実現場への導入を加速する新たなAI（2）
～発見科学と機械学習を融合した説明可能なAI Wide Learning

後藤啓介（ごとう　けいすけ）
株式会社富士通研究所 人工知能研究所

1．はじめに

本稿では富士通研究所が開発した発見科学と機械学習を融合した説明可能なAI Wide Learningのコンセプトと技術内容の解説を行う．

本稿の内容は経営情報学会2019年秋季全国研究発表大会の特別セミナー（大堀，2019）[1]の内容をまとめたものであり，全3回にわたる記事の第2回にあたる．第1回ではデジタルトランスフォーメーション（DX）に向けた取り組みが現在重要視されており，実現場では判断結果を説明する説明可能なAIが求められていることを述べた．第2回の本稿では説明可能なAI Wide Learningのコンセプトと技術内容の解説を行い，続く第3回ではWide Learningの実応用例について紹介する予定である．

2．Wide Learningのコンセプト

Wide Learningは発見科学と機械学習という異なる分野の技術を融合した技術である（図1）．

発見科学の分野では異なるデータ群の中から両者を区別する特徴的なパターンを発見する問題が盛んに研究されてきた．この問題はアイテム集合 $I=\{a_1, \ldots, a_n\}$ 上の正負ラベル付きデータベース D_+, $D_- \subseteq 2^I$ と制約 f が与えられたとき，f を満たすすべてのパターン（アイテム集合）$P \subseteq I$ を列挙する制約パターンマイニング問題として定式化される．様々な制約が考えられるが，例えば顕在パターン（Dong and Li, 1999）[2]は制約 $f(P)=sup(D_+, P)/sup(D_-, P) \geqq \theta$ を満たすパターンであり，正データベース D_+ に多く出現し，かつ負データベース D_- にはあまり出現しないようなパターンを表す（ここで $sup(D, P)$ はデータベース D 中のパターン P を含む要素数である）．

Wide Learningはこの制約パターンを知識発見ではなく分類モデルの構築に利用する．制約パターンは1つ1つがデータベースを特徴づける解釈可能な仮説（ナレッジチャンク）と見なすことができる．これらの重要な仮説を分類モデルに用いることで解釈性を維持したまま高精度な分類を可能にする．

3．否定アイテムを含んだパターンの高速発見

Wide Learningにおいて表現力の高いパターンを用いることは説明能力や分類精度の向上に直結する．Wide Learningでは否定アイテムをデータベースに追加し表現力の高い制約パターンの発見を図る．ここでアイテム a の否定アイテム $\neg a$ とはアイテム a を含まない要素に必ず出現するアイテムのことである．「あるアイテムが出現しなかった」という情報は「あるアイテムが出現した」ことと同様に重要な場合があり，このような否定アイテムの追加により説明能力と分類精度の向上を図っている．

否定アイテムの追加はマイニングに要する計算時間が増大する課題を引き起こす．否定アイテムは既存アイテムの排他表現であるため，追加したデータベースの各要素は否定アイテムを含む $2|I|$ 個のアイテムの内必ず $|I|$ 個のアイテムを持つことになる．つまり否定アイテムが追加されたデータベースは密なデータベースとなる．しかしながら，従来手法の

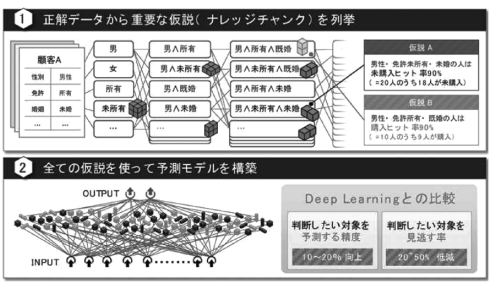

図1　Wide Learningの概要

多くは疎なデータベースを対象としており，密なデータベースに対するパターンの高速な列挙が困難である．

我々はパターンマイニングの過程でアイテムの探索順序を動的に決定することで効果的な枝刈りを可能とするマイニングの高速化手法を提案する．これにより従来手法では困難であった密なデータベースに対する高速な制約パターンマイニングが可能となり，否定アイテムを追加した制約パターンマイニングも同様に高速に列挙可能となる．

4. 計算機実験

前節までで述べたとおり，Wide Learningの技術の核となるのは密なデータベースに対する制約パターンの高速列挙と否定アイテムを含んだ分類モデルの構築である．本節では計算機実験により両者の効果を検証する．

まずCP4IMデータセット[1]を対象としてパターンマイニングに要する時間を提案手法と従来手法LCM（Uno et al., 2003）[3][2]，CP-tree（Fan and Ramamohanarao, 2006）[4]との比較を行った．比較に用いる制約には公開されている実装で計算可能な顕在パターンの亜種である，ジャンピング顕在パターン（JEP），極小ジャンピング顕在パターン（SJEP）を用いた．結果を図2に示す．図より密なデータベースAnnealについて，提案手法はLCMより約100倍高速であり，CP-treeは3600秒以上時間がかかり比較不能な結果となった．一方比較的疎なデータベースMushroomについて，提案手法はLCMより10倍ほど低速，CP-treeより5–10倍高速である結果となった．以上より提案手法は疎なデータベースに対して低速になる場合があるものの，従来技術では計算に時間を要していた密なデータベースに対して，高速にパターンマイニング可能であることが確認できた．

次に極小顕在パターンを基にしたWide Learningと既存の分類モデルとの精度比較を行う．具体的には，入力データベースに否定アイテムを加えた場合，加えなかった場合，パターン長を5以下に制限したパターン抽出を行い，抽出したパターンを説明変数としてもつロジスティック回帰をL1正則化したモデルを考える．このモデルを用いて表1に示すUCIリポジトリ[3]で公開されているデータセットに対して二値分類問題を実行し，F値を既存手法（ロジスティック回帰，決定木，ランダムフォレスト）と五分割交差検証で比較した．提案手法の学習

