北海道科学大学研究紀要
第49号（2021年）
研究報告

Bulletin of Hokkaido University
of Science, No.49(2021)
Research Reports

DNNによる大学入学生の情報スキルタイプの分類

Classification of Information Skill Types for New Students using DNN

深井裕二*

Yuji Fukai

概要

多様化および高度化する情報技術の社会活用に応じ，教育機関では早期の情報スキル習得が求められるなか，大学入学生の情報スキル修得度には大きな個人差が見られる．その原因把握のために学習背景の調査に基づくクラスター分析による入学生の分類を実施してきた．本研究では，今日の人工知能の基盤技術あるディープニューラルネットワークを用いて入学生の情報スキルタイプを分類し，その有効性について検討した．

1. はじめに

今日の情報社会において，様々な場面でのコンピュータ利用や人工知能(Artificial Intelligence, AI）などの高度な情報技術，それに応じた技術者ニーズなど多様化および高度化が進むなか，教育の場でも早期の情報スキル習得が求められている．新学習指導要領に基づく小学校におけるプログラミング教育の必修化（2020年度）や，高等学校における「情報Ⅰ」（必修）および「情報Ⅱ」（選択）への科目再編[(1)]（2022年度）は，誰もが情報を高度に活用できる社会構築のための新たな教育体制である．このようななか，大学入学生の情報スキル修得度には大きな個人差が見られる．そのため上位のスキル学習に移行しづらく，基礎知識の復習や再教育，一定の基礎的PC経験時間などを要するのが現状である．

先行研究[(2)]では，入学生の情報スキル差に対し，学習背景などをアンケート調査し，情報スキルタイプを分類して状況を把握してきた．その分析手法にはクラスター分析(Cluster Analysis, CA)を用い，似たような学習環境で情報スキルを修得してきた学生を分類し，知識理解度などを比較し学習環境による影響を分析した．CAは複数の説明変数から多クラス分類を行う分析手法であるが，同様の多クラス分類の手法として，ニューラルネットワーク（Neural Network, NN）がある．その発展形でもあるディープニューラルネットワーク（Deep Neural Network, DNN）は，今日のAIを大きく技術革新させた深層学習（Deep Learning）の基盤技術として高い認識精度を有する．本研究では，入学生の情報スキルタイプの分類に対し，分析作業がCAと同等以上に簡素にできる手段としてDNNを適用した．DNNの高精度な分類機能を用いて情報スキルタイプの分類を実施し，その有効性について検討した．

2. CAによる入学生の情報スキルタイプの分類

本学の情報基礎教育科目では入学生対象の情報スキル調査を2年間にわたり実施してきた．調査は高等学校での情報科目の授業内容について，79テーマにおける知識理解度，8種類の実習テーマの実習時間と得意度，高等学校の学科，情報科目数，PC使用開始時期，自宅でのPC用途，高校で役立った授業形態などをWebアンケートシステムで多岐選択式回答するものである．調査件数は，昨年度で661件，今年度で670件であり，各調査結果をもとにCAによって入学生を分類した．CAには非階層的手法のk-means法（Hartigan-Wong アルゴリズム使用）を用いた．説明変数として，実習時間，情報系科目数，PC使用開始時期，テーマ別実習時間（タイピング，Office，プログラミング）の6つ学習環境に関する回答値をそれぞれ平均0，分散1で標準化した値を用いた．なお，実習時間にはテーマ別実習時間の3テーマ以外にも，インターネット，制作（デザイン系），組み込みなどの実習時間を含む．クラスターの分割数は4～5で試行し，得られたクラスター平

