4G1-5

# 決定木及び決定ネットワークによるデータ分類過程の説明文の自動生成

Automatic Generation for Sentences which Explain Classification Processes by Decision Tree and Decision Network

崎津　実穂[*1]　土屋　大樹[*1]　菅沼　雅徳[*1]　齊藤　航太[*1]　長尾　智晴[*1]
Miho Sakitsu　Daiki Tsuchiya　Masanori Suganuma　Kota Saito　Tomoharu Nagao

[*1]横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences, Yokohama National University

Automatic construction for image classification algorithms using machine learning has been studied. Although the classifiers constructed by such methods are very effective, these structures and classification processes are hard to understand for human and brought into black boxes. In order to solve this problem, we proposed Evolutionary Decision Network (EDEN) that extends decision tree which has readable structure among these classifiers and puts emphasis on easy to understand for human. In this report, we propose a method which generates sentences to explain classification processes by decision tree and EDEN. The proposed method converts the classification processes into explanatory sentences by using databases of word-feature pairs and word-threshold pairs. In experiments, we apply this method to several image classification problems. As a result, we found that the proposed method generates explanatory sentences which is easy to understand for human.

## 1. はじめに

近年，機械学習を用いて分類器を自動構築する研究が盛んに行われ，その有効性を示している．しかしながら，分類器の構造や使用される特徴量は分類精度が高くなるにつれて複雑になり，なぜそのような分類が行われたのかという分類の過程はブラックボックス化されてしまうことが多い．

分類器の中でも比較的可読性があり，人に理解しやすいといわれている if-then ルールを用いて分類を行う分類器として，決定木が挙げられる．決定木は，条件判断を行うノードが木構造状に配置された構造の分類器である．各ノードで入力特徴量を用いた分岐を繰り返し，最終的に到達した終端ノードに対応するクラスにデータを分類する．また，筆者らのグループでは，高精度かつコンパクトで人に理解し易い構造をもつ分類器を自動構築する手法である，進化的条件判断ネットワーク (Evolutionary Decision Network; EDEN)[*1][中山 13] を提案している．EDEN は決定木を拡張した分類手法であり，入力データに対して単純な条件判断を行うノードを，進化計算法を用いてネットワーク状に自動構築する．EDEN を用いることで，決定木に比べてノード数を大幅に削減した構造で分類を行うことができる．

これらの手法は if-then ルールを用いて分類を行うため，人にとっては分類過程が比較的わかりやすく，可読性があるといえる．しかしながら，我々情報工学の専門家からするとわかりやすく可読性のある構造であっても，情報工学の専門家でない一般の利用者からするとその構造はわかりやすいとはいえない．機械学習を用いて構築された分類器を実用化するにあたり，高い分類精度が求められる一方で，医療の現場などのようになぜそのように分類が行われたのか説明が求められる場合がある．しかし，作成した分類器に対して，分類プロセスをわかりやすく説明するといった研究はこれまでにほとんどなされていない．分類器が高精度なだけでなく，その分類過程を利用者に理解してもらうことで，安心して計算機による知能情報処理を利用してもらうことができると考えられる．そこで本稿では，構築された分類器の処理アルゴリズムに対する説明責任を果たすことを目的とし，比較的説明し易いと考えられる決定木および決定ネットワーク EDEN の 2 つの手法を対象に，データの分類過程を説明する文章を自動生成する手法を提案する[*2]．

連絡先: 崎津実穂，横浜国立大学大学院環境情報学府，〒 240-8501 神奈川県横浜市保土ヶ谷区常盤台 79-7，sakitsu-miho-jx@ynu.jp

*1　特許番号 : 5548990 号

## 2. 決定木および決定ネットワーク

### 2.1 決定木

決定木は，機械学習を用いて条件判断を行うノードを木構造状に配置し，分類を行う手法である．入力ノードから分類対象データの特徴量を入力し，中間ノードでの条件判断によって分岐先を決定する．これを終端ノードに到達するまで繰り返し，終端ノードに対応するクラスにデータを分類する．決定木を生成する代表的なアルゴリズムとして，ID3[Quinlan 86] や C4.5[Quinlan 96] が提案されている．ID3 は，空の木から始めて，各ノードにおいて最もデータ集合のエントロピーを削減する，すなわち最大の情報利得をもつ属性を選択し，元のデータ集合を部分集合に分割していく．これを繰り返し，徐々にノードを付け加えてデータ集合を分割させていくことで，最終的な木構造が得られる．

