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の構築

情報数理応用研究　　5212C037-9　早川真央
指導教員　後藤正幸

A Prediction Model of Finish Dates of Job Hunting Based on Stratification Tree and Mixture Weibull Distribution

HAYAKAWA Mao

1　はじめに

近年，日本では多くの学生が就職ポータルサイトを利用することで就職活動を行っている．しかし，就職活動を支援するサービスが充実するかたわら，その長期化が問題となっている．これは企業の多くが優秀な学生を採用したい一方で，学生の多くも有名な企業に就職したいため，人気が双方の一部に集中し，これら以外の企業や学生がマッチングするまでに時間がかかることがその一因としてあげられる．特に学生は就職期間中，学業に専念できないことに加え，精神的な負荷も高いことが社会問題化している．この問題の解決策の一つとして，就職活動が長期化しそうな学生に対し集中的に就職活動の支援を行うことなどが考えられる．そのためには，早期の段階で就職活動終了時期が遅くなる可能性の高い学生を発見する必要があり，就職活動終了時期の予測モデルの構築が望まれる．

モデル構築を行うために活用できるデータとして，就職活動に関する様々なデータが就職ポータルサイトには蓄積されている．これらのデータには，学生の基本的な情報や，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．これらのデータを活用し，適切な統計的予測モデルを構築できれば，学生の就職活動終了時期の予測が行えると考えられる．しかしながら，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，1) 就職活動終了時期に関する分析を行い，2) 対象問題の特性を踏まえた就職活動終了時期の予測モデルを構築することにより，3) 就職活動終了時期に関する知識発見を遂行するための方法論を提案することを目的とする．具体的には，まず実データに対し，就職活動の終了時期の従う確率モデルの検討を行い，信頼性工学などで扱われるワイブル分布 [1], [2] による推定を行う．就職活動終了時期の予測問題に対し，ワイブル分布を直接適用して予測を試みた結果，予測精度が高くないこと，並びに学生の属性によってワイブル分布の概形が大きく変化してしまうことを示す．これらの問題点を解決するため，まず学生の属性による層別木を構築し，さらに層別木 [3] の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を当てはめることによりその解決を行う．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の提案も行う．混合ワイブル分布を葉に割り当てた層別木モデルの有効性を実データによるシミュレーション実験により示す．

2　実データの基本分析

2.1　学生の基本情報による層別

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．調査には，2013 年度入社の学生を対象とした就職ポータルサイトのデータを使用した．ユーザ数は約 140,000 件であり，学生がポータルサイトに登録する際に入力する基本情報と，そこでの行動情報の 2 種類の情報が蓄積されている．これらのデータを用いて，就職活動終了時期との関係を調査した．調査項目を以下の表 1 にまとめる．

表 1:調査項目

調査番号	調査要素	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	エントリー数	行動情報
Ⅳ	人気企業へのエントリー率	行動情報

事前調査の一例として，調査Ⅱ（大学の偏差値）の調査内容の詳細と結果を示す．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有名な大学を 20 校選びそこ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 1 には調査結果の一例として，4 種類の学生の属性における，就職活動が終了した学生の割合とその終了時期の関係を表している．4 種類の学生とは，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，全ての（学部生も含む）学生である．A 大学は偏差値が 50 近辺の大学であり，B 大学は比較的偏差値が高いとされる大学である．

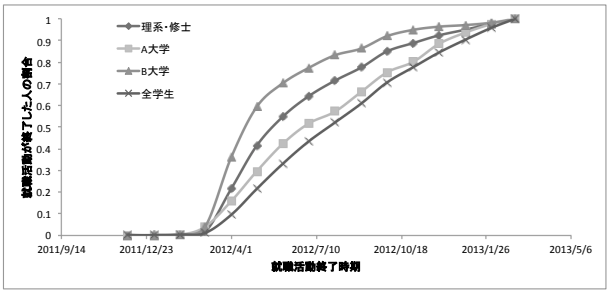

図 1:大学別の就職活動終了時期の差

図 1 の結果から，大学の偏差値における就職活動終了時期の差は顕著に現れることが分かる．

上記に加え，その他の層別分析の結果，調査番号Ⅰと Ⅱ は就職活動終了時期との相関が見られたが，調査番号

Ⅲ と Ⅳ に関しては，関係性を見出すことができなかった．これらの分析から，就職活動終了時期が学生の基本情報に依存していること，理／文系などの学生の属性の違いにより，就職活動終了時期が大きく異なるという点が明らかになった．

