情報処理学会論文誌 Vol.58 No.5 1189–1206 (May 2017)

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

早川 真央[1] 三川 健太[2,a)] 荻原 大陸[3] 後藤 正幸[1,b)]

受付日 2016年9月6日, 採録日 2017年2月9日

概要：近年，多くの大学生は就職ポータルサイトを利用して就職活動を行うようになった．このような就職支援を行うサービスが充実していくとともに，多くの学生の属性情報と就職活動に関する行動履歴などの情報が蓄積され，その活用が望まれるようになっている．就職ポータルサイトの運営企業は，単に学生や企業に対して，就職活動のためのシステムを提供するだけでなく，大学に対しては就職活動ガイダンスのようなサービスも実施しており，そのような場で有用となる情報を分析して提供することも重要である．そのためのツールとして，就職活動終了時期の影響要因を分析するためのモデルは有用であり，過去の実績データに基づくモデル構築が望まれている．本研究ではまず，就職ポータルサイトの実データを様々な角度から分析することで，就職活動終了時期の要因分析モデル構築のための糸口を探求する．その結果より，推定精度を向上させるため，複数のワイブル分布を混合した混合ワイブル分布を導入する．また，学生の属性により，分布の形状が大きく変化することも明らかになったため，学生属性によって層別を行う層別木モデルを構築する．以上の組合せを行い，葉に混合ワイブル分布をあてはめた層別木モデルによる分析手法を提案し，実データを用いた実験により，本提案モデルが高い精度で就職活動終了時期を説明可能であることを示す．本研究で行った分析により，就職活動における新たな知識発見を行うことができ，今後の就職活動支援に有益な情報を提供可能である．

キーワード：就職ポータルサイト，就職活動，ビジネスアナリティクス，混合ワイブル分布，分類木

A Statistical Analysis Model of Students' Success on Job Hunting by Stratification Tree and Mixed Weibull Distribution

Mao Hayakawa[1] Kenta Mikawa[2,a)] Tairiku Ogihara[3] Masayuki Goto[1,b)]

Received: September 6, 2016, Accepted: February 9, 2017

Abstract: Due to the increase of university students and recent recession, it is not an easy work for a university student to get a job in Japan. Recently, many students use internet portal sites for job hunting which help them to find suitable job offers easily. Under this background, various kinds of access log data of student users are now saved in a database and it is desirable to make use of such big data to support universities and students. The companies managing an internet portal site provide not only a web system on the internet but sevices such as a job hunting seminar for universities and students. Therefore, it is important for administrators of internet portal sites for job hunting to construct a model to analyze the relation between the finishing date of job hunting and influence factors. This study proposes a statistical model to analyze when job hunting is expected to be finished for each user. Because the period of job hunting may depend on the students' personality and the access pattern in a web site for job hunting, we show a statistical model estimated by using user's attributes and their access log data on a database. However a simple stochastic model cannot approximate the given empirical distribution. Therefore, a tree model with the mixed Weibull distribution is introduced to predict the pattern of finish of job hunting. Through a simulation experiment by using an actual data set, the effectiveness of the proposed predictive model is clarified.

Keywords: internet portal site, job hunting, business analytics, mixed Weibull distribution, stratification tree

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–0072, Japan
[2] 湘南工科大学
Shonan Institute of Technology, Fujisawa, Kanagawa 251–8511, 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)] mikawa@info.shonan-it.ac.jp
[b)] masagoto@waseda.jp

1. はじめに

近年，多くの大学生は就職ポータルサイトを利用して就職活動[*1]を行うようになった．このような就職支援を行うサービスが充実していくとともに，多くの学生の属性情報と就職活動に関する行動履歴などの情報が蓄積され，その活用が望まれるようになっている．就職ポータルサイトの運営企業は，単に学生や企業に対して就職活動のためのシステムを提供するだけでなく，大学に対しては就職活動ガイダンスのようなサービスも実施している．そのような場で有用となる情報を分析して提供することが必要であり，就職活動終了時期に関する分析は1つの重要な課題となっている．このような就職活動終了時期の分析は「学生や所属大学・学部への適切な就職活動支援を可能とする」，「予測される就職活動終了日を提示することにより学生の行動変化のきっかけとなる」，「就職ポータルサイトの利便性，信頼性の向上に寄与する」など，就職活動中の学生，就職ポータルサイトの運営企業双方にとって有益である．したがって，就職ポータルサイトに蓄積されたデータを最大限活用し，学生の属性情報と就職活動終了時期の関係性を説明する統計モデルを構築することは大変意義深い．

そのため，本研究では就職活動終了時期の分析に焦点を当てる．一般に，学生の就職活動終了時期は，所属している大学や文系/理系の種別，修士/学部の種別といった学生の属性に強く影響を受けることが想定される．そこで，本研究では実データからこれらに対する新たな知見を得るとともに，学生の属性情報と就職活動終了時期の関係分析モデルを構築することを目的とする．この就職活動終了時期の要因分析モデルを構築することにより，たとえば，所属学部や文系/理系の種別によって適切な就職活動支援を講じるための判断材料を与えることができる．ここで，モデル構築を行うために活用できるデータとして，各年の学生たちの就職活動に関する様々なデータが就職ポータルサイトのデータベースに蓄積されている．これらのデータには，学生の基本的な属性情報のほか，ポータルサイト上での行動履歴のログデータなどがある．これらの属性情報や行動履歴データを活用し，適切な統計的予測モデルを構築することにより，学生の就職活動終了時期の予測が可能になると考えられる．

一方，就職活動に対する既存の研究の多くは，社会科学的な観点による就職活動のあり方に対する定性的な議論，あるいは，学生の就職活動への意欲，取り組み内容などに関するアンケート調査に基づくものなどがほとんどである（2.2節参照）．このため，就職活動の終了時期に対する統計的予測モデルの研究事例はなく，重要な要因や予測に適したモデルに関する知見など，明確になっていないことが多い．

そこで本研究では，学生の新卒時の就職活動終了時期の分析モデル構築を目的とし，以下の観点から，この問題にアプローチする[*2]．

(1) 就職活動終了時期に関する基本分析
(2) 対象問題の特性をふまえた就職活動終了時期の分析モデルの構築

具体的には，まず実データに対し，どのような属性情報がそのモデル化に寄与するかを明らかにするため，就職活動終了日と学生の属性の関係性について層別分析を行う．これに加え，就職活動終了日の従う確率モデルの検討を行う．これらの結果から，学生の属性による層別木を構築し，層別木の各葉ノードに複数のワイブル分布を混合した混合ワイブル分布 [2] をあてはめることで統計モデルを構築する．ここでいう層別木とは，決定木 [3], [4] の葉ノードに，ある種の確率モデルを付与した統計モデルを指す．この際，層別木の分岐を混合割合の分布の情報量で決定する学習アルゴリズムを提案し，この方法を用いて統計モデルの学習を行う．以上の手順により，層別木によって層別された各クラスタに対して，混合ワイブル分布を適用した統計モデルにより，学生の基本的な属性情報と就職活動終了時期の関係モデルが構築される．さらに，ある1年間の学生の履歴データ（実データ）を学習データとして混合ワイブル分布を葉に割り当てた層別木モデルを構築し，その翌年1年間のデータを予測することで，構築された分析モデルの推定精度を検証する．その結果，構築された統計モデルは就職活動終了時期の分布を適切に表現できており，様々な角度から考察することで分析の結果を活用できることを示す．

2. 準備

新卒学生の就職活動終了時期に対して，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事例はなく，どのようなモデルが有効であるかの知見は明らかになっていない．そこで本研究ではまず，就職活動，ならびに就職ポータルサイトを対象とした先行研究をまとめる．

さらに，本研究で目的とする就職活動終了時期の分析モデル構築に先立ち，就職ポータルサイトに蓄積された過去のログデータに対して就職活動終了時期に影響を与える要因を探り出すため，いくつかの項目を基に層別分析を行う．これにより，本研究で提案する就職活動終了時期の分析モデルの構築に寄与するであろう要因を探り，本研究における分析モデル構築の指針を示す．

2.1 日本における就職活動の概要

日本における新卒時の就職活動は，アメリカなどと異な

[*1] 日本の大学生の就職活動については，概要を付録に示す．

[*2] 本研究は文献 [1] で発表した内容を発展させたものである．

り，大学の卒業時期に合わせた新卒一括採用という慣行に従っていっせいに行われる．

毎年，5〜7 月になると，多くの企業が学部 3 年生，修士 1 年生を対象とした夏期インターンシップの選考を開始し，8〜9 月に学生が各企業において就業体験を行う．2013 年度までの就職活動は，12 月から多くの企業で採用募集が開始され，学生は企業に興味があることの意思表示を行う[*3]．この行動は「プレエントリ」と呼ばれる．それにともない，後述する就職ポータルサイトのサービスも同時に運用が始まり，多くの学生が就職ポータルサイトから，応募企業へのプレエントリを行っている．学部 4 年生，修士 2 年生に進級した直後の 4 月に，大部分の企業が採用試験をスタートさせ，次第に就職内定者が増加してくる．そのため，4 月〜6 月において就職活動終了のピークが現れるが，この期間に内定を獲得できない学生を対象とし，8 月〜10 月にかけて秋採用を行う企業も多く存在するため，この時期に就職活動の 2 回目のピークを迎えることになる．

就職ポータルサイトの運営企業では，この時期に継続的にサイトに来訪する学生を就職活動が長引いている学生として認識しており，早い段階における適切な対応をとることが望まれている．

この長期化学生の傾向を見るために，学生の就職活動終了日とその累積割合について調査を行った．その結果を**図 1** に示す．図 1 より，約半数の学生が大学 4 年次の 8 月中に就職活動を終えていることが分かる一方で，約 25%の学生が 11 月末にも就職活動を終えていないことが分かる．この点からも，就職活動は長期化しており，大きな問題であることが示唆される．

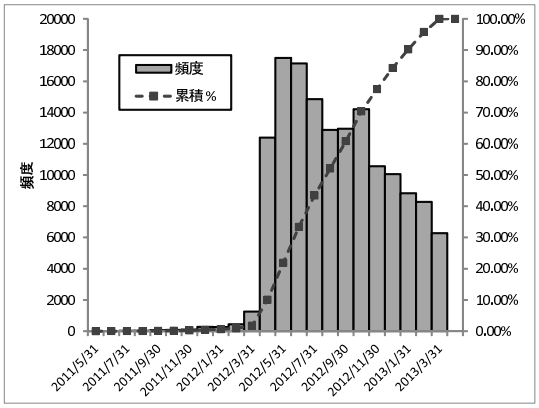

図 1 就職活動終了日とその累積割合

Fig. 1 Finishing dates of job-hunting and its cumulative percentage curve.

