2008 年度卒業研究概要

データマイニングの演習問題自動作成

大谷　紀子　研究室
0532194 森　健太郎

1. 研究の背景・目的

データマイニングとは，大量のデータから知識を抽出すること，または抽出する技術のことである．近年の記憶装置の低コスト化などで増加・多様化した情報を，最大限に活用することを目的としている．

データマイニング手法をより深く理解するには，実際に具体的なデータを使って知識を抽出する練習を行うのが有効である．しかし，具体的なデータを使う方法が高い学習効果を得るためには，いくつかの条件を満たしている必要がある．条件の1つとして，用意された具体的なデータが少量であることが挙げられる．初学者が大量のデータから知識を抽出しようとすると，作業量が膨大になり学習の妨げになるためである．用意された具体的なデータから，最終的に何らかの意味のある知識が得られることも必要である．実際のデータマイニングでは何も知識が得られない場合もあるが，何らかの知識が得られるほうが達成感を持つことができ，学習意欲が向上する．

しかし，高い学習効果を持つ条件を全て満たした具体的なデータを作成するのは大変難しい．扱うデータを少なくすると，データから導くことのできる知識の量も減り，有用な知識が抽出できなくなる可能性が高くなる．

本研究ではデータマイニング初学者の学習支援を目的とする．学習対象は「アプリオリアルゴリズム[1]」「C4.5[2]による決定木生成」の2つの手法とする．知識抽出の練習に用いる具体的なデータを演習問題という形で自動生成するシステムを構築する．最後に評価実験により本システムの有用性を示す．

2. システムの構成

表 1:演習問題の評価ポイント

アプリオリアルゴリズム	C4.5 による決定木生成
枝刈り	ラベル付け
再帰処理	全属性の使用
頻出アイテム集合の生成	分岐に適した属性決定時の迷いの有無
結論部要素数	ノード数
相関ルール個数	与えられる対数の値の数

本システムでは自動生成する演習問題の学習効果の度合いを計る基準として評価値を用いる.評価値は 0～1 の値をとり，1 に近づくほど学習効果が高いことを意味する．最終的な演習問題の評価値を計るために，最初に表 1 にある各評価ポイントにおける個別評価値を計算する．例えば，枝刈りの個別評価値は，n を枝刈り前のアイテム集合数，x を枝刈り後のアイテム集合数としたとき式(1)で求める．

$$\text{評価値} = 1 - \frac{|0.5n - x|}{0.5n} \tag{1}$$

式(1)では枝刈り後のアイテム集合数が枝刈り前のちょうど半分になったとき,個別評価値が最大値の 1 になり，1 つも枝刈りされなかったとき個別評価値が 0 になる．枝刈り量が 0 または少ないと枝

刈りに対する理解に繋がらないためである．最終的な演習問題の評価値は，各個別評価値を全て足し合わせたものとなる．演習問題は数十個同時に生成し，評価値が最も高い演習問題を出力する．

3．評価実験

「データマイニング」の講義を受講したことのある武蔵工業大学の学生 10 人を被験者として評価実験を行った．「アプリオリアルゴリズム」と「C4.5 による決定木生成」の 2 つの手法について，それぞれ本システムで評価値が低いとされた低評価値問題と，高いとされた高評価値問題の 2 種類を学生に解かせる．問題を解く順番によって結果に偏りが出ないように，低評価値問題を先に解く学生と高評価値問題を先に解く学生を 5 人ずつに分ける．問題を解いた後，学生に対してアンケートを実施した．アンケートは 1 問解くごとに行うアンケートと，2 問解いたうえで両者を比較するアンケートの 2 種類である．

4．結果・考察

「アプリオリアルゴリズム」のアンケートでは，問題を解く際に悩んだ箇所に違いが現れた．高評価値問題の場合は「枝刈り」「頻出アイテム集合が 3 つのときの処理」などアプリオリアルゴリズムの重要な処理についての回答が多かった．一方，低評価値問題の場合は「データ数が 11 個なので計算しづらい」「確率の計算」などの回答が多く，重要な処理以外の箇所で被験者が悩んでいる．重要な処理について考えさせるほうがアプリオリアルゴリズムの理解につながるため，高評価値問題のほうが学習効果は高いと考えられる．難易度については，高評価値問題は 6 人が「やや難しい」と答えたのに対し，低評価値問題は 6 人が「やや易しい」と答えている．簡単に問題が解けると学習にならないので，問題として適しているのは高評価値問題と考えられる．解きやすかったのはどちらの設問かという質問には「高評価値問題は量が多くて解くのが大変だったが，いろいろなパターンがあったので理解が深まったという点では高評価値問題のほうが解きやすかった．低評価値問題は量的に楽だったが，解いていてあっているかどうか不安だった」という回答があり，高評価値問題のほうがアプリオリアルゴリズムに対する理解が深まると考えられる．

「C4.5 による決定木生成」のアンケートでは設問を解く際に悩んだ箇所に，低評価値問題・高評価値問題のどちらの場合も「対数を用いた計算」をあげた被験者が多かった．他に悩んだ箇所がほとんどなかったため，どの被験者も対数を用いた計算で悩んだと考えられる．対数を用いた計算は「C4.5 による決定木生成」において，特に重要な処理ではない．したがって解答者の計算の負担を軽減する方法を考える必要がある．

「アプリオリアルゴリズム」「C4.5 による決定木生成」の両方とも，「枝刈り」「再帰処理」「ラベル付け」など重要な処理についての理解度を問う質問では低評価値問題と高評価値問題の間に大きな差が見られなかった．また，問題の正解率は「アプリオリアルゴリズム」「C4.5 による決定木生成」の低評価値問題と高評価値問題それぞれにおいて 5 割以下となっている．今後の課題として，各処理の理解を高めるための各評価ポイントの評価基準の改良や難易度の調整があげられる．

参考文献

[1] R.Agrawal,R.Srikant,“Fast algorithms for mining association rules,”Proc. 20th VLDB Conf., pp.487-499, 1994.

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