情報処理学会論文誌 Vol.59 No.4 1273–1285 (Apr. 2018)

混合回帰に基づく就職ポータルサイトの被エントリ数分析モデルに関する一考察

永森 誠矢[1] 山下 遥[2] 荻原 大陸[3] 後藤 正幸[1,a)]

受付日 2017年4月13日, 採録日 2018年1月15日

概要：近年，企業は就職ポータルサイトを用いて学生に採用情報を提供している．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への影響とその予測値に関心がある．そこで本研究では，就職ポータルサイトに蓄積されている履歴データを活用し，新規企業が獲得できる被エントリ数の予測と被エントリ数の影響要因分析のためのモデルを構築する方法について検討する．具体的には，精度の高い予測とともに，影響要因の効果を分析可能とするモデルとして，各企業が持つ潜在的要因を考慮した混合回帰モデルを提案する．提案したモデルを就職ポータルサイト上の実データに適用し，企業の採用活動と学生の被エントリ数の関係性を解析し，その有効性を示す．

キーワード：回帰モデル，予測モデル，潜在クラスモデル，確率モデル，就職ポータルサイト

A Study of Analysis Model of Number of Students' Applications on Internet Portal Site for Job-hunting Based on Mixture Regression

Seiya Nagamori[1] Haruka Yamashita[2] Tairiku Ogihara[3] Masayuki Goto[1,a)]

Received: April 13, 2017, Accepted: January 15, 2018

Abstract: In recent years, many Japanese companies use Internet portal sites for job-hunting for efficient recruitment. From the companies' viewpoints, there are mainly two interests for using Internet portal sites for job hunting: the predicted number of application from students through the sites and the effect of recruitment activities on the sites. In this study, for the prediction of number of students' applications, we propose a new predictable regression model considering each company's potential factor. In order to verify the effectiveness of proposed model, we demonstrate an analysis of the relation between the companies' recruitment activities and the number of students' application.

Keywords: regression model, prediction model, latent class model, probabilistic model, Internet portal site for job-hunting

1. 研究背景・目的

日本の大学生の就職活動において，近年，就職ポータルサイトの活用が一般的なものとなりつつある．企業は，効率的な採用活動のために就職ポータルサイトを通じて学生に求人情報を提供している．就職ポータルサイトを用いて採用活動を行うことで多くの学生に情報を提供できるようになり，その結果多くの学生からのエントリを期待できることが企業側のメリットとなっている．そのため，就職ポータルサイトを活用しようとする企業は，就職ポータルサイト上での採用活動に対して獲得できる被エントリ数とその変動要因に関心があるといえる．

ここで，企業が獲得できる被エントリ数と，採用活動における行動の関係性を分析する最も基本的な手法として線形回帰モデル [1], [2] が考えられる．しかしながら，被エン

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–0072, Japan
[2] 上智大学
Sophia University, Chiyoda, Tokyo 102–0081, Japan
[3] 株式会社リクルートキャリア
Recruit Career Co., Ltd., Chiyoda, Tokyo 100–6640, Japan
[a)] masagoto@waseda.jp

トリ数と企業が行う採用活動の関係には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業界などの企業の潜在的要因によって，統計的特徴が異なることが考えられる．すなわち，個々の企業特性や認知度などの特徴の異なる企業グループが混在しており，これらに対し単一の回帰モデルを当てはめると，精度の高いモデルの推定が困難となることが考えられる．

一方，単なる予測モデルとしては，近年，様々な機械学習手法に基づく予測モデルが提案されており，たとえば，回帰木やニューラルネットなど [1], [3], [4] の予測モデルを適用可能である．しかしながら，これらの手法では，被エントリ数とその要因との関係，すなわち企業が行う採用活動が与える被エントリ数への影響について明示的に分析することが困難である．本研究で対象とする企業が獲得できる被エントリ数の予測モデル構築では，単に予測が行えればよいだけではなく，「各企業はどのような採用活動を行えば，被エントリ数を効果的に増やせるのか」についても分析できることが望ましい．

このような単一な回帰モデルの適用が難しい問題に対し，混合回帰モデル [5], [6], [7] を用いた予測モデルの構築が可能である．しかしながら，説明変数と目的変数のみを用いた回帰モデルを単純に混合したモデルでは，新たに就職ポータルサイトを利用しようとする企業の被エントリ数や採用活動のインパクトの予測に用いることができない．一方で，本研究で対象としている問題では，業種や従業員規模などの企業の基本情報は上記の関係性に影響を与える要因の 1 つであり，新規掲載企業に対し予測可能なモデルへの拡張に補助変数として貢献するものと考えられる．

本研究では，新たに就職ポータルサイトを利用する企業に対する被エントリ数の予測と採用活動のインパクト評価を可能とする，混合回帰モデルをベースとした新たな分析–予測モデルの提案を行う．具体的には，企業の基本情報の背後に潜在クラスを仮定することで，新規企業に対しても被エントリ数の予測が可能となり，かつ汎化能力の高いモデルが得られると期待できる．この提案モデルにより，新規掲載企業に対しても被エントリ数や採用活動のインパクトの予測が可能となり，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性を表現し，獲得可能な被エントリ数と採用活動における行動情報のインパクトの予測を可能とする．混合回帰モデルの拡張として回帰モデルの偏回帰変数に直接影響する説明変数以外の補助変数を用いたモデル化が存在する [8], [9], [10]．本研究の提案モデルは，補助変数を用いた混合回帰モデルに関する研究の枠組みの中で，文献 [8] で示されている一般モデルの具体的なケースを与えている．文献 [8] で示されている混合回帰モデルでは，一般的な混合回帰モデルの関数形が示されているものの，具体的な構成例としては限られた一部の具体例が示されているのみであるが，本研究で提案しているモデルでは，問題設定を考慮したそのきわめて具体的な回帰モデルの構成法について示しているといえる．さらに本研究では，就職ポータルサイトに蓄積されている実データに対して提案するモデルを適用し，予測精度の観点で混合回帰モデルを導入することの有効性を確認するとともに，推定されたモデルを分析することで有用な知見が得られることを示す．

2. 準備

2.1 日本の就職活動および就職ポータルサイト

日本の多くの学生は就職活動の際に就職ポータルサイトを利用している．就職ポータルサイトとは登録することで，就職活動における包括的なサービスを受けることができる Web サービスである．

学生は就職ポータルサイトのサービスを利用することにより，学生は就職活動の進め方や就職活動の基礎知識など就職活動を行うにあたり必要な情報を得ることができる．また，自己分析，企業分析，インターンシップへのエントリ，本選考へのエントリなど，このサービスによってユーザは包括的な就職活動を行うことができる．また現在日本には様々な就職ポータルサイトが存在し，一般的に学生は複数の就職ポータルサイトに登録を行い，並行して用いることで多角的な視点から情報を入手するとともに，様々な企業の検索を行っている．

学生の本選考をサポートすることを目的とした就職ポータルサイトは 2015 年までに卒業した学生に対するサービスは 12 月に，2016 年以降に卒業する学生に対するサービスは 3 月に開始される[*1]．多くの学生は就職ポータルサイトを通じて企業の検索や企業の説明会の予約，エントリなどを行う．その後 2015 年までに卒業した学生は 4 月に，2016 年に卒業した学生は 8 月，それ以降に卒業する学生は 6 月に会社との面接を開始し[*1]，学生は就職先が決まった後に就職ポータルサイトの利用を終了することとなる．

一方，企業にとっての就職ポータルサイトの利用には，多くの学生が企業の情報を得やすくなることによるエントリ数の向上，またエントリなどの採用活動の一部プロセスを就職ポータルサイトで行うことによる採用活動の効率化，などといった狙いがある．就職ポータルにおける企業の採用活動とは学生の本選考におけるエントリの管理に限らず，インターンシップのエントリの管理，説明会の実施など学生への情報の提供を含め様々な採用活動が存在する．現在日本には多くの就職ポータルサイトが存在するため企業は複数の就職ポータルサイトへの情報掲載を行っている．このことにより，多くの学生に企業を知ってもらう，またエントリへの手間を減らすという効果がある．

以上のように企業は複数の就職ポータルサイトへの情報

[*1] 日本経済団体連合会の採用選考に関する企業の倫理憲章や採用選考に関する指針 [11], [12], [13] により採用選考活動開始時期の取り決めが行われている．

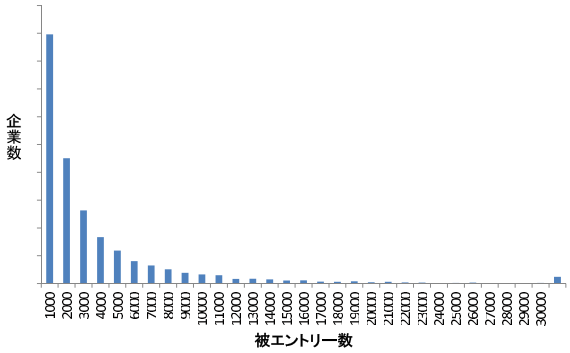

図 1 2014 年卒業学生の企業の被エントリ数と企業数
Fig. 1 Number of entries from students who graduated in 2014 and number of companies.