[*3] 経団連は，加盟企業に対し，2015 年度卒の新卒学生までは，就職活動解禁を大学 3 年 12 月解禁，採用試験は大学 4 年 4 月 1 日以降に行うことを紳士協定として要請している．また，安倍首相の要請により 2016 年度卒以降の就職解禁時期を 3 カ月後倒しとし，大学 3 年生 3 月の解禁，大学 4 年 8 月 1 日以降に採用試験と決定しており，新卒学生の就職活動時期は大きく変化した．しかしやや混乱がみられたことから，2017 年卒の就職活動は，これから 2 カ月前倒しとなっている．これらの 2016 年度以降は就職活動の構造変化が落ち着いておらず，前年度の現象が再現されないため本研究の解析対象とはしないが，今後，経団連の方針が安定する際には本稿のように前年度の統計情報から翌年度を予測するモデルは利用可能である．

2.2 就職活動に関する従来研究

すでに述べているとおり，新卒学生の就職活動終了時期に対して，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なされていない．その一方で，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検討を行っている研究が存在している．以下では，就職活動に関連する意識調査や学生のメンタルに関する分析，学生の就職活動サポートに対する検討など，就職活動に関連する従来研究についてまとめる．

前者の就職活動そのものに対する分析については，たとえば大学生を対象とした就職活動への意識調査に関する研究 [5] や，どのように学生自身がプレエントリを行う企業を見つけているのかということに関する社会心理学的研究 [6]，就職活動をマーケティングの観点から分析した研究などがある [7]．また，輕部ら [8] は，就職活動において不可避な不採用経験に着目し，不採用経験を乗り越えていく過程についてモデルを構築している．そのほかにもキャリア教育という観点からの就職活動支援に関する議論 [9] や就職活動が与えるメンタルヘルスへの影響 [10], [11] など，様々な観点から多くの社会学的な研究が行われている．

後者の就職活動支援に関する研究としては，学生が就職活動を行うにあたり，煩雑化する情報管理を簡素化するとともに，プレエントリ時に学生が作成することが多いプレエントリシートの作成作業を支援するシステムを構築した研究 [12] など，様々な就職活動支援システムに関する研究 [13], [14] がある．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多い [15], [16]．従来研究における分析の多くはアンケートやインタビューにより，就職活動に対する学生の行動調査や就職活動支援のあり方について検討を行っており，就職ポータルサイトのように非常に多くの新卒学生の履歴データを分析し，工学的な観点から就職活動終了日をモデル化している研究はなされていない．

一方，離職者の再就職活動における就業までの時間に関する分析としては，永瀬ら [17] の研究がある．この研究では，労働市場の需給状況や離職理由，学歴，前職，性別によって再就職までの時間分布がどのように異なるかについて，指数分布，ワイブル分布，Cox の比例ハザードモデルによる生存時間分布を用いた分析を行っている[*4]．この研

[*4] 永瀬ら [18] は，これらの研究に先んじて，性別や年齢，学歴，家族人数や所得といった要因と失業の関係について分析も行っている．

究は，再就職までの時間をモデル化しているという意味で本研究と類似性があるが，主に失業から再就職までのジョブサーチ期間と有効求人倍率の関連性に興味があり，地域や年代，前職，家庭環境などの変数を説明変数とした回帰型のモデルによって分析を行っている．回帰型のモデルである場合，説明変数（独立変数）がジョブサーチ期間に与える影響は独立であることが前提であるが，本研究で以下に示す提案手法では，層別によるモデル化であるため，属性データ間で目的変数に与える影響が独立ではないことを許容している点で立場の差異がある．

また，同様の生存時間分析を高等教育の問題に適用した研究としては，田尻ら [19] による成績と中退の分析などがある．一方，市川 [20] は，大卒女性を対象に，労働市場における学歴ミスマッチによる離職への影響について分析を行っており，その中で，カプラン・マイヤー法による就職活動終了確率の推移について示している．ここで分析に用いられているデータは，2000 年度に大学・大学院を卒業・修了した者に対して 2006 年に実施された質問調査であり，サンプル数は 1,171（男性 625，女性 546）である．

以上のように，ジョブサーチ期間に対する分析モデルはいくつか存在するものの，離職から再就職までの要因に主眼があったり，質問調査によって得られたデータに対する分析であるなど，本研究で対象としている就職ポータルサイトの履歴データを活用し，就職活動終了時期に対するモデル化を行った研究はない．

2.3 層別分析

就職活動の終了時期という事象を確率モデルで表現するため，就職活動終了時期について実データを用いた分析を行った．分析には，2013 年度入社の学生を対象とした就職ポータルサイト A のデータベースに蓄積されたデータのうち，「退会処理」によって就職内定時期が明らかとなっている約 15 万件の学生データを使用した[*5,*6]．この就職ポータルサイトには，大きく 2 種類のデータが保存されている．1 つは，年齢，学種（理系/文系，修士/学部）のように，ポータルサイト登録時に入力される学生の基本情報である．2 つ目は，いつポータルサイトにログインしたか，どの企業のページを閲覧したか，どの企業にプレエントリを行ったか，といったポータルサイトを使用した行動のログデータである．このデータは，学生の行動情報と呼ばれている．これらの 2 種類のデータの中から，就職活動が終了する時期に影響を及ぼすであろうと考えられる要素と実際の就職活動終了時期の関係性について層別分析を行った．行った分析の項目（層別に用いる変数）を**表 1** にまとめる．

表 1 基本層別分析項目
Table 1 Basic variables for stratification analysis.

分析番号	分析項目（層別変数）	情報の種類
I	理系/文系・修士/学部	基本情報
II	大学の偏差値	基本情報
III	学部	基本情報
IV	早期プレエントリ数	行動情報
V	人気企業へのプレエントリ率	行動情報

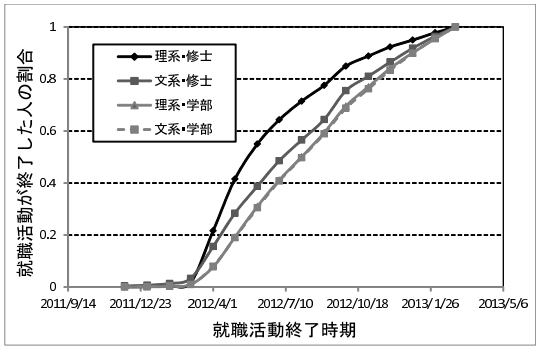

図 2 学種別の就職活動終了時期の差
Fig. 2 The difference of finishing dates of job hunting between undergraduate/graduate and science fields.

これらを基に，層別分析を行うことで，各項目をモデルに取り込むことが妥当であるか否かの検討を行う．

2.3.1 理系/文系，修士/学部による層別分析の結果

理系/文系，修士/学部の属性を持つ学生の差異と就職活動の終了時期の関係性を把握するため，これらの属性と就職活動終了時期の関係について分析した．学生は理系修士，文系修士，理系学部，文系学部の全 4 属性で層別して差異を分析する．**図 2** は 4 種類の学生の属性別の就職活動が終了した学生の割合の時間推移を表している．横軸は時間（日），縦軸はその時点までに就職活動が終了した人数の割合[*7]である．図 2 より，理系修士は就職活動終了時期が他属性より早期に訪れる割合が高いという傾向がみられる．また，学部生については，理系と文系の差に大きな違いがないことも明らかになった．

2.3.2 大学別による層別分析の結果

大学の偏差値と就職活動終了時期の関係性を知るため，ポータルサイトに登録している人数が 100 名以上の大学の中から著名な大学を 20 校選び，それらに属する学生の就職活動終了時期について分析を行った．**図 3** には分析結果の一例として，理系修士の学生全員，A 大学に所属する理系修士学生，B 大学に所属する理系修士学生，すべての（学部生も含む）学生の就職活動終了割合の推移を示す．ただ

*5 ポータルサイト A に登録される 1 年間の全ユーザ数は 60 万程度になるが，実際に「内定獲得」を理由に退会処理する学生はその一部であり，例年 13 万〜15 万件程度となる．
*6 学生数，大学数に関する精緻な値については，ポータルサイト A の運用会社における機密事項に値するため，本稿では厳密な数値の公開は控え，丸めた値を用いるものとする．
*7 本稿における就職活動終了日は，就職ポータルサイト A に登録した学生のうち，内定獲得による退会処理を行った学生が登録した日付としている．

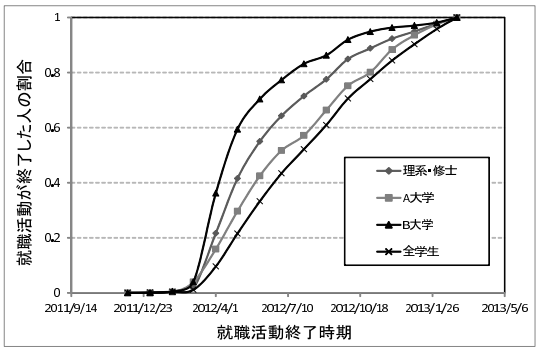

図 3 大学別の就職活動終了時期の差

Fig. 3 The difference of finishing dates of job hunting between universities.

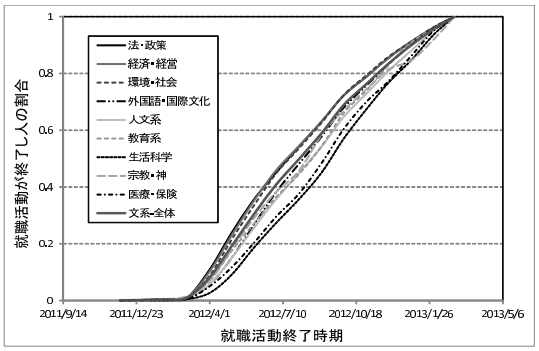

図 4 学部別の就職活動終了時期の差

Fig. 4 The difference of finishing dates of job hunting between undergraduate field.

し，記載を簡潔にするため，これらの 20 大学のうち，典型的な 2 大学を示すこととした．A 大学は偏差値が 50 近辺の私立大学であり，B 大学は比較的偏差値が高いとされる私立大学である．なお，その他の大学に関しても同様の傾向を示している．

図 3 の両軸は，図 2 と同様である．図 3 のように，大学間で就職活動終了時期の差は顕著に表れる．すなわち，これらの情報を用いることは，就職活動終了日を予測するうえで有効な情報となりうることが示唆された．

2.3.3 学部による層別分析の結果

学生の所属する学部と，就職活動終了時に何らかの相関があるかを検証するため，学部と終了時期に関する分析を行った．文系の学生全員を対象とし，そこからポータルサイトにおいて分けられている学部の区分により学生を層別し，各学部における就職活動終了時期を分析した．学部は 9 種類に分割される．**図 4** には，9 つの学部と，文系の学生全員の結果を示した．

図 4 の両軸は，図 2 と同様である．図 4 の結果から，法学部や経済学部といった学部は比較的就職活動終了時期が早いことが分かった．一方，生活科学や医療・保険[*8]といった学部は他の学部と比べ終了時期が遅くなる傾向にあった．しかしながら，理系/文系，修士/学部，大学の偏差値による差に比べるとそれほど大きな差ではないと考えられる．

[*8] 文部科学省の学校基本調査での分類とは異なるが，分析の関係上このような表記を用いることとした．

2.3.4 層別分析のまとめ

これらの分析の結果から，就職活動の終了時期は学生の基本情報，理系/文系，学部/修士や学校名といった学生の属性に依存していることが分かる．すなわち，これらの属性別に層別を行って統計モデルを構築することは有効であると考えられる．しかし，すべての大学で層別すると非常に多くの統計モデルを構築しなければならず，またサンプル数が極端に少ない大学や学部も多く出てきてしまうため，何らかの方法でグルーピングする必要がある．