＊北海道科学大学工学部情報工学科

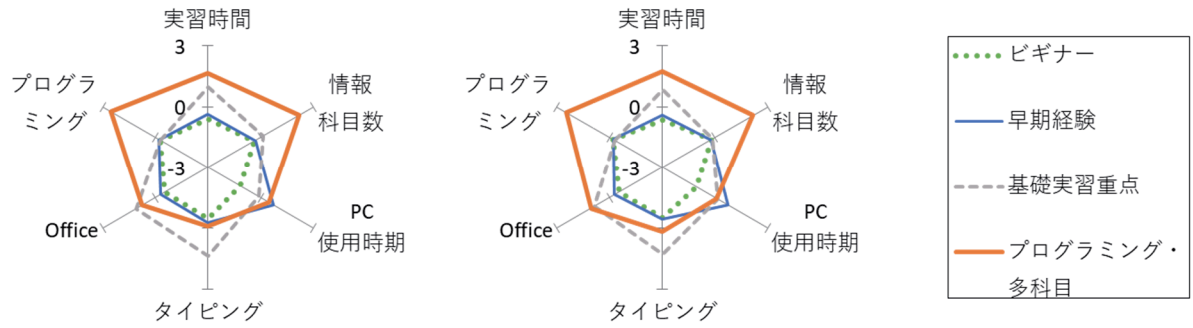

図 1　クラスターの特性（左：昨年度，右：今年度）

均の特徴差の明確な 4 分割のケースを採用した．

図 1 は CA で得られた各クラスターの特性を表すものであり，説明変数ごとにクラスター平均をプロットしている．この特性から各クラスターを情報スキルタイプの分類として以下のように命名した．

(1) **ビギナー**：実習時間・情報科目数が少なく，PC 使用開始が大学入学時など遅い時期である．
(2) **早期経験**：ビギナーに対し PC 使用開始が小学校などの早い時期である．
(3) **基礎実習重点**：タイピングや Office の実習時間が多い．
(4) **プログラミング・多科目**：プログラミング実習時間や情報系科目数が多い．

クラスターの構成率，特性（図 1）および高校での情報科目の知識理解度（全テーマの平均）は昨年度とよく似た傾向であった．表 1 は，構成率および知識理解度の平均である．ビギナーは情報科目数や実習時間が少なく PC 使用開始時期も非常に遅いのが特徴であり，他者よりも PC 操作がスムーズにできない場合が考えられ，知識理解度が最も低い．早期経験は人数が最も多く，ビギナーに対して，PC 使用開始時期の早さを除き他の学習環境に大きな違いが見られないのが特徴であり，PC 操作に十分慣れているものと推察される．また知識理解度も中程度で，実習授業が少なくても早期利用経験があればスキル習得や理解度が良好となる可能性がうかがえる．プログラミング・多科目は少数派で知識理解度が高く，大学での多様かつ高度な情報教育にもスムーズに馴染めるのではないかと思われる．

表 1　各クラスターの構成率および知識理解度
（昨年度と今年度の平均）

クラスター	構成率	知識理解度
ビギナー	29.9%	3.50
早期経験	40.5%	3.74
基礎実習重点	21.1%	3.82
プログラミング・多科目	8.5%	3.91

3. NN による CA の分類の学習

CA では統計ツールの R を用いてデータ処理した．分析に用いた k-means 法は，サンプルがクラスター中心と最小距離になるようクラスターを割り当て直す処理を繰り返すものである．結果的に各サンプルは最も近いクラスター中心のものに所属することで，似たようなデータにグループ分けされることとなる．NN も多クラス分類によって似たようなデータにグループ分けされる．これらは機械学習の手法であり，学習処理を必要とせず入力データのみで分類する教師なし学習の CA に対し，教師あり学習の NN は入力データと対応する出力データ（教師信号）を用いた学習によって未知のデータに対する分類能力を獲得する．例えば人間は手書き数字の形を学習しており，書くたびに微妙に異なる形の数字を認識できる汎化能力を持つ．NN も汎化能力によって学習データと微妙に異なる未知のデータでも似たようなグループに分類可能である．NN は画像をはじめ多様な情報に対し人間の認識能力を模倣するかのような高度な機能を有する．その適用事例には，商品購入情報とアンケート結果から NN を用いて消費者のライフスタイルを推定した研究[3]などがある．

本研究では，図 2 に示すように昨年度の情報スキ

ル調査結果と CA による分類結果をもとに NN を用いて CA の分類機能を学習し，CA に代わって NN によって今年度の情報スキル調査結果を分類した．このような NN による分類結果の妥当性については，昨年度と今年度において，学習環境の数値と情報スキルタイプの関係性がほぼ同様であるという仮定において，NN と CA の分類結果は近いものであるという考えに基づく．この場合，クラスター境界付近のサンプルでは CA を実施するたびに結果が異なる可能性があり，さらに昨年度と今年度ではクラスター中心の変動によって同様のゆらぎが想定される．NN でも特徴が似たサンプルの誤認識は知られており，異なる分類になる可能性がある．このような状況により CA と NN で分類結果が一致しないケースが含まれると考えられる．よって，分類結果を活用する際，個人の情報スキルタイプの分類精度は 100%ではないため，統計的なデータ活用が適当であろう．

