平成２５年度　　修士論文概要					
主査	舟橋　健司	副査	岩田　彰	研究室	舟橋研究室
入学年度	平成２４年度	学籍番号	24417505	氏名	伊藤　圭佑
論文題目	データマイニングによる要注意学生の発見に関する研究 Study about Heuristics Anxious Student by Data Mining				

1　はじめに

近年、教育現場における ICT の発展により、学生に関する情報が記載された膨大な量のデータがサーバに蓄積されている。本研究では、これらのデータを有効活用するため、データマイニングを応用した修学指導を考案している。今までは過去の情報のみを頼りに修学指導を行っていたため、学生の成績や修学状況が悪化したときには既に手遅れとなっていた。しかし、データマイニングを活用し未来予測を行うことで、今後修学に問題を抱える学生を早期発見することができるため、より適切な指導が期待できる。しかし、学生に対する修学指導は、時間的コストが多大だという問題がある。例えば本学では教員 1 人に対し 15 人以上の学生が在籍しており、教員が 1 人 1 人の学生の特性や修学状況を把握し適切な指導を与えるのは困難である。そこで、本研究では「今後指導を与えるべき学生」を『要注意学生』と命名し、データマイニングを応用した分析により『要注意学生』の指摘を試み、さらに、ベイジアンネットワークによる『要注意学生』の発見を提案している。全学生を対象に指導を与えるのではなく『要注意学生』を指導対象者とすることで、指導の時間的コストを軽減することができる。

2　本研究で用いた手法

発見の手法として提案しているベイジアンネットワークは、確率変数、有向グラフ構造、条件付き確率で定義される確率モデルである [1]。また、本研究ではデータマイニングに分類される手法を多く活用している。分析の際に、主成分分析とクラスタリングの K-means 法、発見モデルの構築の際に、クラスタリングのウォード法と CFS を用いている。

3　本研究で用いたデータ

本研究では本学に在籍していた学生 338 名に関するデータを利用した。データの種類は、講義別成績データ、学生の教室への入退室の時間が記録されている打刻データ、卒業研究着手と卒業の時期が記録されている学生修学データの 3 種類である。これらをそのまま分析や予測に適用するのは困難なため、講義別成績データから、各学生の GPA（学期別と科目別）と獲得成績数を、打刻データから、月別の打刻回数を取得した。

4　修学状況の分析

先述したデータを用いて修学状況の分析を行い、『要注意学生』の指摘及びその傾向の調査を試みた。

4.1　修学傾向の分析

全学生 338 名の修学状況を、学生修学状況を用いて調査した。その結果、338 名中 70 名の学生が卒業までに留年または退学していることが判明した。また、GPA データを併用して、1 年前期と 1 年後期の GPA 値域別の留年・退学者数を調査したところ、1 年前期と 1 年後期の両方において、各 GPA が 1.0 未満の学生は 93.5 %という高い割合で留年・退学していたことが確認できた。

表 1: 1 年次の GPA に関する人数

1 年前期と後期の GPA	総数	留・退	割合
どちらかが 1.0 未満	31	29	93.5 %
両方とも 1.0 以上	307	41	13.4 %
合計	338	70	

4.2　データマイニングを応用した分析

修学傾向をより詳細に調査するため、データマイニングの技術を応用し更なる分析を行った。分析の手法は、まず主成分分析により情報を縮約し、その後にクラスタリングを実行しクラスタ別の傾向を読み取ることで全体の傾向を概観した。

科目別 GPA データを用いた分析により、「理系教科の強み」が今後の修学状況に大きく寄与していることが分かった。獲得成績数データを用いた分析により、獲得する成績が両極端な学生ほど今後の修学状況が悪化することが確認できた。月別打刻回数データを用いた分析により、打刻回数少ない学生ほど修学状況が悪化していることが確認できた。全体の結果として、様々な観点から学生を評価することの有用性を示すことができた。

5　『要注意学生』の発見

本研究では『要注意学生』を事前に発見することを試みており、その手法としてベイジアンネットワークを提案している。その有意性を示すため、GPA のみを基準とした発見法との比較を行った。

