

企業の採用活動と被エントリ数の関係性に着目した分析モデルに関する一考察
 An Analytic Model of Relation between
 Company's Recruitment Activities and Number of Students' Application

永森 誠矢[†] 山下 遙[†]
 Seiya Nagamori Haruka Yamashita

後藤 正幸[†] 荻原 大陸[‡]
 Masayuki Goto Ogiwara tairiku

1 研究背景・目的

近年、就職活動において学生は就職ポータルサイトを利用することが一般的となっており、企業もまたそれを用いて学生に求人情報を提供している。企業は就職ポータルサイトを用いて採用活動を行なうことで多くの学生に情報を提供でき、被エントリ数が向上することが期待できる。その際、就職ポータルサイトを活用しようとする企業は、採用活動の被エントリ数の予測値とその変動要因に関心がある。

企業が獲得できる被エントリ数を予測する際、就職ポータルサイトに顕在化している情報のみならず、企業の学生からの認知度や景気などの外的要因が影響すると考えられる。また、企業の業種や従業員規模などの基本情報も企業の採用における行動情報と被エントリ数の関係に影響していると考えられる。そこで本研究では、混合回帰モデルを用いて、企業の採用活動における行動情報と被エントリ数の関係モデルの構築を行なう。さらに混合回帰モデルのパラメータ推定において企業の基本情報をモデルに取り込むことで、より高精度なモデルを構築する方法の検討を行なう。また構築したモデルを用いて企業の採用活動と学生からの被エントリ数の関係性を解析し、その有効性を示す。

2 混合回帰モデル

本研究では企業の行動情報と被エントリ数の関係性を構築することを目的としている。ここでいう行動情報とは、就職ポータルサイト上で行われる企業のインターンシップの募集やインタビューの掲載の有無などの企業が選択実施可能な採用活動オプションのことを指す。上記の目的を達成するための最も基本的なモデルは重回帰モデルであるが、単一の重回帰モデルでは企業ごとの採用における行動情報と被エントリ数の関係性の差異を考慮できない。しかし、例えば2つの企業が同じ採用行動を行ったとしても被エントリ数の増減に影響する効果は異なる。これは企業の持っている基本情報や外的要因によって決められると考えられる。

そこで本研究では、企業の行動情報と被エントリ数の関係性が類似した企業群は同じ潜在クラスに所属することを仮定した混合回帰モデル [1] を導入する。これにより個々の企業の混在的特徴を考慮して関係性を分析することが可能となる。また、適切な企業群の潜在クラスを推定することで、より推定精度が高く、説明能力の高い回帰モデルが構築され、より正確な分析が行えるようになる。

混合回帰モデルの学習は、EM アルゴリズムを用いて行われ、重みを用いた回帰モデルのパラメータ推定および重みの更新からなる。

2.1 モデル式

混合回帰モデルは K 個の潜在クラス集合を $Z = \{z_k : 1 \leq k \leq K\}$ としたとき、各潜在クラスにおける確率密度関数 $P_k(y_l | \mathbf{x}_l)$ の線形結合によりモデル化される。いま、 L 社の企業のうち l 番目の企業の行動情報を表す説明変数ベクトルを $\mathbf{x}_l = (1, x_{l1}, x_{l2}, \dots, x_{lI})^T$ 、目的変数である被エントリ数を y_l 、また潜在クラス z_k における $I + 1$ 個の回帰モ

デルのパラメータを $\beta_k = (\beta_{0k}, \beta_{1k}, \beta_{2k}, \dots, \beta_{Ik})^T$ としたとき、混合回帰モデルは以下の式 (1) で示される。

$$P(y_l | \mathbf{x}_l) = \sum_{k=1}^K w_{lk} P_k(y_l | \mathbf{x}_l) \quad (1)$$

ここで、 w_{lk} は k に関する和が 1 となる企業の潜在クラスへの重みであり、潜在クラス z_k ごとに仮定される正規分布の確率密度関数は次式で表される。

$$P_k(y_l | \mathbf{x}_l) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_l - f_k(\mathbf{x}_l))^2}{2\sigma_k^2}\right) \quad (2)$$

$$f_k(\mathbf{x}_l) = \sum_{i=0}^I \beta_{ik} x_{li} \quad (3)$$

2.2 パラメータの推定方法

混合回帰モデルのパラメータである $\beta_k, w_{lk}, \sigma_k^2$ を EM アルゴリズムを用いて推定する方法を示す。ここでは、「 β_k, σ_k^2 の推定」と「推定した σ_k^2 を用いた w_{lk} の推定」を繰り返し学習を行なう。