本研究における NN の構造には，一般に認識精度が高いとされる多層構造を持つ DNN を用いた．今回は入力層および 3 つの全結合層から成る 4 層構造とした．入力層は CA の説明変数に用いた 6 つの学習環境に関する値が入力される．2 つの中間層は汎化能力を高めるために 0.2 の確率でランダムにユニットを無効化するドロップアウトを行っている．なお，入力数が小規模なので学習の乱れを抑えるためにユニット数は 2 倍にした．活性化関数には，中間層では ReLU（Rectified Linear Unit）関数を用いた．これは勾配消失問題の回避や高速処理ができ，DNN で用いられる代表的な関数である．出力層では 4 つのクラスターに多クラス分類するために Softmax 関数を用いた．また，損失関数として学習速度低下を防ぐことで知られる交差エントロピー誤差を用い，最適化手法として学習率を適応させて収束が速い Adam（Adaptive Moment Estimation）を用いた．

学習に用いたデータとして，昨年度の情報スキル調査結果のうち無作為抽出した 80%を学習用データ，残り 20%を評価用データとした．学習回数は学習状況のモニタリングによって，損失が十分収束しつつ学習と評価の認識率が乖離して過学習とならない状態を判断して設定した．学習終了時の評価用データの認識率は 99.24%，損失は 0.02 となった．

学習および分類プログラムは基礎言語に Python，DNN のフレームワーク（ライブラリ）に Keras を用いて記述した．入力データである学習データには昨年度の情報スキル調査において CA で用いた説明変数の値を用い，教師データには昨年度の CA の結果であるクラスター番号を用いた．これらのデータを CSV ファイルにしたものを学習プログラムで処理した．学習プログラムでは学習後の DNN の重みを学習済みの学習モデルとしてファイルに保存する．一方，分類プログラムは，学習済みの昨年度の学習モデルをもとに，今年度の情報スキル調査結果から入学生の情報スキルタイプを分類する．

リスト 1 に分類プログラムを示す．1～3 行目はライブラリ使用宣言であり，4 行目は情報スキル調査結果の読み込み，5 行目は学習済み学習モデルの読み込みである．6 行目で認識処理を実行し，変数 out には各サンプルにおける各クラスの判定確率である出力ベクトルが得られる．7 行目では出力ベクトルのうち最大確率のインデックスを求めており，これはどのクラスに分類したかを意味するものである．8 行目で分類結果を Excel でも開ける CSV ファイルとして保存する．このように Keras ライブラリを用いることで AI プログラムが簡潔に記述できる．

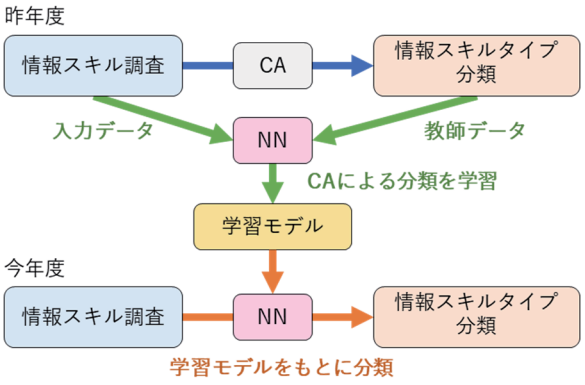

図 2　CA の分類結果を利用した NN による分類

リスト 1　Keras を用いた分類プログラム

```
1: import numpy as np
2: from keras.models import load_model
3: from keras.utils import np_utils

4: d = np.loadtxt('data.csv', delimiter=',')
5: model = load_model('model.dat')
6: out = model.predict(d)
7: result = np.argmax(out, axis=1)
8: np.savetxt('result.csv', result, delimiter=',')
```

4. NN 方式の評価

今年度の情報スキル調査結果に対し，CA と NN で分類結果が一致したサンプルは 95.97%であった．両手法の結果の差異を調べるために，説明変数についてサンプル値とその分類の平均値との二乗平均平方根誤差（Root Mean Square Error，RMSE）を式(1)で求めた．RMSE は分類の中心からの乖離を表す．ここで x はあるサンプルの説明変数ベクトル，$\hat{x}$ は分類平均ベクトル，N は説明変数の個数である．

