令和５年度
人工知能技術等を活用した先行技術文献調査事業の案件選定業務における高度化・効率化に関する実証的研究事業

報告書

第 1.0 版　令和 6 年 3 月 22 日

株式会社 NTT データ

目次

1. 事業概要

1.1. 背景

昨今、産業財産権を取り巻く環境が多様化・複雑化する中、制度の複雑化や先行技術調査における調査対象資料の増加等に起因する業務量の増加が、我が国特許庁における課題の一つとなっている。

こうした環境変化とそれに伴う業務量の増加に適切に対応していくためには、最新の技術を取り入れ、事務の高度化・効率化を図っていくことが必要である。

特許庁では、特許行政の高度化・効率化に資することを目的に、特許庁業務における AI の活用に向けて、2017 年 4 月 27 日に「特許庁における人工知能（AI）の活用に向けたアクション・プラン」を公表しているところ、2022 年には改訂版として今後 5 年程度の AI 活用のあり方を検討した「特許庁における人工知能（AI）技術の活用に向けたアクション・プラン（令和 4～8 年度版）」が策定され、引き続き特許庁業務への AI 技術の適用可能性に関して検証を進めている。当該アクション・プランは、近年の特許行政事務を巡る状況の変化や AI 技術の急速な進展をふまえて改定されており、特許審査業務に関連する新たな取組としては、特許審査管理業務が対象領域として選定されている。

一方、特許庁では、特許審査官の行う先行技術調査の一部を登録調査機関へ外注する先行技術文献調査事業を実施し、審査迅速化の推進を図っている。登録調査機関に先行技術文献調査を外注するにあたっては、審査対象となる案件の中から登録調査機関に外注することが適当と考えられる案件を抽出するために、審査官によって本願書類の確認等を行うことで、外注対象案件が選定されている。また、外注対象案件選定の際には、案件ごとに先行技術文献調査の検索対象となるデータベース等を指定するための検索オプションの指定がなされている。

加えて、選定された外注対象案件については、登録調査機関にて案件の特性等を考慮した上で各検索者に配布されている。先行技術文献調査事業における調査案件数は年間約１４万件にものぼり、これらの案件の選定から配布に至るまでの業務（以下「案件選定等業務」という。）の総量は膨大なものとなっている。

1.2. 目的

本実証的研究事業は、先行技術文献調査事業における案件選定等業務を前述のアクション・プランの特許審査管理業務の一部として位置づけ、案件選定等業務について現状の業務課題を特定し、業務課題の解決手法として適用可能性のある AI 等の技術について調査を行い、その精度分析を通じて案件選定等業務の高度化・効率化に最適な技術・手法について検討することを目的とする。

1.3. 事業全体の流れ

事業全体の流れを**エラー! 参照元が見つかりません。**に示す。

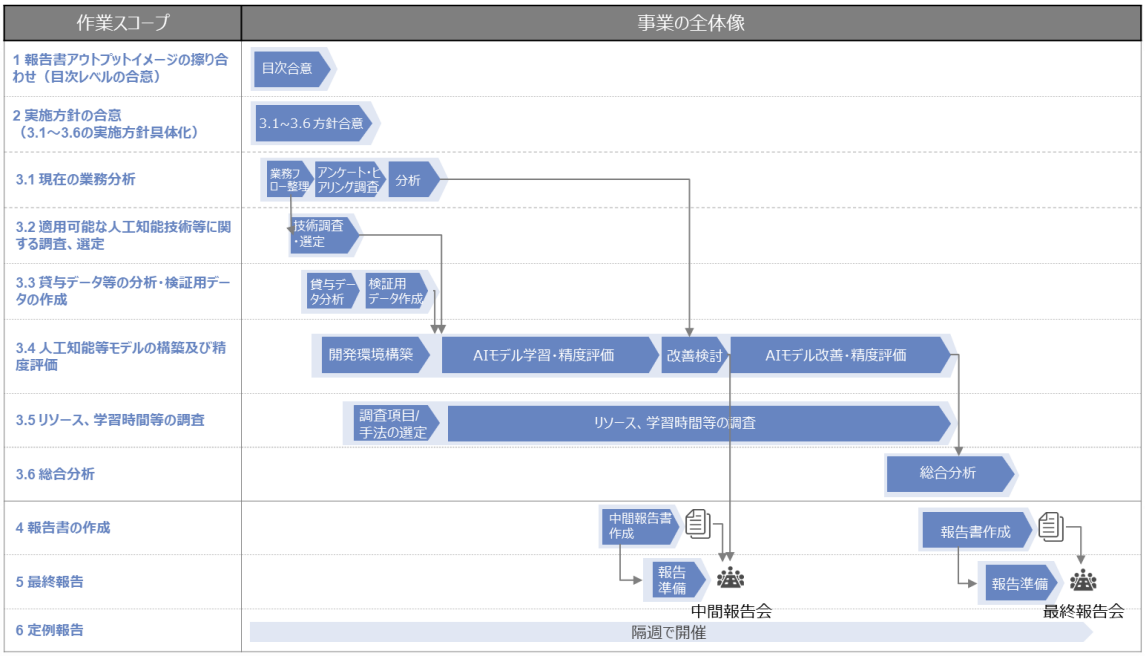


図 1.3-1 事業全体の流れ

事業開始にあたって、特許庁と「報告書アウトプットイメージの摺り合わせ（目次レベルの合意）」及び「実施方針の合意（2.4.1～2.4.6 の実施方針具体化）」を実施することで、手戻りなく作業を進めた。

