# AutoEncoder を導入した Decision Forests によるノイズ発見

EP12106 日比野眞也　　指導教授：藤吉弘亘，山下隆義

## 1. はじめに

ノイズを含んだデータを用いて分類や回帰を行うと，その精度は大きく低下する．若山等が提案した Regression Forests は，分岐ノードにおいてノイズを参照した場合，複数パスを考慮することでノイズの影響を低減している [1]．しかし，この手法はノイズが既知の場合のみ有効である．実問題ではノイズが既知であるデータは少ないため，ノイズを発見する必要がある．そこで本研究では，Regression Forests のトラバーサル時にサンプルが到達したノードパターンからノイズの発見を行い，ノイズの影響を低減した回帰の実現を目的とする．

## 2. 複数パスを考慮した Regression Forests

図 1 に示すように，分岐ノードでサンプルに含まれるノイズを参照するとしきい値により分岐が反転し回帰精度が低下することがある．文献 [1] では，サンプルに含まれるノイズが既知の場合，複数パスを考慮した Regression Forests により，末端ノードまでにノイズを特徴次元を参照した回数により出力される値に重み付けして回帰を行う．これにより，ノイズの影響を低減することが可能である．

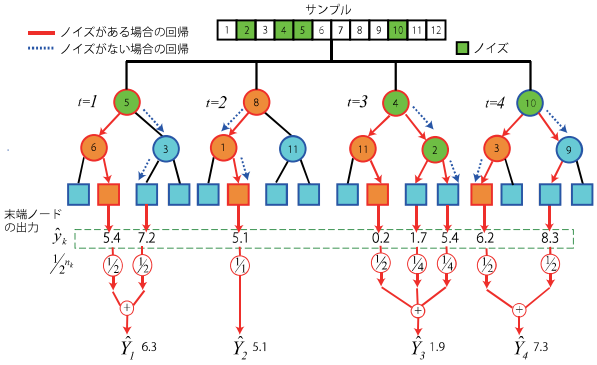


図 1：複数パスを考慮した Regression Forests

## 3. 提案手法

提案手法は，決定木間の相関を利用して，トラバーサル結果を AutoEncoder により想起してノイズの発見を行う．

### 3.1. Regression Forests と AutoEncoder の学習

Regression Forests はノイズが含まれないサンプルを入力して回帰木を学習する．学習した回帰木に未知サンプルを入力して回帰を行う．次に，回帰木にノイズが含まれていないサンプルを入力し，トラバーサルパターンを AutoEncoder を学習する．パターンは図 2 に示すようにサンプルが到達したノードを 1 とし，サンプルが到達していないノードは到達したノードからの距離から式 1 により算出する

$$D = t_u + t_d \tag{1}$$

ここで，$t_u$ は，子ノードから親ノードへ，$t_d$ は親ノードから子ノードへ移動した回数である．なお，パターンは各階層で作成する．

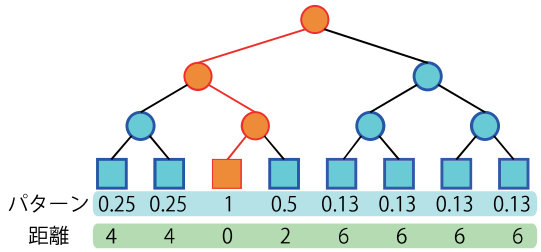


図 2：ノード間の距離を用いたパターン

### 3.2. ノイズ発見法

図 3 にノイズ発見の流れを示し，詳細を以下に述べる．

**Step1:トラバーサルパターン作成**

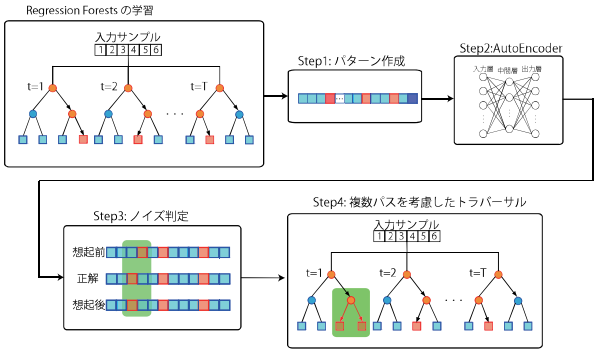


図 3：ノードパターンの想起

未知入力サンプルを Regression Forests に入力しトラバーサル結果からパターンを作成する．

**Step2:AutoEncoder によるパターンの想起**

作成されたトラバーサルパターンを AutoEncoder を用いて正解パターンに想起を行う．

**Step3:ノイズ判定**

各パターンについて，最大値をとる特徴次元の位置を比較し，想起前後で最大値をとる特徴次元の位置が変化していた場合に，ノイズと判定する．また，最大値に近い値が 2 つ以上存在した場合特徴次元の位置の変化にかかわらずノイズパターンと判定する．Regression Forsts の最下層からノイズ判定を行い，ノイズと判定されたらさらに浅い階層のノイズ判定を行う．順番に判定を行い，ノイズと判定されない層の特徴次元をノイズを参照したノードとする．

**Step4:複数パスを考慮した Regression Forests による回帰**

Step3 によりノイズを参照したノードと判定されるため，ノイズを参照したノードについて複数パスを考慮した分岐を行う．

## 4. 評価実験

評価実験では，AutoEncoder によりパターンを想起しノイズを参照したノードを発見し，発見結果より複数パスを考慮した Regression Forests を用いて回帰を行う．サンプルにノイズが含まれない場合とノイズ発見前と後での回帰精度の比較を行い提案手法の評価をする．本実験では，中部大学東キャンパスの電力使用量のデータを使用する．土日や長期休暇などのような大学の規則的な休日ではなく不定期な祝日などをノイズと定義する．Regression Forests の学習と評価には，特徴次元に祝日のデータを含まない平日のデータ 465 サンプルを用いる．Regression Forests のパラメータは木の数 20，木の深さ 4，特徴次元選択回数 7 とする．

評価実験の結果，ノイズ発見率は 75.2 %である．回帰精度の比較結果を表 1 に示す．回帰精度が 2.2 %向上した．ノイズパターンを正解パターンに想起することにより複数パスを考慮できノイズの影響を低減できたといえる．

表 1：回帰誤差

| | 回帰誤差 [kW] | 回帰誤差 [%] |
|---|---|---|
| Regression Forests | 111.9 | 5.0 |
| 複数パス | 103.9 | 4.7 |
| 提案手法 | 62.4 | 2.8 |

## 5. おわりに

ノイズが含まれるパターンから正常なパターンを想起しノイズを参照したノードを発見することにより回帰精度が向上した．今後は，更なる精度向上を目標とする．

## 参考文献

[1] 若山涼至, 藤吉弘亘,“複数パスを考慮した Regression Forests によるカメラのヨー角の推定”，パターン認識・メディア理解研究会, 2013.