一方，詳細を付録に示すように，早期プレエントリ数や人気企業へのプレエントリ数と就職活動終了時期との間には相関がみられなかった．この結果は，個々の学生の就職活動の成否は，単に就職ポータルサイト上の行動履歴のみで完全に予測できるものではなく，就職情報誌などの他の情報の活用，インターンシップへの参加度，OB/OG 訪問の有無といった現実の行動が大きく影響を与えている可能性を示唆する．

2.4 本研究の着眼点

層別分析の結果より，就職活動終了日は学生の属性情報により影響を受けることが示唆された．ここから，就職活動終了日予測モデルを構築する際に学生の層別を行うことでよりあてはまりのよいモデルを構築することができる可能性がある．また，図 2，図 3 に示した曲線は信頼性工学の分野で故障割合の推移などで描かれるものであり，このような場合の統計的推定と故障予測にはしばしばワイブル分布 [21], [22] が用いられている．本研究でもこの点にのっとり，ワイブル分布を援用した学生の就職活動終了日のモデル化を検討する．

一方，実際の就職活動終了日は就職活動が開始される大学 4 年次の 4 月から 5 月中旬にかけて 1 度目のピークを迎え，その後，いくつかの小さなピークがあり，収束していくと考えることができる（図 1 参照）．ここから，就職活動終了日は単純な単峰性の確率分布で表現することは難しく，その混合をとることによる多峰性の確率モデルを用いて表現することが適切であると考えられる．

3. 層別木と混合ワイブル分布に基づく就職活動終了時期予測モデル

前章の実データの分析により，学生の属性により分布の形状が大きく変化すること，就職活動終了日はワイブル分布により表現できる可能性があることを示すとともに，単独のワイブル分布では実データの特性を表現することが難しいであろうことを示した．実データに関するこれらの特徴をふまえ，以下ではよりあてはまりの良いモデルの構築

を検討する．このため，本研究では，混合ワイブル分布による予測モデルを導入するとともに，前章での予備実験の結果を用いることにより学生属性による層別を木を用いたモデルで表現し，層別木の葉ノードに混合ワイブル分布を割り当てたモデルを提案することにより就職活動データの分析を行う．混合ワイブル分布自体は新しいモデルではないが，その混合比を用いて層別木を構成し，混合比の類似性によって自動的に良い層別を生成している点が，本研究で示す新たな工夫である．

以下に，本提案で用いる手法を述べる．

(1) 混合ワイブル分布：複数のワイブル分布を混合した混合ワイブル分布により，単一のワイブル分布では表現できない複雑な確率構造を学習可能となることが期待できる．そのための学習法として，EM アルゴリズムを用いた混合ワイブル分布のパラメータ推定法を構成する．

(2) 層別木：複数の学生属性による層別を見通しよくモデル化するため，木構造の層別を導入する[*9]．さらに，層別後の葉ノードに，混合ワイブル分布を仮定するモデルの特徴を生かした，層別木の構築アルゴリズムを提案する．

以上の議論を組み合わせることにより，**図 5** のような混合ワイブル分布を葉ノードに割り当てた層別木モデルを提案する．ルートノードから，修士/学部の種別，文系/理系の種別などで分岐を行うことで，階層的な層別の構造を表しており，葉ノードに 1 つの混合ワイブル分布を仮定する．図 5 の例では，9 つのグループに層別し，それらのグループごとに混合ワイブル分布を推定していることを表している．以下ではまず，提案手法の一部である混合ワイブル分布のパラメータ推定法について説明し，加えて，層別木の構築アルゴリズムを示す．

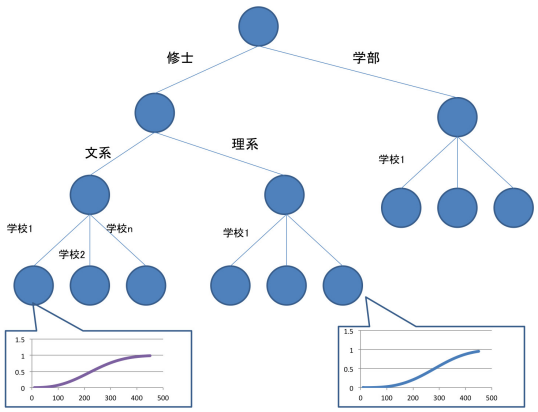

図 5 提案モデルのイメージ図

Fig. 5 The image diagram of the proposed model.

[*9] これは精度の良いモデルを構築するためには，学生属性によって学生を複数のノードに分割し，葉ノードごとに経験分布に対してあてはまりの良いモデルを構築することが望ましいと考えられるためである．

3.1 混合ワイブル分布の導入

就職活動の解禁日から就職活動終了までの日数は学生によって異なる．本稿では，無作為に学生を選択することによって与えられる就職活動終了までの日数 X を確率変数と考え，その実現値を x とする[*10]．x の定義域は $x \geq 0$ である．いま，K を混合するワイブル分布の数，π_k は k 番目のワイブル分布の混合比，m_k，η_k を k 番目のワイブル分布のパラメータとすれば，混合ワイブル分布の確率密度関数 $p(x)$ は以下で表すことができる．

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|m_k, \eta_k) \tag{1}$$

3.1.1 混合ワイブル分布のパラメータ推定

混合ワイブル分布のパラメータ π_k，m_k，η_k を推定する方法として，ワイブル確率紙にプロットして求める方法 [23] などがある．しかし，この方法は近似手法であり，コンピュータによる実装にも不適であるため，本研究では，EM アルゴリズムによるパラメータの推定法を用いるものとする．

いま，$w_\alpha^{(k)}$ は α 番目のデータがクラス k に所属する確率を示している．N はデータ数であり，N_k はクラス k に所属するデータの個数である．混合ワイブル分布の混合比は式 (2) によって推定される．

$$\pi_k = \frac{N_k}{N} \tag{2}$$

ただし，

$$N_k = \sum_{\alpha=1}^{N} w_\alpha^{(k)} \tag{3}$$

$$w_\alpha^{(k)} = \frac{\pi_k p(x_\alpha|m_k, \eta_k)}{\sum_{l=1}^{K} \pi_l p(x_\alpha|m_l, \eta_l)} \tag{4}$$

$$\sum_{k=1}^{K} w_\alpha^{(k)} = 1 \tag{5}$$

とする．また，パラメータ m_k，η_k は以下の式 (6)，(7) によって推定される．

$$m_k = b_k \tag{6}$$

$$\eta_k = \left(\frac{1}{a_k}\right)^{\frac{1}{b_k}} \tag{7}$$

ただし，

$$a_k = \frac{N_k}{\sum_{\alpha=1}^{N} w_\alpha^{(k)} (x_\alpha)^{b_k}} \tag{8}$$

[*10] 標本 ω を学生で定義し，全学生からなる加算有限集合を標本空間 Ω とする（$\omega \in \Omega$）．集合 Ω 上に定義された一様の確率測度 $P(\omega) = 1/|\Omega|$ のもとで，$X = X(\omega)$ によって確率変数 X を定義することができる．ただし，このような定義をすると X の実現値が加算有限集合になってしまうため，厳密には連続分布であるワイブル分布をあてはめることに疑問が生ずる．しかし，十分大きなサイズの Ω に対しては，人間の身長や体重の分布が正規分布で近似できるのと同様に，実用的には連続分布で表現する方が有用である．

$$b_k = \frac{N_k}{\sum_{\alpha=1}^{N} w_\alpha^{(k)} \log(x_\alpha)^{b_k}\{a_k(x_\alpha)^{b_k} - 1\}} \tag{9}$$

とする．

混合ワイブル分布のパラメータと混合比を推定するために，以下のようにEMアルゴリズムを構成した．

【混合ワイブル分布のEMアルゴリズム】

Step 0 [初期化] $w_\alpha^{(k)}$ に初期値を与え，式(3)により，N_k を計算する．

Step 1 [Mステップ] Step1-1～Step1-4で，m_k，η_k を計算する．

Step1-1) a_k，b_k に初期値を与える．

Step1-2) 式(8)，(9)により，a_k，b_k を更新する．

Step1-3) a_k, b_k が収束条件を満たさなければ，Step1-1へ戻る．

Step1-4) 式(6)，(7)により，m_k，η_k を計算する．

Step 2 [Eステップ] 式(A.2)により $p(x|m_k, \eta_k)$ を計算し，式(2)，(4)により π_k，$w_\alpha^{(k)}$ を計算する．式(3)により，N_k を計算する．

Step 3 $w_\alpha^{(k)}$ が収束するまで，Step1～Step2を繰り返す．

□

3.1.2 混合ワイブル分布による推定結果

本研究では，予測モデルを利用する現場レベルでの解釈容易性を確保するため，混合分布の混合数は就職活動終了時期が「早期に終わるグループ」，「平均的なグループ」，「就職活動が長引くグループ」の3つに分かれると仮定し，$K = 3$ とする．この混合数 K については，$K = 5$ や $K = 10$ などの設定も可能であるが，以下の理由により，実務上の結果の利用価値を考慮して，$K = 3$ を採用する[*11]．

(1) 各大学で層別した際のサンプル数は，数件から多くても3,000件程度であり，修士・学部別，理系・文系別まで考慮して層別すると低頻度のグループが生じてしまう．K を大きくとりすぎると，多くの大学において推定精度の問題が起きる．

(2) 就職ポータルサイトの運営側や大学の就職支援部門において，「早期に終わるグループ」，「平均的なグループ」，「就職活動が長引くグループ」の3グループで理解するという考え方は，実務上受け入れられやすく，結果の解釈も比較的容易である．

以上の理由から，$K = 3$ の場合に対して，3.1.1項において提案した手法に基づき，就職ポータルサイトAに登録している学生数が100名以上の約500校の大学に対し，理系/文系，学部/修士別にそれぞれ混合ワイブル分布のパラメータと混合比を計算した．**表2**に一例として，理系・修士の全体と学生数が多く，それぞれの偏差値が異なるように選定した首都圏私立4大学の理系修士に対して推定された各パラメータを，**表3**には混合比を示す．

[*11] 異なる K で全データを学習させた際のパラメータ推定値については，参考のため付録に示したので参照のこと．

表2 理系・修士に対する推定されたパラメータの一例

Table 2 An example of estimated parameters for the group of science and master course.

大学名	η_1	m_1	η_2	m_2	η_3	m_3
A大学	129.82	4.64	202.69	8.67	372.68	5.79
B大学	136.00	10.30	190.90	7.76	347.51	5.03
C大学	139.14	12.20	193.92	7.66	360.38	5.18
D大学	138.88	10.55	195.15	8.05	354.04	5.01
理系・修士全体	138.93	9.35	198.56	8.10	362.73	5.22

表3 理系・修士に対する推定された混合比の一例

Table 3 An example of estimated mixed rates for the group of science and master course.