1.4. 本事業の内容

本事業で実施した内容は、以下のとおり。

1.4.1. 現在の業務分析

案件選定等業務に関して業務フローや業務仮説を整理した上で、以下の 4 つの観点で特許庁（審査推進室及び審査室）及び登録調査機関（全 9 機関）に対してアンケート調査、登録調査機関(全 9 機関)に対してヒアリングを実施し、案件選定等業務の作業内容や工夫点、課題等について調査・分析を行った。

① 外注対象案件の選定
② 選定案件の検索結果の報告形態（対話型（口頭）・補充型（書面））の決定
③ 選定案件への検索オプションの付与
④ 選定案件の登録調査機関の検索者への割り振り

1.4.2. 適用可能な人工知能技術等に関する調査、選定

業務分析で明らかになった課題に対して、国内外の論文や弊社の過去の調査資産を活用して、適用可能と想定される人工知能技術等（学習を要しない技術も含む）を調査し、ユースケースとして整理し、概要をまとめた。

調査にあたっては 0 であげた①～④に対応した以下の業務課題を解決する技術を含めた。また、④の課題解決にあたっては、人材育成的な観点（例：新人検索者に難易度の高い案件も一定数割り振るなど）も考慮した。

① 外注対象案件の自動選定
② 選定案件の報告形態（対話型・補充型）の自動決定
③ 選定案件への検索オプションの自動付与
④ 検索者への案件割り振りの高度化・効率化

また、上述の調査結果をふまえ、1.4.4 で精度検証を実施する技術について、特許庁と協議の上で選定した。

1.4.3. 貸与データ等の分析、検証用データの作成

特許庁より貸与されるデータを分析し、検証用データ（学習、評価データ）を作成した。1.4.2 にて選定した技術要素を対象に、学習及び評価に用いる検証用データを抽出・作成し、整備・加工を行った。

1.4.4. 人工知能技術等モデルの構築及び精度評価

1.4.2 で選定した人工知能技術等に対して、1.4.3 で作成した学習用データを用いて学習を行うことで、本事業に適用可能な人工知能技術等モデルを構築した。

また、構築した人工知能技術等モデルについて、1.4.3 で作成した評価データを用いて精度評価を実施した。精度評価結果をふまえ、精度を改善する手法について

検討し、特許庁担当者に承認を得たうえで構築した人工知能技術等モデルの精度改善を行った。

1.4.5. リソース、学習時間等の調査

将来的に特許庁業務での人工知能技術等の活用を想定し、人工知能技術等モデルの構築及び精度評価に必要となるリソース等について調査を行った。調査対象には、学習時に用いた計算機等ハードウェア構成、学習に要した時間、推論に要した時間等を調査した。

1.4.6. 総合分析

0~1.4.5 の結果をふまえ、今後必要となる技術や課題を総合的に分析した。なお、1.4.2 で挙げた業務課題「④検索者への案件割り振りの高度化・効率化」については、今後、「審査官への案件割り振り」へ応用することを想定した分析も行った。

1.4.7. 報告書の作成

2.4.1~2.4.6 にて実施した内容及びその結果について、報告書として本書に記載した。

1.5. 事業の運営体制

本事業の目的・目標を達成する運営体制にて事業を遂行した。具体的には、AI等先端技術に関する高度な専門性を有する技術調査/モデル構築/基盤担当と、特許庁や登録調査機関業務に関する知見や幅広いネットワークを有する業務分析担当を設置し、各分野のエキスパートを責任者に配置した。

また、このように多岐にわたる専門性を有したチームで構成されることから、プロジェクト運営及びリスク管理の対策を行うプロジェクト管理担当を別途設置し、確実な事業遂行を考慮した。

さらに、業務従事者とは別に、人工知能技術に関する学会等役員を歴任する専門家による技術監修者を設置し、事業に対する客観的かつ専門的な評価・助言を得る体制とするとともに、必要に応じて人員補助等を担うバックアップ体制を構築した。

2. 現在の業務分析

2.1. 調査の目的と位置づけ

現状の業務分析において、顕在化している課題に加え、特許庁や登録調査機関が気づいていない潜在課題について明らかにすることを目的として、これまでの特許庁や登録調査機関へのサービス提供を通じて得た知見やノウハウを活かして、業務仮説を設定のもと、アンケート項目やヒアリング項目を設計し、現在の業務を把握することとした。

また、アンケート及びヒアリングを通じて現状の案件選定等業務の実態を定量（人数・件数・所要時間等）と定性（難易度・心理的負担等）で調査・分析した。

2.2. 調査範囲と調査方法

2.2.1. 調査範囲

アンケート及びヒアリング調査範囲は以下のとおり。

表 2.2.1-1　調査範囲とした登録調査機関一覧

#	組織名
1	A 社
2	B 社
3	C 社
4	D 社
5	E 社
6	F 社
7	G 社
8	H 社
9	I 社

表 2.2.1-2　調査範囲とした特許庁一覧

#	組織名
1	審査推進室
2	審査室

2.2.2. 調査方法

業務フロー及び業務仮説に基づき、特許庁（審査推進室）及び登録調査機関（全９機関）に対してアンケート調査を、登録調査機関(全９機関)に対してヒアリングを実施し、案件選定等業務の作業内容や工夫点、課題等について調査・分析を

行った。

(1) アンケートの方法

先行技術文献調査業務における登録調査機関の工夫点、課題やニーズ等の実態を把握するため、3.2.1 記載の登録調査機関に対して、アンケート調査を実施した。アンケートは MS フォームズを用いて作成し、実施期間は、10/11(水)～10/18(水)とした。全アンケート回答者 42 名のうち、有効回答は検索オプション付与を担当する 10 名、案件割り振りを担当している 17 名であった。

