2013 年度卒業研究概要

受講人数予測における決定木と ILP の比較

大谷紀子研究室
1032203　星　翔太

1. 研究の背景・目的

授業時間割を編成する際,各授業を適切な規模の教室で開講できるように教室の配当を決定する必要がある．本学部では基本的に前年と同規模の教室を割り当てているが，教室と受講人数が見合わずに教室を変更するケースが多い．学生支援センター職員は初回授業の際に授業が開講されている全教室に赴いて，教室変更の有無を確認し，変更の必要がある場合には，教室を手配するとともに，教室変更の情報を提示しなくてはならない．また,掲示情報の増加により,掲示板の利便性は低下する．2012 年には,帰納論理プログラミング（Inductive Logic Programming;ILP）を用いて受講人数を予測する手法が提案された[1]．ILP とは，観察された多くの事例から，論理プログラミングにより帰納推論を行って，目標概念を得るための機械学習の手法である．評価実験により，おおよそ 8 割の予測正解率が得られることが示されたが，実運用するために十分な正解率とはいえない．また，受講人数そのものではなく，受講人数の範囲を表す 5 つのクラスを設定して，各授業の受講人数を予測している．さらに，機械学習の手法としてより一般的で，ILP と比較して連続値の扱いに長けている決定木を用いた手法に関しては検討されていない．本研究では，授業の受講人数予測に関して，ILP と決定木により得られたルールや予測正解率を比較することで，両者の特徴や制度の相違などを明確にする．

2. 決定木

本研究で受講人数の予測に使用する決定木とは,分類や予測を行う際に広く使われている手法であり,機械学習における帰納推論の一手法である．属性と属性値の組からなる事例を入力として受け取り，事例の属するクラスを判定する．決定木を用いる利点として，直観的に理解しやすい，分析結果が明示的で ILP と比較して連続値の扱いに長けていることが挙げられる．そのため，授業の条件と受講人数の関係がより詳細に把握できるようなルールが得られると考えられる．本研究では，決定木生成アルゴリズムとして C4.5 を利用する．C4.5 は，ロス・キンランが開発した最も一般的な決定木生成アルゴリズムであり，連続値と離散値の双方の取り扱いが可能である．

3. 適用手法

予測正解率の高い決定木を生成するために枝刈りをする．枝刈りとは，訓練事例に基づいて生成した決定木から過学習の状態になった部分木を取り去ることである．決定木を構築する際の属性は，授業の開講時間，曜日，同時開講科目，必須・選択，担当学年，担当教員とする．クラスとしては，受講人数が 300 人以上であることを表す huge，200 人以上 300 人未満であることを表す large などの 5 クラス

表 1：受講人数の離散化

クラス	閾値(単位:人)
Huge	300 以上
Large	200 以上 300 未満
Medium	100 以上 200 未満
Small	50 以上 100 未満
Few	50 未満

を表 1 のように設定する．また先行研究と同様に，クラス指定や抽選などによって学生の意思に関係なく受講が決定される事例については，ノイズとなるため例外事例として扱い，あらかじめ学習の対象から除外している．さらに，ILP での扱いが難しい連続値も扱う．本研究ではフリーのデータマイニングツールである weka と独自に構築した C4.5 のシステムで決定木を生成する．後者では，枝刈りの手法などさまざまな条件を変えて結果を検証する．

4.評価実験

決定木と ILP により得られたルールや予測正解率を比較するために，評価実験を行った．平成 19～23 年度の 5 カ年分の時間割データのうち，4 年分の事例データを訓練事例，残りの 1 年分をテスト事例とするクロスバリデーションにより予測正解率を評価する．予測正解率とは，あるルールが未知事例を分類する際の正しさの程度を示すものである．提案手法の比較として，表 2 に WEKA と C4.5，先行研究の ILP の予測正解率の最高値と最低値，平均値をそれぞれ示す．

表 2：適用手法の比較

	weka	C4.5	ILP
最高値	75.0%	74.0%	89.4%
最低値	58.9%	70.6%	71.3%
平均値	66.1%	72.8%	78.4%

C4.5 については，枝刈りの手法を事前枝刈りまたは事後枝刈り，目的変数を離散値または連続値に設定したケースについて実行し，最も分類精度が高い場合の結果を示した．また，C4.5 で抽出したルールのうち高い予測正解率を得ることができたルールと予測正解率を表 3 に示す．

表 3：抽出したルール（一部）と予測正解率

ルール	予測正解率
教員 A が担当する科目は，クラス large に分類される	85.0%
教員 B が 3 限に担当する科目は，クラス medium に分類される	72.6%
5 限目開講する科目は，クラス few に分類される	66.7%

5.考察

評価実験においては，決定木の予測正解率は，ILP よりもやや劣る結果となった．決定木の生成過程では，各属性を選択する時点で局所的にみて最も効果的に分類できるような属性を選択する．生成された決定木の根ノードに担当教員という属性が配置されたことから，授業を開講する担当教員によって学生は受講する授業を決定するといえる．しかし，上記の方法で属性を選択した決定木は大域的にみると，必ずしも最適解とはいえない．また，決定木は ILP より連続値の扱いには長けてはいるが，今回のような目的変数が連続値の場合にはあまり適さないと考えられる．本研究でもおおよその事例においては正しく分類できたことから，受講人数規模の予測手法としてはある程度の有用性があることを示唆していると考える．ILP と比較したところ決定木による特徴的なルールを発見することもできた．また，枝刈りの手法や属性選択基準の見直しなどをすることにより予測正解率と実用性の両方の向上が期待されると考える．

参考文献

[1]加藤 由人，“授業時間割編成における教室配当に関する研究”，東京都市大学環境情報学部情報メディア学科卒業論文，2012.