エデュケーショナルデータマイニングシステムの構築

郷　卓倫
Takamichi GOU

1　はじめに

近年，大量のデータの中から有用な知識を見つけるデータマイニングが，注目されている．例えば，マーケティングや株価予測，天気予報や地震予知などの自然災害シミュレーション，医療での臨床データに基づいた病気の経過や薬の効果の予測などさまざまな応用例が挙げられる．

一方，教育の分野，とりわけ大学教育の現場では，一人の指導者が指導する学生の数が非常に多いために，一人一人の学生にまで，指導が行き届いていないのが現状である．例えば，小テストを行ったとしても，採点されたデータはそのまま成績になるだけであり，その結果が個々の学生にフィードバックされることはまれである..

そこで本研究では，教育支援システムにデータマイニングを応用することにより，小テストの結果などの大量のデータの中から特徴的なデータを抽出し，そのデータをもとに，学生一人一人に対して，よりきめ細かな指導を可能とするシステムを構築する．

2　提案システムの概要

大量のデータから有益な情報を抽出するために，提案システムでは，データに対して様々な補正を行う補正処理と，処理を行ったデータに対して，データマイニングを行い，特徴的なデータを導き出す処理を行う二つの過程が存在する．

最初の過程は，個々の学生のデータが，得られた全データの平均の傾向とどれほど差があるのかということを視覚的に掴むことが目的である．一方，二つ目の過程では，得られたデータを傾向ごとに分類することが目的となる．

今回用いたデータは，8 月に研究室で行われた，4 回生が各自調べた IT 用語の発表会で，その発表をマークカードを用いて評価したデータである．今回は，6 人の発表に対して，38 人が評価を行っている．

補正処理を行う理由としては，今回用いた IT 月例発表会のデータでは，評価の甘い人や辛い人，さらに評価に散らばりがある人やまとまっている人が存在する．この違いにより分類することも意味はあるが，今回は個々のデータの傾向をつかむことに注目しているので，その個々の学生の傾向によりデータを分類し，データの傾向が如実に表れるようにするために，補正処理を行う．

補正処理では，折れ線グラフで表される時系列データに対して，平均値補正，分散補正，異常値除去を行い，データの比較を行えるようにする．また，データマイニングとして，今回はクラスタリング処理により，あらかじめ定めたクラスター数 (クラス数) にデータを分類することにより，特徴的なデータの抽出を行う．

以下の Fig. 1 でシステム全体の概要を示す．

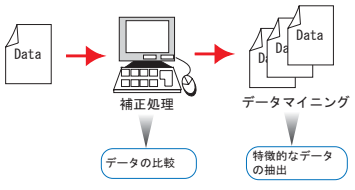


Fig. 1　システムの概要図

2.1　データの補正処理

IT 月例発表会で得られた 38 個のデータでは，評価の基準や，得点の基準が各データによって異なっている．そこで各々のデータに統計的な補正を加えることにより，集められた大量のデータの比較を行えるようにする．各データと比較する対象としては，得られた 38 個の全ての元データの平均値を用いる．具体的な統計処理は以下の 3 つである．

- 平均値補正

  各データにより，平均値が高い，あるいは低いといったばらつきが存在する．そこで各データの平均値をそろえることによって，各データの傾向が如実に現れるように補正を行う．この時，得られた 38 個の全ての元データの平均値を基準とする．具体的な処理を以下の Fig. 2 で示す．

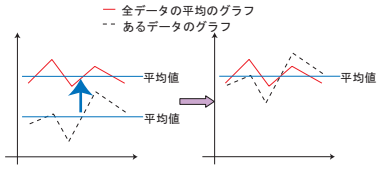


Fig. 2　平均値補正

この補正を式で表すと，

$$E_i = E_i + (A_0 - A_1) \tag{1}$$

E_i は，各発表者に対する評価であり，A_0 は 38 個からなる全データの平均の平均値，A_1 はあるデータの平均値を表している．

- 分散補正

平均値補正を行っても，各データにより，データの散らばり具合というのが異なる．そこで平均値補正と同様に，データの分散をそろえることによって，各データの傾向を探る．この時，得られた 38 個からなる全ての元データの分散値を基準とする．具体的な処理を以下の Fig．3 で示す．

