特徴間の交互作用を考慮した学生ユーザの企業エントリー行動分析モデルに関する研究

1X14C064-0　杉崎 智哉
指導教員　後藤 正幸

1　研究背景・目的

企業の採用活動と学生の就職活動において，就職ポータルサイトの利用頻度が増えている．企業は，これらのサイトの個社ページ上で，説明会やインターンシップ，エントリーの情報を記載する一方，学生ユーザ (以下，ユーザ) は記載された情報をもとに就職ポータルサイトを通し，説明会やインターンシップへの参加，企業へのエントリーを行う．このように，就職ポータルサイトに蓄積されたユーザの行動履歴やエントリー履歴のデータを活用することで，企業はエントリー数増加の施策を打つことが可能となる．一方，就職ポータルサイト上のシステムに，ユーザが興味のある企業を登録し，メールを通じて登録企業の採用情報を受信可能な「気になるリスト」という機能がある．しかし，多くの企業は，ユーザのリストに登録されたにも関わらず，エントリーに結びついていないのが現状である．すなわち，企業にとっては被エントリーに対する機会損失が発生している可能性が示唆される．

そこで本研究では，エントリー数増加策を検討するため，ユーザの属性情報や行動情報，企業情報を入力変数とし，エントリーされるか否かを目的変数とした要因分析モデルの構築を検討する．その際，ある行動は，ある属性を有するユーザに対してのみ被エントリーへの影響を持つといった交互作用が考えられる．これらの交互作用を明らかにすることで，企業はユーザのターゲット層や行う施策を絞ることが可能となる．例えば，ユーザ属性情報である文系とユーザ行動情報である説明会へ参加の特徴間の交互作用が，被エントリーに大きく影響すると明らかになった場合，文系のユーザに絞って説明会を促進するといった施策を打つことができる．

一方，特徴間の交互作用を考慮可能な 2 値分類モデルとして Factorization Machines [1](以下，FM) が知られている．FM では，比較的少ないパラメータによって，交互作用を表現できるため，本研究で対象とする要因分析モデルとしても有用であると期待できる．そこで本研究では，ユーザの属性情報や行動情報，企業情報を入力変数とし，ユーザが企業へエントリーするか否かの 2 値分類に FM を適用し，得られた特徴間の交互作用を分析することで，エントリー数増加のための施策を検討する．以上の分析モデルの有効性を示すために，大手就職ポータルサイト (以下，サイト A) が保有する実データに FM を適用し，2 値分類が正しくできていることを確認する．さらに，得られた特徴間の関係から，エントリーに有効な施策の分析を行う．

2　Factorization Machines

FM はデータの特徴間の交互作用を考慮可能なモデルであり，高い予測精度を示すことが知られている．入力データの特徴量数を I としたとき，2 変数間の交互作用を全て考慮するためのパラメータ数は I^2 に比例するため，I が増加した際にはパラメータ推定に必要なデータ数が大幅に増加してしまう．そこで，FM では $I \times K$ ($K \ll I$) の交互作用行列と呼ばれる低次元の行列の各行の内積を取ることによって，特徴間の交互作用を表現している．

いま，説明変数ベクトル $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \cdots, x_{nI})^{\mathrm{T}}$, ($x_{ni} \in \{0,1\}$, $i = 1, 2, \cdots, I$) と目的変数 $y_n \in \mathbb{R}$ の N 個のペアを考え，バイアス項を w_0，重みベクトルを $\boldsymbol{w} = (w_1, w_2, \cdots, w_I)^{\mathrm{T}}$ とする．交互作用行列は K 次元ベクトル $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \cdots, v_{iK})^{\mathrm{T}}$ を要素とする $I \times K$ の行列 $V = [\boldsymbol{v}_1^{\mathrm{T}}, \boldsymbol{v}_2^{\mathrm{T}}, \cdots, \boldsymbol{v}_I^{\mathrm{T}}]^{\mathrm{T}}$ とかける．これらを用いると $\boldsymbol{x}_n$ に対する FM のモデルは式 (1) で表される．

$$f(\boldsymbol{x}_n) = w_0 + \sum_{i=1}^{I} w_i x_{ni} + \sum_{i=1}^{I} \sum_{j=i+1}^{I} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_{ni} x_{nj} \quad (1)$$

ただし，特徴間の交互作用を表す右辺第三項の内積計算は以下の式 (2) で表される．

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{k=1}^{K} v_{ik} \cdot v_{jk} \quad (2)$$

ここで，$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ の値が大きい i と j 番目の特徴間には強い交互作用があると解釈することができるため，FM はその大小を定量的かつリーズナブルな計算量で把握することができる．しかし，FM に関する研究の多くは予測モデルの性能にのみ言及し，副次的に得られるパラメータ推定値の活用法に対して着目していない．