大学名	π_1	π_2	π_3
A大学	0.20	0.32	0.48
B大学	0.33	0.41	0.26
C大学	0.39	0.38	0.23
D大学	0.15	0.42	0.26
理系・修士全体	0.20	0.40	0.40

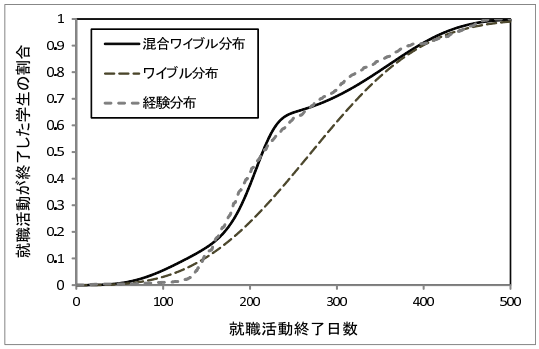

図6 混合ワイブル分布のあてはまり

Fig. 6 Adaptability of the mixed Weibull distribution to the data.

表2に示すとおり，大学間における混合される3つのワイブル分布でパラメータの値に過度に大きな差異はみられない[*12]．一方，表3から，各大学の混合比の値は比較的差異がみられることが分かる．この傾向は，表2，表3に示した例以外にも，就職ポータルサイトに100名以上のユーザがいる約500の大学すべてに対してみられた．この結果から，各大学の平均的な就職活動終了時期は，混合ワイブル分布の混合比によって特徴づけられると考えられる．

図6に表2におけるA大学を対象として，データから推定したパラメータを使用した混合ワイブル分布と従来の単一のワイブル分布を示す．図6より単純なワイブル分布よりも，混合ワイブル分布の方が経験分布に対するあてはまりが良いことが分かる．これらから，混合ワイブル分

[*12] ポアソン分布の形状は，形状パラメータ m の値によって変化する．本研究にて推定した形状パラメータはすべて $m > 1$ であり，その概形には大きな変化はないと考えることができる．したがって，以下では形状パラメータについての議論は行わないものとする．

表 4 要素ワイブル分布の平均値の一例（理系・修士）
Table 4 An example of averages of element Weibull distributions for the group of science and master course.

大学名	第一ワイブル	第二ワイブル	第三ワイブル
A 大学	118.68	191.61	345.06
B 大学	129.54	179.51	319.18
C 大学	133.42	182.33	331.57
D 大学	132.42	183.84	325.11
理系・修士全体	131.79	187.10	333.87

布を導入することにより，学習データの統計的構造を適切に表現することが可能となり，結果として学習データへのフィッティングが良くなるとともに，テストデータへの予測性能も高まることが期待される．

以上の事実は，大学間で就職活動終了時期に平均的な差がみられるものの，平均的に終了時期の遅い大学に属する全学生が遅くなるわけではないことを示している．どの大学にも，就職活動が早々に終了する学生が存在している．逆に，偏差値の高い大学にも就職活動がなかなか終了しない学生が存在するのが実態である．大学間での平均的な就職活動終了時期の差異は，それらのグループの比率が変わることによって生じるものであり，大学や文系/理系などの基本属性のみによって画一的に決まるものではないといえる．

例に示した 4 大学の各ワイブル分布における就職活動終了日の平均値を**表 4** に示す．表 4 におけるそれぞれの値は各ワイブル分布に従う学生の就職活動が終了するまでの日数を表現しており，第一ワイブル分布は「就職活動が早期に終了するであろう学生」，第二ワイブル分布は同様に「平均的な学生」，第三ワイブル分布は「就職活動が長引いてしまう学生」が従う分布とそれぞれ解釈することができる．表 2，表 3 を用いることで，就職活動の終了時期は混合ワイブル分布の混合比により定量的に評価可能となる．π_1 は就職活動が早期に終了する学生の割合，π_2 は平均的な学生の割合，π_3 は就職活動が長引いてしまう学生の割合を意味している．たとえば，A 大学は中堅私立大学であるが，π_2 の平均的な学生の割合が理系・修士全体よりもやや少ないものの，就職活動を早々に終了する学生の割合 π_1 は全体と変わらない平均的な大学であることをうかがわせる．B 大学と C 大学は，私立大学の上位校といわれる 2 大学であるが，全体よりも就職活動を早々に終了する学生の割合 π_1 が高いことが分かる．しかし，差があったとしても，A 大学にも就職活動が早期に終了する学生が一定割合存在しており，すべての大学の学生にとってある意味で勇気付けられる結果となっている．

以上は，理系・修士に対するいくつかの大学の推定結果の例であるが，**表 5**，**表 6** に文系や学部も含めた全学生に対する 4 つの大学の推定結果を示す．

これらの結果から，文系や学部生も含めた全学生に対しても，混合比の差異は見られるものの，各要素ワイブル分布のパラメータ推定値は似通っており，ほぼ同様の傾向を見ることができる．すなわち，大学間での就職活動終了時期の差異は，主に「早期に終わるグループ」，「平均的なグループ」，「就職活動が長引くグループ」の 3 つのグループの割合の差異によるものと考えることが可能である．

表 5 各大学に対して推定されたパラメータ（全学生）
Table 5 An example of the estimated parameters for each university (all students).

大学名	η_1	m_1	η_2	m_2	η_3	m_3
A 大学	134.39	5.58	206.78	8.77	373.02	5.69
B 大学	136.98	7.31	199.19	8.14	350.08	5.23
C 大学	134.77	5.13	195.80	8.19	339.98	5.23
D 大学	136.68	6.96	203.23	8.47	367.47	5.44
全体	134.63	5.36	205.58	8.84	376.42	5.56

表 6 各大学に対して推定された混合比（全学生）
Table 6 An example of the estimated mixed rates for each university (all students).

大学名	π_1	π_2	π_3
A 大学	0.07	0.32	0.61
B 大学	0.21	0.38	0.42
C 大学	0.25	0.42	0.33
D 大学	0.11	0.38	0.51
全体	0.09	0.33	0.58

3.2 層別木の構築アルゴリズム

3.2.1 仮定

層別木は，図 5 で示したように，離散変数を用いてデータを階層的に層別し，木構造の葉ノードに層別されたデータ集合とそこから推定された確率モデルを割り当てたものである．

本研究では，層別木を作成する際に分岐を行う変数の決定として，混合ワイブル分布の混合比を用いる学習アルゴリズムを提案する．3.1.2 項で示したとおり，混合される個々のワイブル分布のパラメータは各大学でほとんど差がないが，混合比が大きく異なっている．そのため，分布の違いを説明するためには，混合比を用いることが効果的であると考えられる．本研究では，混合される分布のパラメータに変化はないが，混合比が大きく変化するようなモデルを仮定し，層別木モデルの学習アルゴリズムを提案する．

3.2.2 層別木モデル

ここでは，提案する層別木モデルの作成アルゴリズムについて説明する．一般的な統計学における層別では，群間で統計的な特徴の差が出るか否かによって層別に用いるための変数を決める．本稿で扱う問題においては，就職活動終了時の分布の統計的特徴に差が生じるように層別を行う必要がある．一方，先に示したように累積度数分布の概形

は混合比に大きく依存するため，木を作成する際に分岐を行う変数の決定として，混合ワイブル分布の混合比を用いることができる．

広く様々な分野で適用がなされている決定木モデルでは，目的変数である離散ラベルの頻度分布が，なるべく偏るように分岐を繰り返していく．その分岐の際に用いられる基準がエントロピーや Gini 係数などの不純度であり，これは目的変数である離散ラベルの度数分布の偏りを測る尺度となっている．本稿で示すアルゴリズムでは，この不純度を混合ワイブル分布の混合比から計算することで，その他は決定木の構成アルゴリズムとほぼ同様の手順で，層別木を構成する．

ノードを層別する際にまだ層別に用いていない J 個の層別因子を $F_1, F_2, \cdots, F_j, \cdots, F_J$ とする．層別因子 F_j の持つ M_j 個の水準を $L_1^j, L_2^j, \cdots, L_m^j, \cdots, L_{M_j}^j$ と表す．あるノード s に割り当てられているデータ数を N_s とする．ノード s に割り当てた混合ワイブル分布の混合確率を π_k^s と表したとき（$k = 1, 2, \cdots, K$），ノード s における混合確率のエントロピー $I(s)$ は式 (10) で計算される．

$$I_s = -\sum_{k=1}^{K} \pi_k^s \log_2 \pi_k^s \tag{10}$$

これは混合ワイブル分布の混合確率の曖昧性を表す尺度となっており，大きいほど一様分布に近く，小さいほど 1 つのワイブル分布の混合確率が高く他が 0 に近いような分布となっている．いま，層別木としては，なるべく分岐された子ノード間で混合ワイブル分布の混合確率が変わるような層別が望ましいと考えることができる[*13]．

因子 F_j で層別した後の水準 L_m^j で割り当てたノードのデータ数を N_m^j とする．また，このノードに割り当てた混合ワイブル分布の混合比を $\pi_k^{j,m}$ と表現する．因子 F_j により層別された後のノードの情報量 $I_s(F_j)$ を式 (11) で計算する．

$$I_s(F_j) = -\sum_{m=1}^{M_j} \frac{N_m^j}{N_s} \sum_{k=1}^{K} \pi_k^{j,m} \log_2 \pi_k^{j,m} \tag{11}$$

これらの情報量を用いて，決定木における学習アルゴリズムと類似した方法で，逐次，層別木を成長させていくことが可能である．

以下に，層別木の生成アルゴリズムを示す．

【層別木の生成アルゴリズム】

Step1) ルートノード s のみからなる木を初期値とする．全データをルートノード s に割り当て，$N_s = N$ とする．

Step2) 式 (11) により，すべての葉ノード s に対して I_s と各層別因子の $I_s(F_j)$ を計算する（$j = 1, \ldots, J$）．

Step3) $(I_s - I_s(F_j))N_s$ が最大となる葉ノード s を選択する．

Step4) Step3) で選択された葉ノード s において，$I_s - I_s(F_j)$ が最大となる F_j を選択し，F_{sj}^* とする．

Step5) 選択した層別因子 F_{sj}^* によりノードを分岐し，各子ノードに属するデータを求め，子ノード s' に属するデータ数 $N_{s'}$ を求める．

Step6) 終了条件を満たさなければ，Step2) に戻る．

□

本研究における終了条件は，層別木の葉ノードに属するデータ数が事前に設定する閾値を下回った場合とした[*14]．こうして得られる層別木の各葉ノードには，混合ワイブル分布が付与されている．各葉ノードは，ルートノードから属性値によって階層的層別されてきた学生のグループを表しており，各グループに対して 1 つの混合ワイブル分布が仮定されたモデルとなっている．

4. 実データの分析結果

本章では，提案した分析法によって，ポータルサイトの実データを分析し，分析モデルを構築した結果を示す．

4.1 分析データ

提案手法の有効性を示すため，ポータルサイト上の実データを用いて分析を行い，モデルの予測精度について検証を行う．具体的には，2012 年度のポータルサイト利用学生のうち，就職活動終了時期を入力した全学生のデータを学習データとして予測モデルを構築し，翌年である 2013 年度のデータ（テストデータ）を予測し，その予測精度を用いて分析モデルを評価する．学習に用いていないテストデータに対しても予測精度が高まるのであれば，分析モデルとしての推定精度も高いと考えることができる．そのうえで，得られたモデルについて考察を与える．

