高次元データの回帰分析結果検証のための可視化手法

理学専攻 情報科学コース 鈴木千絵　(指導教員：伊藤貴之)

1　はじめに

回帰分析は，自然科学や社会科学に関する学術分野や産業分野で活用されている統計的手法の 1 つである．1 つ以上の説明変数と 1 つの目的変数の関係を数式化し，説明変数から目的変数を推測，予測する．回帰分析を用いた予測は健康状態予測などの医療問題，天災予測やエネルギー需要予測などの環境問題，経済予測や販売予測などの社会問題などが対象となり，その用途は非常に広い．

しかし，複数の説明変数を入力情報とする重回帰分析や，複数の回帰式を導入した混合モデルなどの導入により，その分析工程はより複雑になってきている．特に重回帰分析において，予測に大きく寄与する説明変数と大きく寄与しない説明変数，また予測値と実測値の誤差につながる説明変数を特定することが，回帰分析の性能を向上させるために重要である．

以下，小売店での商品販売を例題として議論する．日常的に販売される商品の売り上げは，その日の気温や曜日，また周辺でのイベント開催など，様々な要因に左右される．販売競争の激しい近年において，過剰発注による廃棄・処分を減らしたり，過剰在庫や完売を防いだりするため，適切な在庫数を保つことが必要不可欠である．そのために商品の販売数やその日の気象情報等の販売データを毎日入力・蓄積している企業は少なくない．取得したデータを解析することで将来の販売数をある程度予測することができるからである．そしてその解析方法の一つが，回帰分析である．

例として予測対象となる売り上げ (f) を目的変数とし，最高気温 (x_1) と最低気温 (x_2) と湿度 (x_3) を説明変数として線形回帰分析すると，

$$f = a \cdot x_1 + b \cdot x_2 + c \cdot x_3 + d$$

と表せるとする．この式に別日の気象情報を代入すればその日の売り上げを予測できる，というのが回帰分析の考え方である．

しかし取得するデータは膨大になってきており，それらの中には予測結果にほとんど影響しない要因や，逆に予測するために不可欠な要因が同時に存在している．予測結果に影響しない要因を予測に用いることが，逆にノイズを生むことになり，予測値と実測値との誤差の要因になることがある．予測のための入力情報と予測値に寄与度を理解することは重要であるが，情報の複雑化によってその理解が難しい場合も多い．本論文では各変数が予測値にどの程度影響を与えるのかを理解するため，回帰分析による予測値と実績値との誤差を可視化する一手法を提案する．

2　回帰分析結果の既存の可視化手法

回帰分析や予測問題の研究ツールとして可視化は有用であると考えられる．実際の問題とデータセットと回帰分析結果との関係を理解することが重要であるにも関わらず，それを目的として新しい可視化システムを開発した研究事例はまだ少ない．

代表的な例として Thomas ら [1] は，回帰モデルの双方向的な構築を支援するため，複雑さが最小限に抑えられる数学的モデルと相関の高い説明変数の組み合わせをより効果的に推薦し，精度の高い回帰分析につながることを表現する可視化ツールを提案した．Krause ら [2] は，回帰モデルを含む予測モデルのための対話型の特徴選択を目的として，グリフベースの可視化ツールを提案した．

それに対して本手法では，回帰分析対象の可視化に 3 次元散布図を採用している．さらに予測値と実測値の誤差をプロットの色に割り当てることで，誤差が大きくなる標本が集中する 3 次元空間中の位置を視認しやすくする．

3　前処理とユーザインタフェース

3.1　データ構造

本論文では販売個数や販売日の気温など，実数で表現される情報を実数値型説明変数，曜日や物品属性など，実数値で表されない情報をカテゴリ型説明変数とし，以下のデータ構造を想定する．

$$X = \{x_1, x_2, \ldots, x_n\}$$
$$x_i = \{v_{i1}, \ldots, v_{im}, c_{i1}, \ldots, c_{il}, p_i, a_i\}$$

ここで X は標本群，n は標本数，x_i は i 番目の標本を表す．また m は実数値型説明変数の個数，v_{ij} は i 番目の標本における j 番目の変数値，l はカテゴリ型説明変数の個数，c_{ij} は i 番目の標本における j 番目の変数値，p_i は i 番目の標本における予測値，a_i は i 番目の標本における実測値である．

3.2　3 次元散布図による可視化

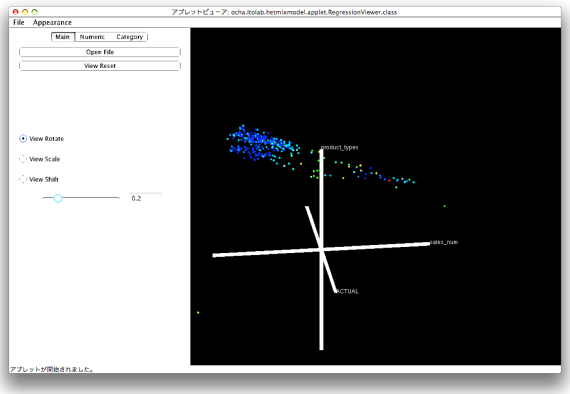

図 1: 提案する可視化ツール

本手法では前節で示したデータ構造の可視化に 3 次元散布図を採用している (図 1)．この可視化では，m 個の実数値型説明変数群の中から 2 個を選んで x 軸および y 軸に割り当て，実績値または予測値を z 軸に割り当てる．また，各標本における予測値と実測値の差の絶対値を色で表現しておる．差の絶対値が大きい標

