

データマイニングアルゴリズム「アプリアリ」と「ID3」の比較

阿久津 忍

倉野 奈央子

1. はじめに

現在の情報技術の発達に伴い、大量のデータの中から有益な情報、知識を抽出する技術データマイニングが社会的に必要とされている。データマイニングではアプリアリ^[1]とID3^[2]という2つの主要なアルゴリズムが用いられることが多いが、両者で抽出される知識にどのような違いがあるか不明であった。

そこで本研究では、実際のアンケートデータに適用した出力結果を比較し、出現属性や論理関係の相関について確かめることにした。また、この手順を一部自動化した。

2. 実験評価

2.1 アンケートデータと実験方法

アンケートデータは、東海大学生 100 人を対象に、2006 年 5 月下旬から 7 月上旬にかけて実施したペーパーアンケートによる。内容は恋愛に関するもので、20 の質問項目を属性として用い、全て二択になっている。

2.2 出現属性の比較

ID3 とアプリアリによる出現属性の違いを確かめるため、同一のアンケートデータを 2 つのアルゴリズムに適用し、あるターゲット属性について、相関の高い属性を求める。ID3 の決定木でルートに近い位置に出現した属性、アプリアリの出力したルールに高頻度で出現した属性を重要属性と定義する。また、ID3 で得られたノードとアプリアリで得られた属性から両者の適合率を求めた。

2.1 のデータについて調べたところ、ID3 で高い位置に出現した属性はアプリアリでも出現回数が多かった。

3 つのターゲット属性についての ID3 の平均適合率は 0.75、アプリアリは 0.37 であった。ここで、ID3 の適合率は両者に出現した属性数/ID3 が出力した属性数、アプリアリの適合率は両者に出現した属性数/アプリアリが出力した属性数と各々定めている。

2.3 論理関係の比較

アプリアリと ID3 の論理関係を求めるため、20 件の属性の中から値の比率が均等な 4 つの属性を選び、それらを両アルゴリズムにかけた。アプリアリの出力結果を表 1 に、ID3 の出力結果を図 1 に示す。

比較すると、アプリアリで $X \wedge Y \Rightarrow Z$ と出力されたルール(表 1 のルール 1)が、ID3 の出力(図 1)で $X \wedge Y \Rightarrow \neg Z$ と出力されることはなかった。つまり、アプリアリに出力されたルールと ID3 の出力の葉に矛盾がないことが確認された。

表 1. 海外旅行経験有無の相関ルール (アプリアリ)

1	感情出す $\wedge$ 染まらない $\Rightarrow$ 海外ある
2	自分嫌い $\wedge$ 感情出さない $\Rightarrow$ 海外ない
3	染まる $\Rightarrow$ 海外ない
4	感情出す $\Rightarrow$ 海外ある

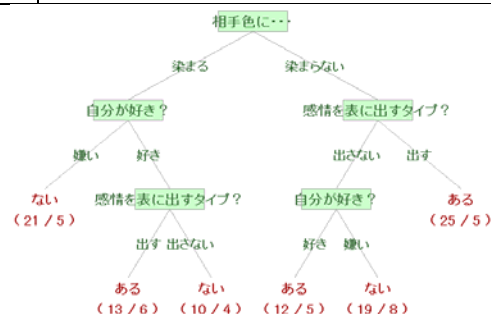


図 1. 海外旅行経験有無の決定木 (ID3)

2.4 考察

2.2 節より、ID3 で高い位置に出現した属性はアプリアリでも出現回数が多く、適合率が高いことから、両者の出力する属性には強い相関がある。

2.3 節より、アプリアリに出力されたルールと ID3 の出力の葉に矛盾がないことが分かった。ID3 の葉に一致するアプリアリのルールは少なかったのは、決定木から抽出される規則のサポートはアプリアリよりも低いためである。これは、アプリアリの相関ルールがそれぞれ独立なのに対し、決定木は全体で一つの矛盾のない木を構成しなければならないからだと考えられる。

3. おわりに

アプリアリと ID3 の比較・評価を行った。両者の出現属性には相関があること、両者の出力するルールに論理的な矛盾がないことが分かった。

参考文献

- [1] 福田, 他, 「データサイエンス・シリーズ(3)データマイニング」, pp. 22-23, 2001.
- [2] 松岡, 他, 「データマイニングによる授業評価アンケートの解析」, 日本知能情報ファジィ学会, 曖昧な気持ちに挑むワークショップ, pp. 10-10, 2003.
- [3] 本多, 「アプリアリアルゴリズムに基づいたデータマイニングツール“Gather”の開発」, 2005 年度東海大学 卒業論文, 2005.
- [4] 阿久津, 倉野, 「データマイニングアルゴリズム“アプリアリ”と“ID3”の比較」, 日本知能情報ファジィ学会, 曖昧な気持ちに挑むワークショップ, pp. 1-5, 2006.