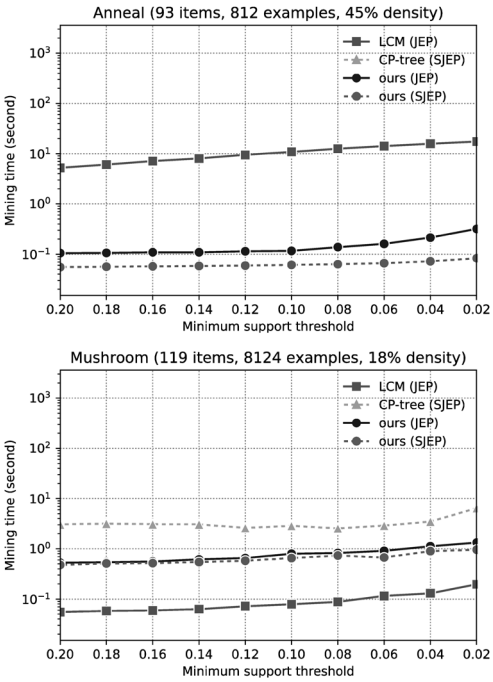

図2　制約パターンマイニングの計算時間

表1　分類精度の比較に用いるデータセット

name	#sample	#feature (not binarized)	#target class
Banknote Authentication	1372	5	1
Breast Tissue	106	10	6
Glass Identification	214	10	6
Iris	150	4	3
Wireless Indoor Localization (Wifi)	2000	7	4
Yeast	1484	8	9

表2　分類精度の比較

	提案手法（否定アイテムあり）	提案手法（否定アイテムなし）	決定木	ロジスティック回帰	ランダムフォレスト
banknote-class-1	**0.998**	0.992	0.989	0.991	0.994
breast-tissue-class-adi	**1.000**	**1.000**	0.935	0.931	0.971
breast-tissue-class-car	**0.937**	0.863	0.891	0.894	0.931
breast-tissue-class-con	**1.000**	**1.000**	0.891	0.740	0.900
breast-tissue-class-fad	**0.806**	0.722	0.599	0.673	0.535
breast-tissue-class-gla	**0.881**	**0.881**	0.771	0.700	0.727
breast-tissue-class-mas	**0.832**	0.770	0.523	0.482	0.474
glass-class-1	0.802	0.800	0.733	0.716	**0.830**
glass-class-2	**0.867**	0.863	0.777	0.599	0.799
glass-class-3	**0.697**	0.625	0.558	0.231	0.371
glass-class-5	0.920	**0.960**	0.777	0.658	0.865
glass-class-6	**1.000**	**1.000**	0.960	0.920	**1.000**
glass-class-7	0.942	**0.966**	0.900	0.915	0.915
iris-class-setosa	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
iris-class-versicolor	**0.962**	0.916	0.948	0.730	0.949
iris-class-virginica	0.949	0.943	0.952	**0.971**	0.952
wifi-class-1	0.993	0.993	0.989	0.989	**0.997**
wifi-class-2	**0.982**	0.961	0.979	0.977	0.978
wifi-class-3	0.974	0.943	0.953	0.598	**0.975**
wifi-class-4	**0.995**	0.956	0.991	0.994	**0.995**
yeast-class-CYT	0.632	0.604	0.604	0.606	**0.650**
yeast-class-ERL	**1.000**	**1.000**	0.647	0.867	0.167
yeast-class-EXC	**0.661**	0.536	0.589	0.530	0.654
yeast-class-ME1	**0.785**	0.734	0.761	0.641	0.779
yeast-class-ME2	**0.591**	0.420	0.485	0.430	0.483
yeast-class-ME3	**0.823**	0.810	0.793	0.768	0.811
yeast-class-MIT	0.634	0.589	0.614	0.590	**0.645**
yeast-class-NUC	0.630	0.545	0.605	0.590	**0.634**
yeast-class-POX	**0.628**	0.614	0.614	0.614	0.560

には最小記述長原理を用いて離散化したデータを使用し，既存手法の学習には元の実数値データを使用した．また，すべての手法はライブラリ Optuna[4] を用いてハイパーパラメータを最適化した．

実験結果を表 2 に示す．結果より，Wide Learning はほとんどのデータセットで高い F 値を示した．さらに，否定アイテムを導入することでより高い精度を達成することがわかった．

注

1） https://dtai.cs.kuleuven.be/CP4IM/datasets/
2） http://research.nii.ac.jp/~uno/codes.htm で公開される LCM バージョン 5.3 を使用した．
3） http://archive.ics.uci.edu/ml/datasets.php
4） https://optuna.org

参考文献

［1］ 大堀耕太郎（2019），2019 年秋季全国研究発表大会特別セミナーのご案内「実現場への導入を加速する新たな AI」経営情報フォーラム，Vol. 28, No. 2, September.

［2］ Dong, G., and Li, J., “Efficient mining of emerging patterns: Discovering trends and differences,” In Proc, KDD, 1999.

［3］ Uno, T., Asai, T., Uchida, Y., and Arimura, H, “LCM: An efficient algorithm for enumerating frequent closed item sets,” In Proc. FIMI '03, IEEE, 2003.

［4］ Fan, H., and Ramamohanarao, K., “Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers,” *IEEE TKDE*, Vol. 18, No. 06, 2006, pp. 721–737.

略歴

後藤啓介（ごとう　けいすけ）

2014 年（株）富士通研究所入社．データマイニングと機械学習アルゴリズムの研究開発に従事．博士（理学）．