C4.5 は，ID3 のアルゴリズムを拡張した手法である．C4.5 は，連続値を扱うことができる点や，学習データの属性値が欠損している場合でも学習データに用いることができる点が ID3 と異なっている．また，ID3 のように情報利得を用いて分岐ノードに利用する属性を決定する場合，分岐数が多い属性が選択され易く，良好な結果が得られない場合がある．そのため C4.5 では，情報利得ではなく，情報利得を分割情報量（ある属性で分割を行った際のエントロピー）を用いて正規化した情報利得比を利用して，各ノードで分岐に利用する属性値を決定する．また，過学習を防ぐために木構造の枝刈りを行うことも特徴の 1 つである．

*2　特願 2015-041313

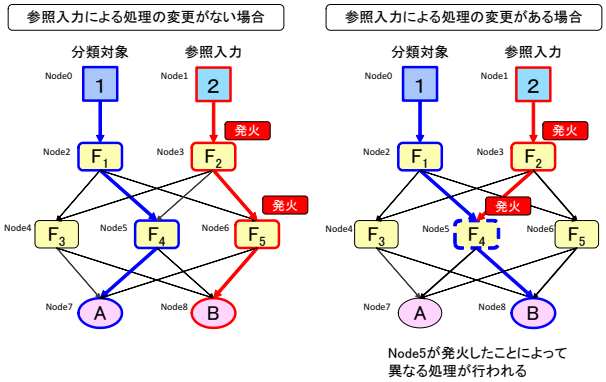


図 1: 参照入力による処理の変更例

### 2.2 EDEN

筆者らのグループでは，高精度かつコンパクトで可読性のある構造をもつ分類器を自動構築する手法である，EDEN[中山 13]を提案している．EDEN は決定木を拡張した手法であり，分類対象の入力データに対して単一の特徴量を用いて単純な条件判断を行うノードを，進化計算法によってフィードフォワード型のネットワーク状に自動構築する．各条件判断ノードは，分岐条件とする特徴量とその分岐の判断基準となるしきい値，および出力先で構成され，EDEN はこれらの組み合わせを最適化する．分類を行う場合は，決定木と同様に入力ノードから分類対象データの特徴量を入力し，条件判断ノードで大小判定を行い分岐先を決定する．最終的に到達した出力ノードは，分類対象データがどのクラスに属するかを表している．このため，EDEN のネットワークの分類過程も if-then ルールで表すことができ，可読性のある構造となっている．

また，EDEN では分類対象となるデータだけでなく，関連する参照データを入力することで，より複雑な処理を表現することができる．参照データの入力が分類対象データの処理に影響を及ぼさない例と影響を及ぼす例を図 1 に示す．まず，参照データを対応する入力ノードから入力し，条件分岐によって通過したノードを発火状態とする．発火状態となったノードでは，分類対象データが通過した際には発火していない状態と異なる処理が行われる．具体的には，分岐の判断をする際のしきい値を，発火前と発火後で異なる値を用いることで処理の変更を行う．図 1 では，参照データが通過した Node5 が発火したことで，Node5 での処理が変更され，分類対象データの分類結果が変化している．これによって，分類対象データだけでは分類が困難なデータに対しても，参照データを入力したことによる処理の変更を利用し，少ないノードでも適切な分類ができるようになることを期待している．

## 3. 提案手法

### 3.1 語句データベース

提案手法では，分類に使用する特徴量やしきい値に対応する語句のデータベースをあらかじめ用意しておく．具体的には，次のような例が考えられる．

- 色特徴量 HSV の V の平均値 :「明るさ」
- 色特徴量 HSV の H の分散 :「色数」
- 値 0.8 以上 :「とても多い」
- 値 0.2 以下 :「ほとんどない」

後に文章が続く場合は,「とても多くて」のように語尾を変化させることで対応させる．しかし，使用する特徴量によっては語句による説明が困難な場合がある．適切な表現方法がない場合は「特徴量 $F_x$」と表現しておき，この語句を説明文中で用いることとする．この場合は一般の利用者にとってのわかりやすさは低下してしまうため，極力わかりやすい語句との対応づけを行う必要がある．

また，EDEN のネットワークでは参照入力によってノードのしきい値が変更される場合がある．この場合は，参照入力の影響を付帯条件として，語句表現に反映させる．具体的には，参照入力が通り発火したノードに接続する，1 つ前のノードについて説明する語句を追加する．図 1 の参照入力による処理の変更がある場合の図では，発火した Node5 について説明する場合に Node3 についての説明を追加する．具体的には，参照入力として分類対象データの周囲の特徴量を利用する場合，付帯条件として追加する語句には次のような例が考えられる．