なお、特許庁に対しては、2022 年 4 月～5 月に審査室外注推進担当官を対象として、審査推進室にて実施したアンケート調査結果から、外注対象選定の選定業務の作業内容や大まかな課題を確認することができたため、特許庁と合意の上で、本事業におけるアンケート調査は実施せず、登録調査機関へのヒアリング調査において工夫点や課題等の深堀を行う方針とした。

(2) ヒアリングの方法

アンケート調査における分析結果をもとに、不明点の確認及び工夫点や課題の深堀を行うため、調査対象組織に対してヒアリングを実施した。ヒアリングの形式は対面とした。

2.3. 調査設計の観点

2.3.1. アンケート設計

特許庁合意の上、アンケート設計を実施した。登録調査機関毎の特性をふまえた業務の実態を把握することを目的として、1.4.1 に記載した観点のうち、登録調査機関にて実施していると想定した観点③選定案件への検索オプションの付与、観点④選定案件の登録調査機関の検索者への割り振りに関する項目を設定した。

アンケート設計方針としては、定性的・定量的な観点でアンケート項目を作成することで、網羅的な情報収集ができるよう工夫した。

定性的な観点では、業務フローや業務仮説の整理結果をふまえ、仮説との乖離やアンケート対象者が抱える課題感等を収集するアンケート項目を設計した。

定量的な観点では、業務量の実態を確認するため、概ねの実際の作業時間や実施件数等に関するアンケート項目を設計した。

2.3.2. ヒアリング設計

アンケート設計の過程で、「2.4.1.現在の業務分析」記載の 4 観点は、それぞれ対応する組織が異なることが判明したため、ヒアリング設計は（1）審査推進室向けと（2）登録調査機関向けに観点を分けて確認することとした。なお、ヒアリング

設計は、特許庁合意の上、実施した。

ヒアリングの形式は、より現場の声を正確に反映するためオンラインではなく対面を基本とした。開始冒頭は業務経験の把握に関するような先行技術文献調査に直接関係のない質問から入ることで、率直な意見を言いやすい雰囲気作りに努め、オープンクエスチョンでより自由な意見を聞き、クローズドクエスチョンにて具体的に聞くという設問の流れを取った。現状の課題感を聞いてから将来像についても聞くことで、差分である深層課題についても発掘できるよう工夫した。

（1）審査推進室向け

審査推進向けのヒアリングにあたっては、プレアンケートで確認した回答の情報を短期間で深堀すべく、確認事項をプレアンケートの設問ごとに網羅的に整理のうえで実施した。

目的としては、⑴プレアンケートの結果をふまえた審査室における現状業務の実態把握　⑵先行技術文献調査業務における具体的な判断要素・基準の把握の 2 点を設定した。

これらの目的を達成するため、プレアンケートをもとに業務仮説を立案し、定性的な観点として、仮説との乖離や具体的な判断要素・基準を収集する項目を設計した。

定量的な観点では、業務量の実態を確認するため、概ねの実際の作業時間や実施件数等に関する項目を設計した。

（2）登録調査機関向け

登録調査機関向けのヒアリングにあたっては、開始冒頭にてヒアリングの目的や意図を説明し、ヒアリング対象者に趣旨を理解いただいた上で実施した。

目的としては、（1）課題原因や本質的な課題の特定、（2）先行技術文献調査業務における具体的な判断要素・基準の把握の 2 点を設定した。

これらの目的を達成するため、業務フロー仮説やアンケート結果等に基づいてヒアリング項目を設計し、後続作業である人工知能技術等モデルの構築に向けて必要となる要素・基準を網羅的かつ具体的に把握できるよう工夫した。

2.4. 調査結果

2.4.1. アンケート結果

アンケート結果について、観点①・②はプレアンケート結果から作業内容・工夫点・課題を記載し、観点③・④は「3.3.1.アンケート設計」から実施したアンケート結果を記載し、ヒアリングを含めて作業内容・工夫点・課題を記載することとする。

- 観点①：外注対象案件の選定

(1) 作業内容

外注対象案件の選定は、特許庁審査推進室にて規定された案件選定スケジュールに基づき、年間全 12 回、毎月 1 週間の期間で標準納品月と早期納品分標準納品月を設定のうえ、各審査室の外注推進担当官が実施している。

(2) 工夫点

費用対効果を含め、工夫点は 3 点あった。

・ 外注することによる費用対効果(以降、外注効果)を考慮

審査請求の中で、外注することによる費用対効果を考慮して、案件の抽出ルールとしてカタログを作成し、選定作業を効率化している。

(3) 課題

外注効果判断の属人化を含め、3 つの課題があった。

・ 外注効果判断の属人化

外注効果について統一的な基準が設定されておらず、カタログに該当しない案件でも各審査室の属人的な外注効果の判断から外注対象案件として選定する場合があるため、現状でのルール化が難しい。

- 観点②：選定案件の検索結果の報告形態(対話型（口頭)・補充型（書面）)の決定

(1) 作業内容

補充型案件は書面のみで審査を完結するため、簡易な審査案件のみを補充型案件として選別する作業を実施。

(2) 工夫点

報告形態の決定において以下の 3 観点を設定している。

・ 頁数や請求項数が少ないか

頁数や請求工数が少ないと書面での判断が比較的容易なため、定量的な観点として採用している。

・ 分割案件か否か

分割案件は中止となる可能性が高いことから、補充型案件の規定件数達成のため、補充型案件に採用しない。(分割案件ではないものを補充

型案件として選定する）

・ 発明内容を理解しやすいか

発明内容が理解しやすいと口頭での説明がなくても審査官が発明内容の理解が可能なため、定性的な観点として採用している。

（3）課題

頁数や請求項数の少なさや発明内容の理解しやすさに関する具体的な基準の定義が確認できなかった。(ヒアリングにて具体的な基準や定義があるかを課題として深堀することとした)

