 階層性のある分類を行うにあたり, 相応の情報量を持つデータが必要となる. これに対して, Web 上から商品名に紐づく画像を自動的に収集してすることで, 低コストで大量にデータを取得する. また商品名によるテキストデータおよび検索結果の画像データという非構造データを用いて, モデリングを行うにあたり, 特徴抽出のため形態素解析や深層学習を用いて総合的な解析を行っている.

2. 従来研究

2.1. クラスタリングとクラス分類

従来, 分類問題に対してクラスタリングとクラス分類によって対応されている. 両者の違いは分類の目的によって決定され, まずクラスタリングでは, 分類対象から類似した集合を見つけるという点に目的が置かれることが一般的である. 代表的な手法として, k-means クラスタリングや階層クラスタリングがある. しかし集合を探索するという目的は構造が所与である本研究の取り組みの意図とは異なる. 一方, クラス分類では分類することに目的を置く. そこで本研究ではクラス分類として問題を捉える. クラス分類では, 識別モデル, 生成モデルによるものに分けて考えることができる. 識別モデルの代表的な手法として, 決定木, サポートベクターマシン, 深層学習などに代表されるニューラルネットワークなどがある. しかし, 問題点として, クラス C_k, 入力ベクトル x が与えられたとき, $f(C_k|\mathrm{x})$ を直接モデリングするため, 学習によって設計された識別系以外の分類に対する柔軟性が低く, 与えられたクラス C_k に対して無理に分類してしまう. そのため「その他」に属されるべき曖昧な対象物が出現した際, 誤分類の可能性は著しく高い. 次に生成モデルに注目すると, これはクラス自体に対して確率分布 $p(C_k)$ を仮定するため, クラスの出現確率を陽に示すことができ, リジェクト機能を付加することは比較的容易である. つまり「その他」のような曖昧な対象において, どのクラスにも所属しないものとしてリジェクトすることが可能となる.

2.2. 従来の生成モデル

従来の生成モデルの中で, 階層構造を生成するものとして Christopher[3] が挙げられる. Christopher[3] では, 構造のモデリングにあたりポアソン分布を用いて階層ごとのデータ数を決め, 正規分布によりその階層における葉を決めている. その結果, 同じ階層でも特定のカテゴリでは子のカテゴリが多く, 他のカテゴリでは単一的な形になっている場面のある現実に存在する非定形的な構造に対して対応が難しいと考えられる. 次に, Radford[4] の研究が挙げられるが, この取り組みでは, データが中間のカテゴリにおいて内在せず末端のカテゴリにのみ内在する. 加えて, 新たな対象を分類する際に生じる新たなカテゴリへの許容性においてどの手法も要件を満たさない. そこで本研究ではこの点に注目し, ノンパラメトリックベイズ法を用いて構造のモデリングを行う. 中でも木構造を生成することが可能な木構造 Stick-Breaking 過程 [2] をベースに解析を行うことで構造におけるデータの出現確率を導出する.

3. ノンパラメトリックベイズ法

ノンパラメトリックベイズ法は, データを表現するパラメータを無限に仮定することで現象に対する表現力が高い手法である. 起源は, Ferguson[5] まで遡り, 1970 年代には確立された枠組みであったが, 計算機の高精度化や探索アルゴリズムの発展により, 今日まで自然言語処理やバイオインフォマティクス等, 様々な分野への応用利用が進められている.

3.1. Stick-breaking 過程と中華料理店過程

ノンパラメトリックベイズ法を実現させる重要な要素としてディリクレ過程がある. ディリクレ過程とは, 基底分布 H, パラメータ α が与えられた時, H から無限離散分布 G を生成する確率過程である. G がディリクレ過程に従っていることを次のように表す;

$$G \sim \mathrm{DP}(\alpha, \mathrm{H}). \tag{1}$$

ここで問題となるのは, $G \sim \mathrm{DP}(\alpha, \mathrm{H})$ をどのように実現するかであり, そのディリクレ過程の構成法の一つとして Stick-breaking 過程があり, 次に示す;