- 1 つ前のノードで V の平均値が値 0.9 で分岐を行った場合 :「周囲がとても明るくて」
- 1 つ前のノードで H の分散が値 0.1 で分岐を行った場合 :「周囲の色数がほとんどなくて」

### 3.2 説明文の生成

次に，データベースに登録されている特徴量を用いて，決定木と EDEN のネットワークを自動構築する．提案手法では，分類データを流した際に通過したパス上に存在するノードで使用されている特徴量および流したデータの特徴量の平均値を，データベースを参照して対応する語句に変換して説明文を自動生成する．また，分類器を利用するにあたり，できるだけ詳しく分類プロセスを説明してほしいと考える利用者から，おおまかにどのような分類を行っているのかを知りたいという利用者まで様々な利用者が存在すると考えられる．そこで，提案手法ではどの程度分類器について詳細に説明するのかを 2 つのパラメータで表す．

まず 1 つ目は，説明するパスの数 $N$ である．本稿では，分類器にデータを流した際に多くのデータが通るパスほど，分類クラスを表すのに重要なパスであると考える．このため，構築された決定木や EDEN のネットワークに対して，全分類対象データを流した際のノードおよびパスの通過頻度を計測する．その後，指定された必要なパスの数 $N$ だけ，データの通過確率が高いパスから順に説明文を生成する．

2 つ目は文章の長さ $n_{\max}$ である．決定木ではルートノードからエントロピーが最も削減される順にノードが配置されているため，ルートノードに近いノードの方がデータの分類において重要であると考えられる．EDEN においては，必ずしもルートノードに近いノードほど重要であるとは一概に言えない場合もあるが，データフローの観点からは同様に考えることができる．そこで，提案手法では説明文を生成する際に，ルートノードから指定された $n_{\max}$ 個のノードを利用する．全てのノードを利用する場合は，ノードの最大値 max とすれば入力ノードから出力ノードまで全てのノードを説明する文章を生成することができる．また，1 本のパスの中で同一の特徴量が複数使用された場合は，冗長な部分を削除した文章を生成する．

## 4. 説明文生成実験

### 4.1 実験概要

提案手法を用いて，一般画像と医用画像を対象とした 2 種類の 2 クラス分類問題について実験を行った．

決定木については，オープンソースソフトウェアである Weka（ver. 3.7.9）[*3]を用い，C4.5 アルゴリズムによって決定木を生成した．パラメータは，デフォルトである信頼度 0.25，リーフノードにおける最小データ数 2 を用いた．

EDEN については，各クラスの再現率の積と，ノード数を抑制する項の和を適応度 fitness として利用した．再現率 recall は，データの総数を $D$，正しく分類したデータ数を $D_{\text{correct}}$ とすると，式（1）で表される．

$$\text{recall} = \frac{D_{\text{correct}}}{D}, \tag{1}$$

各クラスの再現率の積を RC，構築されたネットワークの条件判断ノードの総数を $C$ とすると，fitness は式（2）で表される．

$$\text{fitness} = \alpha \times \text{RC} + \beta \times \frac{1}{C}, \tag{2}$$

ここで，$\alpha$ および $\beta$ は変更可能なパラメータであり，本稿の実験においては $\alpha = 0.9$，$\beta = 0.001$ とした．

### 4.2 一般画像分類

一般画像の 2 クラス分類として，フラミンゴと豹，ひまわりと蓮，バイクとグランドピアノを分類する 3 種類の実験を行い，学習データに対して説明文の生成を行った．画像セットは Caltech-101[*4]から，各クラスそれぞれ 50 枚の画像を用い，画像 1 枚単位での判定を行った．使用した画像の例を図 2 に示す．C4.5 については画像全体から特徴量を算出し，EDEN では画像の中心から画像サイズの 0.5 倍部分を判定対象，それ以外の周囲の部分を参照入力として利用した．EDEN の条件判断ノードとしては，入力特徴量としきい値による 2 分岐のノードだけを利用した．使用した特徴量 14 種類と語句の対応表を表 1 に示す．

また，C4.5 と EDEN による一般画像分類の精度と構築された構造の総ノード数を表 2 に示す．これらの学習結果を用いて，説明文の生成を行った．指定したパラメータは $N = 1$，$n_{\max} = \max$ である．生成された説明文の例を次に示す．