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \qquad (1)$$

図 3 は，今年度の情報スキル調査結果をもとにした CA と NN における RMSE 分布を分類別に示しており，横軸は RMSE，縦軸は出現頻度である．また，表 2 は分類ごとの RMSE 平均，標準偏差および CA に対する NN による分類の一致率を示している．図 3 および表 2 から，特にプログラミング・多科目は RMSE のばらつきが大きく凝集性と一致率が低いことがわかる．また，表 3 は CA と NN 間での分類結果の合致性に着目した RMSE 平均である．両手法とも一致群の約 0.6 に対し，不一致群は約 0.9 以上と，分類の中心から離れている傾向がうかがえる．

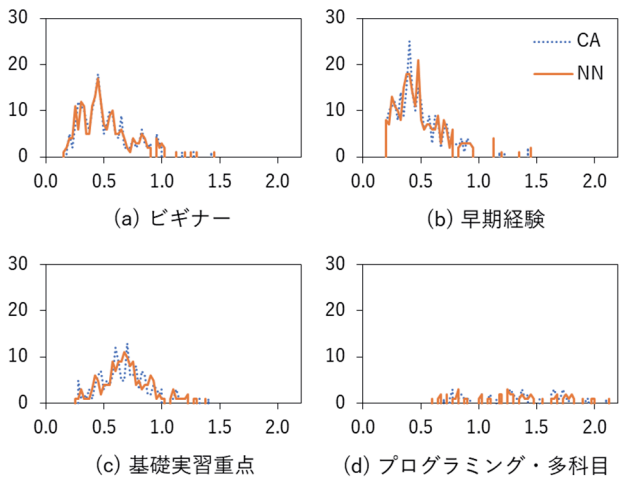

図 3　RSME 分布（今年度）

表 2　分類別 RMSE 平均（今年度）

分類	RMSE		一致率
	CA	NN	
ビギナー	0.54±0.23	0.54±0.24	96.2%
早期経験	0.51±0.22	0.51±0.23	99.2%
基礎実習重点	0.71±0.23	0.72±0.22	92.4%
プログラミング・多科目	1.31±0.36	1.31±0.40	90.6%

さらに，NN による分類の妥当性を検討するにあたり，学習モデル構築のもととなる昨年度の CA の分類結果に対して，式(2)に示すシルエット係数[4]を求めてクラスターの分離性を調べた．

$$s_i = \frac{b_i - a_i}{max(a_i,\ b_i)} \qquad (2)$$

各サンプルについて，a_i は所属クラスターの他のサンプルまでの平均距離であり凝集度を表す．また b_i は近隣クラスターに属するサンプルまでの平均距離であり乖離度を表す．シルエット係数 s_i は所属クラスター中心に近い程 1，近隣クラスターに近い程 -1 となる．図 4 は全サンプルのシルエット係数をクラスターごとに降順に整列し横棒グラフにしたシルエットプロット[5]である．シルエット係数は全体的に低めでクラスターの分離性は低いが，これは個人の学習環境が多様であるという性質によるものと考えられる．特に基礎実習重点とプログラミ

