6ZC-04

プログラミング教育における Convex Factorization Machines を用いた学生モデリングの有用性の検証

Verification of Student Modeling using Convex Factorization Machines in Programming Education

清水 大幹 † Daiki SHIMIZU
大枝 真一 ‡ Shinichi OEDA

1. はじめに

Educational Data Mining（EDM）と呼ばれる教育を専門としたデータマイニング分野がある．EDM は，教育システムによって収集された大量の電子データから有用な情報を発見することを目的としている．

近年，教育現場において学習中のログを収集し，機械学習による分類や予測を行う試みが増えている．Intelligent Tutoring Systems（ITS）は個々の学習者に適した設問を推薦することを目標としているが，教育効果の高い ITS を提供するためには，学習者のスキル状態の把握が必要である．その方法として学生モデリングと呼ばれる手法があり，Convex Factorization Machines（CFM）は有用な手法であることが示されている [1]．

本研究では，プログラミング教育においても CFM が有用であることを検証する．具体的には，プログラミングソースコードを用いて，学生モデリングを行う．

2. 手法

2.1.Factorization Machines

FM は，Support Vector Machines（SVM）の長所と Matrix Factorization（MF）のような factorization models を組み合わせたモデルクラスである [2]．FM は，SVM と同様に教師あり学習を行う予測モデルであり，MF の様に未知の値を推測することが可能である．

2-way FM のモデル式は以下で表される．

$$\hat{y}(\boldsymbol{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle x_i x_j. \quad (1)$$

このとき，$\langle \cdot, \cdot \rangle$ は以下に示すように，k 次元の 2 つのベクトルの内積を表す．

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}. \quad (2)$$

推定すべきモデルパラメータは以下となる．

$$w_0 \in \mathbb{R}, \quad \boldsymbol{w} \in \mathbb{R}^n, \quad \boldsymbol{V} \in \mathbb{R}^{n \times k}. \quad (3)$$

w_0 は全体のバイアス，w_i は i 番目の変数の強度，$\hat{w}_{i,j} := \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ は i 番目と j 番目の変数間の強度，$\boldsymbol{v}_i$ は $\boldsymbol{V}$ の i 行目の k 次元ベクトルを表す．また $k \in \mathbb{N}_0^+$ は，分解の次元数を定義するハイパーパラメータである．

†Advanced Course of Control and Information Engineering, National Institute of Technology, Kisarazu College
‡Department of Information and Computer Engineering, National Institute of Technology, Kisarazu College

2-way FM では，二次の相互作用を $w_{i,j}$ という独立したパラメータではなく，$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$ というベクトルの内積に分解して表現している．これにより，SVM では対応できない非常にスパースなデータに対しても，高い精度で予測を行うことが可能となっている．また，学習時間が線形となるため，Stochastic Gradient Descent を使用した直接最適化が実行可能となる．

2.2.Convex Factorization Machines

CFM では，FM の式 (1) を以下のように書き換える．

$$\hat{y}(\boldsymbol{x}) := \boldsymbol{w}^\top \boldsymbol{x} + \sum_{i=1}^{d} \sum_{j=1}^{d} z_{ij} x_i x_j = \boldsymbol{w}^\top \boldsymbol{x} + \left\langle \boldsymbol{Z}, \boldsymbol{x}\boldsymbol{x}^\top \right\rangle. \quad (4)$$

ここで，$\boldsymbol{x} = \{1, \boldsymbol{x}\}^\top$，$\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{V}^\top \in \mathbb{S}^{d \times d}$ とした．また，z_{ij} は $\boldsymbol{Z}$ の ij 要素を表す．

FM は，非凸最適化問題を含むため，局所的極小値を取得してしまう．これは初期値に依存している．また，ランクハイパーパラメータ k の選択が必要であり，予測精度は k によって大きく変化してしまうという欠点がある．CFM は，$\mathrm{rank}(\boldsymbol{Z}) \ll n$ となるような低ランク行列 $\boldsymbol{Z}$ を学習することを目的としており，これは nuclear norm を用いて $\boldsymbol{Z}$ を正則化することにより達成することができる．

つまり，CFM は FM を凸定式化したものであり，上記した欠点を克服したモデルとなっている．

3. 先行研究

先行研究 [3] では，学習者，設問，設問を解いた時間帯，その設問を解く前に解いた設問を特徴ベクトルに採用し，その設問が解けたかどうかを教師データとした FM による予測を行っている．表 1 に，先行研究で用いられた特徴ベクトルと教師データの例を示す．

先行研究 [4] では，FM と CFM による予測精度の比較実験を行っている．実データには Movielens のデータセットが使用され，特徴量として利用者と視聴映画，教師データとして利用者がその映画に何点（1〜5）を付けたかが用いられている．精度の評価指標は RMSE のみであるが，実データの 1 つを除き，CFM の方が良い精度を示している．

4. 提案手法

本研究では，先行研究により提案されている FM を用いた学習者の解答結果予測に対し，FM を凸最適化した CFM を用いることで精度向上を目指す．このとき，特徴量による精度の比較を行い，有用な特徴量を探索する．

表 1: Example of feature vector and target.