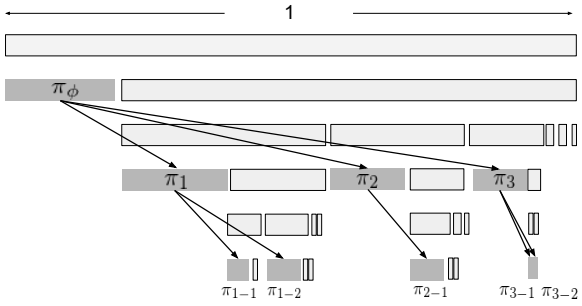
$$v_i \sim \mathrm{Be}(1, \alpha), \tag{2}$$

$$\pi_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \qquad \pi_1 = v_1, \tag{3}$$

$$G = \sum_{i=1}^{\infty} \pi_i \, \delta(\theta_i), \qquad \theta_i \sim \mathrm{H}. \tag{4}$$

ここで, $v_i \sim \mathrm{Be}(1, \alpha)$ はパラメータ $(1, \alpha)$ のベータ分布である. $\delta(\theta_i)$ はデルタ測度を表し, $\theta = \theta_i$ に位置する大きさ 1 の点を意味する. その挙動は, 長さ 1 の棒を用意し, $\{v_k : k = 1, 2, ...\}$ の割合で折り続け, その折った棒の左端を基底分布 H からサンプリングしてきた θ_i の場所に立てる動作に等しい. また Stick-breaking 過程の他にディリクレ過程を構成する代表的な方法として中華料理店過程を使って表すことができると知られている. この構成法は, 中華料理店での客の動きに対するメタファーであり, 同じテーブルに着席した客を一つのまとまりと考え, 客をテーブルごとに分割していく. 初めの客は, 任意のテーブルに座り, それ以降の客は, 先客がいるテーブル i を $n_i/(n-1+\alpha)$ の確率で選択し, 誰も客のいない新たなテーブルに対し $\alpha/(n-1+\alpha)$ の確率で着席する. この表現は, G からのサンプル $\{x_1, x_2, ...x_{n-1}\}$ が既知の下, x_n を生成することに等しく, 条件付き確率分布 $p(x_n | x_1, x_2, ..., x_{n-1})$ において, θ を積分消去することで導出が可能である. α は, Stick-breaking 過程におけるベータ分布 $\mathrm{Be}(1, \alpha)$ の α と等価である.

図 1. 木構造 Stick-Breaking 過程の挙動



3.2. 木構造 Stick-Breaking 過程

ディリクレ過程は分割の確率過程である. 上述の通り Stick-breaking 過程は長さ 1 の棒を折っていく状況を表しており, 分割の確率過程であることが容易に想像できる. 木構造 Stick-breaking 過程 [2] では, この Stick-breaking 過程を縦方向と横方向に適用することにより, 木構造及び階層性を生成し, 木構造における各ノードがカテゴリを表している. その挙動は図 1 に示す. また次の式に表される;

$$\begin{aligned} \epsilon &= (\epsilon_1, \epsilon_2, ..., \epsilon_K), \qquad \epsilon_k \in \mathbb{N}, \\ v_\epsilon &\sim \mathrm{Be}(1,\ \alpha(|\epsilon|)), \\ \psi_\epsilon &\sim \mathrm{Be}(1,\ \gamma), \end{aligned} \tag{5}$$

$$\begin{aligned} \pi_\epsilon &= v_\epsilon \varphi_\epsilon \prod_{\epsilon' \prec \epsilon} \varphi_{\epsilon'} (1 - v_{\epsilon'}) \\ &= v_\epsilon \prod_{\epsilon' \prec \epsilon} (1 - v_{\epsilon'}) \prod_{\epsilon' \preceq \epsilon} \varphi_{\epsilon'} \end{aligned} \tag{6}$$

$$\varphi_{\epsilon\epsilon_i} = \psi_{\epsilon\epsilon_i} \prod_{j=1}^{\epsilon_i - 1} (1 - \psi_{\epsilon_j}). \tag{7}$$