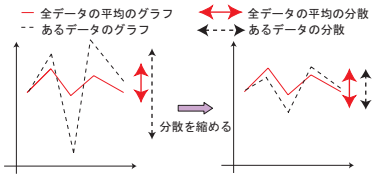


Fig. 3　分散補正

この補正を式で表すと，

$$V = V_0/V_1 \tag{2}$$

$$D_i = (C_i - E_i) * V \tag{3}$$

$$E_i = C_i + D_i \tag{4}$$

V は，38 個からなる全データの平均の分散値 V_0 を，あるデータの分散値 V_1 で割った値である．さらに D_i は，38 個からなる全データの平均の値 C_i から，あるデータの値 E_i を引いた値に V をかけた値である．そして得られた値に，それぞれ C_i を加えた値が，補正後のあるデータの値になる．このとき $_i$ は 1 から 6 までである．

- 異常値除去

各データは，それぞれの要素において，元データの平均のデータとの差の二乗和の値を評価値として持っている．その際，最も平均のデータとの差が大きい値を異常値として捉え，評価の対象から除去する．具体的な処理を以下の Fig．4 で示す．

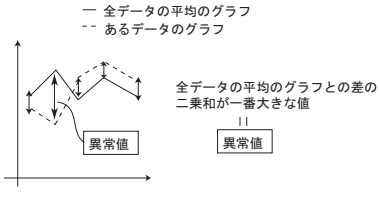


Fig. 4　異常値除去

異常値を除去する理由として，今回集められた IT 発表会のデータには，発表者自身が自分にも評価を付けている．そのために自分自身に必要以上に低い評価を与えてしまう場合が多く存在する．また，記入者がマークミスをしている可能性も得られた 38 個からなる全ての元データとの差が大きいデータには存在する．そのようなデータは評価としてあまり信頼性が無いために，異常値として除去する．

以下の Fig．5 において，データの補正処理を行うシステムを示す．

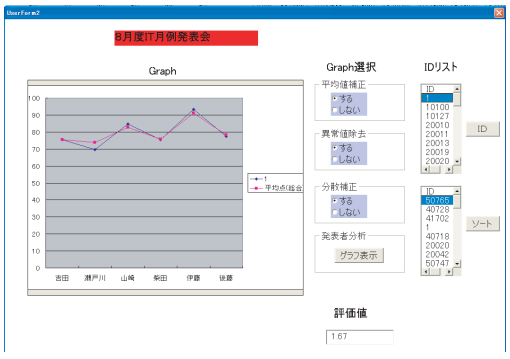


Fig. 5　補正処理システム

Fig．5 のシステムでは，IT 用語で集められた，それぞれの発表者に対して評価を行ったデータを使用している．それぞれのデータは，ID をキーとして持っている．システムの利用者が，平均値補正，異常値除去，分散補正を組み合わせることが可能で，さらに ID リストから ID を選択すると，それぞれのデータに対して補正を行ったグラフと得られた全ての元データの平均値のグラフを表示し，また平均との差の 2 乗和を評価値として表示させることができる．このとき評価値が低ければ低いほどその評価者は，全体の平均と傾向が似ているということが言える．一方，評価値が高ければ高いほど，全体の傾向からはずれ，特徴的なデータである可能性が高いと言える．

2.2　クラスタリング処理

クラスタリングとは，教師無し分類の代表的な手法で，まだ分類されていない対象を似ているもの同士からなるいくつかのグループに分類することを目的とする．本研究では，特徴的なデータを抽出するために，weka と呼ばれるデータマイニングツールを使用している．weka とは，ワイカト大学 (ニュージーランド) が中心に開発し，世界で最も使われているフリーのデータマイニングツールである．今回は，この weka で用いられているクラスタリング手法の中でも，代表的なクラスタリング手法である K-means 法を用いて，データを分類している．K-means 法とは，非階層型のクラスタリング手法で，事前にクラスター数を指定しなければならない．クラスター作成プロセスとしては，以下のようになる．

1. クラスター数の指定
2. 各クラスターの重心 (seed 値) の座標を乱数により決定
3. 各要素を最も近い seed のクラスターに分類
4. それぞれのクラスター内で重心を計算し，その点を seed とする
5. 旧 seed と新 seed との距離の移動がなくなるまで 3 と 4 を繰り返す