掲載を行っているため，多くの就職ポータルサイトでエントリを募ることによって被エントリ数の向上も期待できる．また，メディアへの露出による学生人気の向上などは被エントリ数の増加に大きく寄与すると考えられる．しかしながら，過大な広告や PR 活動はコストがかかるため，各々の企業はコストや被エントリ数向上の効果など多くの要素を考慮して採用計画を策定していくことが求められている．**図 1** に 2014 年卒業の学生からの企業の被エントリ数と企業数の関係を示した．

図 1 より，企業によって獲得できる被エントリ数に差があることが明らかになった[*2]．特に，被エントリ数が少ない企業が多く存在することが分かる．このような企業にとって，多くの学生のエントリを獲得することは 1 つの大きな課題である．また，多くのエントリを獲得できている企業も割合としては少数であるが存在している．

また，就職ポータルサイトのメリットとして企業が採用対象とする多くの学生がユーザとして存在することがいえる．このように学生が集まる当該サービスにおいて企業が採用行動を起こすことは大きな効果を生むことも考えられる．

2.2 関連研究

新卒学生の就職活動そのものに対する分析や，就職活動支援の方法については社会学的なアプローチによる様々な研究がなされている [14], [15], [16], [17], [18], [19]．これらの研究では，学生の就職活動に対する意識や就職活動が与えるメンタルヘルスへの影響など，就職活動に関わる問題に対して，対象者への意識調査や公的統計データに基づく社会学的な研究をベースとした議論がなされている．一方で，就職活動支援の方法について検討を行っている研究もなされている [20], [21], [22], [23], [24], [25]．しかし，就職活動支援という意味では，学生相談室や就職サポート部門のあり方など，学生へのキャリア教育のあり方や組織体制に関する調査研究が主たる議論の対象となっている．

一方で，就職ポータルサイトのデータベースに蓄積される学生の履歴データを用いた統計的分析モデルに関する研究は始められた段階といえる．早川ら [26] は学生の履歴データを用いて，学生の属性情報に基づく就職活動の終了時期を予測するモデルを提案している．これは，層別木と混合ワイブル分布を併用したモデルであり，ある種の混合モデルの有用性を指摘している．Yamagami ら [27] は，早川らのモデルとは異なり，よりシンプルな形の潜在クラスモデルを提案し，実データ分析の結果を示している．潜在クラスモデルに基づく就職活動に関するデータの分析方法については，他の観点からの分析モデルも議論されている．坂元ら [28] は企業のアピールするポイントと学生の企業に対する志望理由の関係性に着目し，マッチング分析モデルを構築している．この研究では，学生と企業の双方に着目した観点から学生と企業のマッチング分析を行っている．これに対し，Sugiyama ら [29] は，個社ページの閲覧行動とエントリ行動の関係性をベクトルで表し，企業と学生ユーザとの共起を表現する潜在クラスモデルを提案している．

一方，本論文と関連性の深い，就職ポータルサイトにおける企業の被エントリ数の予測を目的とした分析としては，ポワソン混合効果モデルを用いた研究がある [30]．この研究では「就職ポータルサイト以外の顕在化されていない外部要因が予測に悪影響を及ぼす可能性がある」，また「変数選択が非常に重要な問題である」という被エントリ数の予測問題に関する 2 点の課題について言及している．前述の課題に対してはポワソン混合効果モデルの導入，後述の課題に対しては説明変数が異なるモデルの混合，というアプローチで解決を図っている．しかし，モデルが複雑化しており，新規企業に対する予測精度については，既存企業よりも精度差が大きい点が課題とされていた．また，Nagamori ら [31] は，混合回帰モデルを用いて企業の被エントリ数と企業の行動情報の関係性を分析し，企業の基本情報を用いて混合回帰モデルの初期値の与え方について言及している．しかし，新規掲載企業の被エントリ数の予測には着目しておらず，関係性の分析に研究の重点が置かれていた．すなわち，本研究で対象としている新規掲載企業も含む企業の採用活動が被エントリ数に与える影響を分析するモデルとはなっていない．

2.3 混合回帰モデル

混合回帰モデル [5], [6], [7] とは目的変数 y と説明変数 $\boldsymbol{x} = (x_0, x_1, x_2, \ldots, x_d)^{\mathrm{T}}$ の線形構造の背後に潜在クラス [3] を仮定したモデルである．このモデルはそれぞれの潜在クラスに対し回帰モデルを仮定しており，それらの混合 [4], [32] により表現される．K 個の潜在クラスを仮定したとき，混合回帰モデルのモデル式は以下の式 (1)，(2) で表現される．

[*2] 対象事例の都合上，詳細の企業数に関しては伏せている．

$$h(y|\boldsymbol{x},\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k g_k(y|\boldsymbol{x},\boldsymbol{\theta}_k) \tag{1}$$

$$\pi_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

ここで，$\boldsymbol{\psi}$ は混合回帰モデルのすべてのパラメータを表すベクトルであり，π_k は混合割合，$g_k(\cdot)$ は平均 $\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{x}$，分散 σ_k^2 の正規分布を示し，$\boldsymbol{\theta}_k$ は $\boldsymbol{\beta}_k^{\mathrm{T}}$ および σ_k^2 の値を表すベクトルである．なお，$\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \ldots, \beta_{kd})^{\mathrm{T}}$ は回帰パラメータを示す．

このモデルでは新たな $\boldsymbol{x}$ に対する目的変数が入力されれば，y の予測は可能である．しかし，どのような $\boldsymbol{x}$ が入力されても，混合重みである π_k は不変であり，固定パラメータとなる．したがって，複数の回帰式を混合しているものの，その混合は 1 つの回帰式で表現できてしまうため[*3]，1 つの線形回帰式によるモデルの表現能力と同等の性能しか出すことができない．

一方，混合回帰モデルの拡張モデルとして，補助変数を用いたモデル [8], [9], [10] が提案されている．このモデルでは混合回帰モデルの混合割合が補助変数 $\boldsymbol{v}$ に依存するモデルとなっている．補助変数を用いた混合回帰モデルは以下の式 (3)，(4) で表現される．

$$h(y|\boldsymbol{x},\boldsymbol{v},\boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{v},\boldsymbol{\alpha}) g_k(y|\boldsymbol{x},\boldsymbol{\theta}_k) \tag{3}$$

$$\pi_k(\boldsymbol{v},\boldsymbol{\alpha}) \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k(\boldsymbol{v},\boldsymbol{\alpha}) = 1 \tag{4}$$

ここで，$\boldsymbol{\phi}$ は補助変数を用いた混合回帰モデルのすべてのパラメータを表すベクトルであり，$\boldsymbol{\alpha}$ は補助変数に対するパラメータである．このモデルの混合割合 $\pi_k(\boldsymbol{v},\boldsymbol{\alpha})$ は制約式 (4) を満たすように関数を設定すればよい．

2.4 被エントリ数に対する企業行動モデルの定式化

本研究では企業の行動情報と被エントリ数の関係性をモデル化している．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募集の有無など，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．

上記のモデル化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．しかし被エントリ数と行動情報の関係性には企業の特徴により異なる構造が混在していると考えられる．そのため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違いを表現することができない．被エントリ数の増減に影響する効果は異なるものと考えられ，これは企業の持っている基本情報や外的要因などの潜在的要因によって決められると考えられる．

[*3] たとえば，$y = 10 + 30x_1 + 30x_2$ という回帰式と $y = 20 + 10x_1 + 50x_2$ という回帰式を固定の重み 0.5 ずつで混合したとすると，y の平均値は $y = 15 + 20x_1 + 40x_2$ という単一の式で記述できてしまう．複数の多項式の重み付き平均は，やはり多項式になるため，より複雑な入出力関係を表現できるわけではない．

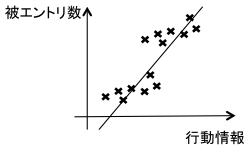

図 2 単一の重回帰モデルのイメージ
Fig. 2 Image of normal regression model.