上式 (6),(7) が表すように Stick-breaking 過程が掛け合わされたような構造になっている. ϵ は各ノードを表し, ϵ_k は, ノードに至るまでのインデックスを示す. $\{\epsilon\epsilon_i : \epsilon_i \in 1, 2, ...\}$ は ϵ に対する子ノードを表し, $\epsilon' \prec \epsilon$ は ϵ に対する親ノードを表す. $\epsilon' \preceq \epsilon$ は, 親ノードに現在のノードを含めた集合を意味する. 木構造 Stick-breaking 過程は, 初期のノード (根ノード) から各ノードに下向きに降りていき, それぞれのノードに対し, 対象となるデータが止まる確率を導入することで, 階層的な構造に対するカテゴリの事前分布を構成することができる. ノード ϵ に対して, そのノードで止まる確率を v_ϵ で定め, 止まらなかった場合の子ノードの選択確率を $\psi_{\epsilon\epsilon_i}$ が定めている. また階層的な構造において, 各ノードが浅い位置 (根ノードに近い) ほど, 入力されたデータが通過しやすく, 逆に深ければ深いほど, データが止まりやすくなることが理想的である. そのため縦の動きに対するデータ集中度を決める α に対して, 深さによって通過しやすさを制御するための深さパラメータを導入する. 定式は次の通りである;

$$\alpha(\epsilon) \ = \ \alpha_0 \lambda^{|\epsilon|}. \tag{8}$$

ここで, 上式 (9) において, $\alpha_0 > 0$ を満たし, また $\lambda \in (0, 1]$ を満たす. ここで $|\epsilon|$ はノード ϵ における深さを表す自然数である. したがって, 木構造 Stick-breaking 過程 [2] におけるパラメータは $(\alpha_0, \gamma, \lambda)$ となる. 次に v_ϵ はベータ分布における確率変数であるから, データ $\mathrm{X} = \mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}$ が与えれた時, ベータ分布の事後確率分布は, N_ϵ: ノード ϵ で止まったデータの数, $N_{\epsilon\prec\cdot}$: ノード ϵ で止まらずに通過したデータの数とした時, $\mathrm{Be}(\mathrm{N}_\epsilon + 1, \mathrm{N}_{\epsilon\prec\cdot} + \alpha(|\epsilon|))$ となる. その期待

値は次のように示すことができる;

$$\mathbb{E}[v_\epsilon|\ \mathrm{X}] = \frac{\mathrm{N}_\epsilon + 1}{\mathrm{N}_\epsilon + \mathrm{N}_{\epsilon\prec\cdot} + \alpha(|\epsilon|) + 1} \tag{9}$$

また入力されたデータがノード ϵ において, $1 - v_\epsilon$ の確率で, 子ノードに降りた場合, 子ノードの選択確率は, 中華料理店過程により次のように定式化される;

$$\begin{cases} \dfrac{N_{\epsilon\epsilon_i} + N_{\epsilon\epsilon_i\prec\cdot}}{N_{\epsilon\prec\cdot} + \gamma}, & i \in \{1, 2, ..., k\} \\ \dfrac{\gamma}{N_{\epsilon\prec\cdot} + \gamma}, & i = k + 1 \end{cases} \tag{10}$$

上記の計算を入力されたデータに対して再帰的に繰り返すことで, 木構造 Stick-breaking 過程において階層的な事前分布を構成できる. 階層的な構造を得ることができるノンパラメトリックベイズ法では, 木構造 Stick-breaking 過程以外に, David et al[6] が存在する. しかし, この手法では深さに対して, 別のディリクレ過程を事前分布とすることで, 無限の深さと広さを持つ階層構造を構成しているが, ノードの広がりが定型的である. 例えば, あるカテゴリにおいて, その子カテゴリが無数に多く, 一方ではあるカテゴリとその子カテゴリが一対一というような実際によく見られるカテゴリ間のばらつきが表現できない. また Stick-breaking 過程を再帰的に繰り返すことで階層を生成する方法として Polya 木 [7] があるが, この方法だと, 途中のカテゴリにデータが存在せず, カテゴリの最も深い部分にのみ存在してしまうため本研究の意図に沿わない. したがって本研究では, 木構造 Stick-Breaking 過程を用いて構造のモデリングを行う.

4. 本研究の提案

4.1. 使用データについて

本研究の取り組みにあたり, 小売店の保有するデータを代理で分析を行っている企業で用いられている商品分類を対象とする. 表 1 はその分類構造におけるカテゴリの構成例を表す. 階層 1 は, 8 項目があるが, 今回は, そのうちの 1 項目である「農産」を対象に分類を行う. また階層 2 には, 野菜, 野菜加工品などの粒度の分類がなされる. 階層 3 には, 果菜, 根菜等, 階層 4 には, トマト, ごぼう等の粒度で分類されている. 商品の分類にあたり, 全対象商品 (7,329 個) のうち, 各カテゴリに属する商品の 2/3 個分 (4,832 個) の所属情報を学習データとして与える. 残りのデータ (2497 個) をテストデータとして用いる.