表 3　一致状況における RMS 平均の比較（今年度）

分類手法	RMSE の平均	
	一致群	不一致群
CA	0.62	0.91
NN	0.62	0.99

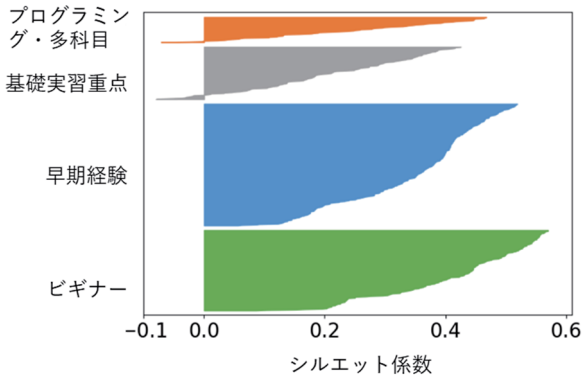

図 4　シルエットプロット（昨年度，CA）

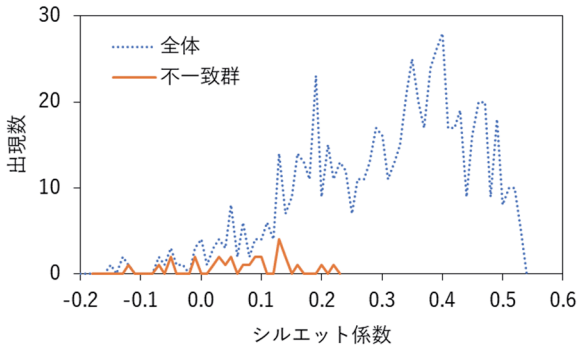

図 5　シルエット係数分布（今年度，CA）

ング・多科目においてスコアが 0 に近いサンプルが比較的多く見られ，それらは近隣クラスターとの分類境界付近に位置している可能性があり．さらに，スコアが負の値となるサンプルについては分類が適当でない可能性がある．このように NN の学習モデル構築に用いた CA の分類結果には信頼性の低いデータが若干含まれ，NN の認識に影響している可能性がある．例えば，CA で間違って分類されたサンプルとほぼ等しい値のサンプルが NN の入力データに出現した場合，NN では間違った分類の学習から間違った結果が導かれる可能性も否定できない．

図 5 は，今年度の CA による分類におけるシルエット係数の分布を示したものである．不一致群は NN の認識結果と異なるサンプルを示しており，シルエット係数が低い値に分布する傾向が見られる．つまり，不一致群は凝集度において密集部分から外れたもの，あるいは乖離度において他の分類に近いものなどに含まれる可能性を意味する．

以上の分析や CA および NN の性質から，NN による分類の妥当性を考察する．CA はアルゴリズムや試行によって結果が異なり，サンプルの分類結果が変動する場合がある．これは，クラスター中心の変動が小さいという前提において，特に分類境界付近のサンプルに起こる可能性が高いと考えられる．また，分類の凝集性が低いとクラスター中心の変動が大きいため分類結果の変動も大きい．凝集性を高めるためには，情報スキル調査における学習環境における項目を追加・削除する方法が考えられ，調査内容設定に検討の余地がある．本研究では，NN の学習モデルを CA の分類結果から構築しており，CA の分類結果を NN が模倣するものと言えよう．このとき CA のアルゴリズムによるゆらぎの影響が NN にも伝搬する可能性が挙げられる．また NN の性質上，未知のサンプルに対する誤認識が起こる場合があり，特に分類境界付近のサンプルが該当すると考えられる．一方で，NN の学習で実施したドロップアウトにより，学習データへの過剰な適合を避けて汎化能力を高めているため，サンプルのずれに対する許容が大きくなる．これによって CA のゆらぎを吸収するような状況も仮説として考えられる．CA のゆらぎと NN のへの影響については検証が必要である．また，分類境界付近の不一致を減らすために，CA の分割数を高めて分類を詳細化する方法や，CA におけるシルエット係数の低いサンプルを除去して学習モデルを構築する方法などの発展が考えられる．

5. まとめ

大学入学生に見られる情報スキルの個人差は授業の構成や進行を難しくしており，とりわけ大人数教室では，少なからず個々の学生へのストレスや不利益が生じているものと思われる．個人差の原因を分析し改善策を授業に取り入れるためには，情報スキル調査の実施と分析は重要であろうと思われる．今回，DNN を用いて入学生の情報スキルタイプの分類を行った．CA の分類変動および NN の誤認識により分類境界付近のサンプルで分類精度が低くなるが，CA と NN の分類結果の一致率は約 96%と高く，NN の分類結果を様々な統計分析に十分活用できるものと思われる．

参考文献

(1) 文部科学省：高等学校学習指導要領（平成 30 年告示）解説　情報編，2018.
(2) 深井裕二：多変量解析による大学入学生の情報スキル分析，令和２年度電気情報関係学会北海道支部連合大会，pp. 159-160，Vol. 2020，2020.
(3) 土井千章，片桐雅二，太田賢，重野寛：購入商品レベルでの購買行動に着目したライフスタイルの推定，情報処理学会論文誌，pp. 298-307，Vol. 58，No. 2，2017.
(4) Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, pp.53-65，20，1987.
(5) scikit-learn developers: Selecting the number of clusters with silhouette analysis on KMeans clustering，https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html，参照日：2021-1-15