➢ 観点③：選定案件への検索オプション付与

（1）経験

全アンケート回答者 42 名のうち、検索オプション付与を担当する 10 名から（2）以降の回答を確認した。

（2）判断観点

判断に必要な観点は、“テーマコード”が最も多く 7 名で、次いで“出願人の国籍”5 名だった。なお、”その他”で最も回答が多い観点は、審査室からの指示であった。

（3）形式知の有無

形式知が”ある”と答えた回答者は 10 名中 7 名であった。回答の形式知はいずれも”特許庁審査室からの指示に基づくもの”となっていた。

➢ 観点④：選定案件の登録調査機関の検索者への割り振り

（1）経験

全アンケート回答者 42 名のうち、案件割り振りを担当する 17 名から（2）以降の回答を確認した。

（2）判断観点

案件割り振りの判断で最も多い観点は、”これまでの経験年数などの検索者のケイパビリティ（経験）”で 13 名、次いで”難易度の公平さ”の 9 名だった。

また、「人材育成」として具体的に考慮している観点では、”幅広い分野の経験”、“少し経験を積んだ検索者（1 年程度）で審査官からの評価が高い検索者”、”本願の本質を掴む力”が挙げられた。

（3）形式知の有無

形式知が”ある”と答えた回答者は 17 名中 4 名であった。回答の形式知では、”検索者の得意分野と、出願のテーマコードや難易度に基づく”等が挙げられた。

2.4.2. ヒアリング結果

前述のヒアリング設計に準じ、“(1) 審査推進室向け”は、観点①～③のヒアリング結果を記載し、“(2) 登録調査機関向け”は、観点②～④ヒアリング結果を記載する。

(1) 審査推進室向け

- 観点①：外注対象案件の選定

  外注可能な案件の抽出においては既定のカタログを使用しており、カタログを用いて抽出した外注選定可能リストから、各審査室の判断によって、外注不適案件を除外する傾向であることを確認した。

  選定案件の検索結果の報告形態においては、選定案件のうち、補充型とするものをいかに選別するかという作業方針であることを確認した。

- 観点③：選定案件への検索オプション付与

  各審査室において、検索オプション付与の判断基準や優先順位が存在することを確認した。

(2) 登録調査機関向け

- 観点②：選定案件の検索結果の報告形態(対話型（口頭）・補充型（書面）)の決定

  総論として、共通する判断要素は、”頁数や請求項数が少ないか”、”発明内容を理解し易いか”、”分割案件か否か”であることを確認。また、判断にあたっての基準やルールは存在しないことを確認した。

- 観点③：選定案件への検索オプション付与

  総論として、観点は、審査室からの指示に準じた内容であることを確認。また、基準やルールは、検索オプションごとにそれぞれ存在することを確認した。

- 観点④：選定案件の登録調査機関の検索者への割り振り

  総論として、定量的には、一部組織で”新人には請求項数が少ない案件を割り振る”等の実態を確認。定性的には、一部組織で”検索者の職歴・経験・実績等から割り振る”等の実態を確認。

  観点の標準化は、一部組織で”検索者のレベルに合わせて案件を割り振る“等の実態を確認。

  現状の課題は、一部組織で”審査官の評価基準が一定でなく、審査官の評価基準をそのまま検索者の評価に適用することへの難しさがある”等の実態を確認。

  人材育成は、一部組織で”低難易度から高難易度へ徐々に割り振り検索者に慣れてもらう”等の実態を確認した。

2.4.3. 業務分析結果

アンケート結果、ヒアリング結果をふまえた業務分析結果として、各観点における考察と人工知能技術等モデルへの適用に向けた方向性を記載する。

まず、業務分析結果の概要を述べる。

- 観点①：外注対象案件の選定

  各審査室における独自ツールも含めて業務を標準化している実態が主であるが、一部に外注効果を意識した外注案件選定という暗黙知(属人的判断)による標準化されていない対応の実態も確認した。また、外注対象案件の選定業務アプローチとして、外注案件を選定するというよりは、外注不適案件をいかに除くかという側面が強いことも確認した。

  そのため、人工知能技術等モデルへの適用に向けた方向性としては、現状の審査室における業務実態を意識しながら、外注不適案件の選定に伴うルールベース処理と AI モデルによる分類を並行して適用することができないか可能性を模索することとした。

- 観点②：選定案件の検索結果の報告形態(対話型（口頭）・補充型（書面）)の決定

  決定判断における共通観点として、「頁数や請求項数量」、「分割案件か否か」、「発明内容の理解しやすさ」等を確認したものの、各観点の基準は、曖昧な暗黙知(属人的判断)による対応という実態を確認した。

  そのため、人工知能技術等モデルへの適用に向けた方向性としては、少ない観点・基準のデータの中で、AI モデルによる分類の適用可能性を模索しつつも、定性的な整理として、書面で審査を完結するという補充型案件の特性をふまえて、補充型案件の決定におけるあるべき姿とは何かを検討することとした。

- 観点③：選定案件への検索オプション付与

  各審査室で、外国語特許文献に関する検索オプション、非特許文献に関する検索オプションともに具体的な観点と基準を詳細に確認することができた。また、一部の検索オプションは、観点や基準に関する具体的な情報が少なかった。