3　Factorization Machines の 2 値分類への適用

3.1　概要

本研究では，あるユーザが「気になるリスト」へ登録した企業に対し，最終的にエントリーをするか否かの 2 値分類に FM を適用し，得られた分類器から有効な特徴間の関係を分析する．

式 (1) で表される FM は一般的な回帰モデル（以下，回帰 FM）で，出力は実数値となる．このため，出力の値域を (0,1) とし，2 値分類に適したモデルにするため，回帰 FM により得られた出力をロジスティック関数の入力とするモデルへと変更する．ここで，出力値が 0.5 以上の場合をエントリー，それ未満の場合を非エントリーとすると，FM によって得られた $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ が大きい場合，これらの i と j の組み合わせによりエントリーにつながりやすいという解釈を与えることができる．

3.2　定式化

回帰 FM を 2 値分類に対応させるため，ロジスティック関数を用いる．このとき，モデル式は以下の式 (3) で表される．

$$g(\boldsymbol{x}_n) = \frac{1}{1 + \exp\Big(-f(\boldsymbol{x}_n)\Big)} \quad (3)$$

式 (3) の値域は (0,1) となり，$g(\boldsymbol{x}_n)$ の出力はエントリー確率を表していると解釈することができる．予測の際には，この値が 0.5 以上の場合をエントリー，それ未満の場合を非エントリーとして扱う．

3.3　確率的降下法によるパラメータ推定

学習データ数を N 件としたとき，負の対数尤度関数を最小化するようにパラメータ更新することを考える．さらに，過学習を防ぐための正則化パラメータを $\lambda (\geq 0)$, 正則化項として l_2 ノルムを用いれば，目的関数は以下の式 (4) で表される．

$$\min \sum_{n=1}^{N} -\Big\{ y_n \log g(\boldsymbol{x}_n) + (1 - y_n) \log\Big(1 - g(\boldsymbol{x}_n)\Big) \Big\} + \frac{1}{2}\lambda ||\boldsymbol{w}||_2^2 + \frac{1}{2}\lambda \sum_{i=1}^{I} ||\boldsymbol{v}_i||_2^2 \quad (4)$$

本研究では，式 (4) を最小化する際に確率的降下法 (以下，SGD) を導入する．学習率を α としたとき，式 (4) の値が収束するまで全 N 件の学習データのうち 1 件をランダムに用い，以下の式 (5)–(7) によりパラメータ更新を行う．ただし，更新前のパラメータを w_0^{old}, w_i^{old}, v_{ik}^{old}，更新後のパラメータを w_0^{new}, w_i^{new}, v_{ik}^{new} とする．

$$w_0^{new} \leftarrow w_0^{old} - \alpha\Big(y_n - g(\boldsymbol{x}_n)\Big) \quad (5)$$

$$w_i^{new} \leftarrow w_i^{old} - \alpha\Big\{\Big(y_n - g(\boldsymbol{x}_n)\Big)x_{ni} + \lambda w_i\Big\} \quad (6)$$

$$v_{ik}^{new} \leftarrow v_{ik}^{old} - \alpha\Big[\Big(y_n - g(\boldsymbol{x}_n)\Big)x_{ni} \cdot \Big\{\sum_{j=1}^{I} v_{jk}x_{nj} - v_{ik}x_{ni}^2 + \lambda v_{ik}\Big\}\Big] \quad (7)$$

4 分析

4.1 分析概要

ロジスティック関数を用いて定式化した FM（以下，ロジスティック FM）が，エントリーに有効な施策を評価するモデルとして有効であることを示すため，サイト A の実データを用いて分析を行う．まず，対象問題に対するモデルの当てはまりを検証するため，一般的な 2 値分類の手法である線形ロジスティック回帰との予測結果を比較する．データの対象期間を 2015 年 6 月〜2017 年 3 月とし，外れ値の影響を防ぐために，その期間において気になるリストに登録した企業数が 6〜19 個のユーザを対象とした．入力データの特徴量をユーザ属性情報（文理区分，出身地，現住所，大学区分），企業情報（株式公開，従業員規模，企業種，本社所在地），ユーザ行動情報（説明会予約，インターンシップ説明会予約，インターンシップエントリー）の 3 種類とし，総特徴量数は 286 となった ($I = 286$)．また予備実験の結果，分解後の交互作用行列の次元数を $K = 13$，SGD の学習率を $\alpha = 0.001$，正則化パラメータ λ=0.00008 とした．

4.2 分析結果と考察

上述の実データを用いて，5-fold cross validation を行い，それぞれの手法におけるテストデータに対する正答率，再現率，適合率および F 値の結果を表 1 に示す．