学習データは 2012 年度利用学生のポータルサイトの約 15 万名分の学生データ，予測精度を測るためのテストデータは 2013 年度利用学生のポータルサイト約 15 万名分の学生データである．すなわち，2012 年度版のデータを使用し，2013 年度版の学生の就職活動終了日を予測するものとする．なお，ここで用いている 2013 年度版のデータについては，基本分析に使用したデータと同一のものとなっていることに注意されたい．

また，層別を行うための説明変数を $\boldsymbol{u}$ =（学種，文理，学校クラスタ）とした．学種は，学部生か大学院生かを表す変数，文理は文系か理系かを表す変数である．学校クラ

[*13] 混合ワイブル分布において，混合される要素ワイブル分布が固定の場合，混合割合が変わらないということは，その分岐（層別）によって分布が変化しないことを意味する．群間でなるべく統計的特徴が異なるような層別が望ましい．

[*14] 決定木などの分岐条件で使われている MDL（Minimum Description Length）基準や AIC（Akaike Information Criterion），BIC（Bayesian Information Criterion）などのモデル選択基準を用いることも可能である．

スタは，混合ワイブル分布の混合比をもとに k-means 法によって学校をクラスタリングしたクラスタとし，クラスタ数は予備実験を通じて実験的に決定した．層別木の構築方法として，まず（学種，文理）の 2 つの説明変数を用いて，木を分岐させる．その後，各葉ノードにおいて，混合比をもとに学校を複数のクラスタへ分割して，層別木を構築する．評価指標として，葉ノードに割り当てた分布の中央値とテストデータにおける中央値との平均絶対誤差，ならびに，各葉ノードで推定した分布とテストデータの分布との KL 情報量の平均を使用する．いま，経験分布の確率密度関数を $g(x)$，各葉ノードで推定した混合ワイブル分布を $\hat{g}(x)$ としたときの，KL 情報量の計算式を式 (12) に示す．

$$KL = \sum_{i=1}^{N} g(x_i) \log \frac{g(x_i)}{\hat{g}(x_i)} \tag{12}$$

平均絶対誤差を用いることで，各属性の学生の半数が就職活動を終了する時期を予測する．KL 情報量により，推定した分布とテストデータの類似度を測る．本研究では，就職ポータルサイトに保存されたデータの分析を対象とするため，比較対象とする従来の予測モデルが明確ではない．このため，ベースラインとなる予測性能を明確とするための比較手法として，単一のワイブル分布による予測値を使用する[*15]．

4.2 実験結果・考察

まず，層別木を生成するために推定した全データに対する混合ワイブル分布の推定結果を**表 7**，**表 8** に示す．また，表 7 の要素ワイブル分布を用いて提案した層別木生成アルゴリズムによって得られた層別木を**図 7** に，この層別木と混合ワイブル分布による推定モデルを用いて翌年度のテストデータを予測した結果を**表 9** に示す．

表 7 全データに対して推定されたパラメータ
Table 7 The estimated parameters for all data (no stratification).

	第一ワイブル		第二ワイブル		第三ワイブル	
パラメータ	η_1	m_1	η_2	m_2	η_3	m_3
推定値	134.63	5.36	205.58	8.84	376.42	5.56
平均値	124.11		194.52		347.73	

表 8 全データに対して推定された混合比
Table 8 The estimated mixed rates for all data (no stratification).

パラメータ	π_1	π_2	π_3
推定値	0.09	0.33	0.58

[*15] 本研究では，混合ワイブル分布を用いて学生の就職活動終了時期の分析モデルを構築しているため，混合モデルを導入することの妥当性を検証するためのベースラインとして，混合をしない単一のワイブル分布の精度を取り上げる．混合ワイブル分布では，パラメータ数が増え，モデルが複雑化しているため，これが無意味であれば，予測精度が劣化するはずである．

表 9 より，中央値の平均絶対誤差，KL 情報量ともに提案手法がより小さい値を示している．中央値の平均絶対誤差がより小さいことから，提案手法の予測した中央値がテストデータの中央値により近い値であることが分かる．また，KL 情報量の値が比較手法よりも小さいので，予測した分布がよりテストデータの分布と近似していると考えられる．ここから，本提案手法の予測精度は単一のワイブル分布を用いたものと比較して優れていることが分かる．

さらに，提案手法の予測精度が高まった理由について検討を行うため，層別木の各葉ノードごとに分析を行った．**図 8** に各葉ノードにおける中央値の平均絶対誤差，ならびに**図 9** にその KL 情報量を示す．図 8 の結果より，（理系，修士，Cluster2）の属性以外では，提案手法の予測値が比較手法の予測値に比べ誤差が小さいことが示される．また，図 9 の結果より，（理系，修士，Cluster2）の属性以外では，提案手法の KL 情報量がより小さいことが分かる．

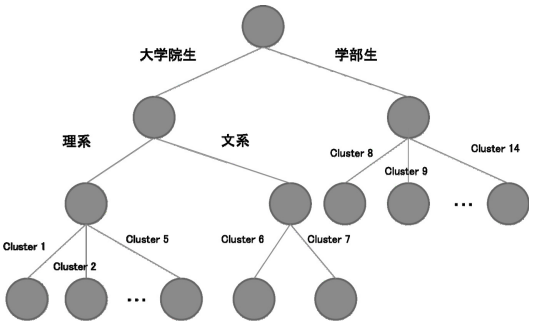

図 7 2012 年度の全データから構築された層別木
Fig. 7 Stratification tree structure estimated by the data of the 2012 academic year.

表 9 予測精度
Table 9 Prediction accuracy.

評価指標	提案手法	比較手法
平均絶対誤差	25.07	53.12
KL 情報量	120.48	547.24

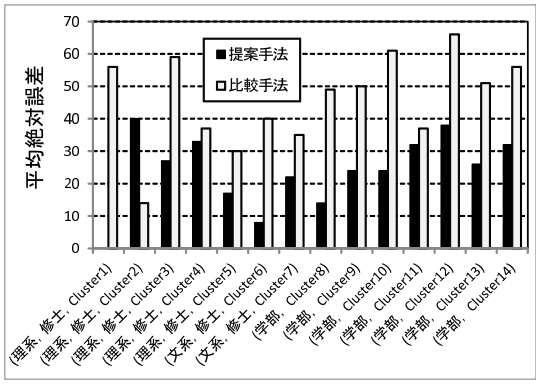

図 8 各葉ノードの中央値の平均絶対誤差
Fig. 8 Mean absolute error of the median on each leaf node.

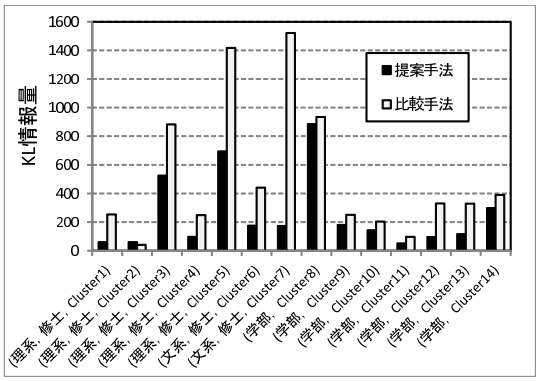

図 9 各葉ノードの KL 情報量
Fig. 9 The KL divergence on each leaf node.

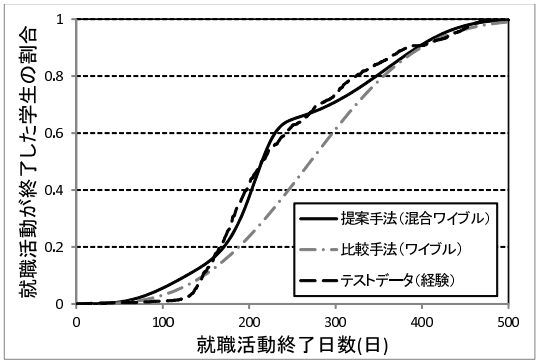

図 10 （理系，修士，Cluster1）の混合ワイブル分布
Fig. 10 The estimated mixed Weibull distribution for the group of Science, Master course, and Cluster1.

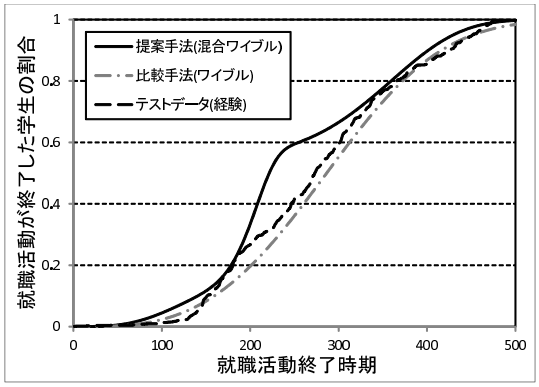

図 11 （理系，修士，Cluster2）の混合ワイブル分布
Fig. 11 The estimated mixed Weibull distribution for the group of Science, Master course, and Cluster2.

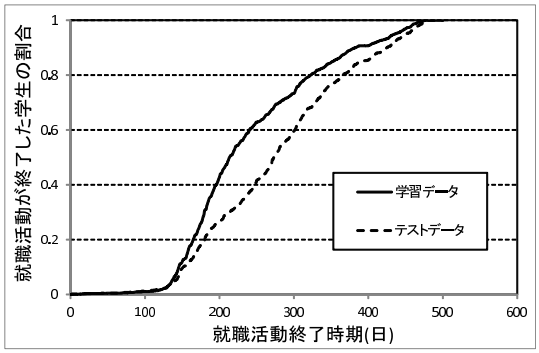

図 12 （理系，修士，Cluster2）の学習データとテストデータの差
Fig. 12 The difference between training and test data (Science, Master course, and Cluster2).

これらの結果のうち，提案手法の特徴について考察を行うため，提案手法が優れていた（理系，修士，Cluster1），ならびに，提案手法が比較手法よりも劣っていた（理系，修士，Cluster2）の両者に焦点を絞り，分析を行う．

図 10 に（理系，修士，Cluster1）の属性で推定した結果を示す．就職活動の終了時期は，早期の段階で内定を獲得できる人数の割合によって決まるという特徴から，経験分布の形状が非対称となるという特徴が見て取れる．混合ワイブル分布により，この非対称性を表現できたため，高い精度で推定を行うことができ，予測した中央値の値もテストデータにより近い値だったと考えられる．図 10 より，提案手法がより経験分布と近似していることが見て取れる．

うまく予測を行うことができなかった，（理系，修士，Cluster2）の属性で推定した結果も**図 11** に示す．図 11 より，（理系，修士，Cluster2）のノードでは，うまく予測を行うことができなかったことが分かる．一般的に，多くの企業では，4〜5 月に内定を出す企業が多い．そのため，4〜5 月にかけて就職活動が終了する人が多くなり，グラフの傾きが急になる傾向がある．しかし，（理系，修士，Cluster2）の属性を持つ学生には，この傾向が顕著にみられず，時間経過において一定量で就職活動が終了した学生が増加していることが分かる．このことから通常のワイブル分布によっても十分に予測ができたものだと考えられる．その他の要因として，この属性のデータの傾向が学習データとテストデータで大幅に異なっていることも考えられる．**図 12** に（理系，修士，Cluster2）の学習データとテストデータの経験分布を示す．

図 12 より，学習データとテストデータで傾向が大きく異なっていることが分かる．このことから，属性によっては年ごとに就職活動の傾向が大きく異なってしまうと考えられる．データの詳細を調べてみると，（理系，修士，Cluster2）の属性は比較的就職活動終了時期が遅い学生であることが分かった．これは景気などの外的要因により，その年の就職活動の傾向が決定され，就職活動が苦手な学生はその影響を大きく受けることが考えられる．

5. 考察

5.1 提案モデルから得られた知見

表 2，表 3 に示したとおり，大学間の各混合分布のパラメータに比較的大きな変化はなく，混合比のみが変化するという結果から，各大学において就職活動終了時期に何らかの差はあるが，どの大学にも早期の段階で内定を獲得できる学生が存在していることが分かる．これらの人数の比

表 10 学部生の各クラスタの混合比と特徴
Table 10 The mixed rates and characteristics of each cluster (undergraduate students).