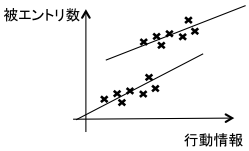

図 3 混合回帰モデルのイメージ
Fig. 3 Image of mixture regression model.

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し，同じ回帰式が当てはまることを仮定した混合回帰モデル [5], [6], [7] 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被エントリ数と採用活動における行動との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確率的に推定することで，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，より正確な解釈を与えることを可能とする．

単一の回帰モデルのイメージと混合回帰モデルのイメージを**図 2**，**図 3** に示す．

データの線形構造が複数存在し，単一の回帰モデルでは表現が困難な場合，任意の潜在クラス数を設定し混合することで，よりデータに適したモデルが推定可能となる．上記の例ではデータの線形構造が 2 つ存在し，2 つの潜在クラスを仮定した例となっている．

いま，K 個の潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ としたとき，混合回帰モデルは各潜在クラスにおける確率密度関数 $P_k(y_l|\boldsymbol{x}_l)$ の線形結合によりモデル化される．ここで，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は $\boldsymbol{x}_l = (x_{l0}, x_{l1}, x_{l2}, \ldots, x_{lI})^{\mathrm{T}}$，目的変数である被エントリ数は y_l である．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 $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \ldots, \beta_{kI})^{\mathrm{T}}$ としたとき，混合回帰モデルは以下の式 (5) で示される．ただし，$x_{l0} = 1$ とする．

$$P(y_l|\boldsymbol{x}_l) = \sum_{k=1}^{K} w_{lk} P_k(y_l|\boldsymbol{x}_l) \tag{5}$$

ここで，w_{lk} は k に関しての和が 1 となる第 l 企業の各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 y_l の確率密度関数は分散を σ_k^2 としたとき，次式で表さ

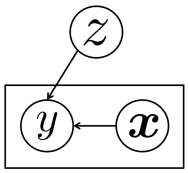

図 4 混合回帰モデルのグラフィカルモデル
Fig. 4 Graphical representation of the mixture regression model.

れる．

$$P_k(y_l|\boldsymbol{x}_l) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(y_l - f_k(\boldsymbol{x}_l))^2}{2\sigma_k^2}\right\} \tag{6}$$

$$f_k(\boldsymbol{x}_l) = \sum_{i=0}^{I} \beta_{ki} x_{li} \tag{7}$$

混合回帰モデルのグラフィカルモデルは**図 4** で表される．

混合回帰モデルは，EM アルゴリズム [33], [34], [35] によって学習可能である．混合回帰モデルの $\boldsymbol{\beta}_k$ のパラメータ推定は潜在クラス z_k に対して大きい重みを持つ企業を重点的に学習し，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．これは学習データの目的変数に対しての推定精度を向上させるようなパラメータ推定となっている．すなわち，行動情報 $\boldsymbol{x}_l$ と被エントリ数 y_l の組合せで潜在クラスが構築されるため，これから就職ポータルサイトを用いようとしている（被エントリ数 y_l のデータがない）新規企業に対しては単純に予測を行うことができない．

この問題に対し，補助変数を用いた手法が適用可能である．文献 [8] では補助変数を潜在クラスに反映させるモデルの一般式が与えられている．しかし具体的なモデルの例としては，式 (4) の混合割合を補助変数を用いたロジットモデルで与えるケースが示されているのみである．一方，補助変数として複数の離散データの基本情報を想定する本研究では，企業の特徴は基本情報の組合せに依存すると考えられる．すなわち，それぞれの潜在クラスの特徴は基本情報の組合せに依存して表現されるものと考えることができる．そこで，本提案モデルは基本情報それぞれに対して多項分布を仮定し，基本情報間の交互作用を考慮したモデル化を行う．次章以降，混合回帰モデルを基礎とし，扱う問題に適した形で被エントリ数の予測モデルへの拡張を行う．

3. 基本情報を考慮した予測モデル（提案モデル）

企業の行動情報と獲得できる被エントリ数の関係性をモデル化する際に，同時に新たな企業が就職ポータルサイトを用いることで獲得できる被エントリ数を予測できることが望ましい．また，就職ポータルサイトを利用しようとする企業は被エントリ数を向上させるうえで効果的な採用活動における行動に関心があり，その効果を定量的に判断できることが求められる．

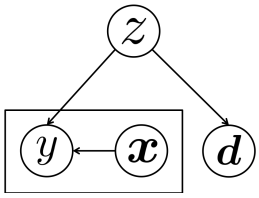

図 5 提案モデルのグラフィカルモデル
Fig. 5 Graphical representation of the proposed model.

そこで，本章では，学習データにおける行動情報 $\boldsymbol{x}_l$ と被エントリ数 y_l の関係性を表す回帰モデルと企業の基本情報 $\boldsymbol{d}_l$ により潜在クラスを構築するモデルを提案する．このモデルは回帰式により表現される被エントリ数と行動情報の関係性と，企業の基本情報の共起を表しており，基本情報からその企業の潜在クラスへの所属確率を推定することができる．すなわち，潜在クラスに基本情報の特徴を反映させることで，被エントリ数が学習データにない新規企業に対しても，潜在クラスへの重みを推定することが可能となる．

これにより，就職ポータルサイトを利用する新たな企業に対して，その企業の基本情報を用いることで被エントリ数を予測することが可能となる．さらに，混合回帰モデルのパラメータを分析することで，採用活動における行動情報が与える被エントリ数への影響も定量的に把握することができる．

3.1 定式化

いま，企業に関する j 番目の基本情報（$1 \le j \le J$）の要素集合を $\mathcal{D}^j = \{d^j_{n_j} : 1 \le n_j \le N_j\}$，$d^j_{n_j}$ を j 番目の基本情報の n_j 番目の要素，N_j を j 番目の基本情報の要素数とし，l 番目の企業の基本情報を表す変数ベクトルを $\boldsymbol{d}_l = (d_{l1}, d_{l2}, \ldots, d_{lJ})^{\mathrm{T}}$，$d_{lj}$ を l 番目の企業の j 番目の基本情報の要素とする．このとき提案する確率モデルは，式 (8) で示される．

$$P(y_l, \boldsymbol{x}_l, \boldsymbol{d}_l) = \sum_{k=1}^{K} P(z_k) P_k(y_l|\boldsymbol{x}_l) \prod_{j=1}^{J} \prod_{n_j=1}^{N_j} P(d^j_{n_j}|z_k)^{\delta(d^j_{n_j}, d_{lj})} \tag{8}$$

ここで，$\delta(a, b)$ は a と b が一致していれば 1，さもなくば 0 とする指示関数とする．提案モデルのグラフィカルモデルは**図 5** で表される[*4]．

[*4] 図 5 に示したグラフィカルモデルにおいて，潜在クラス z から，基本情報ベクトル $\boldsymbol{d}$ に矢印があるが，これは式 (8) にあるように，潜在クラスのもとでの基本情報の条件付確率を用いてモデル式を記述していることに対応している．条件付確率はベイズの定理で，条件部を反転できるため，矢印が逆向きのモデルを考えることもできるが，本研究では以後の学習アルゴリズムの構築や潜在クラスの解釈という観点から，式 (8) と図 5 で表されるモデルを考える．

3.2 パラメータの推定方法

提案モデルのパラメータを EM アルゴリズムを用いて推定する方法を示す．学習データに対する対数尤度関数 LL は以下の式 (9) のように示される．

$$LL = \sum_{l=1}^{L} \log P(y_l, \boldsymbol{x}_l, \boldsymbol{d}_l) \tag{9}$$

EM アルゴリズムは対数尤度を最大化するパラメータを E-step と M-step の繰り返し計算を行うことによって求める．以下に，提案モデルのパラメータである w_{lk}，$P(z_k)$，σ_k^2，$\boldsymbol{\beta}_k$，$P(d_{n_j}^j|z_k)$ 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，「w_{lk} の推定」と「$P(z_k)$，σ_k^2，$\boldsymbol{\beta}_k$，$P(d_{n_j}^j|z_k)$ の推定」を繰り返すことでパラメータの学習を行う．

【E-step】

まず E-step では以下の式 (10) で w_{lk} が計算され更新される．

$$w_{lk} = \frac{P(z_k)P_k(y_l|\boldsymbol{x}_l)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j}P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})}}{\sum_{k=1}^{K}P(z_k)P_k(y_l|\boldsymbol{x}_l)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j}P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})}} \tag{10}$$