	Feature Vector $\boldsymbol{x}$													Target y
$\boldsymbol{x}^{(1)}$	1	0	0	⋯	1	0	0	⋯	14	0	0	0	⋯	1
$\boldsymbol{x}^{(2)}$	1	0	0	⋯	0	1	0	⋯	15	1	0	0	⋯	1
$\boldsymbol{x}^{(3)}$	0	1	0	⋯	0	1	0	⋯	18	0	0	0	⋯	0
$\boldsymbol{x}^{(4)}$	0	0	1	⋯	0	1	0	⋯	10	0	0	0	⋯	1
$\boldsymbol{x}^{(5)}$	0	0	1	⋯	0	0	1	⋯	17	0	1	0	⋯	0
	Student				Item				Hour	Last Item				

表 2: Details of Synthetic and Algebra dataset.

	Records	Students	Items	Skills
Synthetic	10,000,000	5,000	10,000	1,000
Algebra	8,918,054	3,310	781,620	1,070

また，プログラミングソースコードを特徴ベクトルに利用した実験を行うことで，プログラミング教育おける CFM での学生モデリングの有用性を検証する．

5. 実験

5.1.FM と CFM の性能比較

本実験では，人工データと Algebra I 2008-2009[5] の 2 種類のデータセットを用いて FM と CFM の予測精度の比較実験を行う．各データの概要を表 2 に示す．このとき，各データセットで特徴量の組み合わせを変え，精度の比較を行う．

5.1.1. 人工データ

人工データの生成には項目反応理論を用いる．学習者が設問を 1 回解くと，その設問を解くために必要なスキルに対する学習者の能力値が上昇する．また，忘却を加えるために，学習者がある設問を最後に解答した時間からの経過時間を算出し，経過時間が長いほどその設問を解くために必要なスキルに対する学習者の能力値が減少する．さらに，各設問を解くためには 1 つ以上の複数スキルが必要になるように設定する．

5.1.2.Algebra I 2008-2009

実際の e-Learning システムから取得されたデータとして，KDD Cup 2010 Educational Datamining Challenge で提供されたデータセットである Algebra I 2008-2009 を使用する．

5.2. プログラムソースコードでの学生モデリング

本実験では，2018 年度 木更津工業高等専門学校 情報工学科 第 2 学年の受講科目であるプログラミング演習 I の定期試験で出題された問題に関し，学生が提出したプログラムソースコードを用いて CFM での学生モデリングを行う．

データセットの概要を表 3 に示す．設問は，大問 1 つに対し小問が 3 つ程度で構成されているが，大問を Items として扱う．

ソースコードを特徴量に変換し，それを CFM の入力として用いる．変換する特徴量は，ソースコード著者を特定する実験 [6] で有効だったものを採用する．

表 3: Details of Programing dataset.

	Records	Students	Items
前期中間	134	42	4
前期期末	112	42	4
後期中間	76	41	3
後期期末	143	41	4

教師データには，大問あたりの点数取得割合を用いる．

6. まとめ

本研究では，プログラミングソースコードを特徴ベクトルに利用した実験を行うことで，プログラミング教育おける CFM による学生モデリングの有用性を検証する．

謝辞

本研究は，JSPS 科研費 19H01728 の助成を受けたものです．

参考文献

[1] 清水大幹, 大枝真一. “convex factorization machine を用いた学生モデリングの提案と有効性の検証”. 情報処理学会第 82 回全国大会, 2020.

[2] Steffen Rendle. “factorization machines”. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 995–1000, 2010.

[3] Nguyen Thai-Nghe, Lucas Drumond, Tomas Horvath, and Lars Schmidt-Thieme. “using factorization machines for student modeling”. *UMAP Workshops, FactMod*, 2012.

[4] Mathieu Blondel, Akinori Fujino, and Naonori Ueda. “convex factorization machines”. In *ECML PKDD 2015*, pp. 19–35, 2015.

[5] Algebra I 2008-2009. `http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp`.

[6] Edwin Dauber et al. “git blame who?: Stylistic authorship attribution of small, incomplete source code fragments”. *Proceedings on Privacy Enhancing Technologies*, Vol. 2019, pp. 389–408, 2019.