クラスタ番号	π_1	π_2	π_3	所属大学の主な特徴
8	0.08	0.37	0.55	地方私立大学
9	0.11	0.43	0.46	偏差値 50 付近 私立大学
10	0.17	0.56	0.27	偏差値 55 付近 国立・私立大学
11	0.04	0.11	0.85	偏差値 40 付近 私立大学
12	0.30	0.48	0.22	偏差値 60 付近 国立・私立大学
13	0.11	0.51	0.38	偏差値 50 付近 国立・私立大学
14	0.09	0.27	0.64	地方国立・私立大学

表 11 学部生の各クラスタの平均値
Table 11 The mean of each cluster (undergraduate students).

クラスタ番号	平均値
8	273.15
9	257.25
10	223.91
11	321.93
12	207.10
13	244.99
14	286.23

率によって，各大学の就職活動終了時期が異なっていると考えられるため，混合比の詳しい分析を行うことにより，就職活動に有利な大学の発見や類似傾向のある大学どうしのクラスタリングが可能になる．**表 10** に実験で作られた図 7 の層別木の学部生の属性を持つ 7 個のクラスタの混合比を，**表 11** にそれらの平均値を示す[*16]．

クラスタ 8 に所属する大学は，π_1 が比較的小さく，π_2，π_3 とだんだん値が大きくなっていることが分かる．そのため，早期に就職活動が終了する学生はとても少ないが，時間経過につれて終了する学生が増加していく傾向がある．クラスタ 9 の大学は，π_2，π_3 の値がほぼ同じという傾向が見て取れる．この傾向から，一般的な春採用および，秋採用において大多数の就職活動が終了したと考えられる．クラスタ 10 の大学は，π_2 が大きいことから，比較的に就職活動が早く終了する大学の集団であることを示している．クラスタ 11 に含まれる大学では，他と比較しても，π_3 の値が非常に大きいことが分かる．このことから，非常に多くの学生の就職活動が長期化する傾向があると推測される．クラスタ 12 に所属する大学は，π_1，π_2 の値が比較的大きく，最も就職に強い大学であることが分かる．特に π_3 の値は他のクラスタと比べ，大変小さく，非常に優秀な学生が集まっている．クラスタ 13 は π_2 の値が大きいがクラスタ 9 と類似したパターンである．クラスタ 14 の傾向は，π_3 の値が大きいがクラスタ 8 と同じ傾向を持つと考えられる．各クラスタにおけるこれらの就職活動のパターンから，π_3 の値が大きいクラスタ 11 に所属する大学は就職活動の終了時期が遅く，就職ポータルサイトの運営企業や，大学は積極的に就職活動支援を行う必要がある．

[*16] 3 つの要素ワイブル分布のパラメータ推定値と平均値は，表 7 に示したとおりで，すべてのクラスタで同一である．各クラスタの混合ワイブル分布の平均値は，混合比に依存して決まっている．

5.2 分析モデルの特徴について

本研究で示した分析モデルにおける「どの大学にも就職活動が早期に終了する学生と長期化する学生が存在しており，大学や学部間の差はそれらの比によって生じている」という結果は，個々の大学や学部に対して就職支援活動を行う際に大変受け入れてもらいやすい結果といえる．すべての学生の就職活動が在籍する大学や学部のみで決まってしまうわけではなく，実際にはどの大学にも就職活動が早期に終了する学生は存在している．一方で，そのような学生グループの割合について大学間や学部間，理系・文系別での差異を把握できることは，各大学や学部における就職活動指導などにおいても有益な情報を与えていると考えられる．

一方，混合比により定量的に判断するのではなく，各大学や学部別に就職活動終了日の累積割合をグラフ化することで定性的な分析を行う方法も考えられる．しかし，日本には大学が 700 以上存在するため（文部省の 2015 年度学校基本調査によれば，大学数は 779），大学や学部による層別分析には次のような問題が生じてしまう．

(1) 700 以上の大学別に層別することの手間は非常に大きい．
(2) データ数が 10 万以上であるものの，大学で層別した場合，データ数が不十分な大学がほとんどになってしまう．

したがって，累積割合グラフを描くためにはある程度のデータ数を必要とするため，比較的学生数の多い総合大学などでは分析が可能であるが，多くの大学・学部では分析ができないことになってしまう．これに対し，本研究の分析モデルでは全体を属性で層別していくため，データ数の少ない大学や学部はクラスタにまとめられて混合ワイブル分布のパラメータが推定され，おおよその推定が可能となる．

本研究のモデルでは，大学や学部などが混合ワイブル分布の混合割合という特徴量を有し，その就職活動終了時期に与える影響度を考慮した類似性によってクラスタリングされていると考えることもできる．その意味で類似した大学などのグループが生成されているため，これらの各クラスタを分析の結果として活用することも可能である．

また，本研究のモデルは，1 年間の学習データでモデルを構築し，翌年のテストデータに対するあてはまりによって精度を検証できているため，ある程度の再現性を有する統計的特徴をモデル化できていると考えられる．一方，就職活動に関する統計的傾向の分析は，経済状況のほか，業

界の人気の動向など，動的で複雑な社会問題を対象としている．そのため，1年間の学習データで得られた統計モデルが，将来にわたって永続的に利用可能であるとはいえない．ある年度の1年間の学習データの分析から得られる統計的特徴が，翌年度もおおよそ保持されることを仮定した結果の利用となるが，構造変化の可能性をふまえた結果の取扱いや考察も必要である．一方で長期的な傾向の変化に関しては，就職ポータルサイトのデータは，毎年60万件規模の履歴データが積み上がっていくため，毎年，データを更新してモデルを学習し直し，逐次最新の統計モデルを用意して対応することは比較的容易である．この点は，アンケート調査などのデータ取得にコストがかかる方法と比較した際のメリットの1つと考えることができる．

5.3 予測モデルとしての解釈について

本研究は，個々の学生の就職活動終了時期の予測の精度が高ければ良しとする予測モデルを構築することが目的ではなく，就職活動終了時期の差異と影響要因を分析することを目指した分析モデルを提案した．一方で，提案する分析モデルが翌年のデータを精度良く推測できているか否かは，学習データから統計的特徴を正しくモデル化できているかどうかを判断するための基準となる．

そのため本稿ではテストデータに対する予測精度という側面から，提案した分析モデルの評価を行った．本提案モデルが高い精度で就職活動終了時期を説明可能であることは，学習データの統計的特徴を正しく学習できていることを示している．

一方，個々の学生に対する就職活動終了時期の予測モデルという立場に立てば，連続変数の予測モデルである線形回帰モデルやニューラルネットなどによるモデル化も考えられる．このような予測モデルを構築した場合，2015年度からの経団連指針による就職採用活動の構造変化といった問題には対処し難いという点に注意が必要である．また，就職活動終了時期の分布は，図1に示したように正規分布とはいえないため，最小二乗法をベースとする回帰モデルの妥当性については入念な検証が必要である．ニューラルネットも最小二乗誤差を最小化するモデルであることに加え，非線形モデルであり，学習されたモデルの内部構造が分かり難いという解釈上の問題がある．

これに対して，本研究で示した混合ワイブル分布による分析モデルでは，各属性別での混合ワイブル分布の混合割合の差異によって，就職活動終了時期の差異を説明するモデルとなっており，実務的に理解がしやすいという特徴を有している．

5.4 就職ポータルサイトA運営企業との議論から得られた経験的知見

これまでに述べてきたとおり，混合比により就職活動終了時期が遅い大学を定量的に判断できるため，就職ポータルサイトの運営企業はこれらの大学に積極的にアプローチすることで就職活動終了時期を早めることができる可能性がある．大学ごとの混合比の差異により，その大学に適した就職活動の支援策を立案することもできると考えられ，今後の就職活動支援への貢献が期待できる．また，就職活動終了時期の予測モデルが示す予測ラインとの差異をモニタリングすることにより，当該年度の就職活動がどの程度順調に進んでいるのかを確認することができる．新たに実施した就職活動支援策の有効性を測るためにも有用と考えられる．

一般に，就職ポータルサイトの運営企業には，データ解析の専門部門のほか，各大学の就職課や就職担当教職員とパイプを持ち，実際の就職活動を支援する営業部門が設置されていることが多い．本研究の成果は，そのような営業部門が活用し，各大学・学部の方と該当年度の就職活動の見通しを予測し，実際との差異をモニタリングしながら対応を進めていくような用途に活用することができる．

5.5 提案モデルの限界について

本研究の提案モデルは，1年間の学習データでモデルを構築し，翌年のテストデータに対するあてはまりによって精度が検証されている．したがって，学習データの統計的性質を説明するためだけのモデルではなく，ある程度の再現性を有する統計的特徴をモデル化できていると考えられる．

しかしながら，2015年の経団連指針による就職活動時期の変更のような大きな構造変化が起こる場合，学生の就職活動や企業の採用活動が前年度とはまったく異なり，再現性が保持されなくなる．過去のデータに基づく分析モデルである以上，このような構造変化が生じた際に，何が起こるのかを正確に予測することは困難と考えられる．このような状況で，過去のデータを学習して構築した分析モデルがどの程度役立つのかについては明らかではない．

一方で，企業の採用活動のルールや時期に大きな変化が生じない場合には，ある程度の統計的構造は翌年も保持されると仮定できるため，提案モデルによる分析は翌年の就職活動支援に対しても有用な知見を与えると考えられる．また，全学生に対する就職活動終了時期に変化が生じた場合であっても，学生の属性が与える就職活動時期への影響についてはある程度は保存されると考えられる．したがって，学生の就職活動終了時期がどのように変化するのかは正確に予測できなかったとしても，学生の属性と就職活動終了時期との関係性という知見を現実場面に役立てることは可能である．ただし，就職活動市場に大きな構造変化が起きた場合には，実際に結果を観測してみるまで何が起こるかが分からないということも事実であり，そのような場合に提案した分析モデルがどのように活用されうるのかについては今後の検討が必要である．