  そのため、人工知能技術等モデルへの適用に向けた方向性としては、現状の審査室における業務実態を意識しながら、外注不適案件の選定に伴うルールベース処理と AI モデルによる分類を並行して適用することができないか可能性を模索することとした。

- 観点④：選定案件の登録調査機関の検索者への割り振り

検索者へ案件を割り振る検索指導者の独自性が強く、暗黙知(属人的判断)による対応、かつ、各観点や基準が具体的にデータ化されていない実態を確認した。また、人材育成的な観点では、検索者の調査業務における対応範囲を広げるために「未経験テーマ群をあえて割り振ること」や、「低難易度から高難易度の案件を徐々に割り振ること」で検索者の調査業務の習熟度を上げていく工夫をしている実態を確認した。

そのため、人工知能技術等モデルへの適用に向けた方向性としては、現時点でデータ化されている要素を多面的に捉えてデータの分解やデータ同士の新たな組み合わせを検討するデータドリブンで、検索者への割り振り判断要素を実現できないか模索するとともに、現時点でデータされていない要素をいかにデータ化・活用していくのか必要な仕組みや定義を検討することとした。

各観点における業務分析結果の詳細について述べる。

- 観点①：外注対象案件の選定

(1) 作業内容

特許庁審査推進室にて規定された案件選定スケジュールに基づき、年間全 12 回、毎月 1 週間の期間で標準納品月と早期納品分標準納品月を設定のうえ、各審査室の外注推進担当官が実施している。

実態上の外注対象案件の選定業務アプローチとしては、外注案件を選定するというよりは、外注不適案件をいかに除くかという側面が強い。

(2) 工夫点

外注不適案件の除外では、各審査室で観点や基準をマニュアル化・Excel マクロといった独自ツール化して、業務ノウハウを標準化し、業務を効率化している。

(3) 課題

外注不適案件の除外作業の標準化は、現状、技術単位で個別に対応(サイロ化)している実態から、標準化による業務効率化の実現が一部の技術単位組織に偏っており、特許庁組織全体あるいは審査部全体として最適化できていないことが課題といえる。

以上の課題をふまえて、本事業で研究した人工知能技術等モデルの成果を各審査部全体へ展開することで、特許庁組織全体としての業務効率化を実現できるのではないかと推察する。

- 観点②：選定案件の検索結果の報告形態(対話型（口頭）・補充型（書面）)の決定

(1) 作業内容

「3.4.1.アンケート結果 観点② (1)」に記載のとおり、補充型案件は書面のみで審査を完結するため、簡易な審査案件のみを補充型案件として選別する作業を実施。

(2) 工夫点

「3.4.1.アンケート結果 観点② (2)」に記載の 3 点に加えて、一部の技術単位組織では、特定技術分野に該当する時点で補充型案件として選定することを業務ノウハウとして標準化し、業務効率化をしている。

(3) 課題

現状、補充型案件の選定における各観点の基準が、審査官の属人的な判断基準によることや外注年度計画の補充型件数を達成するために対応件数を調整していることから、報告形態(対話型（口頭）・補充型（書面）)の決定は、現場による運用対処が中心となっている実態を確認している。

このことから書面で審査を完結するという補充型案件としての特性は各組織に認識されていても、具体的な報告形態の決定方法が曖昧になっており、業務の標準化、それに基づく作業の効率化が実現できていないことが課題といえる。

以上の課題をふまえて、本事業で研究した人工知能技術等モデルの成果の適用可能性を考慮しつつ、審査を完結するという補充型案件の特性をふまえた補充型案件の決定におけるあるべき姿を検討し、各組織で共通認識を持ったうえで、業務の標準化や効率化を実現する必要があると推察する。

- 観点③：選定案件への検索オプション付与

(1) 作業内容

案件ごとに先行技術文献調査の検索対象となるデータベース等を指定するために審査室または登録調査機関のいずれかで検索オプション(外国語特許文献・非特許文献)を付与している。

(2) 工夫点

外国語特許文献、非特許文献ともに、特定のテーマコードに該当する案件であれば全件オプション付与、または、優先度をつけたオプション付与をするといった条件を規定したマニュアルの整備や独自ツールを利用したオプション付与を行って、業務を標準化し、作業を効率化している。

なお、各検索オプションの付与数は、年度で計画件数として決まっているため、計画件数を達成するために運用対処として調整する場合があり、規定条件は、原則として準じるものである。

また、連番が多い案件は、代表的な案件のみにオプション付与を行い、作業を効率化している。

(3) 課題

観点①と同様、検索オプション付与は、現状、標準化のレベルが少量のメモ程度のマニュアルから独自ツール化まで様々で、技術単位で個別に対応(サイロ化)している実態がある。

このことから、標準化による業務効率化の実現が一部の技術単位組織に偏っており、特許庁組織全体あるいは審査部全体として最適化できていないことが課題といえる。

そのため、本事業で研究した人工知能技術等モデルの成果を各審査部全体へ展開することで、特許庁組織全体としての業務効率化を実現できるのではないかと推察する。

- 観点④：選定案件の登録調査機関の検索者への割り振り

(1) 作業内容

登録調査機関の検索指導者が、特許庁からの外注案件を対象に調査担当となる検索者へ案件を割り振る作業を実施。

(2) 工夫点

請求項数等の定量的な指標から難易度設定して条件を絞り込んだ割り振り、独自ツールを利用した割り振り、連番案件を同一検索者に割り振りを行うことで、業務を標準化し、作業を効率化している。

また、検索者の人材育成の観点では、未経験テーマ群や低難易度から高難易度の案件を徐々に割り振っている。一部では、人材育成観点を考慮せずに経験年数だけで割り振る実態もある。

