講演番号： 155

欠損値を含む学習履歴データを用いた学習予測の検討

Prediction of learning tendency using learning history data including missing values

D－15

前波 夏織† 小川 賀代†
Kaori Maenami † Kayo Ogawa†
† 日本女子大学理学部
† Faculty of Science, Japan Women's University

1. はじめに

近年、eラーニングやタブレット端末の普及により、教育ビッグデータに対する関心が高まっている。更に、教育の情報化に伴い、得られる教育データは個人単位の情報として益々膨大化することが想定される。収集された教育ビッグデータの分析を行うことで、科学的な側面から教育サービスの向上を図ることが期待されている。一方、データが増加すれば、データ内に含む欠損値も増加することは必然であり、正確な統計分析を行う上で欠損値への対応は必要不可欠である。本学の先行研究として教育ビッグデータの一部である Learning Management System(LMS)に蓄積された学習履歴データの可視化と学習予測を Hyper Self-Organizing Maps(HyperSOM)によって行ってきた[1]。しかし欠損値を含むデータは排除していたため、情報量の減少が懸念されてきた。そこで HyperSOM における学習履歴データの欠損値補完方法の有効性について検討する。

2. 欠損値補完方法

欠損値は MCAR(Missing Completely At Random)、MAR(Missing At Random)、MNAR(Missing Not At Random)の 3 種類に分けられる。学習履歴データにおける欠損値は学習者個人の意思や体調不良等、学習者の成績と相関のないデータを取り扱うため、MCAR であると仮定する。MCAR の欠損値補完方法には大きく削除法(リストワイズ法、ペアワイズ法)、最尤法(完全情報最尤法)、代入法(単一代入法、多重代入法)がある。HyperSOM を用いるには、欠損値と補完値が一対一対応である必要がある。そこで欠損値補完方法として単一代入法を採用した。また、単一代入法の中でも扱いが容易である平均値代入法を用いた。平均値代入法とは欠損値を含む 1 変数に対し、欠損値を含まないデータの平均値を欠損値に代入する手法である。

3. シミュレーション実験および結果

3.1 データの概要

解析に用いた完全データ(欠損率 0%)は、資格試験のための e ラーニングにおける学習履歴データであり、受講者 221 名分のデータである。コンテンツは全 10 章からなり、2-3、4-6、7-10 章の学習後に章末テストとアンケート、全章学習後に修了テストが設けられている。使用したデータ変数は、講座学習回数、アンケートの回答、章末テストの点数(初回・最高点)、章末テストの受講・受験回数、第 1 回模擬試験の受験回数、修了テストの受講・受験回数、修了テストの点数(初回・最高点)の 31 変数である。属性の分類には修了テストの初回点と最高点の平均を用いた。また擬似完全データは、完全データの内、属性予測に使用する 29 変数それぞれに対しランダムに 10,20,30,40%の欠損を起こし、平均値代入法による欠損値補完処理を施したものである。

3.2 解析結果

完全データ、擬似完全データに対し 10 分割交差検定を行った結果、平均識別率は表 1 のようになった。擬似完全データの正解率は完全データに対し低下するものの、データ欠損値数による情報量の減少を抑制出来ることを考慮すれば、平均値代入法による欠損値補完は有効であると言える。更に、入力データそれぞれに対する識別結果が完全データと擬似完全データとで一致するか調べたところ、表 2 のようになった。欠損率が上がった場合においても識別結果一致率に大きな違いが現れないことから、平均識別率の低下は欠損値補完処理に因らないことがわかる。以上より、欠損値を含む学習者に対しても、目標に向けての学習方法の提示を行える可能性が得られた。

シミュレーション諸元は、マップサイズ 8×8、学習時の繰り返し回数(エポック)1000 回、学習率係数 0.01、再学習時のエポック 1000 回、学習率係数 0.0001 とした。今回の属性は修了テストの結果を点数 x で表し、80≦x 、80＞x で分類し、それぞれクラス 1,2 とした。各コードベクトルの属性は、マップされた教師データから決定し、属性の予測は、コードベクトルの属性とマップされたテストデータの属性が一致する場合を正解、異なる場合を不正解、コードベクトルに教師データが入らず、属性決定ができなかった所にテストデータがマップされた場合を判定なしとしてマップ上に表示した。HyperSOM の表示結果は図 1 に示す。

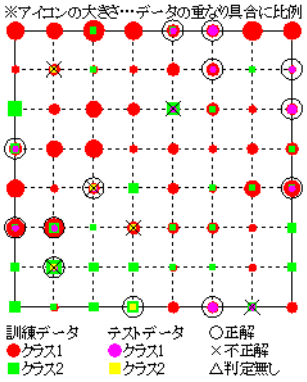

図 1 HyperSOM 出力結果
擬似完全データ(欠損率 20%)

表 1 平均識別率

欠損率 [%]	HyperSOM識別率[%] 正解	不正解	判定無し
0	69.68	30.32	0.00
10	66.13	33.87	0.00
20	63.81	36.19	0.00
30	62.00	38.00	0.00
40	65.14	34.86	0.00

表 2 識別結果一致率

欠損率 [%]	識別結果一致率	識別結果一致数/データ数
10	76.92	170/221
20	73.76	163/221
30	76.02	168/221
40	71.04	157/221

4. まとめ

HyperSOM における学習履歴データの欠損値補完方法として、平均値代入法が有効であることがわかった。更に、学習者の母集団における位置を可視化できるため、欠損値を含むデータにおいても HyperSOM を用いて個人に適した学習支援を行える可能性が得られた。

参考文献

[1]秋本尚美 他、“HyperSOM による可視化を用いた学習支援への適用”、平成 26 年度電子情報通信学会東京支部学生会研究発表会講演論文集、158 頁