K-039

データマイニング手法を用いたマークシートテストの分析

橋本　一成†　　小柳　滋†
立命館大学大学院　理工学研究科†

1 はじめに

本稿では、教育という分野においてどのようにデータマイニングを活用していけるのかを考察する。学習とは授業や講義、教材により得た知識をテストなどによりいかに身についているかを確認し、理解が出来ていないところを復習することや新しい課題にステップアップすることを繰り返す。この流れにデータマイニングで得られる結果を取り入れることにより、問題（コンテンツ）作成者側には問題や教材の改善、評価基準、理解度に応じたクラス分け、などを行える。一方ユーザー側には、自分のペースに合ったより良い学習の方針を得ることが期待できる。

本研究では、データマイニングの異なる手法であるクラスタリングと相関ルールの２つの手法を用いて、それぞれの特徴の分析を行なった。異なる手法を比較することで、活用により役立てられるように幅を持たせる。また、データとして今回は実際に大学で行なわれた３種類のマークシートテストの結果を用いた。マークシートテストを用いたのは、解答が数値化されデータマイニングに用いやすい点と、e-Learningにも取り扱いやすい初歩の段階と思われるからである。

2 分析方法

本節では今回分析に用いた２つのマイニング手法、相関ルールとクラスタリング[1][2]について簡単に説明する。

2.1 相関ルール

今回相関ルールの発見には、Agrawalによって提案され広く普及しているアプリオリアルゴリズムを用いた。アプリオリアルゴリズムではユーザが確信度のしきい値として最小確信度、サポートのしきい値として最小サポートを与えることで、しきい値以上である相関ルールを重要な相関ルールとして、すべて発見することができる。今回は正答の問題番号をアイテムとし、個人の答案をトランザクションとする分析と、誤答の問題番号をアイテムとし、個人の答案をトランザクションとする分析の２通りを行った。データベースＤ中のすべてのトランザクションのうち、アイテム集合Ｘを含むトランザクションの割合をＸのサポートといい、しきい値 support(X)としてその値を与え、アイテム集合を抽出し分析を行う。

実際に出力される相関ルールとは

[1,5,32,33]⇒[41]　確信度９８％　サポート４５％

と表され 1，5，32，33 番の問題を正解した(間違った)ならば 41 番の問題も正解する(間違う)といったような結果となる。またここでいう確信度とは、相関ルールＸ⇒Ｙにおいて、support(X∪Y)／supportX からなる確からしさのことである。

An analysis of mark sheet tests by data mining
† Kazunari Hashimoto　† Shigeru Oyanagi
† Department of Computer Science, Graduate School of Science and Engineering, Ritsumeikan University

2.2 クラスタリング

クラスタ分析では互いに似ているレコードを見いだすようなモデルを構築する手法である。見いだされる類似するレコードが、同じように行動をするような似通った集まりであろうという期待にもとづいて探索するものであり、この類似するレコードの塊のことをクラスタと呼ぶ。クラスタリングを行うさい、データレコードの類似度が何らかの数値として計算できている必要がある。その指標として距離が用いられ、数値データに対して用いられるユークリッド距離などがある。今回の分析では、解答者に関するクラスタリングと、問題に関するクラスタリングの２通りを行った。ユークリッド距離は解答の正誤を１と０で表わして計算した。

またクラスタリングとして凝集法を使った。この手法では、各レコードがそれ自身のクラスタを形成することから始め、すべてのレコードが１つの大きなクラスタに集められるまで、しだいにクラスタを合併していくことが行われる。クラスタの合併の履歴により木構造が作られる。

分析のステップとしてまず、すべてのレコード間のユークリッド距離を算出しこれを関連度として“類似度行列”をつくる。次にその類似度行列の中で最も小さい値をみつける。これにより互いに最も類似する２つのクラスタを特定する。そしてそれらの２つのクラスタを合併し、対応する２つの行を合併した新しい行に置き換え、距離を算出し、類似度行列を更新する。出力される木構造より、有意義なクラスタを抽出して分析した。

3　取り扱ったマークシートテストデータ

今回取り扱ったマークシートテストのデータは実際に大学で行われたテストの結果である。以下の３つのテスト

① 計算機に関わる問題
② ①と同じ教科でテスト問題を変えたもの
③ ①、②と異なる情報全般の実力テスト

を取り扱った。

4　分析結果

4.1 相関ルールによる結果

①、②、③それぞれのテストにおいて正解、不正解の問題の相関ルールをアプリオリアルゴリズムにより分析を行った。このアプリオリアルゴリズムでは、まずどのような相関ルールが生成されるかをみるため、結論部の要素数が１であるルールだけを生成したので、確信度による枝刈りはせずサポート値による枝刈りのみで相関ルールを生成した。

テスト①は９つの大問から構成され、それぞれはいくつかの小問に分かれている。ここで５つのアイテムからなる相関ルールが見つかった。これはサポートが９０％以上のルールであることからほとんどの解答者が出来た問題である。この５つのアイテムそれぞれが９つの大問のうち５つ

の大問の最初の小問であった。問題を作る側からこの点をみると、問題を作成する上でほとんどの解答者が解ける問題を最初に配置し、順にレベルを変えると解答者のレベルをみることも出来、問題としていいものであるといえる。この面からもう少し深く分析をするなら、サポートを下げてアイテム数の多いルールをもっと検出し分析することによってさらに深く問題の評価に近づける。しかしそれにはルール数の爆発を防ぐため何らかの対策が必要である。