まず、重み w_{lk} については、回帰モデルの推定精度が高い潜在クラスに対し大きい重みを持つことが望ましい。そこで精度良く推定されやすい潜在クラスへの重みが大きくなるように更新を行なう。各企業は複数の潜在クラスに確率的に所属することが可能であり、回帰式への当てはまりが良い潜在クラスへの重みが大きくなることで、各企業に対する被エントリ数の推定精度の向上が期待できる。重み w_{lk} は以下の式 (4) で計算され更新される。

$$w_{lk} = \frac{P(z_k) P_k(y_l | \mathbf{x}_l)}{\sum_{k'=1}^K P(z_{k'}) P_{k'}(y_l | \mathbf{x}_l)} \quad (4)$$

ここで、各潜在クラスにおける分散 σ_k^2 および混合割合 $P(z_k)$ はそれぞれ (5)-(6) で更新される。

$$\sigma_k^2 = \frac{\sum_{l=1}^L w_{lk} (y_l - f_k(\mathbf{x}_l))^2}{\sum_{l=1}^L w_{lk}} \quad (5)$$

$$P(z_k) = \frac{\sum_{l=1}^L w_{lk}}{L} \quad (6)$$

これらの更新式では各企業を K 個の潜在クラスに確率的に所属させ、その重みを用いて各潜在クラスで回帰モデルを構築することを考えている。潜在クラス z_k における各パラメータ β_k は重み付き残差の二乗和 S_k^2 を最小化することで推定する。

$$S_k^2 = \sum_{l=1}^L w_{lk} (y_l - f_k(\mathbf{x}_l))^2 \quad (7)$$

上記の β_k のパラメータ推定を用いることにより大きい重みをもつ企業に適應するようにモデルが作用し、企業の特徴を回帰モデルのパラメータに反映させることができる。

[†]早稲田大学

[‡]株式会社 リクルートキャリア

2.3 混合回帰モデルのアルゴリズム

混合回帰モデルは以下のアルゴリズムで構築される。

- Step1 各企業の各潜在クラスへの所属確率 (重み) w_{ik} の初期値をランダムに設定する。
- Step2 それぞれの潜在クラスに対し式 (7) を最小化する回帰モデルのパラメータ β_k を推定する。
- Step3 式 (4)-(6) を計算し、各企業の重み w_{ik} を更新する。
- Step4 収束条件を満たしていれば終了。さもなければ Step2 にもどる。

□

3 実データへの適応の際のパラメータ初期値の検討

混合回帰モデルでは推定精度が Step1 で設定する初期値に依存する。そこで、効率的な初期値設定のために、企業を確率的にクラスタリングする手法である Aspect Model (以下 AM) [2] を導入する。ここで企業の採用活動における行動情報と被エン트리数の関係性に影響を与える一つの要素として、企業が持つ業種や従業員規模などの基本情報が考えられる。この基本情報により AM を用いて企業を確率的にクラスタリングを行い、その潜在クラスに対する所属確率を混合回帰モデルの重み w_{ik} の初期値とすることで、安定した推定精度かつより高精度なモデルとなることが期待される。

いま、 J 個の企業の基本情報のうち j 番目の基本情報の要素を表す集合を $D^j = \{d_{n_j}^j : 1 \leq n_j \leq N_j\}$ とする。ここで $d_{n_j}^j$, N_j はそれぞれ j 番目の基本情報の n_j 番目の要素、要素数を表す。このとき、AM の確率モデルは以下の式 (8) で示される。

$$P(d_{n_1}^1, d_{n_2}^2, \dots, d_{n_J}^J) = \sum_{k=1}^K P(z_k) \prod_{j=1}^J P(d_{n_j}^j | z_k) \quad (8)$$

ただし、式 (8) における各パラメータは EM アルゴリズムにより推定する。推定されたパラメータを用いて企業の潜在クラスに対する所属確率を計算し、混合回帰モデルの重みの初期値とする。

