No. F-13

会津大学

理解可能な多変数決定木による データ分類と解析

～全体像を見れば隠されている規則がよくわかる～

教授　趙　強福

概要

○機械学習のモデルには、ニューラルネットのような非記号的モデルと決定木のような記号的モデルがある。前者は、データに含まれる情報を効率よく取り入れることができるが、学習結果はブラックボックスであり、人間も機械も理解できない。後者は理解しやすいモデルとされている。しかし、記号的モデルの学習結果は、機械的に形式的には解釈できるが、人間が理解できないものが多い。

○様々な応用において、コンピュータは補助的に使用され、人間が最終決断を行う。故に、学習結果を「人間に理解しやすく」する必要がある。多変数決定木技術は前記二つのモデルを融合したもので、一つのソリューションを提供する。本技術には以下の特徴がある。

- 類似度を基にした多変数決定木を利用しているので、人間にも理解可能なルールを学習結果として提供することができる。
- 忘却学習、注意学習、次元圧縮などいくつかの技術を採用しているので、コンパクトな多変数決定木を効率よく構築することができる。
- データ間の位相関係を階層的に可視化し、学習結果が直感的に理解できる。

実用化の可能性

○情報化社会において、データはモノであり、資産である。データの価値を高めるために、様々なデータマイニング技術が開発されている。その中で最も重要視されているのが機械学習に基づく技術である。

○本技術は、我々によって確立された新しい機械学習技術である。その有効性は様々な公開データベースで実証されている。本講座では、この技術を基に、文書解析システム、画像認識システムなどを開発している。商品化を視野に企業との共同研究も進めている。

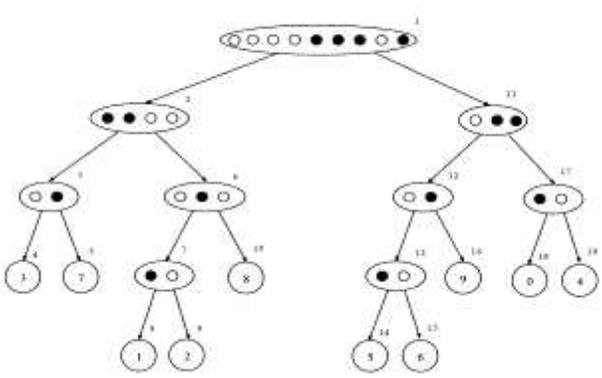
UBICからのメッセージ

○多変数決定木とは、中間ノードに多変数判別関数を利用した決定木の一種である。複雑なデータを効率よく分類できるのがこのモデルの特徴である。

○本技術に採用している多変数決定木はデータ間の類似度を基にしたものであり、学習結果を可視化して見ることもできれば、ルールに直して読むこともできる。

○本技術は、データを分類し、その結果を人間にわかる形で提供できるので、情報検索、セキュリティシステムなどの分野への応用が期待できる。

研究概要図



左図は手書き数字認識のための多変数決定木の例である。白丸は左子ノードに割り当てたデータの代表点、黒丸は右子ノードに割り当てたデータの代表点である。代表点はデータの典型的なパターンであり、データの内部表現である。これらの代表点との類似度を測ることによって、入力データの分類を行う。代表点を可視化すれば、典型的なデータパターンが一目でわかる。また、決定木をルールの形にも直せるので、興味のあるデータだけを分析することもできる。

生のデータを見えるように、読めるようにしましょう