テスト②については問題作成者が①の問題よりも偶然の正解がなくなるよう解答選択肢を多くしたり問題構成も少し工夫された問題であった。この相関ルールとして生成されたものの中には、基本知識と応用知識を複合した大問題を解けたならば、基本知識の問題も解けるといったものがある。知識の複合による問題において解答者の理解度を発見できる。

テスト③においては絶対的に難しい問題群と簡単な問題群があり、正解、誤答のそれぞれの相関ルールとして抽出されたため有益な結果は得られなかった。

いくつかの発見はあったが相関ルールのアプリオリアルゴリズムそのままがマークシートテストの分析には満足できるものではなかった。ルールの絞込み、またその為の何が有益なルールであるかの定義付けなどが必要である。

4.2 クラスタリングによる結果

①のテストでは解答者側のクラスタリングを行うことにより図１のような木構造結果を得た。

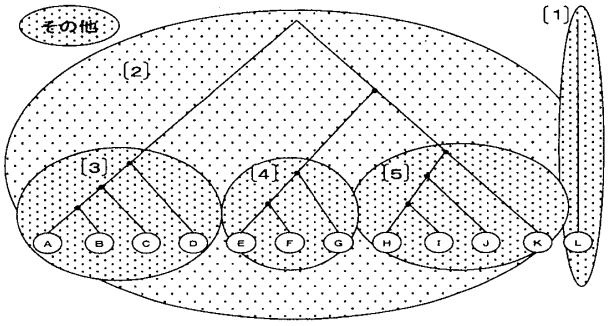

図１　テスト①解答者クラスタ形成図

それぞれ１０人以上の集まりであるクラスタがＡからＬまで１２個生成された。図１はそれぞれが合併していく様子である。この木構造にも階層によるクラスタの特徴があり、まず１２個のクラスタの段階で、それぞれ細かく解答の傾向が似た集まりとなっている。次に〔3〕、〔4〕、〔5〕の階層ではまたそれぞれのクラスタに含まれる要素の解答の傾向をもつグループとなる。そしてさらに要素規模の大きくなる〔その他〕、〔1〕、〔2〕の階層があり、例えばそれぞれの特徴として、クラスタ〔その他〕は安易なテストだが低得点者の集まりで、クラスタ〔１〕は得点が特にいいわけではないが悪くもないグループ、クラスタ〔2〕はクラスタ〔3〕、〔4〕、〔5〕の集まりであり要素がかなり大きく、高得点者たちの集まりである。このテストは全体的に簡単であったので、高得点者のクラスタが大きくなった。大きなクラスタになるほど関係は希薄になり特徴がなくなる。

またテスト②においてもさらに解答者の解答傾向によるクラスタ分けが結果として出た。

テスト③においては相関ルール同様有益な結果は得られなかった。やはり問題の質によるものと思われる。

テスト①や②においての結果からは、解答者への教え方への反映として、解答者のクラスタごとのグループへ、弱点や理解不足の内容を伝え復習の目安となる。例えば、一人の解答者にその解答者の属するクラスタの弱点の傾向の問題を復習させる。たとえそのテストにおいてこの解答者が正解していたとしても、同類の問題を復習させることにより以後のミスをなくさせる可能性を秘めている。

問題側からのクラスタリングとしては②のテストにおいてそれぞれ大問題ごとにクラスタを形成した。要素としてそれぞれの小問題を持つ。その大問題同士のクラスタが合併していく中で解答者が分かりづらい問題の傾向が分かる。

5　おわりに

本論文では実際に行われたマークシートテストの結果をデータマイニングにより分析し考察した。分析手法として相関ルールとクラスタリングを行い、それぞれともに特徴を見ることができた。

相関ルールは膨大なルールが生成されるため、その分析にかなりの労力がいる。しかし評価につながる糸口をみつけることができる。

クラスタリングでは解答者の関連性や、問題の関連性がわかる。そこからどういった理解の傾向を持っているかを個人から似通ったグループまでみてとれることができる。

この２つの手法は一般的なアルゴリズムのままでは不十分であり、今後はこのそれぞれの手法の目標にあった改良や組み合わせが必要である。相関ルールでは従来のサポート率の高いものが有益とされるアルゴリズムではなく、なんらかのルールの絞込みのアルゴリズムが必要となる。またその為には何が有益なルールであるかの定義付け、を考えていかなければならない。クラスタリングでは発見されたクラスタ内の特徴の明確な分析方法が必要である。

また今回はテストを複数用いたが、問題の形式などにより分析にも大きく差が出た。このことからマイニングに向いた問題形式の作成も必要である。

今後は問題作成者の意図を利用したマイニングツールの開発により、問題作成に有効なツールを目指す。

参考文献

[1] マイケル J.A ベリー、ゴードン・リノフ　著　SAS インスティチュート　ジャパン、江原　淳、佐藤　栄作　共訳:“ データマイニング手法” 、海文堂出版、1999.

[2] 福本　剛志、森本　康彦、徳山　豪:“ データマイニング” 、共立出版、2001.