【M-step】

次に，M-step では w_{lk} を固定した元で，各パラメータを更新する．混合割合 $P(z_k)$ および各潜在クラス z_k における分散 σ_k^2 はそれぞれ式 (11) および式 (12) で更新される．

$$P(z_k) = \frac{\sum_{l=1}^{L} w_{lk}}{L} \tag{11}$$

$$\sigma_k^2 = \frac{\sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2}{\sum_{l=1}^{L} w_{lk}} \tag{12}$$

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 におけるパラメータ $\boldsymbol{\beta}_k$ は，式 (13) を用いて更新する．

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{\beta}_k} \sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2 \tag{13}$$

企業の基本情報に関するパラメータについては以下の式 (14) で更新する．

$$P(d_{n_j}^j|z_k) = \frac{\sum_{l=1}^{L}\delta(d_{n_j}^j, d_{lj})w_{lk}}{\sum_{l=1}^{L} w_{lk}} \tag{14}$$

3.3 新規データに対する被エントリ数の予測

本モデルは新規企業に対して被エントリ数の予測が可能である．企業の特徴は基本情報の組合せで表現できると考えられる．ここで，$\boldsymbol{d}'_t = (d'_{t1}, d'_{t2}, \ldots, d'_{tJ})^{\mathrm{T}}$ を t 番目 $(t = 1, 2, \ldots, T)$ の予測対象企業の基本情報を表す変数ベクトル，$\hat{y}_t$ を t 番目の予測対象企業の被エントリ数の予測値とすると，基本情報から潜在クラスへの所属確率が計算可能である．よって，それぞれの潜在クラスの回帰モデルを混合することで新規企業に対して被エントリ数を予測することができる．ここで，$\hat{w}_{tk}$ を予測対象企業の潜在クラスへの所属確率の予測値，$\hat{\beta}_{ik}$ を回帰パラメータの推定値とすると，$\hat{w}_{tk}$ および被エントリ数の予測値 $\hat{y}_t$ は，以下の式 (15) および式 (16) で推定される．

$$\hat{w}_{tk} = \frac{P(z_k)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j}P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d'_{tj})}}{\sum_{k=1}^{K}P(z_k)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j}P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d'_{tj})}} \tag{15}$$

$$\hat{y}_t = \sum_{k=1}^{K}\left(\hat{w}_{tk}\sum_{i=0}^{I}\hat{\beta}_{ki}x'_{ti}\right) \tag{16}$$

ここで，T 社の予測対象企業のうち t 番目の企業の行動情報を表す説明変数ベクトルを $\boldsymbol{x}'_t = (x'_{t0}, x'_{t1}, x'_{t2}, \ldots, x'_{tI})^{\mathrm{T}}$ とし，$x'_{t0} = 1$ とする．

3.4 提案モデルのアルゴリズム

提案モデルは以下のアルゴリズムで構築される．

Step1 各パラメータの初期値をランダムに設定する．

Step2 E-step：w_{lk} を式 (10) を用いて推定する．

Step3 M-step：$P(z_k)$，σ_k^2，$\boldsymbol{\beta}_k$ および $P(d_{n_j}^j|z_k)$ を式 (11)〜式 (14) を用いて更新する．

Step4 収束条件を満たしていれば Step5 へ．さもなければ Step2 にもどる．

Step5 新規データに対して式 (15) を用いて潜在クラスへの重みを推定し，式 (16) を用いて被エントリ数の予測を行う．

□

4. 提案モデルの実データへの適用

4.1 提案モデルの評価実験

本節では，実データを用いて提案モデルの推定を行い，学習データに対しての当てはまりと予測対象企業に対する予測精度の 2 つの観点から結果を考察する．学習データへの当てはまりが良いほど表現能力が高いモデルといえるが，一方，予測精度が高いほど汎化能力の高いモデルが得られていると判断できる．

4.1.1 実験条件

実験データとして，2014 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 1,000 件以下の被エントリを獲得した企業約 5,000 社（$L \approx 5000$）を学習データ，2015 年度卒業の学生に対する就職ポータルサイトで同様の被エントリを獲得した企業約 5,000 社（$T \approx 5000$）を予測対象データとして用いた．目的変数を各企業の被エントリ数とし，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 4 つの行動オプショ

表 1 説明変数間の相関係数
Table 1 Correlation coefficients between explanatory variables.

	変数 1	変数 2	変数 3	変数 4
変数 1	1.0000	0.0024	0.0005	−0.0020
変数 2		1.0000	0.0973	0.0267
変数 3			1.0000	0.0449
変数 4				1.0000

ンを説明変数として用いた（$I = 4$）．これらは，企業がその行動オプションを採用すれば 1，採用しなければ 0 をとるダミー変数である[*5]．念のため，多重共線性の問題が起こらないか否かを確認するために，これらの説明変数間の相関係数行列を求めたものを**表 1** に示す．この結果，説明変数間の相関はほとんど見られず，多重共線性の問題は起きないことが確認できる．

また，基本情報として 4 変数を用いた（$J = 4$）．潜在クラス数 K は 2〜10，15，20 として実験を行った．比較モデルとして潜在クラスモデルの 1 つである Aspect Model [36], [37], [38], [39], [40] を多変量に拡張したモデルを用いて企業の基本情報で潜在クラスを推定し，それぞれに回帰モデルを構築するモデル（AM＋回帰）と潜在クラスや企業の基本情報を用いない単一の回帰モデル（単一回帰）の結果も示す．この実験ではそれぞれのモデルで，各学習データの企業に対しては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二乗誤差により推定精度を評価した．また各予測対象データの企業に対しては被エントリ数の予測を行い，算出された予測値と実測値の平均二乗誤差を評価することにより，予測精度を評価した．また異なる初期値で 100 回実験を繰り返し，それらの平均を用いて評価を行う．

4.1.2 結果と考察

実験結果を**表 2** に示す．

表 2 より複数の回帰モデルを混合することによるモデルの推定精度の向上が確認できる．この結果より被エントリ数と企業の行動情報の関係性の構造は複数存在し，潜在クラスを仮定し混在的要因を考慮しているモデルの方が本データの分析に適していることが示唆される．

また表 2 より，提案モデルでは各潜在クラス数において学習データへの当てはまり，予測精度の観点から優れたモデルが推定されていることが分かる．提案モデルでは学習データに対する当てはまりを良くする混合回帰モデルを推定すると同時に企業の基本情報のクラスタリングが適切に行われたことで予測精度が向上したと考えられる．すなわち，企業の基本情報に加え，行動情報の被エントリ数への効果を考慮した新たな企業クラスタリングが可能となっている．

[*5] 行動オプションの詳細については機密情報を含むため公開することができないが，たとえば「あるキャンペーン企画を実施する」という行動オプションのようなものを想像すればよい．

表 2 平均二乗誤差の比較結果
Table 2 Comparison result of mean squared error.

K	提案モデル		AM+回帰		単一回帰	
	学習	予測	学習	予測	学習	予測
2	54238.0	60010.8	59326.3	60800.8		
3	28091.1	55607.4	58249.6	59759.4		
4	28647.7	55204.6	57909.5	59544.0		
5	27534.0	55017.3	57461.5	59119.1		
6	26169.1	**54943.6**	57160.2	58806.6		
7	25278.0	54947.0	57018.2	58718.5	59836.3	61148.1
8	22386.3	55084.1	56842.5	58612.4		
9	20928.9	55235.1	56694.9	58617.7		
10	20789.2	55612.2	56570.1	58593.2		
15	16444.5	56833.0	55926.2	**58545.9**		
20	14401.6	58090.0	55469.9	58640.6		

表 3 提案モデルによって得られた偏回帰変数
Table 3 A estimated partial regression coefficient by proposed model.

	z_1	z_2	z_3	z_4	z_5	z_6
$\hat{P}(z_k)$	0.10	0.21	0.21	0.19	0.11	0.18
$\hat{\beta}_{k0}$	177.65	218.73	411.70	462.90	513.88	615.85
$\hat{\beta}_{k1}$	12.22	8.26	68.91	25.01	−15.82	40.24
$\hat{\beta}_{k2}$	52.41	19.95	160.83	104.91	61.75	68.24
$\hat{\beta}_{k3}$	31.93	65.28	148.78	129.31	105.59	49.00
$\hat{\beta}_{k4}$	23.93	47.21	99.23	83.26	116.09	58.81

また潜在クラス数の増加にともない，学習データおよび予測対象データへの当てはまりが向上していく一方，ある K 以上では予測精度の低下が見られる．これは潜在クラス数に応じて，パラメータ数が増加することで学習データへの過度なフィッティングが起きていると考えられる．本提案モデルを適用する際には目的に応じてモデルのパラメータ数，学習データ数を考慮し，適切に潜在クラス数を決定する必要がある．