本を赤に近い暖色系の色相で，誤差の小さい要素を青に近い寒色系の色相で描画する．

可視化ツールの画面左側には 4 つのタブがある．1 つ目のタブはファイル操作や描画調節をサポートする．2 つ目のタブには x,y,z 軸に割り当てる変数選択のためのラジオボタンを搭載する．また，後節で示す手法による評価値の高い変数がリストの上位に配置される．3 つ目および 4 つ目のタブではカテゴリ型説明変数の選択をサポートする．3 つ目のタブでユーザが変数を 1 個を選択すると，そのカテゴリ変数の選択肢となりえる変数値を選択するための 4 つ目のタブが表示される．4 つ目のタブに搭載された選択肢群のうち，チェックされているカテゴリ変数値をもつ標本は彩度の高い色で描画され，チェックされていないカテゴリ変数値を持つ標本は灰色で描画される．この機能により，誤差分布とカテゴリ変数値の関係を表現可能にしている．

3.3 説明変数の選択

説明変数が非常に多い場合，どの説明変数を x,y 軸に割り当てるかによって可視化の効果は大きく変わる．また回帰式を作った場合，次元数が多いほど訓練データへの適合率は高まる．しかしノイズなどの異常値にも影響されるため，過剰適合が生じる場合がある．変数を削減することで過剰適合を防ぐことができるが，過度な変数削減はそのデータへの適合率の低下を招く．そこで前処理として，各説明変数と変数組について予測値への寄与や誤差への要因を評価し，各説明変数または変数組の興味深さを定量的にユーザに提示することが有用である．本論文では赤池情報量基準 (Akaike's Information Criterion; AIC)[3] をもとに各実数値型説明変数を評価し，標本の分布をもとに各カテゴリ型変数を評価する．

AIC は訓練データへの当てはまりの悪さと複雑さを数値化したもので，以下の公式で表される．

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot k$$

上式において L は最大尤度，k は自由パラメータの数である．AIC 値が最小のものを選択することで，多くの場合，良質な予測を実現できるモデルを選択できることが知られている．

カテゴリ型説明変数は，実績値と予測値の誤差が大きい点の乱雑さをもとに評価する．標準偏差を基準として任意の数値以上の誤差がある要素を「誤差が大きい」として，x-y 平面におけるエントロピーが小さいカテゴリを高評価値とする．

これらをもとにステップワイズ回帰変数減少法を適用し，変数または変数組を順位づける．

4 販売情報の回帰分析結果への適用

販売情報の回帰分析結果について本手法を適用した．入力データには，実測値と予測値を含む 344 サンプル，12 個の説明変数，および 8 個のカテゴリ変数が含まれている．回帰分析には異種混合モデル [4] [5] が適用されている．説明変数の名前は機密情報であるため，本論文では各実数値型説明変数を A から L のアルファベットで示す．カテゴリ変数には，取得した月，日，曜日，および式番号が含まれている．入力データに対して AIC 値を求め，AIC 値が小さくなる 2 つの実数値型説明変数（説明変数 A,B）を x 軸および y 軸に割り当て (図 2)，誤差の大きい点群のエントロピーが小さくなるカテゴリを選択して可視化した (図 3)．

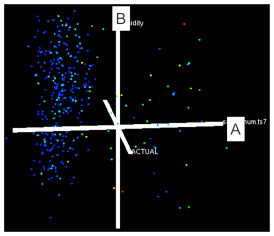

図 2: 実数値型説明変数の選択後の可視化結果

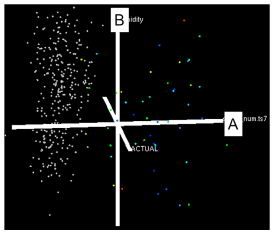

図 3: カテゴリ型説明変数の選択後の可視化結果

続いて曜日カテゴリを選択した．ここでは月曜日のみを選択した．図 3 より，A と B が大きいときに暖色点が多いことがわかる．原因として，月曜日は祝日の代休となることが多く，平日と祝日を同様に扱っていることが考えられる．その日が祝日か否かによって使用する回帰式を使い分けることで，精度向上に繋がると考えられる．

5 まとめ

本研究では説明変数と誤差の関係を 3 次元散布図を用いて可視化することで，誤差が大きくなる説明変数を特定できた．適用例では，実数値型説明変数 A,B が大きいときに誤差が大きく，さらにカテゴリ型説明変数の選択により誤差の詳細な要因の一つ（具体的には曜日）を特定した．今後は可視化結果を最も効果的な方向から表示するための評価基準の検討と，その自動化を目指す．

参考文献

[1] T. Muhlbacher, H. Piringer, A Partition-Based Framework for Building and Validating Regression Models, IEEE Transactions on Visualization and Computer Graphics, 19(12), 1962-1971, 2013.

[2] J, Krause, A. Peter, E. Bertini, INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data, IEEE Transactions on Visualization and Computer Graphics, 20(12), 1614-1623, 2014.

[3] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, Second International Symposium on Information Theory, B. N. Petrov & B. F. Csaki (Eds.), 267-281, 1973.

[4] R. Fujimaki, S. Morinaga, Factorized Asymptotic Bayesian Inference for Mixture Modeling, International Conference on Artificial Intelligence and Statistics (AISTATS), 400-408, 2012.

[5] R. Eto, R. Fujimaki, S. Morinaga, H. Tamano, Fully-Automatic Bayesian Piecewise Sparse Linear Models, International Conference on Artificial Intelligence and Statistics (AISTATS), 238-246, 2014.