本研究では, 階層分類に十分な情報量を低コストで取得するために, 商品名に対する Web ブラウザの検索画像を採用し, 自動的に収集している. このとき直接, 商品名に対して検索を行うと, 分析に見合う検索結果を得ることが難しいため, 商品名に対して形態素解析を行い, 検索に効果的な特徴語を探索し, その単語に絞り画像を収集している.

表 1. 対象とする商品分類のカテゴリ構造

階層1	階層2	階層3	階層4
農産	野菜	果菜	トマト
		根菜	ごぼう
	野菜加工品	⋮	⋮
1分類	4分類	24分類	159分類

4.2. 特徴抽出

本研究では, テキストデータである商品名及び分類構造のカテゴリ情報と検索画像データの二種の非構造データを使用しているため, そのままデータを解析することは難しい. そこで, 特に非構造データである画像データから特徴抽出するために, 本研究では深層学習の一つである自己符号化器 (Auto Encoder) を用いている. 自己符号化器を用いることで, 高次元データの次元削減を行っている.

4.3. モデル・定式化

本研究にあたり, 主として二つのモデリングが必要となる. 構造全体と各カテゴリに対するモデリングである. 構造に対するモデリングとして前述の通り木構造 Stick-Breaking 過程 [2] を採用する. また今回, 既に存在する分類構造を対象にするため, 初期状態として 4.1 に述べたカテゴリ情報を与えている. ここで各カテゴリに対するモデリングについて説明する. 本研究では木構造 Stick-Breaking 過程 [2] の原論を参考に多次元のベルヌーイ分布を仮定したロジスティック関数を使用する;

$$f(x_n \mid \theta_\epsilon) \ = \ \prod_{d=1}^{\dim} (1 + \exp\{-{\theta_\epsilon}^{(d)}\})^{-{x_n}^{(d)}} (1 + \exp\{{\theta_\epsilon}^{(d)}\})^{1-{x_n}^{(d)}}. \tag{11}$$

ここで, θ_ϵ はノードパラメータを示す. また x_n は $x_n \in \{0,1\}^{\dim}$ となる入力データを示す. 木構造 Stick-Breaking 過程 [2] の原論で述べられているようにノードパラメータの事前分布において, 親の確率分布と子の確率分布に類似性があることが自然であると考えられる. そのため親の確率分布の拡散過程として子に対して親の特徴を継承している. 本研究では, ノードパラメータはガウス分布に従うと仮定して, 子は親の平均を期待値とする拡散過程を採用している;

$$T_{norm}(\theta_{\epsilon\epsilon_\mathrm{i}} \leftarrow \theta_\epsilon) \ = \ \mathcal{N}(\theta_{\epsilon\epsilon_\mathrm{i}} \mid \eta\ \theta_\epsilon, \Lambda). \tag{12}$$

表 2. 構造を固定した際の正答率

	階層 1	階層 2	階層 3	階層 4
カテゴリ数	1	4	24	159
accuracy	—-	0.88021	0.20833	0.03.125

ここで $\eta \in [0,1)$ はであり, 親のノードパラメータに対し適当なノイズを乗せていることに等しい.

4.4. 推定方法

構造の推定において, $p(\epsilon \mid \mathrm{X})$ に従ってサンプリングすれば良いが, 無限の木構造を成していることから単純なサンプリング法では難しい. そこで, 木構造 Stick-Breaking 過程 [2] で述べられているサンプリング法を採用し, マルコフ連鎖モンテカルロ法の一種であるスライスサンプリングと Retrospective sampling の併用により効率的にサンプリングする. またハイパーパラメータの推定には, 同様にスライスサンプリングにより推定を行う. ここでハイパーパラメータ $(\alpha_0, \gamma, \lambda)$ に対して, 値の範囲を条件として与えた条件付き事後分布からサンプリングを行う.