4.2 提案モデルを用いた分析

本節では構築された提案モデルの応用として実データを用いて結果の分析を行う．潜在クラスに着目した分析と各企業に対する分析に焦点を当てる．本研究では提案モデルにおいて最も良い予測精度結果となった潜在クラス数 $K = 6$ のときに推定されたパラメータを用いて分析を行うこととする．分析データは前節と同様のデータを用いている．

4.2.1 潜在クラスに着目した分析

提案モデルの各潜在クラスにおける回帰モデルのパラメータ推定値を**表 3** に示す．表 3 における $\hat{P}(z_k)$ は混合割合の推定値を示す．

表 3 に示した各潜在クラスモデルのもとでの偏回帰変数は，各説明変数が目的変数に対して与えるインパクトの

表 4 各潜在クラスに属する企業がそれぞれの行動を実施している割合

Table 4 Ratio of each activity each company takes action in each latent class.

	z_1	z_2	z_3	z_4	z_5	z_6
行動 1	0.798	0.817	0.754	0.822	0.752	0.743
行動 2	0.009	0.037	0.013	0.010	0.044	0.022
行動 3	0.024	0.034	0.032	0.033	0.073	0.050
行動 4	0.573	0.678	0.676	0.659	0.673	0.710

大きさを示している．通常の回帰分析では，回帰係数が 0 であるか否かの検定を行うための t 値や p 値を計算することができるか，ここでは混合回帰モデルを構築しているため，通常の重回帰モデルのように偏回帰係数の統計量分布が明示的に与えられておらず，直接的な有意性の検定が難しい．しかしながら，次に示す理由により，これらの推定された回帰係数には意味があると考えることができる．

(1) 各説明変数のとりうる値は 1（行動オプションを実施）か 0（行動オプションを実施せず）の 2 値であり，極端に分散が大きい変数は含まれない．
(2) 回帰係数の意味は，行動オプションを実施した場合の被エントリ数へのインパクトを表していると考えられる．
(3) 誤差の分散が，学習データに対して 25,000 程度，テストデータに対して 55,000 程度であることから，標準偏差は，学習データに対して 160 程度，テストデータに対して 240 程度である．得られている偏回帰係数の数値は，これらと比較しても小さいとはいえない．

次に，表 3 より，各潜在クラスの特徴について考察を行う．推定されたパラメータは，潜在クラスごと，行動ごとに異なっており，それぞれの行動が効果的な潜在クラスは異なることが分かる．たとえば，行動 1，2，3 に関しては，潜在クラス 3 が最も効果的な潜在クラスであるが，行動 4 に関しては潜在クラス 5 が最も効果的な潜在クラスである．次に，それぞれの行動ごとに被エントリ数への効果の大きさが異なることが分かる．たとえば推定されたパラメータのばらつきから行動 2 は他の行動と比較し，潜在クラスごとの実施の効果が大きく異なるが，行動 1 は潜在クラスごとに実施の効果の変動が小さいといえる．また，各潜在クラスが異なる特徴を持っていることが分かる．たとえば，潜在クラス 5 に関しては，他の潜在クラスと比較して行動 1 の効果が最も低い潜在クラスであるのに対して，行動 4 の効果が最も高い潜在クラスである．以上のような分析からも，企業の混在的特徴を潜在クラスによって表現できていることが示唆される．

次に各潜在クラスにおける行動を起こしている割合に着目し，結果を**表 4** に示す．

表 4 より，潜在クラスごとに行動を起こしている割合が異なることが分かる．また表 3 と合わせて考察することで各潜在クラスにおける行動に効果があるかどうかを解釈することが可能である．たとえば，潜在クラス 6 は他の潜在クラスと比較し行動 1 を起こす割合が最も低い潜在クラスであるが，行動 1 の効果は比較的高い潜在クラスである．よって，潜在クラス 6 に所属する企業は行動 1 に対してより積極的な行動を行うことで被エントリ数の効果的な獲得が期待される．

表 5 各企業の各潜在クラスへの所属確率

Table 5 Belonging probabilities from each company to each latent class.

	$\hat{w}_{t1}$	$\hat{w}_{t2}$	$\hat{w}_{t3}$	$\hat{w}_{t4}$	$\hat{w}_{t5}$	$\hat{w}_{t6}$
企業 A	0.00	0.29	0.00	0.00	0.31	0.40
企業 B	0.76	0.00	0.00	0.24	0.00	0.00

表 6 提案モデルにより計算された各企業の各行動の効果

Table 6 Effect of recruitment activities of each company calculated by proposed model.

	β^*_{t1}	β^*_{t2}	β^*_{t3}	β^*_{t4}
企業 A	13.65	52.31	71.44	73.17
企業 B	15.33	65.17	55.59	38.35

4.2.2 各企業に着目した分析

本モデルでは各企業に対して採用活動における行動の効果を定量化することが可能である．行動に対する t 番目の予測対象企業の効果ベクトルを $\boldsymbol{\beta}^*_t = (\beta^*_{t1}, \beta^*_{t2}, \ldots, \beta^*_{tI})^{\mathrm{T}}$ とすると行動の効果の定量化は以下の式 (17) で推定される．また，学習データの企業に対しても同様に推定可能である．

$$\boldsymbol{\beta}^*_t = \left(\sum_{k=1}^{K} \hat{w}_{tk}\hat{\beta}_{k1}, \sum_{k=1}^{K} \hat{w}_{tk}\hat{\beta}_{k2}, \ldots, \sum_{k=1}^{K} \hat{w}_{tk}\hat{\beta}_{kI} \right)^{\mathrm{T}} \quad (17)$$

この定量化式により，各企業に対して特徴に応じ，個々に行動の効果を推定することができる．

ここで，例として 2 つの企業に着目し分析を行った．着目した 2 企業は提案モデルによって**表 5** のように企業から潜在クラスへの重み $\hat{w}_{tk}$ が推定された．

表 5 より企業 A は潜在クラス 2，5，6 に比較的大きな重みを持っている企業であり，企業 B は潜在クラス 1 に大きな重み，潜在クラス 4 に小さな重みを持っている企業である．これは企業 A は潜在クラス 2，5，6 の特徴を有しており，企業 B は潜在クラス 1，4 の特徴を有した企業であるといえる．この特徴が異なる 2 企業に対し，式 (17) を用いることにより行動の効果を**表 6** のように計算できる．

表 6 より着目した 2 企業はそれぞれ行動に対する効果が異なることが分かる．この 2 企業を比較すると行動 1，2 に関しては企業 B の方が企業 A より効果的な行動である一方で行動 3，4 に関しては企業 A の方が企業 B より効果的な行動であるといえる．

5. 考察

提案手法を実データに適用した結果，単一の回帰モデルや AM と回帰モデルの組合せ手法よりも推定精度の面で優れていることから，複数の潜在クラスを仮定した提案モデルの有効性が示されたといえる．この結果は，規模や業種などによって統計的特徴の異なる企業グループが混在しており，この企業の異質性を考慮したモデル化が必要であることを示唆している．

また，提案した分析モデルの活用により，被エントリ数の推定，また行動を起こした際の効果を定量的に推定可能であり，どのように採用活動を変化させていけば，被エントリ数を向上させることができるかという観点から採用活動の計画を立てるための一助となることが期待される．企業の個々の採用活動にはコストが発生するため，費用対効果を見極めながら策を講じる必要があり，そのために本提案モデルは有用であると考えられる．

今回，モデルを構築するための学習データには，被エントリ数が 100 件以上，1,000 件以下の企業のデータを用いた．これは，エントリ数が極端に多い企業と少ない企業が外れ値となって回帰モデルの推定に大きな影響を与えてしまうためであるが，実際にこの範囲に入らない企業を予測しようとすると外挿になることを意味する．したがって，予測値がこの範囲を超えた場合には，予測モデルの外挿であることを付記して利用すべきである．

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基に基本情報を考慮することで，被エントリ数を予測可能な分析モデルを提案した．提案したモデルの有効性を示すために，就職ポータルサイトに蓄積された実データを用いて分析を行い，予測精度の面から評価を行った．加えて推定されたモデルを用いて実際のエントリデータの分析を行い，有用な知見が得られることを示した．

今後の課題として，予測精度の向上，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，潜在クラス数の決定方法の検討などがあげられる．

