2009 年度卒業研究概要

バイクの落札価格を推定するための決定木生成

大谷　紀子　研究室
0632228　吉田　憲司

1.　研究の背景と目的

オークションとは多くの参加者が商品に対して「この商品ならこのくらいだろう」と価格を付け、最高価格をつけた人が競り落とし、商品を入手することができる商品の販売手法の一つである。有名なものに yahoo オークションがある。バイクの中古車は中古車販売会社などが集まって中古バイク専門のオークションを開いている。中古バイクの値段は出品されたバイクの状態で決まる。価格決定に使用する情報は、車体の色や走行距離、フロントやリアの見た目の状態など項目が多岐にわたる。しかし、実際にすべての情報がバイクの価格に直結しているとは限らない。本研究では中古車の価格決定に用いられた情報を分析し、落札価格を推定するための決定木を生成することを目的とする。

2.　決定木

2.1.　決定木

決定木とは木構造で分類規則を表す方法である。主に、ID3 や C4.5 などのアルゴリズムを用いて生成される。葉が分類を表し、枝が分類に至るまでの特徴の集まりを表す。ID3 では情報利得を用いて分岐に適した属性を決定し決定木を生成する。情報利得とは、ある属性によって分岐することでデータのクラスの偏り具合が増加する量を表す。情報利得が大きいことで正確な決定木を生成できる。よい決定木を生成するためには、分類の精度を向上させることに加え、木のサイズをできるだけ小さくすることが重要である[1]。決定木のサイズを小さくすることによって理解しやすく、大量のデータ処理を高速化することができる。

2.2.　C4.5

C4.5 とは決定木を生成するためのアルゴリズムであり，J.R. Quinlan よって提案された[2]。C4.5 は ID3 を拡張し作られたものである。ID3 では情報利得を用いて分岐に適した属性を決定していたのに対して、C4.5 では情報利得比を用いて決定している。情報利得を用いると多数の属性値をとる属性を高く評価する傾向があるため，C4.5 では情報利得比を利用する． C4.5 ではデータが欠損している部分に対して「？」を入力することにより、欠損値を除いた分析ができる。また、離散値と連続値の両方を使うことができる。

3.　決定木による分析

3.1.　分析に用いるデータと Weka

本研究では、決定木を作成する際に Weka というフリーソフトを利用した。Weka とはニュージーランド Waikato 大学がメインで製作したデータマイニングソフトである。今回用意したオークションでのバイク情報の属性の一部を以下に列挙する。

- オークションの開催日

- オークションの開催場所
- 車種
- 車体番号
- 排気量
- 車体色
- メーカー
- 走行距離車検の有無とバイクの初年度登録
- 総合評価点数
- フロント、リア、外装、エンジン等の評価点数
- 落札価格

以上のほかにも計 21 項目のデータを集計した。

3.2. 分析結果

2000 件以上のデータを入力し決定木を生成したところ､データ数が過剰だったためノードの数が多くなり、決定木が煩雑になった。国内メーカーの車種に限定し、合計 1268 件のデータで生成するとノードの数が減り煩雑さは減少した。金額を 1 万単位から 10 万単位にまとめ直して生成するとさらに煩雑さが解消された。排気量ごとに生成したところ、250cc ではフロント評価点数の条件分岐がトップであり、50cc では車名車種が条件分岐のトップであった。データ数を制限し生成したところ、データが 20 個で生成したときは車体色で分岐し金額に直結した。データが 50 個のときは車体番号で分岐し金額に直結していた。

4. 考察

生成した決定木は多くが車体番号で分けられた。また、フロントの点数や車名で分岐することは多いが電装や車体の分岐は少ないことから、フロントの点数などは前オーナーの転倒歴などがわかるため金額に影響を与えていると考えられる。50cc のバイクでは車種は多いが、出品数が多く、同様の状態のものが多くあり、違いが車名車種のみであると考える。車体番号で分けることで決定木のサイズを小さくすることができているのも事実である。しかし、車体番号とはバイク一台一台で違うものなので本研究で扱うのは不適切だと考える。

バイクのオークション価格を推定するための決定木を生成するためには、排気量とフロントの評価点数が特に重要で、次いで走行距離や総合評価点数などである。項目に違いのない電装の項目や車体番号などは本研究においては不必要な項目であると考える。ノード数が多く、よい決定木としての条件 2 を満たせていない。ノード数を減らすために価格の種類を削減することで条件 2 を満たせると考える。条件 1 に関しては決定木生成の前段階でのノイズの軽減が不十分だったので精度が悪くなることがあった。ノイズ除去の方法を模索し精度が高く、決定木のサイズもコンパクトになるようにすることが課題である。

参考文献

[1] 寺邊正大,片井修,椹木哲夫,鷲尾隆,元田浩,“属性間相関ルールにもとづく決定木改良のためのデータ前処理手法の提案”人工知能学会誌,Vol.12, No.6,pp.1-2,1997.

[2] J.R. Quinlan,“C4.5: Programs for Machine Learning,”Morgan Kaufmann,1993.