2.2　ワイブル分布による当てはまりの検証

前節の分析結果で得られた図 1 のような曲線は，信頼性工学の分野で故障割合の推移などで扱われるものと類似しており，予測にワイブル分布が使用されることが多い．そのため，本研究における就職活動終了時期の予測にもワイブル分布の応用を検討してみる．ワイブル分布は信頼性モデルとしてよく用いられ，寿命のモデルとして使用される．ワイブル分布の確率密度関数を以下の式 (1) に示す．

$$p(x|m,\eta) = \frac{m}{\eta}\left(\frac{x}{\eta}\right)^{m-1}\exp\left\{-\left(\frac{x}{\eta}\right)^{m}\right\} \qquad (1)$$

m は確率密度関数の形を変えるため，形状パラメータと呼ばれる．また，η は横軸の尺度を規定することから，尺度パラメータと呼ばれる．x には一般的に時間データが入力される．

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 [4]-[6] により m, η の 2 つのパラメータを推定した．以下の図 2 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．

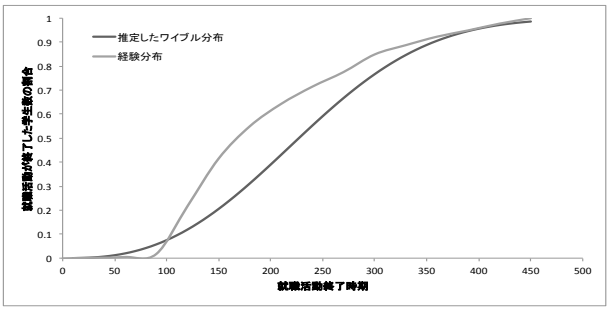

図 2:推定したワイブル分布

図 2 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．

3　就職活動終了時期予測モデルの提案

前節の基本分析により，学生の属性により分布の形状が大きく変化すること，直接ワイブル分布を用いた場合の予測精度が高くないことが明らかになった．実データに関するこれらの特徴を踏まえ，以下ではより当てはまりの良いモデルの構築を図る．このためまず，学生属性による層別木モデルの構築を行い，さらに混合ワイブル分布による予測モデルを導入する．

前者では，学生属性によって層別木を構築するためのアルゴリズムを提案する．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，葉ノード毎に経験分布に対して当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．

後者では，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予測精度の向上を目指す．その理由として，複数の行動パターンが想定される就職活動を単一の分布で表現しても，その特性を捉えきれないと考えられるためである．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提案する．

以上の議論を組み合わせることにより，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．図 3 に提案モデルの全体像を示す．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．

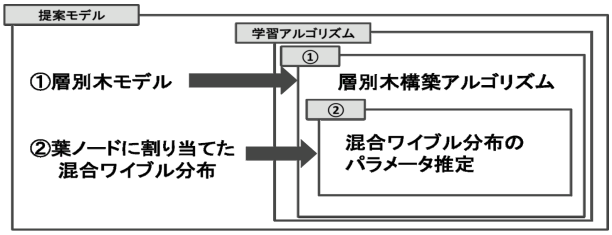

図 3:提案モデルの全体像

3.1　混合ワイブル分布の導入

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．

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|m_k,\eta_k) \qquad (2)$$

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 [7] などがある．しかし，この方法は近似手法であり，PC 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提案する．

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 (3) によって推定される．

$$\pi_k = \frac{N_k}{N} \qquad (3)$$

ただし，

$$N_k = \sum_{\alpha=1}^{N} w_\alpha^{(k)} \qquad (4)$$

$$w_\alpha^{(k)} = \frac{\pi_k p(x_\alpha|m_k,\eta_k)}{\sum_{l=1}^{K}\pi_l p(x_\alpha|m_l,\eta_l)} \qquad (5)$$

$$\sum_{k=1}^{K} w_\alpha^{(k)} = 1 \qquad (6)$$

とする．また，パラメータ m_k，η_k は以下の式 (7), (8) によって推定される．

$$m_k = b_k \qquad (7)$$

$$\eta_k = \left(\frac{1}{a_k}\right)^{\frac{1}{b_k}} \qquad (8)$$

ただし，

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

$$b_k = \frac{N_k}{\sum_{\alpha=1}^{N} w_\alpha^{(k)} \log(x_\alpha)^{b_k} \{a_k (x_\alpha)^{b_k} - 1\}} \tag{10}$$