$$
\begin{aligned}
p(\alpha_0, \lambda | \{v_\epsilon\}) &\propto \prod_{\epsilon} \mathrm{Be}(v_\epsilon | 1, \lambda^{|\epsilon|}\ \alpha_0) \\
p(\gamma | \{\psi_\epsilon\}) &\propto \prod_{\epsilon} \mathrm{Be}(\psi_\epsilon | 1, \gamma).
\end{aligned}
\tag{13}
$$

ノードにおけるパラメータの推定においては, マルコフ連鎖モンテカルロ法の一種であるハミルトンモンテカルロ法により推定を行っている. 構造の決定には, 確率 $p(\{x_n, \epsilon_\mathrm{n}\} \mid \{v_\epsilon\}, \{\psi_\epsilon\}, \alpha_0, \lambda, \gamma)$ の最大化により行っている.

5. 検証と考察

5.1. 検証

サンプリングの試行回数は 10,000 回に設定し, ハイパーパラメータは最頻値を用いる. 検証に関して, 冒頭において述べた特徴の中でも「曖昧なカテゴリの存在」及び「新しいカテゴリへの許容性」に対して上手く適合できているかという観点のもと検証を行う. なお曖昧なカテゴリの表現方法として分類する対象について末端まで分類せず親のカテゴリで対象を止めることで表現する. また全体としての構造の拡張を止めた場合において正答率という観点のもと検証を行う.

5.2. 考察

まず曖昧なカテゴリの存在を示す「その他」という分類を考察する. 対象とする分類構造において,「その他」を表すものは 25 個存在する. そのカテゴリに分類されるべき対象の一部で, 末端カテゴリでない中間のカテゴリに対象が所属していることが確認された. 次に「新しいカテゴリへの許容性」については, 一部のカテゴリにおいて構造の拡張を許して解析を行った. 結果としてデータの投入に対応して新たなカテゴリへの許容性が確認された. したがって現実の特徴に沿った分類を可能としている. 次に構造を固定したときの全体としての正答率に関して, 階層 2 において高い精度を算出している. しかし階層 4 での正答率には課題が残る. これについてデータ, モデリングともに要因があると考えられる. データについて, Web 上から収集したデータは, 特徴語に絞り検索した画像であるが, 一部適切に対象を表現していない画像が収集され, その結果, 効率的な特徴抽出を困難にしていた. これに対して, データを増やしデータの精度を安定させるまたはデータを選別することで対処は可能である. 次にモデリングについて解析結果から, 特定のカテゴリにおいて π_ϵ が 0.9 以上の値を確認した. これはカテゴリごとのデータの所属に偏りがある場合に, $\{v_\epsilon, \psi_\epsilon\}$ がベータ分布の事後確率分布から生成されるためだと考えられる.

6. おわりに

本研究では, 現実にある「3 つ」の分類構造の特徴に上手く適合した手法の確立を目的とし, 木構造を生成するノンパラメトリックベイズ法を用いて解析に取り組んだ. 検証結果から木構造 Stcik-Breaking 過程を用いることでその特徴を表現できた. また今後の課題として, カテゴリごとのデータの偏りに対応する変数の導入などが挙げられる.

参考文献

[1] 持橋大地 : “最近のベイズ理論の進展と応用 (III) ノンパラメトリックベイズ”, 電子情報通信学会 電子情報通信学会論文誌, Vol.93,pp.1-6 (2010)

[2] R. P. Adamas, G.Zoubin and M.I.Jordan:“Tree-Structured Stick Breaking Process for Hierarchical Data”, Advances in Neural Information Processing Systems, Vol.23, pp.19-27 (2010)

[3] Williams, C.:“A MCMC approach to hierarchical mixture modelling”, Advances in Neural Information Processing Systems, Vol. 23, pp. 680-686 (2000)

[4] Neal, R. M.:“Density modeling and clustering using Dirichlet diffusion trees”, *Bayesian Statistics 7*, pp. 619-629 (2003)

[5] T. Ferguson:“A Bayesian analysis of some nonparametric problems”, *The Annals of Statistics*, Vol. 1, pp.201-230 (1973)

[6] D. M. Blei, T. L. Griffiths and M. I. Jordan: “The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies”, *Journal of the ACM*, Vol. 57, pp.1-30(2010)

[7] R. D. Mauldin, W. D. Sudderth, and S. C. Williams: “Polya Trees and Random Distributions”, *The Annals of Statistics*, Vol.20, pp.1203-1221 (1992)