表 1: 実験結果

	線形ロジスティック回帰	ロジスティック FM
正答率	0.664	0.687
再現率	0.685	0.660
適合率	0.683	0.710
F 値	0.684	0.683

表 1 より，線形ロジスティック回帰に対して，ロジスティック FM は正答率で約 2.2%上回り，F 値も同等な値を得た．この結果から，ロジスティック FM は線形ロジスティック回帰に比べ，高い精度を得られ，ユーザのエントリー予測により適したモデルであるといえる．

次に，エントリーに有効な施策の分析を行うため，ロジスティック FM により得られた重みベクトル $\boldsymbol{w}$ と交互作用項 $\boldsymbol{v}_i$ を用いて，特徴分析を行う．まず，w_i の値の上位 5 件と下位 5 件を表 2 に示す．なお，文理区分はユーザが文系であるか理系であるかを表し，株式公開は株式公開の有無を表している．

表 2: w_i の値

Rank	特徴量	w_i の値
1	文理区分 A	+1.182
2	株式公開 A	+1.085
3	文理区分 B	+0.715
4	本社所在地 A	+0.708
5	出身地 A	+0.551
⋮	⋮	⋮
282	従業員規模 A	-0.360
283	業種 A	-0.363
284	業種 B	-0.412
285	業種 C	-0.430
286	業種 D	-0.535

表 2 より，上位に出現している特徴量として本社所在地 A がある．これは，本社所在地 A の企業は,「気になるリスト」に登録された企業の中ではエントリーされやすいことを示している．そのため，本社所在地 A の企業は,「気になるリスト」に登録してもらえるような施策自体が重要である．一方，下位に出現している特徴量の例として業種 D がある．すなわち，業種 D の企業は「気になるリスト」に登録されても，他の業種の企業よりもそれがエントリーに結びついていない傾向がある．そのため，業種 D の企業は，エントリー数増加に向けて,「気になるリスト」に登録された後に，適切な施策を打つ必要があると考えられる．次に，特徴間の関係を得られた交互作用ベクトル $\boldsymbol{v}_i$ と $\boldsymbol{v}_j$ の内積をもとに特徴間の交互作用の定量化を行う．代表的な特徴量としてインターンシップ説明会予約に着目し，式 (2) の計算結果からこの特徴量との内積値の上位 5 件と下位 5 件を表 3 に示す．

表 3: インターンシップ説明会予約ベクトルとの内積値

Rank	特徴量	$\boldsymbol{v}_i$の内積値
1	現住所 A	+1.077
2	現住所 B	+0.930
3	業種 E	+0.887
4	業種 A	+0.807
5	業種 F	+0.805
⋮	⋮	⋮
282	業種 G	-0.910
283	業種 H	-0.920
284	業種 I	-1.040
285	業種 J	-1.126
286	業種 K	-1.467

表 3 より，値が上位の特徴量には現住所 A と現住所 B が出現していることがわかる．すなわち，現住所 A や B のユーザは，インターンシップ説明会に予約した企業に対してエントリーをしやすい傾向にあると解釈することができる．このことから，企業は現住所 A や B のユーザに対して，インターンシップ説明会への参加を促進することで，エントリー数の増加が見込める．逆に下位 5 件に含まれた業種はインターンシップ説明会に予約されるとエントリーされにくいことを示唆していると解釈することができる．

さらに注目すべき点として w_i の大きさでは下位 5 件に含まれていた業種 A が，インターンシップ説明会予約との交互作用の大きさでは上位に含まれていることが挙げられる．すなわち，業種 A は全体的にはエントリーされにくい企業カテゴリであるが，インターンシップ説明会に予約がされると，逆にエントリーされやすい企業になると解釈することができる．この結果から，業種 A の企業はユーザに対してインターンシップ説明会に予約してもらうように働きかけることで，エントリー数増加が見込まれると考えられる．

以上の結果より，ロジスティック FM を適用することでユーザがエントリーするか否かの予測と，ユーザのエントリー行動に対して有効な特徴間の関係の分析を同時に行うことが可能であることを示せた．

5 まとめと今後の課題

本研究では，就職ポータルサイトのデータにロジスティック関数を用いて定式化した FM を適用し，ユーザが「気になるリスト」へ登録した企業へ最終的にエントリーをするか否かの予測モデルを構築した．実データの分析を通じて，ユーザの属性情報または行動情報とエントリーしやすい企業間の特徴が抽出され，得られた結果はエントリー増加のための施策を立案するうえで有効であることが示された．今後の課題としては，行動情報の時系列を考慮したモデルの検討および分析などが考えられる．

参考文献

[1] S. Rendle,“Factorization Machines,” *Proc. 2010 IEEE International Conference on Data Mining,*2010.