6. まとめと今後の課題

本研究では，就職活動終了時期の分析モデルの構築を目的として，就職ポータルサイトのデータの分析を行うとともに，この事例に適合すると考えられる分析モデルの提案を行った．予備的な分析検討により，就職活動終了時期が混合ワイブル分布によって推定できることに加え，学生を層別することにより精度の良い分析モデルの構築ができることが考えられたため，葉ノードに混合ワイブル分布を割り当てた層別木モデルを提案し，実データの分析を通じた検証により予測精度が良いことを示した．また，各大学の就職活動終了時期の分布が，主にワイブル分布の混合比で規定されることが明らかとなり，就職活動終了時期を定量的に判断することができることを明らかにした．以上の結果は就職ポータルサイトに蓄積されたデータの有効活用の方法を示しており，得られた結果を実務に適用していくことで，様々な波及効果が期待できる．

今後の課題としては，混合ワイブル分布の混合数 K の決定方法があげられる．本研究で示した層別木の生成アルゴリズムでは，与えられた K に対して推定された要素ワイブル分布のパラメータを用いて分岐を行うため，あらかじめ K を与える必要がある．一方で，統計的モデル選択の観点から適切な K は，統計的特徴とデータ数の兼ね合いで与えられるため，層別の仕方によって変化する可能性がある．このような K の最適化も含めた層別木の生成アルゴリズムを構成するためには，さらに何らかの改良と工夫が必要であり，今後の課題とする．また，本手法に加えて，学生の行動履歴情報も加味した行動パターンによる予測モデルの構築も今後の課題である．実務上は，個々の学生の行動パターンによって，就職活動時期の予測を行い，学生別のきめ細かなサポートが可能となれば，大変有効なツールとなりうる．また，膨大な量の活動履歴データからプレエントリ数の予測を行うなど，様々な切り口からの分析を積み重ねることも今後の課題とする．

謝辞 本研究にあたり，貴重なデータの提供，ならびに熱心な議論をいただいた就職ポータルサイト A の運営会社の皆様に深く感謝いたします．また，本稿の査読にあたり，大変貴重なコメントをいただいたメタ査読者，ならびに査読者の皆様に深く感謝いたします．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] 早川真央，三川健太，石田 崇，後藤正幸，小川晋一郎：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予測モデル，第 36 回情報理論とその応用シンポジウム予稿集 (2013).

[2] Jiang, S. and Dimitri, K.: Maximum likelihood estimates, from censored data, for mixed-Weibull distributions, *IEEE Trans. Reliability*, Vol.41, No.2, pp.248–255 (1992).

[3] Quinlan, J.R.: Induction of Decision Trees, *Machine Learning*, Vol.1, No.1, pp.81–106 (online) (1986), available form⟨http://dx.doi.org/10.1023/A:1022643204877⟩.

[4] Quinlan, J.R.: Improved Use of Continuous Attributes in C4.5, *Journal of Artificial Intelligence Research*, Vol.4, pp.77–90 (1996).

[5] 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.7, pp.645–665 (2005).

[6] 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).

[7] 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).

[8] 輕部雄輝，佐藤 純，杉江 征：大学生の就職活動維持過程モデルの検討—不採用経験に着目して，筑波大学心理学研究，No.48, pp.71–85 (2014).

[9] 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：Bulletin of the Japanese Society for Study of Career Guidance，No.15, pp.11–19 (1994).

[10] 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：Bulletin of the Japanese Society for Study of Career Guidance，Vol.18, No.1, pp.9–16 (1997).

[11] 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).

[12] 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，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, No.17, pp.1–6 (2009).

[13] 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（教育・学習支援プラットフォーム/一般），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).

[14] 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, No.3, pp.1–6 (2014).

[15] 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，東京女子大学紀要論集，Vol.66, No.1, pp.103–118 (2015).

[16] 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.3_21–3_27 (2014).

[17] 永瀬伸子，縄田和満，水落正明：『労働力調査』を用いた離職者の再就職行動に関する実証的研究，総務省統計局リサーチペーパー，Vol.24 (2011).

[18] 永瀬伸子，水落正明：労働力調査のパネル構造を用いた失業・就業からの推移分析，総務省統計局リサーチペーパー，No.19 (2009).

[19] 田尻慎太郎，白鳥成彦：ビジネス系大学における学習履歴と活動データを用いた生存時間分析，日本教育社会学会大会発表要旨集録，Vol.65, pp.120–121 (2013).

[20] 市川恭子：若年大卒女性の早期離職に関する実証分析，生活社会科学研究，Vol.22, pp.31–46 (2015).

[21] Weibull, W.: *A Statistical Theory of the Strength of Materials*, Ingeniörsvetenskapsakademiens handlingar, Generalstabens litografiska anstalts förlag (1939).

[22] Sekine, M. and Mao, Y.: *Weibull Radar Clutter*, Electromagnetics and Radar Series, P. Peregrinus Limited (1990).

[23] 毛利正光, 塚口博司, 金 甲洙：ワイブル分布を用いた環境騒音の推計に関する研究，騒音制御，Vol.8, No.6, pp.314–320（オンライン），DOI: 10.11372/souonseigyo1977.8.314 (1984).
[24] 株式会社ダイヤモンド・ヒューマンリソース：ダイヤモンド就活ナビ（オンライン），入手先 ⟨https://navi17.shukatsu.jp/17/⟩（参照 2016-08-31）.
[25] 株式会社マイナビ：マイナビ（オンライン），入手先 ⟨http://job.mynavi.jp/⟩（参照 2016-08-31）.
[26] 株式会社リクルートキャリア：リクナビ（オンライン），入手先 ⟨http://www.rikunabi.com/⟩（参照 2016-08-31）.
[27] 株式会社アクセスヒューマネクスト：アクセス就活（オンライン），入手先 ⟨http://www.ac-lab.jp/⟩（参照 2016-08-31）.
[28] 株式会社ディスコ：キャリタス就活（オンライン），入手先 ⟨https://job.career-tasu.jp/⟩（参照 2016-08-31）.
[29] 株式会社ドリームキャリア：理系ナビ（オンライン），入手先 ⟨https://rikeinavi.com/⟩（参照 2016-08-31）.
[30] Castillo, E.: *Extreme Value Theory in Engineering*, Ingeniörsvetenskapsakademiens handlingar, Academic Press (1988).
[31] Nelson, W.B.: *Applied Life Data Analysis*, Wiley (1982).
[32] Cohen, C.A. and Whitten, B.: Modified maximum likelihood and modified moment estimators for the three-parameter Weibull distribution, *Communications in Statistics – Theory and Methods*, Vol.11, No.23, pp.2631–2656 (1982).
[33] Kappenman, R.F.: Estimation for the Three-parameter Weibull, Lognormal, and Gamma Distributions, *Comput. Stat. Data Anal.*, Vol.3, pp.11–23 (1985).
[34] Cohen, A.C.: Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples, *Technometrics*, Vol.7, No.4, pp.579–588 (1965).
[35] Lehmann, E. and Casella, G.: *Theory of Point Estimation*, Springer Verlag (1998).
[36] Nocedal, J. and Wright, S.: *Numerical Optimization, 2nd edition*, Springer Series in Operations Research and Financial Engineering, Springer, New York, NY (2006).

付　　録

A.1 就職活動における基本情報

A.1.1 就職ポータルサイトの機能

就職ポータルサイトは，学生の就職活動を包括的に支援する Web サービスである．就職ポータルサイトにおいて，学生は主に以下のような機能を無料で利用することができる．

- 自己分析
- 業種や地域などの条件を基にした企業の検索
- 企業へのプレエントリ
- 企業の説明会の予約
- 言語・非言語検査

このように，学生は就職活動を行ううえで必要となる活動を就職ポータルサイト上で行うことができる．このようなサイトは現在数多く存在しており，その主なサイトを**表 A·1** に示す．

表 A·1 代表的な就職ポータルサイト（2016 年 8 月現在）
Table A·1 Representative portal sites for job-hunting (as of August 2016).

サイト名	URL
ダイヤモンド就活ナビ [24]	https://navi17.shukatsu.jp/17/
マイナビ [25]	http://job.mynavi.jp/
リクナビ [26]	http://www.rikunabi.com/
就活ラボ [27]	http://www.ac-lab.jp/
キャリタス就活 [28]	https://job.career-tasu.jp/
理系ナビ [29]	http://www.rikeinavi.com/

A.1.2 一部企業へのエントリ集中と就職活動の長期化

近年では，多くの学生がこれらの就職ポータルサイトを利用して就職活動を行っている．就職ポータルサイトを利用することで，学生は手軽に企業にプレエントリを行うことが可能となった．学生は就職ポータルサイトで自分に適合する企業を徹底的に調べることが可能であるが，一方ですでに知っている企業にプレエントリを行う学生も多い．そのため，被プレエントリ数の多い企業と少ない企業の差が非常に大きく，人気企業にプレエントリが集中する傾向がある．そのため，多くの学生のプレエントリが人気の企業へ集中していることが，就職活動の長期化問題の一因となっている可能性がある．学生の多くが人気企業へ入りたいと思っている一方で，人気企業では企業側の希望に適合する学生を欲している．そのため，大部分の学生は人気企業の採用試験に落ちた後で，他の企業を探すことになり，学生の大半がどこかの企業に内定をもらうまでこの行動を何度も繰り返すことになる．すなわち，就職活動が長期化してしまう学生は自分を採用してくれる企業が見つかるまで，多くの時間を就職活動に費やすことになってしまう．

この長期化問題の解決策の 1 つとして，学生が所属する大学・学部やポータルサイトの運営企業が，早期の段階から適切な就職活動支援などの対策を行うことが考えられている．そのため，各大学・学部の例年の標準的な就職活動進行のペースと照らし合わせて状況をモニタリングすることが肝要であり，そのための就職活動終了時期の分析モデルの構築が望まれている．

A.1.3 就職終了曲線のあてはめ

図 2 から図 4 に示したような時間に対する割合の推移を示した曲線は，信頼性工学の分野で故障割合の推移などでも描かれるものである．また，信頼性工学分野における，故障分布の統計的推定と故障予測には，しばしばワイブル分布が利用されている．

そこで，本研究でもまず，就職活動終了時期の予測のため，ワイブル分布の適用を考えてみる．前節の分析結果より，学生の就職活動終了時期の分布は，学生個々の行動履歴よりも基本属性データによって決まっているため，属性別にワイブル分布によって分布のあてはめが可能であれ

ば，前年度までのデータから推定したワイブル分布を用いて，当該年度の学生のおおよその就職活動終了時期を予測できると期待できる．

A.1.3.1 ワイブル分布

ワイブル分布は医学，薬学，工学などで破壊，寿命を取り扱う分野において非常に頻繁に用いられている確率分布である．たとえば，癌の手術をした後のヒトの寿命，装置の寿命，絶縁破壊電圧の分布などはワイブル分布に従うといわれている．ワイブル分布の累積分布関数，密度関数の一般形は以下の式で示される．

$$F(x|m,\eta,\gamma) = 1 - \exp\left\{-\left(\frac{x-\gamma}{\eta}\right)^m\right\} \quad \text{(A.1)}$$

$$f(x|m,\eta,\gamma) = \frac{m}{\eta}\left(\frac{x-\gamma}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{x-\gamma}{\eta}\right)^m\right\} \quad \text{(A.2)}$$