5.1　発見の時期と『要注意学生』の定義

早期に発見を行うため、1 年次終了時を発見の時期とし、2 年次以降に『要注意学生』になるかどうかを判定・予測した。また、発見対象の『要注意学生』を、分析の際に得た知見を参考に、「1 年前期と 1 年後期の GPA が共に 1.0 以上であるが、2 年次以降に留年・退学してしまう学生」と厳密に定義した。これは 4.1 の調査結果より、GPA1.0 未満の学生は指導対象となるのは自明だと分かったため、発見モデルの適用対象から除外した。対象人数は表 1 より、307 名となった。

5.2　理想の発見モデルとモデルの評価方法

理想の発見は「実際の『要注意学生』を全員見つけ出す発見」である。機械学習における評価指標として、正解率 Accuracy、適合率 Precision、再現率 Recall が存在するが、再現率 Recall は本研究における「実際の『要注意学生』のうち、どれぐらいの割合を発見できたか」に相当するため、この指標を最重要視した。また Precision と Recall の調和平均である F-measure もモデル評価の基準の 1 つとして採用した。

5.3　発見モデルの構築

本研究では、科目別 GPA と、全変数から CFS によって取捨選択された変数を確率変数として適用した。グラフ構造は Naive Bayes 構造と Free Network 構造のものを構築した。それぞれの設定毎のモデルを構築し、発見精度の比較を行った。

5.4　ベイジアンネットワークの出力

ベイジアンネットワークの出力は、ある事象の事後確率で表現される。例えば「『要注意学生』である確率は 38 %」といった形式で表される。『要注意学生』になるかどうかの事前確率は、$\frac{41}{307} = 13.4\%$ だと言えるため、『要注意学生』の判定に用いる閾値を、13.4 %（事前確率から確率値が増減したかどうか）、30 %、50 %と設定し、それぞれの発見精度を評価した。

5.5　発見モデルの評価

学生を 1 年次通年 GPA の低い順に並べ、GPA がある閾値を下回る学生を『要注意学生』と認定する手法と、本提案手法の発見精度を比較した。

表 2 に GPA を基準にした発見と、ベイジアンネットワークによる発見モデルの中で最も評価が高かったものの F-measure と Recall を示す。ベイジアンネットワークを用いた方が指標値が高くなることが確認できた。提案手法のモデルでは、『要注意学生』41 名中 28 名（68 %）発見することができた。さらに、図 1 が示す実際に『要注意学生』でありモデルによって発見された学生の 1 年次 GPA の分布を検討すると、1 年次の成績が中間層に属する『要注意学生』も発見できていることが確認できた。ゆえに、提案手法の方が精度が高く、柔軟な発見ができることが示された。

表 2: 各モデルの指標値

モデル	F 値	Recall	発見数
GPA による発見	0.404	54 %	22 名
BN による発見	0.424	68 %	28 名

※ベイジアンネットワークを BN と表記

表 3: 提案手法の発見モデルにおける各人数

	『要注意』	非『要注意』	合計
『要注意』予測	28 名	63 名	91 名
非『要注意』予測	13 名	203 名	216 名
合計	41 名	266 名	307 名

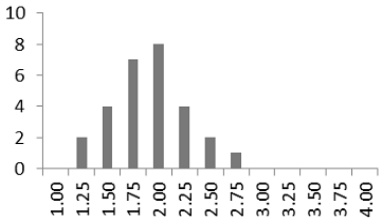

図 1: 『要注意学生』だと認定された学生の GPA

提案手法のモデルを実際の『要注意学生』の発見モデルとして適用すると、表 3 より、モデルによって認定された 91 名と、1 年次 GPA が 1.0 未満の学生 31 名を足した 122 名を『要注意学生』と認定し、さらに、実際の『要注意学生』70 名中、29 名＋ 28 名＝ 57 名（81.4 %）を発見できることが示された。換言すれば、全学生を指導対象とする場合の時間的コストの約 3 分の 1 のコストで、約 8 割の『要注意学生』を発見できたと言える。

6　まとめ

本研究では『要注意学生』の分析、定義、発見を試みた。また、発見の手法としてベイジアンネットワークを提案した。これにより、単純に GPA だけで指導対象を決定するよりも、本提案手法を用いた方がより効率的に指導対象者を見つけ出せることを示すことができた。今後の課題としては、年度による変容にも対応した汎化なモデルの構築と、実用に向けた更なる検証などが挙げられる。

参考文献

[1] 鈴木恒一, “大量データから知識を抽出するベイジアンネットワークの学習技術とその応用”, Softechs, L3765A, 32 巻, 1 号, pp. 14-17 (2011)