- EDEN によるフラミンゴと豹の分類過程の説明文
  - 赤色っぽい部分がある程度多くて，色数がある程度多いため，フラミンゴである．
  - 赤色っぽい部分が少なくて，周囲の色数が少なくて，黄色っぽい部分が多いため，豹である．
- EDEN によるひまわりと蓮の分類過程の説明文
  - 黄色っぽい部分が多くて，ある程度鮮やかで，色数が少ないため，ひまわりである．
  - 黄色っぽい部分が少なくて，鮮やかさの差がある程度あって，あまり鮮やかでないため，蓮である．

*3 http:// www.cs.waikato.ac.nz/ml/weka/
*4 http://www.vision.caltech.edu/Image_Datasets/Caltech101/

図 2: 使用した画像例

表 1: 一般画像分類実験で用いた特徴量と語句の対応表

| 特徴量 | 対応する語句 |
|---|---|
| H が 30° 未満または 330° 以上の画素の割合 | 赤色っぽい部分 |
| H が 30° 以上 90° 未満の画素の割合 | 黄色っぽい部分 |
| H が 90° 以上 150° 未満の画素の割合 | 緑色っぽい部分 |
| H が 150° 以上 210° 未満の画素の割合 | 水色っぽい部分 |
| H が 210° 以上 270° 未満の画素の割合 | 青色っぽい部分 |
| H が 270° 以上 330° 未満の画素の割合 | 紫色っぽい部分 |
| S の平均値 | 鮮やかさ |
| V の平均値 | 明るさ |
| H の標準偏差 | 色数 |
| S の標準偏差 | 鮮やかさの差 |
| V の標準偏差 | 明るさの差 |
| 水平方向の ±20° のエッジをもつ画素の割合 | 横線 |
| 垂直方向の ±20° のエッジをもつ画素の割合 | 縦線 |
| 上記以外の方向のエッジをもつ画素の割合 | 斜線 |

- C4.5 によるバイクとグランドピアノの分類過程の説明文
  - 明るくて，横線が少ないため，バイクである．
  - ある程度明るくて，縦線がとても少ないため，グランドピアノである．

提案手法を適用することで，分類器の構造や流したデータの分類パスをネットワーク上に表示するなどの方法よりも理解し易い表示ができているといえる．図 2 の画像と比較しても，特徴を良好にとらえることができており，構築された分類器による分類が妥当であることがわかる．しかし，バイクとグランドピアノの例のように，分類自体は行うことができているものの，人間の判断プロセスとは異なるような文章が出てくる場合がある．本稿では，分類プロセスを説明するための文章を生成するために 2 クラス分類問題に対して提案手法を適用している．計算機にとっては“2 つが異なるものである”という分類をすることができれば良いため，人間の判断とは異なるプロセスで分類を行っていると考えられ，バイクやグランドピアノの人間が考えるような特徴を抽出する文章にはなっていない．分類器を構築し，ある物体の特徴を抽出したいといった場合には，今回のような 2 クラス分類ではなく，多クラスの中から 1 クラスを抽出するような問題で分類器を構築することで，提案手法によって分類データの特徴を抽出するような文章を生成することも可能であると考えられる．

### 4.3 医用画像分類

医用画像として，カプセル内視鏡によって撮影された小腸画像を用いて異常と正常の 2 クラス分類の実験を行い，学習

表 2: C4.5 と EDEN を用いた一般画像分類の正解率と総ノード数

| | C4.5 | | EDEN | |
|---|---|---|---|---|
| | 分類精度 | ノード数 | 分類精度 | ノード数 |
| フラミンゴ・豹 | 95.0% | 7 | 100.0% | 8 |
| ひまわり・蓮 | 99.0% | 13 | 100.0% | 9 |
| バイク・グランドピアノ | 98.0% | 11 | 100.0% | 9 |

データに対して説明文の生成を行った．使用した小腸画像のサイズは $256 \times 256$ pixel であり，専門家によって $8 \times 8$ pixel のブロックごとに正常，異常のフラグが付与されている．この小腸画像を $32 \times 32$ に分割し，$8 \times 8$ pixel のブロック単位で異常と正常の分類を行った．C4.5 と EDEN の分類対象ノードから入力する特徴量は，分類対象ブロックを中心とした周囲 $16 \times 16$ pixel の領域から算出した．また，EDEN の参照入力ノードから入力する特徴量は，分類対象ブロックを中心とした周囲 $32 \times 32$ pixel の領域から特徴量を算出した．この際，中心の $16 \times 16$ pixel 部分の特徴量は使用しない．EDEN の条件判断ノードとしては，入力特徴量としきい値による 2 分岐のノードに加えて，周囲の特徴量および画像全体の特徴量との比較を用いた 2 分岐のノードを利用した．使用した特徴量 15 種類と語句の対応表を表 3 に示す．