関数の各パラメータ η，m，γ はそれぞれ尺度，形状，位置パラメータと呼ばれている．通常，寿命分布を取り扱うときには $\gamma = 0$ であることが多く，これを 2 パラメータワイブル分布と呼ぶ．それに対し，一般形として式 (A.1) を 3 パラメータワイブル分布と呼ぶ．また，m により分布の形状が大きく変化するため，多様な場面の確率分布に比較的容易に適合させることができる．ワイブル分布が実務者に多く用いられているのは，このような実用的な理由による．

ワイブル分布のパラメータ推定法の研究は精力的に行われてきた [30]．パラメータ推定の方法については，確率紙を使ったグラフィカルな方法 [31]，モーメントを用いた方法 [32]，順序統計量を用いた方法 [33] などもあるが，一般的な方法は最尤推定法であり，本研究でもこれを採用する[*17]．

A.1.3.2 ワイブル分布による推定結果

ポータルサイト上での学生の就職活動は全員 12 月から開始される（2014 年 1 月 27 日現在）ため，位置パラメータ $\gamma = 0$ の 2 パラメータワイブル分布 [2] を用いる[*18]．学生の就職活動終了時期のデータを用いて，一般的に知られている最尤法 [34], [35] により m，η の 2 つのパラメータを推定した．**図 A·1** に推定したパラメータを使用したワイブル分布と実データの経験分布を示す．図 A·1 より，ワイブル分布を用いた場合，その形状は比較的類似しているものの，推定したモデルと経験分布には乖離があることが分かる．これはパラメータが経験分布の両端に過度に適合するように推定されたため，中間において大きく経験分布との差が生じてしまったためであると考えられる．すなわち，就職活動終了時期曲線は，単一のワイブル分布では近似できないため，より表現能力の高い確率モデルを導入する必要がある．

[*17] ワイブル分布の最尤推定については付録 A.3 参照のこと．

[*18] 本稿で対象としている 2013，2014 年卒の学生の就職活動は大学 3 年次の 12 月 1 日から開始されている．すなわち，ワイブル分布の開始位置は大学 3 年次の 12 月を起点とし，終了日も大学 4 年次の 3 月となる．このため，特に位置母数を導入しなくても精度の良い推定が可能となる．一方で，このような開始，終了位置が等しいデータに対して位置母数の推定を行った場合，パラメータ数の増加により推定精度が低下してしまう可能性がある．以上の理由から，位置母数の導入を行わないものとした．

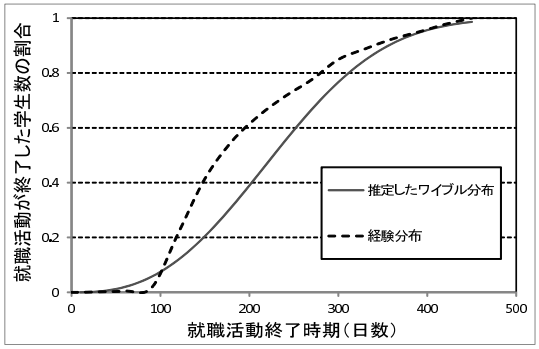

図 A·1 推定されたワイブル分布による就職活動終了時期曲線
Fig. A·1 The curve of finishing dates of job hunting calculated by the estimated single Weibull distribution.

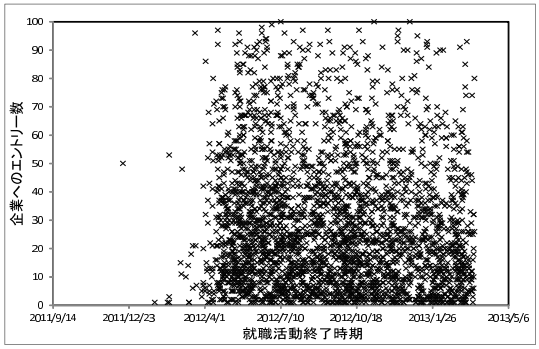

図 A·2 早期のエントリと就職活動終了時期の散布図
Fig. A·2 The scatter plot diagram between early entries and the finishing date of job hunting.

A.2 基本分析の結果

以下では，表 1 に示した，各分析項目のうち，就職活動終了時期に大きな影響がみられなかった項目について，その分析結果を示す．

A.2.1 早期エントリ数による分析の結果

早期の段階で多くの企業にエントリしている学生は，就職活動に熱心であり，就職活動終了時期が早いのではないかという仮説のもと，早期の段階でのエントリ数と就職活動終了時期との関係について分析を行った．本分析では，2012 年 3 月 1 日以前に行ったエントリを早期のエントリと定義した．また，100 件以上のエントリ数は異常値として排除した．**図 A·2** は，A 大学の学生の早期のエントリ数と就職活動終了時期の関係を散布図として表したものである．横軸は就職活動終了時期，縦軸は 2012 年 3 月 1 日までの学生の企業へのエントリ数を示している．図 A·2 より，早期のエントリ数と就職活動終了時期に大きな相関は

みられなかった．このことより，早期のエントリ数は終了時期に影響を与えないことが分かった．

A.2.2 人気企業へのエントリ率による分析の結果

一般的に人気企業へのエントリ割合が著しく高い学生は，大企業志向が強く，就職活動終了時期が遅くなる可能性があると考えられている．その関係性が正しいものかを把握するため，人気企業への早期のエントリ率と就職活動終了時期についての分析を行った．今回の分析の対象は IV の分析対象と同一である．ここで，学生からのエントリ数が10,000 以上の企業を人気企業とし，早期の段階のエントリは，分析 IV と同様に 2012 年 3 月 1 日以前のエントリと定義した．**図 A·3** に A 大学の学生の早期の人気企業へのエントリ率と就職活動終了時期の散布図を示す．横軸は就職活動が終了した時間，縦軸は 2012 年 3 月 1 日までの人気企業へのエントリ率を示している．図 A·3 から相関を見出すことができなかった．これにより，人気企業へのエントリ率は就職活動終了時期には関係がないと考えられ，人気企業に多くのエントリを行ったとしても就職活動が遅くなるとは限らないことが分かる．

A.3 ワイブル分布の最尤推定

最尤推定値は以下の手順で求められる．いま，N 個の観測データ $x_1, x_2, \cdots, x_N$ が得られたときの尤度関数 L を式 (A.3) で定義する．

$$L = \prod_{i=1}^{N} f(x_i|m, \eta, \gamma) \tag{A.3}$$

L が最大となる m，η，γ を探す．このような推定量を $\hat{m}$，$\hat{\eta}$，$\hat{\gamma}$ と表す，これを最尤推定量と呼ぶ．通常は L の最大値を探すことと，$\log L$ の極値を探すことは同値であると仮定して，次の対数尤度方程式

$$\frac{\partial \log L}{\partial \eta} = \sum_{i=1}^{N} \left\{ -\frac{m}{\eta}(1 - {z_i}^m) \right\} = 0 \tag{A.4}$$

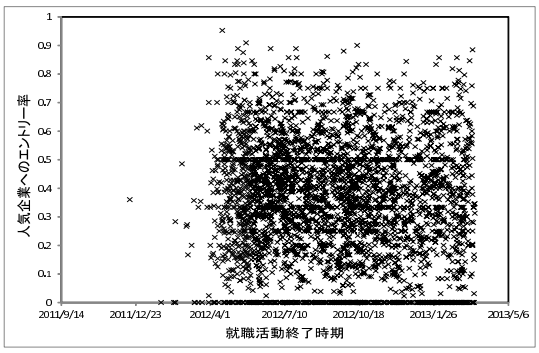

図 A·3 人気企業へのエントリ率と就職活動終了時期の散布図
Fig. A·3 The scatter plot diagram between the entry rates for popular companies and the finishing date of job hunting.

$$\frac{\partial \log L}{\partial m} = \sum_{i=1}^{N} \left\{ \frac{1}{m} + (1 - {z_i}^m) \log z_i \right\} = 0 \tag{A.5}$$

$$\frac{\partial \log L}{\partial \gamma} = \sum_{i=1}^{N} \frac{1}{\eta z_i} \{(1 - m) + m z_i^m\} = 0 \tag{A.6}$$

を解くことによって最尤推定量を求めることができる．ただし，

$$z_i = \frac{x_i - \gamma}{\eta} \tag{A.7}$$

であり，式 (A.4)–(A.6) は非線形の方程式なので，解析的に最尤推定量を式で解くことができず，ニュートン法 [36] などの数値的な繰返し手順が用いられる．

なお，本研究ではすでに述べているとおり，位置パラメータは 0 として推定を行う関係から $\gamma = 0$ としてその他のパラメータ推定を行う．すなわち，式 (A.3)，(A.7) における γ を 0 とし，更新式として式 (A.4)，(A.5) を用いる．

A.4 $K = 3$ 以外の混合ワイブル分布を推定した結果

本稿では，混合ワイブル分布の混合数は $K = 3$ として分析を進めたが，ここでは参考までに，K を変えた場合の全学習データに対する推定結果として，$K = 2$ の場合を**表 A·2** に，$K = 4$ の場合を**表 A·3** に示す．

より K を大きくし，たとえば，$K = 10$ とした場合の各要素ワイブル分布の平均は，小さい方から順に，158.04，177.81，255.72，269.20，269.42，289.78，291.24，296.59，319.17，441.51 となっており，平均値的にはかなり似た要素ワイブル分布が生成されてしまっている（たとえば，269.20 と 269.42 など）．K が大きい場合には，要素ワイブル分布の特徴を解釈することが難しくなるといえる．

表 A·2 全学習データから推定されたパラメータ（$K = 2$）
Table A·2 The estimated parameters learned by all data ($K = 2$).

	第一ワイブル	第二ワイブル
η	289.40	317.64
m	2.96	3.13
平均	258.26	284.19

表 A·3 全学習データから推定されたパラメータ（$K = 4$）
Table A·3 The estimated parameters learned by all data ($K = 4$).

	第一ワイブル	第二ワイブル	第三ワイブル	第四ワイブル
η	183.90	300.07	329.83	351.06
m	6.58	3.24	3.39	4.19
平均	171.47	268.94	296.27	319.04

早川 真央

1988 年生．2014 年早稲田大学大学院修士課程修了．在学時，機械学習に基づく就職ポータルサイトの履歴データの分析手法の研究に従事．2014 年より，株式会社ディー・エヌ・エー入社．ゲーム開発に従事．

三川 健太

1981 年生．2005 年武蔵工業大学環境情報学部環境情報学科卒業．2007 年同大学大学院修士課程修了．2013 年早稲田大学大学院博士後期課程修了．博士（工学）．2013 年早稲田大学助手．2016 年湘南工科大学講師．機械学習とその応用に関する研究に従事．IEEE，電子情報通信学会，日本経営工学会等各会員．

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 （正会員）

1969 年生．1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学大学院博士課程修了．博士（工学）．1997 年早稲田大学理工学部助手．2000 年東京大学大学院工学系研究科助手．2002 年武蔵工業大学環境情報学部助教授．2008 年早稲田大学創造理工学部経営システム工学科准教授．2011 年同大学教授．情報数理応用とデータサイエンスの研究に従事．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，経営情報学会等各会員．