(3) 課題

前述の工夫点はあるものの、実態としては一部に過ぎず、概ねの検索者割り振り作業は、検索指導者の独自性が強く、暗黙知(属人的判断)による対応、かつ、各観点や基準が具体的にデータ化されていない(検索指導者の頭の中だけにある)実態がある。

このことから検索指導者によって、業務効率や人材育成における影響の差が大きく、登録調査機関組織全体で検索者への割り振り業務を最適化できていないことが課題といえる。

そのため、本事業で研究した人工知能技術等モデルの成果の登録調査機関や関連する特許庁審査官へ展開と、現時点でデータされていない要素のデータ化およびデータ活用する仕組みの登録調査機関・特許庁への展開から、案件割り振り業務の効率化と人材育成への寄与を含めた最適化が実現できるのではないかと推察する。

3. 適用可能な人工知能技術等に関する調査、選定

3.1. 調査の目的と位置づけ

本事業の前提となる以下の 4 つの業務課題を解決する技術を検討することを調査の目的として、適用可能と想定される人工知能技術等に関する論文調査を実施する。また、調査した論文の中から、本事業で検証を実施する技術について、4 つの業務課題毎に 1 つ以上選定する。

- 業務課題① 外注対象案件の自動選定
- 業務課題② 選定案件の報告形態（対話型-補充型）の自動決定
- 業務課題③ 選定案件への検索オプションの自動付与
- 業務課題④ 検索者への案件割り振りの高度化・効率化

3.2. 調査範囲と調査方法

本事業では 4 つの業務課題があり、それぞれの業務課題で必要となる技術分野も異なることが想定されるため、技術調査の調査範囲を絞り込むための前段として、各業務課題に対する仮説を立て、アプローチ手法を整理した。

さらに、各業務課題のアプローチ手法（ルールベース処理、AI モデルによる分類、レコメンド）を処理ステップ（データ準備、ロジック作成、判定）に分解し、詳細化することのより各ステップで必要となる技術要素をピックアップし、それを調査範囲として定義する。また、各技術要素に関連する論文を調査するためのキーワードを検索クエリとして、自然言語処理に関連する国際学会（ACL, TACL, EACL, NAACL, ENMLP 等）を含めて、広く論文誌を検索することが可能な論文検索サイト（Google Scholar, Paper With Code 等）を用いて、調査を実施する。

3.3. 技術要素の調査観点

上記の調査方法でヒットした論文の中から、各業務課題への適用性（対象データの構造がマッチしているか 等）及び技術的実現性（利用しやすいライブラリが存在するか 等）を総合的に判断する。（表 4.3-1）

表 4.3-1 技術要素の調査観点

観点	指標
業務適用性	・データ構造がマッチしているか
技術的実現性	・python ライブラリが存在しているか ・ライブラリが頻繁に更新されているか

3.4. 調査結果

3.4.1. 調査結果の概要

4.2 の調査範囲と調査方法で抽出した論文に対して、4.3 の調査観点で総合的な判断を行い、選定した技術要素別の詳細説明論文を表 4.4.1-1 に示す。

表 4.4.1-1 技術要素別の詳細説明論文一覧

#	技術要素	論文タイトル	URL
1	可読性を考慮した難易度の定量化	日本語の難易度に関する特徴分析	https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/C6-1.pdf
2	事前学習済モデルによるテキスト特徴抽出/特徴ベクトルで検索者が経験した案件の傾向を表現	特許庁における AI 技術の活用の現状と最新の取組	https://www.jstage.jst.go.jp/article/jkg/73/7/73_256/_pdf/-char/ja
3	高精度な分類アルゴリズム・ライブラリ	TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED	https://arxiv.org/pdf/2106.03253.pdf
		Benchmarking Multimodal AutoML for Tabular Data with Text Fields	https://arxiv.org/pdf/2111.02705.pdf
4	説明可能な AI	Explainable Artificial Intelligence for Tabular Data: A Survey	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9551946
5	マルチラベル分類	Comprehensive Comparative Study of Multi-Label Classification Methods	https://arxiv.org/pdf/2102.07113.pdf
6	検索者のクラスタリング	A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects	https://bibbase.org/service/mendeley/bfbbf840-4c42-3914-a463-19024f50b30c/file/3bf7d292-fec6-2dc7-20f4-c81abae91b5e/1_s20_S095219762200046X_main.pdf.pdf
7	コンテンツベースフィルタリング/コールドスタート問題の対応/人材育成を考慮したレコメンド	Recent Developments in Recommender Systems: A Survey	https://arxiv.org/pdf/2306.12680.pdf