また，C4.5 と EDEN による医用画像分類の精度と構築された構造の総ノード数を表 4 に示す．これらの学習結果を用いて，説明文の生成を行った．指定したパラメータは $N = 1, n_{\max} = \max$ である．生成された説明文の例を次に示す．

- C4.5 による小腸画像の分類過程の説明文
  - あまり鮮やかでなくて，最も明るい部分がとても明るいため，異常である．
  - 最も鮮やかでない部分がある程度鮮やかで，鮮やかさの差がほとんどなくて，最も明るい部分が明るいため，正常である．
- EDEN による小腸画像分類過程の説明文
  - 黄色っぽい部分の割合が画像全体より多くて，鮮やかさの差がほとんどなくて，最も明るい部分がとても明るくて，最も鮮やかでない部分があまり鮮やかでなくて，赤色っぽい部分がとても多いため，異常である．
  - 黄色っぽい部分の割合が画像全体より少なくて，周囲がある程度明るくて，周囲と対象部の最も鮮やかでない部分がある程度鮮やかで，最も明るい部分が明るいため，正常である．

こちらも一般画像分類と同様に，提案手法を適用することで説明文という新たな方法で結果の表示を行うことができている．特に，C4.5 で構築された決定木は，総ノード数が 485 の非常に複雑な構造になっており，一般の利用者にとって流したデータの分類パスを追跡することや，分類器の構造を見てなぜこのような分類が行われたのかを把握することは非常に難しいと考えられる．このように複雑な構造でどのように分類を行っているかが理解しにくい場合でも，提案手法を適用することで分類過程を従来よりもわかりやすく表示することができているといえる．しかし，まだ冗長な表現や自然でない表現が散見されるため，より自然な文章を作るために改善が必要であると考えられる．また，現在は 1 つのパスを 1 文で表現をしているが，長くなりすぎた場合は 2 文にするなどの改善も合わせて必要であると考えられる．

表 3: 医用画像分類実験で用いた特徴量と語句の対応表

| 特徴量 | 対応する語句 |
|---|---|
| H が 30° 未満または 330° 以上の画素の割合 | 赤色っぽい部分 |
| H が 30° 以上 90° 未満の画素の割合 | 黄色っぽい部分 |
| H が 90° 以上 150° 未満の画素の割合 | 緑色っぽい部分 |
| H が 150° 以上 210° 未満の画素の割合 | 水色っぽい部分 |
| H が 210° 以上 270° 未満の画素の割合 | 青色っぽい部分 |
| H が 270° 以上 330° 未満の画素の割合 | 紫色っぽい部分 |
| S の平均値 | 鮮やかさ |
| V の平均値 | 明るさ |
| H の標準偏差 | 色数 |
| S の標準偏差 | 鮮やかさの差 |
| V の標準偏差 | 明るさの差 |
| S の最大値 | 最も鮮やかな部分 |
| V の最大値 | 最も明るい部分 |
| S の最小値 | 最も鮮やかでない部分 |
| V の最小値 | 最も暗い部分 |

表 4: C4.5 と EDEN を用いた医用画像分類の正解率と総ノード数

| | C4.5 | EDEN |
|---|---|---|
| 正解率 | 95.6% | 92.2% |
| 総ノード数 | 485 | 14 |

## 5. まとめと今後の課題

本稿では，決定木および決定ネットワーク EDEN を用いてデータを分類する際の分類過程を説明する文章を自動で生成する手法を提案した．一般画像分類と医用画像の分類を行う決定木と EDEN のネットワークを構築し，これらに対して提案手法を適用して説明文を生成する実験を行った．結果として分類過程の説明文を生成することができ，今までブラックボックス化されていた分類過程をこれまでよりわかりやすく表示することができた．

今後の課題としては，より直感的に理解し易い文章を生成できるよう手法を改良することや，より複雑な特徴量を利用した際にどのような文章を生成するかといったことが挙げられる．また，客観的な評価によって手法の有効性を評価することも今後の課題である．

## 参考文献

[中山 13] 中山史朗, 穂積知佐, 矢田紀子, 長尾智晴: 進化的条件判断ネットワーク EDEN による画像分類, 映像情報メディア学会誌, Vol. 67, No. 7, pp. J278-J285 (2013).

[Quinlan 86] Quinlan, J. Ross: Induction of decision trees, Machine learning, Vol. 1, No.1, pp. 81-106 (1986).

[Quinlan 96] Quinlan, J. Ross: Improved use of continuous attributes in C4.5, Journal of artificial intelligence research, Vol. 4, pp. 77-90 (1996).