まず予測精度の向上について，就職ポータルサイト上に蓄積したデータのみで被エントリ数の精度の高い予測は非常に難しい問題である．今回の研究のように行動情報と被エントリ数の関係性の構築が目的にある場合，高い予測精度はその分析の正確性を高めるものであると考えられる．しかし行動情報と被エントリ数の明確な因果関係が断定できず，外部要因などに影響されるものと考えられるため，さらなるモデルの高度化には説明変数の適切な選択，外部要因の検討を今後の課題としたい．

次に，具体的な企業の採用活動における計画のサポート手法の検討について，企業に対して情報を提供する際に各企業の要望を考慮することはサービスの拡充につながる．たとえば企業から目標とする被エントリ数を提示してもらうことが可能であれば，その被エントリ数に達する最も効率的な就職ポータルサイト上での行動を提供できる．その際には採用行動にかかるコストと被エントリ数の獲得の両観点から分析することが必要である．また，たとえば「経営を学んだ学生からのエントリを獲得したい」といった要望に対しても対応できるような学生側の特徴を考慮した分析モデルの構築も有用なものであると考えられる．

また，潜在クラス数 K の決定方法について，本研究では予測精度の高いモデルの潜在クラス数を最適なモデルであるとした．この方法は潜在クラス数を変化させてモデルを決定しなければならないため，実問題への適用に時間とコストがかってしまう．このような状況で，データ特性やデータ数，特徴量数などから潜在クラスを決定する手法の考案が求められる．

一方で，ある程度，豊富なパラメータを有する統計モデルであっても，正則化を用いた学習アルゴリズムを導入することで過学習を抑えることができる．たとえば，l_1 正則化を用いれば，不要なパラメータを 0 にする機能を有しているため，よりシンプルなモデルを得られるという意味でモデル選択の機能を有した学習アルゴリズムを構成することも可能である．ただし，潜在クラスの生起確率や混合確率，各回帰パラメータという解釈の異なるパラメータに対して，有効に働く正則化項を見つけるためには，適用可能な様々な正則化項について評価を行う必要がある．これらの正則化学習アルゴリズムの検討についても今後の課題とする．定量的な観点から過学習を評価することで，モデルの妥当性，解釈性を高める一助になると考えられる．

謝辞 本研究にあたり，多くのご助言をいただいた湘南工科大学の三川健太先生，早稲田大学創造理工学部経営システム工学科後藤研究室の方々に深く感謝致します．また，株式会社リクルートキャリアの方々からは，手厚いサポートを賜りました．ここに感謝の意を表します．本研究の一部は科学研究費（26282090，26560167）の助成を受けたものである．

参考文献

[1] Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer (2006).
[2] Conway, D. and White, J.M.：入門機械学習，株式会社オライリー・ジャパン (2012).
[3] 後藤正幸，小林 学：入門パターン認識と機械学習，コロナ社，pp.200–206 (2014).
[4] 平井有三：初めてのパターン認識，森北出版株式会社，pp.175–197 (2012).
[5] Faria, S. and Soromenho, G.: Fitting Mixtures of Linear Regressions, *Journal of Statistical Computation and Simulation*, Vol.80, No.2, pp.201–225 (2010).

[6] De Veaux, R.D.: Mixtures of linear regressions, *Computational Statistics & Data Analysis*, Vol.8, pp.227–245 (1989).
[7] Jones, P.N. and McLachlan, G.J.: Fitting finite mixture models in a regression context, *Australian Journal of Statistics*, Vol.34, No.2, pp.233–240 (1992).
[8] Grun, B. and Leisch, F.: FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters, *Journal of Statistical Software*, Vol.28, No.4, pp.1–35 (2008).
[9] Leisch, F.: FlexMix: A general framework for finite mixture models and latent glass regression in R, *Journal of Statistical Software*, Vol.11, No.8, pp.1–18 (2004).
[10] Grun, B. and Leisch, F.: Fitting finite mixtures of generalized linear regressions in R, *Computational Statistics & Data Analysis*, Vol.51, No.11, pp.5247–5252 (2007).
[11] 採用選考に関する企業の倫理憲章，入手先 ⟨https://www.keidanren.or.jp/policy/2011/015.pdf⟩.
[12] 採用選考に関する指針，入手先 ⟨http://www.keidanren.or.jp/policy/2013/081_shishin.pdf⟩.
[13] 採用選考に関する指針，入手先 ⟨http://www.keidanren.or.jp/policy/2015/112_shishin.pdf⟩.
[14] 永野　仁：就職活動成功要因として就職意義–大学生調査の分析–，政経論議，Vol.73, No.5, pp.645–665 (2005).
[15] 下村英雄，堀　洋元：大学生の就職活動における情報探索行動：情報源の影響に関する検討，社会心理学研究，Vol.20, No.2, pp.93–105 (2004).
[16] 高橋　潔：就職・採用活動におけるマーケティング・モデルからの脱却，国民経済雑誌，Vol.202, No.1, pp.113–128 (2010).
[17] 下村英雄，木村　周：大学生の就職活動における就職関連情報と職業未決定，進路指導研究（日本進路指導学会研究紀要），No.15, pp.11–19 (1994).
[18] 下村英雄，木村　周：大学生の就職活動ストレスとソーシャルサポートの検討，進路指導研究（日本進路指導学会研究紀要），Vol.18, No.1, pp.9–16 (1997).
[19] 北見由奈，茂木俊彦，森　和代：大学生の就職活動ストレスに関する研究：評価尺度の作成と精神的健康に及ぼす影響，学校メンタルヘルス，Vol.12, No.1, pp.43–50 (2009).
[20] 三井所健太郎，藤村直美：WEB インターフェースによる就職活動支援システムに関する研究，情報処理学会研究報告グループウェアとネットワークサービス（GN），Vol.2009-GN-73, No.17, pp.1–6 (2009).
[21] 岡田昌也，長谷川忍：就職活動における企業研究支援システムの開発，電子情報通信学会技術研究報告 ET，教育工学，Vol.112, No.269, pp.77–82 (2012).
[22] 古川達也，森田佐知子，福本尚生：ICT を利用した学生・教職員のための就職活動支援システムの構築，電気学会研究会資料 FIE，Vol.2014, No.25, pp.71–76 (2014).
[23] 垂水春樹，大楠拓也，白川勇気，徐　海燕：就職活動情報登録閲覧 Web システムの開発および利用状況に関する分析，研究報告コンピュータと教育（CE），Vol.2014-CE-127, No,3, pp.1–6 (2014).
[24] 森田慎一郎：大学生の就職活動支援における学生相談部門と就職サポート部門の協働：相談員へのインタビュー調査に基づく期待と課題の探索 q，東京女子大学紀要論集，Vol.66, No.1, pp.103–118 (2015).
[25] 吉田　晋，福田耕治：グループワークを活用した就職活動支援に有効なキャリア教育，工学教育，Vol.62, No.3, pp.21–27 (2014).
[26] 早川真央，三川健太，荻原大陸，後藤正幸：層別木と混合ワイブル分布に基づく就職活動終了時期の分析モデルの構築，情報処理学会論文誌，Vol.58, No.5, pp.1189–1206 (2017).
[27] Yamagami, K., Mikawa, K., Goto, M. and Ogihara, T.: A Statistical Prediction Model of Students' Finishing Date on Job Hunting Using Internet Portal Sites Data, *The 16th Asia Pacific Industrial Engineering and Management Systems Conference* (*APIEMS 2015*), Ho Chi Minh City, Vietnam (2015).
[28] 坂元哲平，山下　遥，荻原大陸，後藤正幸：就職ポータルサイトにおける企業のアピールポイントと志望理由のマッチング分析モデルに関する一考察，情報処理学会論文誌，Vol.58, No.9, pp.1535–1548 (2017).
[29] Sugiyama, Y., Arai,T., Yang, T., Goto, M. and Ogihara, T.: An Analytical Model of Relation Between Browsing and Entry Activities on an Internet Portal Site for Job-hunting, *15th Asian Network for Quality Conference* (*ANQ2017*), Soaltee Crowne Plaza, Kathmandu, Nepal, ICT-02 (2017).
[30] 野津琢登，三川健太，後藤正幸，荻原大陸：就職ポータルサイトにおける被エントリ数の予測モデルに関する一考察，電子情報通信学会技術研究報告人工知能と知識処理研究会（*AI*），Vol.115, No.381, AI2015-34, pp.49–54 (2015).
[31] Nagamori, S., Yamashita, H., Goto, M. and Ogihara, T.: An Analytic Model of Relation between Companies' Recruitment Activities and Number of Students' Application Based on Mixture Regression Model, *The 17th Asia Pacific Industrial Engineering and Management Systems Conference* (*APIEMS 2016*), No.150, Taipei, Taiwan (2016).
[32] Govaert, G. and Nadif, M.: Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data, *Computational Statistics & Data Analysis*, Vol.23, No.1, pp.65–81 (1996).
[33] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Royal Statistical Society, Series B*, Vol.39, No.1, pp.1–38 (1977).
[34] 宮川雅巳：アルゴリズムとその周辺，応用統計学，Vol.16, No.1, pp.1–21 (1987).
[35] Wedel, M. and DeSarbo, W.S.: A mixture likelihood approach for generalized linear models, *Journal of Classification*, Vol.12, No.1, pp.21–55 (1995).
[36] Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. AGIR '99, ACM Press*, pp.50–57 (1999).
[37] Hofmann, T.: Probabilistic Latent Semantic Analysis, *Proc. UAI'99*, pp.289–296 (1999).
[38] Hofmann, T. and Puzicha, J.: Latent Class Models for Collaborative Filtering, *Proc. 16th International Joint Conference on Artificial Intelligence*, Vol.99, pp.688–693 (1999).
[39] Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning Journal*, Vol.42, No.1, pp.177–196 (2001).
[40] Hofmann, T.: Latent Semantic Models for Collaborative Filtering, *ACM Trans. Inf. Syst.*, Vol.22, No.1, pp.89–115 (2004).