なお、論文選定の過程において、詳細説明対象とせずに採用を見送った論文について表 4.4.1-2 に示す。

表 4.4.1-2 詳細説明対象としなかった論文

■可読性を考慮した難易度の定量化

#	論文タイトル/URL	不採用の理由
1	What Can Readability Measures Really Tell Us About Text Complexity? https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f0719d1b9a3a89fb50552403d2dd4b5c5c945fd6#page=19	テキストの複雑さを測定するために可読性指標と言語的特徴の関係について調査を実施した論文である。可読性指標として、Flesh Reading Ease score、Flesh-Kincaid readability formula、Fog Index、SMOG grading が利用されているが、これらの指標は英語を対象としており、日本語への対応が未確認であるため、採用を見送った。
2	Patent Claim Processing for Readability - Structure Analysis and Term Explanation - https://aclanthology.org/W03-2007.pdf	特許請求項の可読性を向上させるための手法について述べた論文である。特許請求項は、一文で長く複雑な構造を持ち、発明固有の用語や分野固有の用語が多く使われるため、一般の人にとって読みにくいという問題がある。これに対して、特許請求項の自動構造解析、用語説明部分の抽出及び言い換え等の手法を提案している。構文の複雑さが可読性に影響するという主張は採用論文と共通しており、採用論文の方がより具体的な手法が提案されていたため、本論文の採用は見送った。
3	日本語教育のための 文章難易度に関する研究 https://waseda.repo.nii.ac.jp/?action=repository_action_common_download&item_id=35923&item_no=1&attribute_id=162&file_no=1	日本語文章の可読性を定量化する手法として、リーダビリティと計量文体論の概念や歴史が紹介された論文である。日本語教育のためのリーダビリティ公式として、5 つの言語的特徴（平均文長、漢語率、和語率、動詞率、助詞率）の重回帰分析した結果が提案されている。採用論文と類似した主張であり、採用論文の方が手法の評価結果が具体的であるため、本論文の採用は見送った。

■事前学習済モデルによるテキスト特徴抽出/特徴ベクトルで検索者が経験した案件の傾向を表現

#	論文タイトル/URL	不採用の理由
1	A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques https://arxiv.org/pdf/1707.02919.pdf	テキストマイニングの基本的なタスクや手法について体系的に紹介した論文である。テキストの情報抽出手法として、固有表現抽出、隠れマルコフモデル、条件付き確率場、関係抽出の手法が述べられている。採用した論文の手法（事前学習済モデルによる特徴抽出）と比較して、隠れマルコフモデルや条件付き確率場等の統計的モデルは、タスク固有のデータからパラメータを学習する必要があるため、データが少ない場合やドメインが異なる場合は性能が低下しやすい。また、固有表現抽出や関係抽出は事前学習済モデルでも対応可能であるという点から本論文の採用は見送った。
2	Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey https://arxiv.org/pdf/2111.01243.pdf	事前学習済モデルを使用して、事前学習、ファインチューニング、プロンプティング、またはテキスト生成のアプローチによって自然言語処理タスクを解決する最近の研究を紹介した論文である。様々な事前学習済モデルが研究されているが、採用論文では特許公報データで事前学習したモデル（特許 DeBERTa）の有用性が見られることから、本論文の採用は見送った。
3	A literature review on the state-of-the-art in patent analysis https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=303568d0ce52c7597dec398218189ce6dfe4b129	特許分析に関するテキストマイニング技術と可視化技術を分類し、最新の研究動向を紹介した論文である。テキストマイニング技術としては、自然言語処理、プロパティ・ファンクション、ルールベース、セマンティック、ニューラルネットワークのアプローチ手法、可視化技術としては、特許マップや特許ネットワーク手法が挙げられている。特許分析に特化したテキストマイニング技術として有用と考えられるが、手法の紹介に留まり、性能比較は実施されていないことから、本論文の採用は見送った。

■高精度な分類アルゴリズム・ライブラリ

#	論文タイトル/URL	不採用の理由
1	Deep Neural Networks and Tabular Data: A Survey https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9998482	表形式データを対象とした深層ニューラルネットワークの手法を体系的に整理し、最新の研究動向を調査した論文である。深層ニューラルネットワークモデルとしては、表形式データを画像データ形式（2 次元行列）に変換する手法（SuperTML, IGTD 等）、古典的な機械学習と深層学習を融合させた手法（NODE、SDR 等）、Transformer ベースの手法（TabNet、TabTransformer 等）、学習パラメータに強力な正則化をかける手法（RLN、STG 等）が提案されている。本論文でもテストデータを用いた手法の性能比較が行われているが、採用論文の方がより広範な比較評価が行われているため、採用を見送った。
2	NCART: Neural Classification and Regression Tree for Tabular Data https://arxiv.org/pdf/2307.12198.pdf	表形式データを対象とした深層ニューラルネットワークの新手法として、Neural Classification and Regression Tree（NCART）が提案された論文である。NCART は、ResNet の全結合層を複数の微分可能な決定木に置き換えたモデルであり、決定木のような解釈可能性を維持する。20 個の表形式データセットに対して、代表的なアルゴリズムとの性能比較評価が行われ、LightGBM 等の決定木モデルと競合する性能を示した。採用論文の深層学習モデル（TabNet）と比較して、精度は同程度であるため、実装の実現性の観点から TabNet を採用し、本論文の採用は見送った。

■説明可能な AI

#	論文タイトル/URL	不採用の理由
1	Argumentative XAI: A Survey https://arxiv.org/pdf/2105.11266.pdf	計算論理学の分野で用いられる議論フレームワーク（AF）を使用して、人工知能の説明可能性（XAI）を支援する方法について調査した論文である。AF を用いた説明は、AI のモデルや手法に対して、議論の構造や関係性を示すことで、その出力や動作の理解を促進する。本論文の手法は、XAI の手法の延長線上の手法であり、まずは XAI を検証する必要性があるため、本論文の採用は見送った。
2	Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey https://arxiv.org/pdf/2006.11371.pdf	深層学習のモデルや手法に対して、解釈可能で直感的な人間の理解を促進するツール、技術、アルゴリズムを提供する説明可能な人工知能（XAI）の現状と課題について調査した論文である。論文の内容は概ね採用論文と重複しており、採用論文の方がより視覚的な説明がなされているため、本論文の採用は見送った。
3	A survey on XAI and natural language explanations http://ww.sentic.net/explainable-artificial-intelligence.pdf	説明可能な人工知能（XAI）と自然言語生成の技術に関する相互作用や連携について調査した論文である。本事業のデータ構造においては、数値型やカテゴリ型などの構造化されたデータが主であり、自然言語生成の技術との連携がなくとも、人間が直感的に理解できる XAI を提供可能と考えられるため、本論文の採用は見送った。