クラスタリングをする際，K-means 法では，要素を多次元空間におけるユークリッド距離で分類している．

以下の Fig. 6 で，K-means 法のイメージ図を示す．

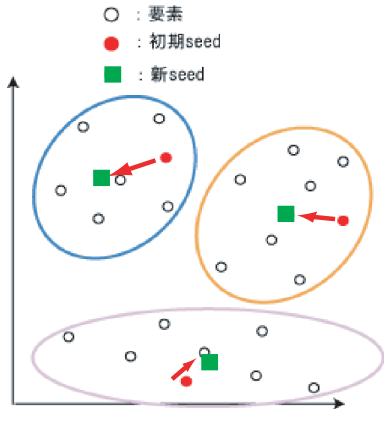


Fig. 6　K-means

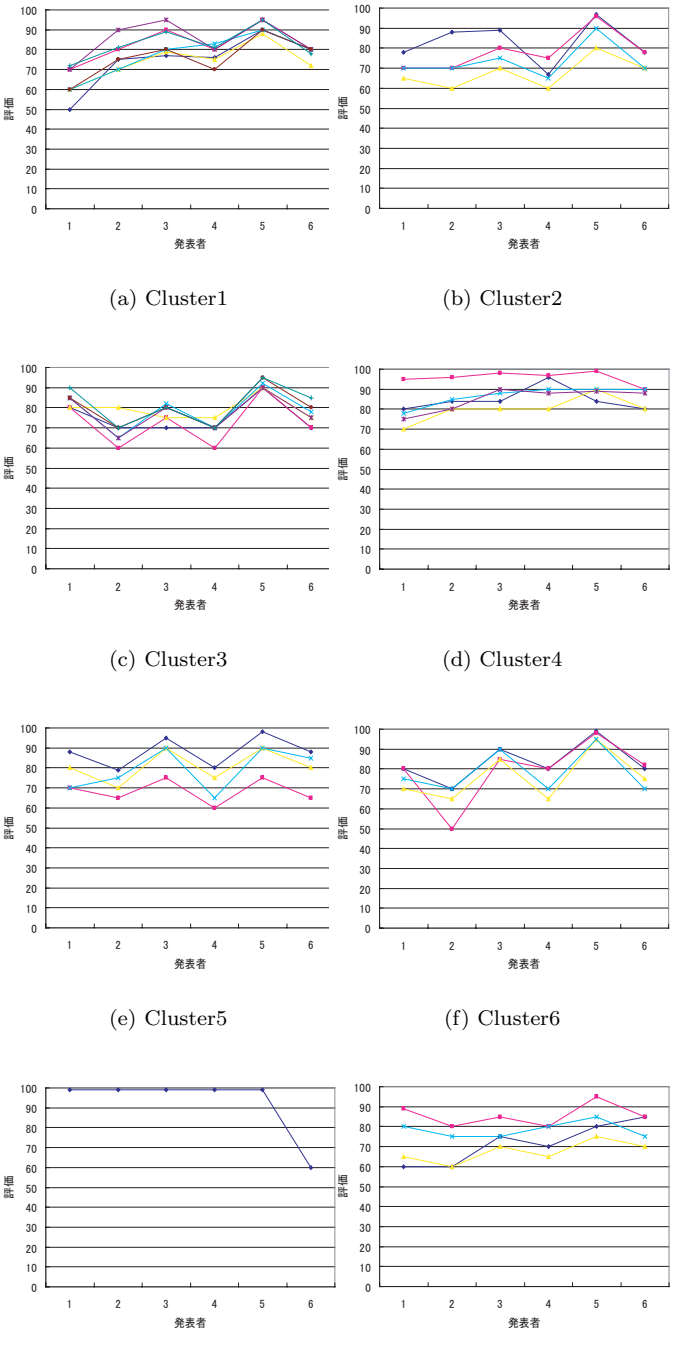
3　問題点

上でも述べたように，K-means 法では，ユークリッド距離で，データを分類しているため，順序などが考慮されないといった問題点が生じる．そのために折れ線グラフで表される時系列データの分類には向いていないと考えられる．以下の Fig. 7 で，補正処理を行ったデータをそのまま K-means 法を用いてクラスタリングした結果を示す．データは IT 用語の 8 月発表のデータで，データ数は 38，クラスター数は 8 に設定した結果である．

Fig. 7 で示しているように，Cluster2 や Cluster5 では，傾向が異なるグラフも同一のクラスターに分類されていることがわかる．また，傾向が一緒のグラフでも，別のクラスターに分類されているのがわかる．この結果からも，元の時系列データをそのまま利用して，K-means 法によりクラスタリングしても，良い結果が得られないということがわかる．