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，以下の二重 EM アルゴリズムを提案する．

【混合ワイブル分布の二重 EM アルゴリズム】

Step1) $w_\alpha^{(k)}$ に初期値を与える．

Step2) 式 (4) により，N_k を計算する．

Step3) Step3-1〜Step3-4 で，m_k, η_k を計算する．

Step3-1) a_k, b_k に初期値を与える．

Step3-2) 式 (9),(10) により，a_k, b_k を更新する．

Step3-3) a_k, b_k が収束するまで，Step3-2 を繰り返す．

Step3-4) 式 (7),(8) により，m_k, η_k を計算する．

Step4) 式 (1) により，$p(x|m_k, \eta_k)$ を計算する．

Step5) 式 (3),(5) により，π_k, $w_\alpha^{(k)}$ を計算する．

Step6) $w_\alpha^{(k)}$ が収束するまで，Step2〜Step5 を繰り返す．

□

3.1.2 混合ワイブル分布による推定結果

3.1.1 節において提案した手法に基づき，約 300 校の大学における混合ワイブル分布のパラメータと混合比を計算した．以下の表に各大学のパラメータと混合比の一例を示す．表 2 には一例として，修士・理系の全体，A 大学，B 大学のパラメータ，表 3 には混合比を示す．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，一般的に終わるグループ，長引くグループの 3 つに分かれると仮定したため $K = 3$ とした．

表 2:理系・修士のパラメータの一例

大学名	η_1	m_1	η_2	m_2	η_3	m_3
理系・修士	138.93	9.35	198.56	8.10	362.73	5.22
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

表 3:理系・修士の混合比の一例

大学名	π_1	π_2	π_3
理系・修士	0.20	0.40	0.40
A 大学	0.20	0.32	0.48
B 大学	0.33	0.41	0.26

表 2 より，大学間における混合ワイブル分布でのパラメータの値にあまり大きな差異はみられない．一方表 3 から，各大学の混合比の値は比較的大きく異なっていることが分かる．この傾向は，100 人以上のユーザがいる約 300 の大学全てに対して見られた．このことから，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．

3.1.3 混合比の解釈

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．データの詳細を確認したところ，表 3 における π_1 が早期に就職活動が終了する学生のクラスの混合比，π_2 が一般的な時期に終了する学生，π_3 が後期に終わるクラスの混合比となることが分かった．このことから，B 大学は π_1 の値が比較的に大きいため，就職活動が早期に終了する学生が多い大学であると考えられる．一方，A 大学は，π_3 の値が大きく，就職活動が長引いてしまう傾向のある学生の比率が多い大学といえる．このように，混合比によって，各大学の就職活動のパターンが決定されることが説明できる．このため本手法の活用方法の一つとして，就職活動終了時期という観点からみた，大学のクラスタリングが混合比を使うことによって可能となる．

3.2 層別木の構築アルゴリズム

3.2.1 前提

本研究では，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 節で示した通り，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変化がないが，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．一般的な層別木モデルでは，データのまとまり具合によって層別に用いるための変数を決める．一方，提案手法では，各クラスにおける分布のグラフが大きく変化するように層別を行う必要がある．前述の通り，グラフ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．

ノードを層別する際にまだ層別に用いていない S 個の層別因子を $\boldsymbol{u}$=$(u_1, u_2, \cdots, u_s, \cdots, u_S)$ とする．その層別因子の持つ M 個の水準を $\boldsymbol{u_s}$= $(u_1^s, u_2^s, \cdots, u_m^s, \cdots, u_M^s)$ と表す．層別する前のノードのデータ数を N_D とし，因子 u_s で層別した後の水準 u_m^s で割り当てたノードのデータ数を N_m^s 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 $\pi_k^{s_m}$ と表現する．因子 u_S により層別された後のノードの情報量 I_s を式 (11) で計算する．

$$I_s = -\sum_{m=1}^{M} \frac{N_m^s}{N_D} \sum_{k=1}^{K} {\pi_k}^{s_m} \log_2 {\pi_k}^{s_m} \tag{11}$$

層別を行う前の親ノードに割り当てた混合ワイブル分布の混合比を π_k^D と表しそのノードの持つ情報量 I^D は式 (12) で計算される．

$$I^D = -\sum_{k=1}^{K} \pi_k^D \log_2 \pi_k^D \tag{12}$$

以下のアルゴリズムにより，層別木の変数決定を行う．

【層別木の変数決定アルゴリズム】

Step1) 式 (12) により，I^D を計算する．

Step2) 式 (11) により，各層別因子の $I_s(s=1,...,S)$ を計算する．

Step3) I^D と各 $I_s(s=1,..,S)$ の差を計り，差が最大となる $\boldsymbol{u_s}$ を選択する．

Step4) Step3 で選択した層別因子 $\boldsymbol{u_s}$ により，ノードを層別する．□

4 実験

4.1 実験条件

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シミュレーション実験を行う．2012 年度版のポータルサイトの 153,535 人数分の学生データを学習データとして使用し，混合ワイブル分布の推定を行う．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) とする．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のことを示す．また 2013 年度版のポータルサイトのデータ，148,571 件をテストデータとする．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二乗誤差，ならびに，葉ノードでの，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．前者を用いることで，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．後者により，推定した分布とテストデータの類似度を計る．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．

