膨大な情報の組合せから楽々学習

2次の多項式回帰では，要因の組合せを考慮することで線形回帰（1次の多項式回帰）よりもデータに適合したモデルを得ることが期待できますが，要因数が多くなると組合せ数が膨大になり実現が困難になります．本稿では，NTTコミュニケーション科学基礎研究所が開発した要因数が多いデータでも組合せを効率的に扱うことで解析を可能にし，初期値非依存な学習アルゴリズムにより解析結果を得るのを容易にするConvex Factorization Machines（CFM）を紹介します．

Mathieu Blondel（ふじの あきのり）/藤野 昭典
上田（うえだ なおのり） 修功
NTTコミュニケーション科学基礎研究所

回帰技術

近年，コンピュータや計測機器の性能向上や，インターネット，ソーシャルメディア，小型端末などの普及により，データの収集，蓄積が容易になり，集めたデータを科学研究やビジネスに活用することが期待されています．機械学習はデータの解析や活用を実現するための技術として注目を集めており，機械学習の一手法である回帰技術はデータ解析や未観測データの予測によく用いられます．

回帰技術を紹介するために，住宅価格の予測タスクを考えます（**図1**）．住宅価格は，都市の規模，駅からの距離，戸建て・集合住宅の種別，住居の広さ，築年数，車庫の有無など，複数の要因によって変動します．回帰技術を用いると，これらの要因と住宅価格の関係を表す式を，過去に売買契約が成立した事例から得ることができます．一般によく用いられる線形回帰技術（1次の多項式回帰技術）では，要因の変数値 $x=(x_1, \ldots, x_d)$ と住宅価格yとの間の関係式を，各要因に対応する重み $w=(w_1, \ldots, w_d)$ を用いて，$y=\sum_{j=1}^{d} w_j x_j = w^T x$ のようにモデル化し，過去に売買契約が成立した事例から重みの値を推定します．得られた重みの値を確認することで，住宅価格に大きく影響する要因を見出すことができます．また，線形回帰技術で得られた関係式を用いて，新規物件の契約価格を予測することができます．

しかし，線形回帰技術はモデルが単純なため，扱いやすい一方で得られる関係式の精度に限界があります．例えば，戸建てと集合住宅ではともに駅からの距離が遠いと住宅価格は下がりますが，集合住宅は戸建てと比べて下がり方が大きいとします．このような場合，駅からの距離と戸建て・集合住宅の種別のそれぞれの要因に対して重みを推定する線形回帰技術では精度が高い関係式を得られません．精度が高い関係式を得るためには，駅からの距離に対する重みを戸建て・集合住宅のそれぞれの場合で分ける，すなわち要因の組合せを関係式に導入する必要があります．このように，各要因に加えて要因の組合せも考慮して関係式を得る技術を2次の多項式回帰技術といいます．