4　改良点

折れ線グラフで表されるような時系列データをクラスタリングにより，特徴が類似している要素をうまく分類するために，各要素を簡単なモデルにモデル化を行って



(a) Cluster1　(b) Cluster2

(c) Cluster3　(d) Cluster4

(e) Cluster5　(f) Cluster6

(g) Cluster7　(h) Cluster8

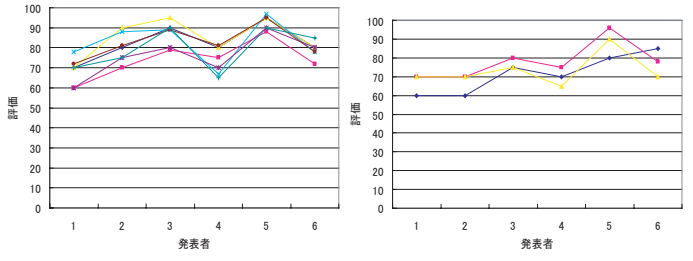
Fig. 7　クラスタリング結果

から，クラスタリングを行う．それにより，各データの特徴の差異をはっきりさせることにより，特徴が類似している要素をうまく分類できると考えられる．モデル化は Table 1 に従う．直前の要素との差，$\triangle D$ が 3 以上の場合は，その要素の値を 1 に，-3 以下の場合は-1 を与える．

モデル化を行ってクラスタリングした結果を Fig. 8 に示す．今回モデル化を行ったデータは，Fig. 7 で用いたデータと同様である．

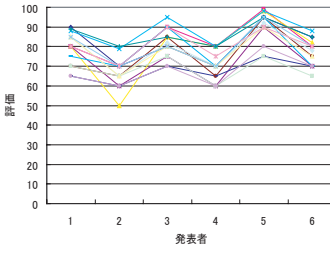
Table 1　モデル化

$\triangle D > 3$	1
$-3 \leq \triangle D \leq 3$	0
$\triangle D < -3$	-1

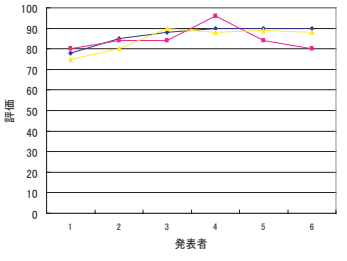


(a) Cluster1

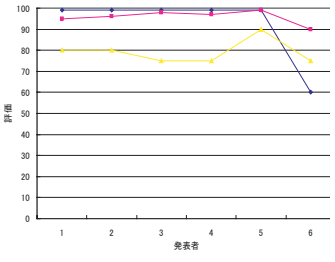
(b) Cluster2



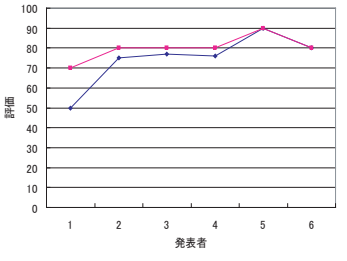
(c) Cluster3



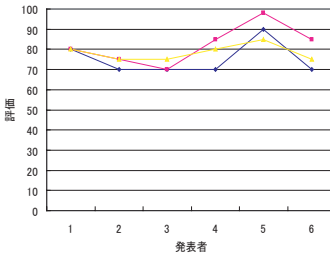
(d) Cluster4



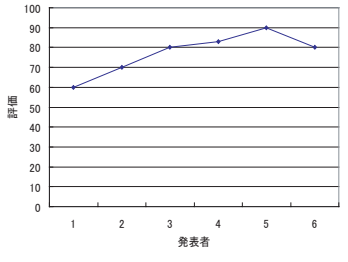
(e) Cluster5



(f) Cluster6



(g) Cluster7



(h) Cluster8

Fig. 8　モデル化クラスタリング結果

Fig. 7 と比較すると，傾向が異なるグラフが同一のクラスターに分類されておらず，全て折れ線グラフで表すと，同じ形をしたデータが同じクラスターに分類されていることがわかる．これにより，元のデータをそのままクラスタリングするのではなく，順序，つまりグラフの形状でモデル化することにより，時系列データを適切なクラスターに，クラスタリングすることが可能であるといえる．

この結果を見て，システムの利用者が，どのクラスターが特徴的かを視覚的に判断し，その結果をもとに個々の学生に対して，よりきめ細かな指導が行えると考えられる．例えば，データに小テストの結果を用いたとすると，多くの学生が前回に比べて点数が上がっているのに対して，下がっているクラスターに分類された学生は，その単元の理解が不足していると考えられ，集中的に指導することができるといった利用方法が考えられる．

5　まとめ

本報告では，エデュケーショナルデータマイニングシステムとして，教育現場で得られたデータから，特徴的なデータの抽出を行い，そのデータをもとにそのシステムの利用者が，生徒一人一人に対して，よりきめ細かな指導が可能となるようなシステムの構築について述べた．

今後の課題としては，データを与えるだけで，全ての処理を自動で行えるようなシステムを目指す．また今回は，クラスタリング手法として K-means 法を用いたが，他のクラスタリング手法が有効であるかどうか比較し，検証を行っていく予定である．

参考文献

1) 石井一夫．よくわかるデータマイニング．日刊工業新聞社．2004.