4.2 実験結果・考察

図 4 に実験で得られた層別木，表 4 に実験の結果を示す．

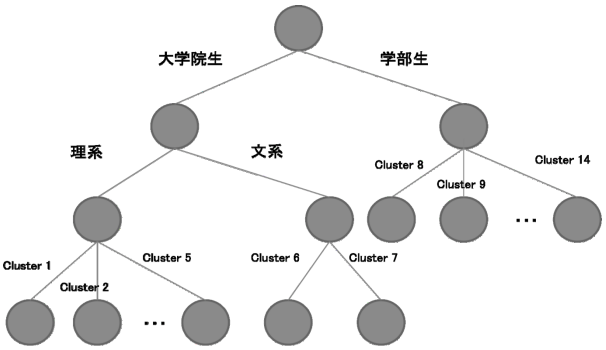

図 4:実験で得られた層別木

表 4:実験結果

評価指標	提案手法	比較手法
平均二乗誤差	25.07	53.12
KL 情報量	120.48	547.24

表 4 より，中央値の平均二乗誤差，KL 情報量共に提案手法がより小さい値を示している．中央値の平均二乗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．一例として，以下の図 5 に (理系，修士，Cluster1) の属性で推定した結果を示す．

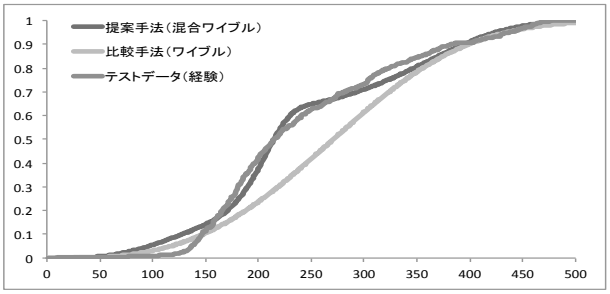

図 5:(理系，修士，Cluster1) の混合ワイブル分布

図 5 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．

5 提案の活用

本研究により，大学間における就職活動終了時期の差異は混合比の違いによって表現できることが分かった．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優秀な学生が存在していることが分かる．これらの人数の比率によって，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同士のクラスタリングが可能になると考えられる．また，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，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．大学毎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．

6 結論と今後の課題

本研究では，就職活動終了時期の予測に適した新しい学習・予測アルゴリズムとして，混合ワイブル分布を葉に割り当てた層別木モデルと層別木作成アルゴリズムを提案した．実験結果より，提案手法の有効性を示すことができた．

今後の課題としては，本手法に加えて，学生の行動情報も加味した予測モデルの構築が考えられる．本研究では，学生の基本情報のみを考慮しているため，基本情報が同一の学生には，同じ結果が出力される．しかし，実務上は行動情報を予測モデルに取り入れることが望ましく，これを今後の課題とする．

参考文献

[1] Weibull, W., “A stastical theory of strength of materials,” Ver. Ak. Handl., No.151, stockholm, 1939.

[2] Sekine, M, and Mao, Y. “Weibull Radar Clutter,” Peter Peregrinus Ltd., London, 1990 .

[3] Quinlan, J. R, “Induction of decision tree,” *Machine learning.*, Vol.1, pp.81-106, 1986.

[4] Cohen, A. C, “Maximum likelihood estimation in the Weibull distribution based on complete and censored samples,” *Technometrics*, Vol.7, pp.579-588, 1965.

[5] Jonson, N. L., Kotz, S. and Balakrishnan, N. “Continuous Univariate Distributions,” Vol.1, 2nd ed., Wiley, 1994.

[6] Lehmann, E. L., Casella, G., “Theory of Point Estination,” 2nd ed., *Springer*, 1998.

[7] 金甲洙，毛利正光，塚口博司, “ワイブル分布モデルに基づく道路交通騒音の予測,” 土木計画学研究発表会講演集, Vol.6, pp. 315-318, 1984.