付　　録

A.1 提案モデルのパラメータ更新式の導出

提案モデルのパラメータを EM アルゴリズムを用いて推定する詳細を紹介する．学習データに対する対数尤度関数 LL は以下の式 (A.1) のように示される．

$$LL = \sum_{l=1}^{L} \log P(y_l, \boldsymbol{x}_l, \boldsymbol{d}_l) \tag{A.1}$$

EM アルゴリズムは対数尤度関数 LL を最大化するパラメータを E-step と M-step の繰返し計算を行うことによって求める．ここでは，「事後確率 w_{lk} の推定」と「事後確率を固定した元での対数尤度関数 LL を最大化する $P(z_k)$，σ_k^2，$\boldsymbol{\beta}_k$，$P(d_{n_j}^j|z_k)$ の推定」を繰り返すことでパラメータの学習を行う．ここでは特に M-step における各パラメータの更新式の導出を示す．

【E-step】

まず，E-step では以下の式 (A.2) で事後確率 w_{lk} が計算され更新される．

$$w_{lk} = \frac{P(z_k)P_k(y_l|\boldsymbol{x}_l)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j} P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})}}{\sum_{k=1}^{K} P(z_k)P_k(y_l|\boldsymbol{x}_l)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j} P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})}} \tag{A.2}$$

【M-step】

M-step では事後確率 w_{lk} を固定した元で，LL を最大化する各パラメータを求める．

$$LL = \sum_{l=1}^{L} \log P(y_l, \boldsymbol{x}_l, \boldsymbol{d}_l) \tag{A.3}$$

$$= \sum_{l=1}^{L} \log\left(\sum_{k=1}^{K} w_{lk}\frac{(*)}{w_{lk}}\right) \tag{A.4}$$

$$\geq \sum_{l=1}^{L}\sum_{k=1}^{K} w_{lk} \log\left(\frac{(*)}{w_{lk}}\right) \tag{A.5}$$

$$= \sum_{l=1}^{L}\left(\sum_{k=1}^{K} w_{lk}\log(*) - \sum_{k=1}^{K} w_{lk}\log w_{lk}\right) \tag{A.6}$$

なお，上記の式における $(*)$ は以下で表される．

$$(*) = P(z_k)P_k(y_l|\boldsymbol{x}_l)\prod_{j=1}^{J}\prod_{n_j=1}^{N_j} P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})} \tag{A.7}$$

式 (A.4) から式 (A.5) の変形は Jensen の不等式から得られる．LL の最大化は LL の下限である式 (A.6) の最大化と置き換えられる．さらに式 (A.6) の最大化に関係のない定数項（M-step で更新する $P(z_k)$，σ_k^2，$\boldsymbol{\beta}_k$，$P(d_{n_j}^j|z_k)$ に関係のない項）を式 (A.6) から除外したものを LL' とすると LL' は以下のように示される．

$$LL' = \sum_{l=1}^{L}\sum_{k=1}^{K} w_{lk}\log P(z_k)P_k(y_l|\boldsymbol{x}_l) \times \prod_{j=1}^{J}\prod_{n_j=1}^{N_j} P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})} \tag{A.8}$$

ここで，$\sum_{k=1}^{K} P(z_k) = 1$，また各潜在クラス z_k において，各基本情報 $\mathcal{D}^j$ に対して $\sum_{n_j=1}^{N_j} P(d_{n_j}^j|z_k) = 1$ という制約式からラグランジュの未定乗数法を用いて LL' の最大化を行う．ラグランジュ関数は以下のように示される．

$$J = LL' - \alpha\left(1 - \sum_{k=1}^{K} P(z_k)\right) - \sum_{j=1}^{J}\sum_{k=1}^{K}\lambda_{kj}\left(1 - \sum_{n_j=1}^{N_j} P(d_{n_j}^j|z_k)\right) \tag{A.9}$$

式 (A.9) における α，λ_{kj} はラグランジュの未定乗数である．この式 (A.9) を $P(z_k)$，$P(d_{n_j}^j|z_k)$ で偏微分を行い 0 と置くことで，それぞれの更新式が導かれる．

A.1.1 $P(z_k)$ の推定式の導出

$$\frac{\partial J}{\partial P(z_k)} = \frac{\sum_{l=1}^{L} w_{lk}}{P(z_k)} - \alpha = 0 \tag{A.10}$$

より，

$$P(z_k) = \frac{\sum_{l=1}^{L} w_{lk}}{\alpha} \tag{A.11}$$

式 (A.11) の両辺を k に関して和をとると，

$$\sum_{k=1}^{K} P(z_k) = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L} w_{lk}}{\alpha} \tag{A.12}$$

$\sum_{k=1}^{K} P(z_k) = 1$，$\sum_{k=1}^{K}\sum_{l=1}^{L} w_{lk} = L$ より，

$$\alpha = L \tag{A.13}$$

よって，式 (A.11)，(A.13) より $P(z_k)$ の推定式は以下のようになる．

$$P(z_k) = \frac{\sum_{l=1}^{L} w_{lk}}{L} \tag{A.14}$$

A.1.2 $P(d_{n_j}^j|z_k)$ の推定式の導出

$$\frac{\partial J}{\partial P(d_{n_j}^j|z_k)} = \frac{\sum_{l=1}^{L}\delta(d_{n_j}^j, d_{lj})w_{lk}}{P(d_{n_j}^j|z_k)} - \lambda_{kj} = 0 \tag{A.15}$$

より，

$$P(d_{n_j}^j|z_k) = \frac{\sum_{l=1}^{L}\delta(d_{n_j}^j, d_{lj})w_{lk}}{\lambda_{kj}} \tag{A.16}$$

式 (A.16) の両辺を n_j に関して和をとると，

$$\sum_{n_j=1}^{N_j} P(d_{n_j}^j|z_k) = \frac{\sum_{n_j=1}^{N_j}\sum_{l=1}^{L}\delta(d_{n_j}^j, d_{lj})w_{lk}}{\lambda_{kj}} \tag{A.17}$$

ここで，$\sum_{n_j=1}^{N_j} P(d_{n_j}^j|z_k) = 1$，$\sum_{n_j=1}^{N_j}\sum_{l=1}^{L}\delta(d_{n_j}^j, d_{lj})w_{lk} = \sum_{l=1}^{L} w_{lk}$ より，

$$\lambda_{kj} = \sum_{l=1}^{L} w_{lk} \tag{A.18}$$

よって，式 (A.16)，(A.18) より $P(d_{n_j}^j|z_k)$ の推定式は以下のようになる．

$$P(d_{n_j}^j|z_k) = \frac{\sum_{l=1}^{L}\delta(d_{n_j}^j, d_{lj})w_{lk}}{\sum_{l=1}^{L} w_{lk}} \tag{A.19}$$

A.1.3 $\boldsymbol{\beta_k}$ の推定式の導出

潜在クラス z_k におけるパラメータ $\boldsymbol{\beta}_k$ は，式 (A.20) を用いて更新する．

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{\beta}_k} \sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2 \tag{A.20}$$