2次の多項式回帰技術では，線形回帰技術よりも推定に用いるデータによく適合する関係式を得ることができま

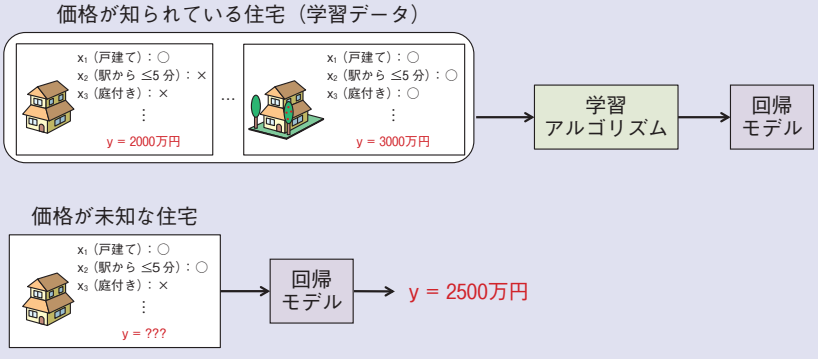

図1　回帰技術を住宅価格予測に適用した例

す．しかし，要因の組合せ数は要因数の2乗オーダーであるため，要因数が多いタスクでは組合せ数が膨大になります．例えば，植物の遺伝子情報と収穫量の関係を解析するタスク（ゲノミックセレクション）では，得られる遺伝子の情報量が多いほど組合せ数が膨大になるため，すべての組合せを考慮して関係式を推定するのが困難になります．この組合せ数の問題に対処するため，近年，Factorization Machines（FM）技術が提案されました[(1)]が，この技術では最適な関係式を得られる保証がなく，試行錯誤を繰り返して良い関係式を見つけ出す必要があるという問題がありました．そこで，NTTコミュニケーション科学基礎研究所では，組合せ数の問題と最適性の保証の問題の両方を解決する新技術Convex Factorization Machines（CFM）を開発しました[(2)]．

CFM

2次の多項式回帰技術では，要因の組合せに対する重みを要素とする行列Wと，各要因に対する重みを与えるベクトルw，要因の変数値を表すベクトルxを用いて，予測対象の変数値yと要因の関係式を二次回帰x^TWxと線形（一次）回帰w^Txの和$y=x^TWx+w^Tx$でモデル化します．行列Wとベクトルwは学習用データから推定するパラメータです．2次回帰の重み行列Wの要素数は要因数dが大きい場合に膨大（d×d）となるため，Wを直接推定するのは困難です．そこで，CFMと従来法であるFMでは，Wが低ランクな行列であると仮定して，推定すべきパラメータ数を大幅に削減します．FMでは，Wを低ランクなd×k（k<<d）の行列Pを用いて，WをPP^Tで置き換えて，Wの代わりにPの要素の値を学習用データから推定します（**図2**）．しかし，学習用データに対して最適なPの値を推定する問題は非凸最適化問題であり，Pの推定を始めるときに初期値をどの値にするかによって推定結果が異なる初期値依存の問題があります．このため，実タスクでFMを用いてデータ解析を行う場合に，Pの初期値を何度も変えて最良の関係式を探す必要があります．

CFM技術は，FMが持つ初期値依存の問題を解決することで，実タスクでのデータ解析を容易します．CFMでは，行と列の数が等しい正方行列は図2のように固有値分解で表現されることに着目し，ランクkの重み行列Wを与えるk個の固有値と固有ベクトルを学習用データから直接推定する学習アルゴリズムを開発しました．このアルゴリズムを用いれば，k個の固有値とd次元の固有ベクトルk個に相当する（1+d）×k個のパラメータ値のみを算出することで2次回帰を得ることができます（**図3**）．また，固有値，固有ベクトルの算出は初期値に依存しないため，重み行列Wの低ランク制約の強さを決めるハイパーパラメータ値を設定すれば，学習用データに対して唯一の解が得られます．このため，CFMを用いたデータ解析でパラメータの初期値をいろいろ変えて最良の関係式を探す必要がありません．

表に，CFMとFM，重み行列Wを直接推定する通常の2次回帰技術を，ゲノミックセレクションのタスク（植物の遺伝子情報から収穫量を予測する）に適用した実験結果を示します．表中の値は，学習用データを用いてそれぞれの技術で推定した関係式を用いて予測したテストデータと真値の相関係数（高いほうが良い）を表します．FMの結果は，初期値を何度か変えてパラメータ推定を行って得られた最良の関係式を用いた場合の結果を示します．実験により，CFMではFMより予測精度が同等以上の結果が得られることを確認しました．この結果は，初期値依存があるFMよりもCFMでは容易に最