4 混合回帰モデルを用いた実データ分析

本節では、就職ポータルサイトに蓄積した実データを用いてモデル推定を行い、得られる結果を分析する。本実験では推定精度に着目している。

4.1 分析条件

実験データとして 2015 年度卒業の学生に対する就職ポータルサイト上で 100 件以上 1000 件以下の被エントリを獲得した企業約 6000 社 ($L \approx 6000$) のデータを用いる。目的変数を各企業の被エントリ数とし、説明変数として就職ポータルサイトに蓄積されているデータから利用可能な企業の採用活動における 6 つの行動オプションを用いた。また、初期値決定のための基本情報として 4 変数を用いた。潜在クラス数 K は $K=5, 10, 15$ として実験を行った。比較モデルとして領域を分割して、それぞれに別の線形式を当てはめる区分線形モデルを用いた。区分線形モデルにおいては、被エントリ数の大きさで企業を等分割するように領域を設定した。本実験の区分線形モデルにおける K は、領域分割する時の分割数を指す。この実験ではそれぞれの手法で各企業に対し被エントリ数の推定を行い、算出された推定値と実測値の平均絶対誤差を評価し、検討を行なうこととした。評価指標は以下の式 (9) で算出される。また、結果は 100 回実験を行った平均とする。

$$\text{絶対平均誤差} = \frac{\sum_{i=1}^L |y_i - \hat{y}_i|}{L} \quad (9)$$

4.2 検証結果と考察

結果を表 1 に示す。

表 1. 絶対平均誤差の比較結果

	$K=5$	$K=10$	$K=15$
区分線形	44.80	22.22	15.13
混合回帰 (ランダム)	43.01	20.91	15.49
混合回帰 (AM)	40.49	22.47	15.58

表 1 より、潜在クラス数 K の数によりデータへの当てはまりの面で優れているモデルが異なることがわかる。潜在クラスが少ない状況下では、混合回帰モデルは企業が基本情報を考慮した上で複数の潜在クラスに確率的に所属できることで、より特徴を表現できていると考えられる。また、混合回帰モデルでは EM アルゴリズムを用いた学習により、推定精度を向上させる上で適切なパラメータが推定され、説明能力の高い被エントリ数と企業の行動情報の関係モデルが構築されていると考えられる。またこの結果から、企業の基本情報が行動情報と被エントリ数の関係性に影響していることが示唆される。一方、潜在クラスが多い状況下では回帰モデルの推定精度が目的変数である被エントリ数の大小に大きく依存するようになり、区分線形モデルの精度が良くなったと考えられる。

5 提案手法を用いた分析

得られた回帰モデルを用いた分析により、有用な知見が得られることを示す。ここでは、混合回帰モデル (AM) において潜在クラス数 $K=5$ の時のパラメータを表 2 に示す。

表 2. 混合回帰モデル (AM) において推定されたパラメータ

潜在クラス k	β_{0k}	β_{1k}	β_{2k}	β_{3k}
1	155.15	-1.35	0.13	-2.90
2	284.42	0.66	-0.16	-4.63
3	302.42	11.40	0.44	40.30
4	580.31	-7.56	1.33	39.65
5	959.75	-8.86	-0.03	12.10
潜在クラス k	β_{4k}	β_{5k}	β_{6k}	混合割合
1	7.82	-14.92	-7.73	0.14
2	-135.97	437.22	-59.43	0.25
3	643.55	-70.73	-27.02	0.21
4	-99.05	-177.96	-66.76	0.34
5	-234.11	-29.68	-5.94	0.06

表 2 より推定されるパラメータは潜在クラスごとに異なることがわかる。このように、企業が所属する潜在クラスによって、被エントリ数向上に有効な就職ポータルサイト上での採用行動が異なることがわかる。この推定されたパラメータを分析することで、企業は各々の特徴に応じて被エントリ数向上に有効な採用活動計画を立てる一つの指針となることが期待できる。

6 まとめと今後の課題

本研究では企業の採用における行動情報と被エントリ数の関係性を混合回帰モデルを用いて分析を行った。また、企業の特徴をより柔軟に表現する混合回帰モデルの効果および初期値依存問題の検討を行った。今後の課題として、推定精度の向上、新規掲載企業に対する被エントリ数の予測、具体的な企業の採用における行動計画のサポート手法の検討などが挙げられる。

参考文献

- [1] Faria, S., and Soromenho, G. "Fitting mixtures of linear regressions." *Journal of Statistical Computation and Simulation*, Vol. 80, No. 2, pp201-225, 2010.
- [2] Hofmann, T. "Probabilistic Latent Semantic Analysis." *Proc. of UAI'99*, pp.289-296, 1999.