$\boldsymbol{\beta}_k$ は以下の導出により求まる．まず，行動情報間の重み付き平方和 S_{abk} を以下のように定義する．ここで a，b は行動情報の番号を指す．

$$S_{abk} = \sum_{l=1}^{L} w_{lk}x_{la}x_{lb} - \frac{(\sum_{l=1}^{L} w_{lk}x_{la})(\sum_{l=1}^{L} w_{lk}x_{lb})}{\sum_{l=1}^{L} w_{lk}} \tag{A.21}$$

また，行動情報と被エントリ数の重み付き偏差積和 S_{ayk} を以下のように定義する．

$$S_{ayk} = \sum_{l=1}^{L} w_{lk}x_{la}y_l - \frac{(\sum_{l=1}^{L} w_{lk}x_{la})(\sum_{l=1}^{L} w_{lk}y_l)}{\sum_{l=1}^{L} w_{lk}} \tag{A.22}$$

このとき，行動情報と被エントリ数の関係性を表す各潜在クラスにおける回帰パラメータは以下の式 (A.23) を解くことによって得られる．

$$\begin{bmatrix} \hat{\beta}_{k1} \\ \vdots \\ \hat{\beta}_{kI} \end{bmatrix} = \begin{bmatrix} S_{11k} & \cdots & S_{1Ik} \\ \vdots & \ddots & \vdots \\ S_{I1k} & \cdots & S_{IIk} \end{bmatrix}^{-1} \begin{bmatrix} S_{1yk} \\ \vdots \\ S_{Iyk} \end{bmatrix} \tag{A.23}$$

また $\hat{\beta}_{k0}$ は以下の式 (A.24) によって得られる．

$$\hat{\beta}_{k0} = \frac{\sum_{l=1}^{L} w_{lk}y_l}{\sum_{l=1}^{L} w_{lk}} - \sum_{i=1}^{I} \hat{\beta}_{ki} \frac{\sum_{l=1}^{L} w_{lk}x_{li}}{\sum_{l=1}^{L} w_{lk}} \tag{A.24}$$

A.1.4 $\boldsymbol{\sigma_k^2}$ の推定式の導出

式 (A.8)（σ_k^2 は式 (6) により定義されている）を σ_k^2 で偏微分をして 0 とおくと，

$$\frac{\partial LL'}{\partial \sigma_k^2} = -\frac{\sum_{l=1}^{L} w_{lk}}{2\sigma_k^2} + \frac{\sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2}{2(\sigma_k^2)^2} = 0 \tag{A.25}$$

より，σ_k^2 について解くと以下の推定式が得られる．

$$\sigma_k^2 = \frac{\sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2}{\sum_{l=1}^{L} w_{lk}} \tag{A.26}$$

A.1.5 M-step における推定式のまとめ

M-step において各パラメータは以下の式で推定される．

$$P(z_k) = \frac{\sum_{l=1}^{L} w_{lk}}{L} \tag{A.27}$$

$$\sigma_k^2 = \frac{\sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2}{\sum_{l=1}^{L} w_{lk}} \tag{A.28}$$

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{\beta}_k} \sum_{l=1}^{L} w_{lk}(y_l - f_k(\boldsymbol{x}_l))^2 \tag{A.29}$$

$$P(d_{n_j}^j|z_k) = \frac{\sum_{l=1}^{L} \delta(d_{n_j}^j, d_{lj})w_{lk}}{\sum_{l=1}^{L} w_{lk}} \tag{A.30}$$

A.2 AM+回帰モデル

ここでは本研究で比較モデルとして用いている AM+回帰モデルを説明し，アルゴリズムを示す．AM は潜在クラスモデルの 1 つである．このモデルを多変量に拡張し用いることで企業の基本情報のみで潜在クラスを推定することが可能である．また，構築した潜在クラスそれぞれに回帰モデルを構築するモデルとなっている．

すなわち，AM+回帰モデルでは，基本情報を用いて企業を確率的にクラスタリングを行った後に，それぞれの潜在クラスに対し回帰モデルを推定するモデルである．

まず，企業の基本情報 $\boldsymbol{d}_l$ を用いて AM を学習する．この際に，企業の基本情報は多変量であるため，AM を拡張し，それぞれの基本情報の要素に対しパラメータを付与する．モデル式は以下のように表現可能である．

$$P(\boldsymbol{d}_l) = \sum_{k=1}^{K} P(z_k) \prod_{j=1}^{J} \prod_{n_j=1}^{N_j} P(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})} \tag{A.31}$$

多変量に拡張した AM は**図 A·1** のグラフィカルモデルで表現される．

それぞれのパラメータは EM アルゴリズム [33], [34] を用いて推定可能である．次に，推定されたパラメータを用いて企業の潜在クラスへの所属確率を計算する．ここで，企業の特徴は基本情報の組合せにより表現可能であると考える．よって企業の所属確率は以下の式で推定することができる．

$$\hat{P}(z_k|\boldsymbol{d}_l) = \frac{\hat{P}(\boldsymbol{d}_l|z_k)\hat{P}(z_k)}{\sum_{k=1}^{K} \hat{P}(\boldsymbol{d}_l|z_k)\hat{P}(z_k)} \tag{A.32}$$

$$= \frac{\prod_{j=1}^{J}\prod_{n_j=1}^{N_j} \hat{P}(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})}\hat{P}(z_k)}{\sum_{k=1}^{K}\prod_{j=1}^{J}\prod_{n_j=1}^{N_j} \hat{P}(d_{n_j}^j|z_k)^{\delta(d_{n_j}^j, d_{lj})}\hat{P}(z_k)} \tag{A.33}$$

この推定した企業の潜在クラスへの重みを用いて，それ

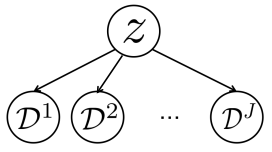

図 A·1 多変量の AspectModel のグラフィカルモデル

Fig. A·1 Graphical representation of the multivariate aspect model.

ぞれの潜在クラスに回帰モデルを構築することを考える．すなわち，$\hat{P}(z_k|\boldsymbol{d}_l)$ を各企業の潜在クラスに対する重み w_{lk} として，式 (13) により各潜在クラスに仮定されている回帰モデルのパラメータ推定を行う．

本比較モデルは企業を基本情報によって確率的にクラスタリングを行い，その後にそれぞれの潜在クラスに回帰モデルを仮定するモデルである．このモデルと提案モデルを比較することにより，回帰モデルの推定と同時にクラスタリングをすることによる有効性を示すことを目的としている．

以下に，比較モデルである AM+回帰モデルのアルゴリズムを示す．

Step1 企業が持つ基本情報 $\boldsymbol{d}_l$ により基本情報を AM を用いて確率的にクラスタリングする．

Step2 基本情報ごとに推定されたパラメータを用いて，企業の潜在クラスへの所属確率 $\hat{P}(z_k|\boldsymbol{d}_l)$ を推定する．

Step3 企業の潜在クラスへの所属確率により，式 (13) を用いて潜在クラスごとに回帰モデルを構築しパラメータを推定する．

Step4 新規データに対して式 (A.32)，(A.33) を用いて潜在クラスへの重みを推定し，式 (16) を用いて被エントリ数の予測を行う．

□

永森 誠矢

1992 年生．2015 年早稲田大学創造理工学部経営システム工学科修了．2017 年同大学大学院修士課程修了．

山下 遥

1987 年生．2010 年東京理科大学理工学部経営工学科卒業．2012 年慶應義塾大学大学院修士課程修了．2015 年慶應義塾大学大学院博士課程修了．博士（工学）．同年早稲田大学創造理工学部助手．2017 年上智大学理工学部情報理工学科助教，品質管理，統計学，情報工学を融合させた新たなデータ解析方法に関する研究に従事．応用統計学会，日本経営工学会，日本品質管理学会等，各会員．

荻原 大陸

1989 年生．2014 年早稲田大学大学院修士課程修了．2015 年より株式会社リクルートキャリア入社．就職支援サイトの企画職として，機械学習を用いたサービス開発に従事．

後藤 正幸

1969 年生．1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学博士課程修了．博士（工学）．1997 年早稲田大学理工学部助手．2000 年東京大学助手．2002 年武蔵工業大学環境情報学部助教授．2008 年早稲田大創造理工学部経営システム工学科准教授．2011 年同大教授．情報数理応用とデータサイエンス，およびパターン認識と機械学習の技術をベースとしたビジネスアナリティクスの研究に従事．著書に，『入門パターン認識と機械学習』，コロナ社 (2014)，『ビジネス統計 統計基礎とエクセル分析』，オデッセイコミュニケーションズ (2015) 等．IEEE，電子情報通信学会，人工知能学会，日本経営工学会，日本オペレーションズ・リサーチ学会，経営情報学会等，各会員．