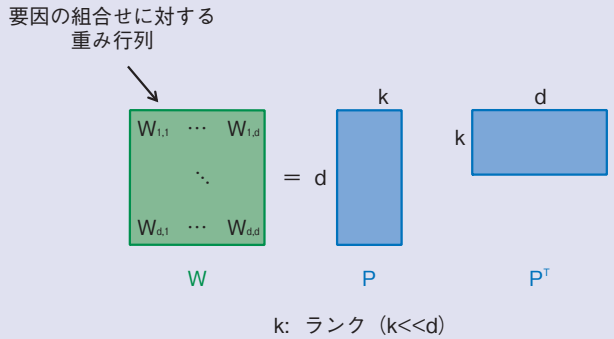

図2　従来のFMで得られる行列分解

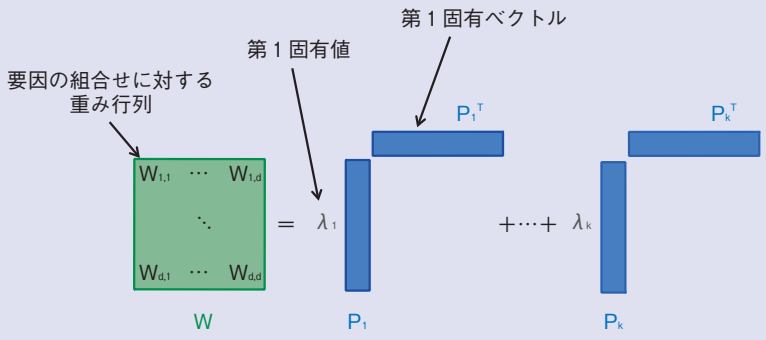

図３　提案法CFMで得られる固有値分解

表　ゲノミックセレクションのタスクに適用した実験結果

	２次多項式回帰	FM	CFM
小麦１	0.397	0.376	0.402
小麦２	0.471	0.501	0.526
稲	0.660	0.656	0.662

良の２次回帰を得られることを示しています．また，CFMでは通常の２次回帰技術と比べて予測誤差が小さい結果が得られました．一般に機械学習技術には，推定すべきパラメータの数が学習用データと比較して多い場合に学習用データにモデルが過剰に適合してしまい，他のデータに対する予測結果が悪くなる過学習の問題があります．CFMでは，重み行列Wを低ランクな行列であると仮定し，通常の２次回帰と比べて大幅にパラメータ数を削減することで過学習の問題を回避できたと考えられます．

要因の組合せに対する重み行列に対して低ランクな解を得るCFMには，学習用データに出現しない要因の組合せに対する重みを推定する特性があります．この特性を活かして，行列で表される関係データの欠損値を補完するタスクにCFMを応用できます．代表的な応用先としてeコマースのサイトでよく用いられる推薦システムがあります．多くの利用者の購買履歴を収集し，CFMを用いて 解析することで，利用者ごとに購入していない商品に対して購入する可能性の高さを予測できます．

今後の展開

２次の多項式回帰を用いて効率的なデータ解析を実現するCFM技術を紹介しました．２次の多項式回帰では要因の組合せを考慮することで線形回帰と比べて高い予測精度を実現する関係式を得られますが，３つ以上の要因の組合せを考慮することでさらに良い関係式を得られる可能性があります．しかし，m次の組合せ数は要因数のm乗オーダーとなるため次数が大きくなるほど実現が困難になります．そこで，高次な多項式回帰を実現するための手法を最近考案[3]するとともに，パラメータ推定を効率化するためにアルゴリズムの改良を検討しています．

■参考文献

(1) S. Rendle: “Factorization machines,” Proc. of ICDM 2010, pp.995-1000, Sydney, Australia, Dec. 2010.
(2) M. Blondel, A. Fujino, and N. Ueda: “Convex Factorization Machines,” Proc. of ECML PKDD 2015, Vol.9285, pp.19-35, Porto, Portugal, Sept. 2015.
(3) M. Blondel, M. Ishihata, A. Fujino, and N. Ueda: “Polynomial Networks and Factorization Machines: New Insights and Efficient Training Algorithms,” Proc. of ICML 2016, pp.850-856, New York, U.S.A., June 2016.

（左から）Mathieu Blondel/
藤野　昭典/ 上田　修功

多項式回帰は汎用的で重要な技術であると考えています．今後は，提案技術CFMを自然言語処理と医学データにも含めて，さまざまなデータに適用していきます．

◆問い合わせ先

NTTコミュニケーション科学基礎研究所
上田特別研究室
TEL　0774-93-5320
FAX　0774-93-5245
E-mail　mathieu.blondel@lab.ntt.co.jp