■マルチラベル分類

#	論文タイトル/URL	不採用の理由
1	A Review on Multi-Label Learning Algorithms https://romisatriawahono.net/lecture/rm/survey/machine%20learning/Zhang%20-%20Multi-Label%20Learning%20Algorithms%20-%202013.pdf	マルチラベル学習アルゴリズムについて、最先端の手法のタイムリーなレビューを提供することを目的とした論文である。本論文では、8 つの代表的な手法について、分析・考察が行われている。ここで述べられた手法は採用論文に包括されているため、本論文の採用は見送った。
2	A Survey on Extreme Multi-label Learning https://arxiv.org/pdf/2210.03968.pdf	マルチラベル学習アルゴリズムの中でも、ラベル空間のサイズが大きい場合の課題やアプローチについて体系的に整理した論文である。本事業においては、マルチラベルの対象となる検索オプションは 11 個であり、ラベル空間のサイズが大きい場合には当てはまらないため、本論文の採用は見送った。

■検索者のクラスタリング

#	論文タイトル/URL	不採用の理由
1	Survey of Clustering Algorithms https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=1763&context=ele_comeng_facwork	クラスタリングアルゴリズムについて体系的に整理した論文である。整理されたアルゴリズムは、採用論文の内容と重複しているものが多く、採用論文の方がより広範な整理がなされているため、本論文の採用は見送った。
2	Data Clustering: A Review https://dl.acm.org/doi/pdf/10.1145/331499.331504	クラスタリング手法の分類法を提示し、横断的なテーマと最新の進歩について整理された論文である。本論文はクラスタリングの歴史を含めて体系的に整理されているものの、1999 年の論文であり、採用論文の方がより最新の手法も提案されているため、本論文の採用は見送った。

■コンテンツベースフィルタリング/コールドスタート問題の対応/人材育成を考慮したレコメンド

#	論文タイトル/URL	不採用の理由
1	A SURVEY ON MODERN RECOMMENDATION SYSTEM BASED ON BIG DATA https://arxiv.org/pdf/2206.02631.pdf	レコメンドシステムの歴史及び最新のアプローチを包括的にレビューし、ビックデータに基づく最新の推薦システムの画期的な進歩について分析した論文である。一とおりの手法について述べられているが、採用論文と比較して、得られる情報が少ないため、本論文の採用は見送った。
2	Recommendation systems: Principles, methods and evaluation https://www.sciencedirect.com/science/article/pii/S1110866515000341	レコメンドシステムにおける様々な予測技術の特徴と可能性について調査した論文である。レコメンドシステムの手法として、コンテンツベースフィルタリング、協調フィルタリング、ハイブリッドフィルタリングの手法が整理されている。採用論文と概ね重複するため、本論文の採用は見送った。

3.4.2. 調査結果の解説

(1) 可読性を考慮した難易度の定量化

論文タイトル：日本語の難易度に関する特徴分析
URL：https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/C6-1.pdf

(a) 論文の概要

日本語文章の難易度を自動推定する手法の研究として、やさしい日本語と通常の日本語において、構文解析から得られる特徴量の統計量の比較がなされている。なお、やさしい日本語は外国人や小中学生向けに平易化した NHK 社による「NEWS WEB EASY[1]」の記事からテキスト抽出した情報が利用され、通常の日本語は NHK 社による「NEWS WEB[2]」の記事からテキスト抽出した情報が利用されている。

図 4.4.2.（1）-1 では、日本語難易度によって、「単語数」「漢字率」「サ変接続名詞率」「出現頻度最大値」「出現頻度平均値」「係り受け最大距離」の基本統計量に差があることが示されている。なお、補足として特徴量の作成方法は表 4.4.2.（1）-1 に示す。

特徴量	やさしい日本語				通常の日本語			
	平均値	最大値	最小値	標準偏差	平均値	最大値	最小値	標準偏差
単語数	24.281	60.000	6.000	8.362	38.471	285.000	3.000	17.096
漢字率	0.261	0.714	0.000	0.091	0.352	0.750	0.000	0.096
外来語率	0.029	0.250	0.000	0.038	0.021	0.429	0.000	0.030
受身率	0.001	0.100	0.000	0.007	0.009	0.125	0.000	0.016
サ変接続名詞率	0.023	0.188	0.000	0.032	0.068	0.500	0.000	0.045
副詞率	0.009	0.167	0.000	0.021	0.010	0.333	0.000	0.020
読点率	0.044	0.308	0.000	0.030	0.052	0.368	0.000	0.031
否定率	0.008	0.182	0.000	0.021	0.006	0.167	0.000	0.015
出現頻度最大値	3055.456	9046.000	8.000	2147.463	4720.032	9046.000	0.000	2127.299
出現頻度平均値	507.638	5145.125	4.333	386.579	707.079	3688.797	0.000	441.815
係り受け平均距離	1.990	4.125	1.000	0.565	2.454	7.324	0.000	0.797
係り受け最大距離	6.312	20.000	1.000	3.526	10.940	120.000	0.000	6.796
係り受け被修飾数	2.754	7.000	1.000	0.930	3.453	11.000	0.000	1.235

引用元：前川絵吏、村尾元(2021).「日本語の難易度に関する特徴分析」『言語処理学会 第 27 回年次大会 発表論文集』, c6-1

図 4.4.2.（1）-1 日本語難易度別の特徴量の統計量の違い

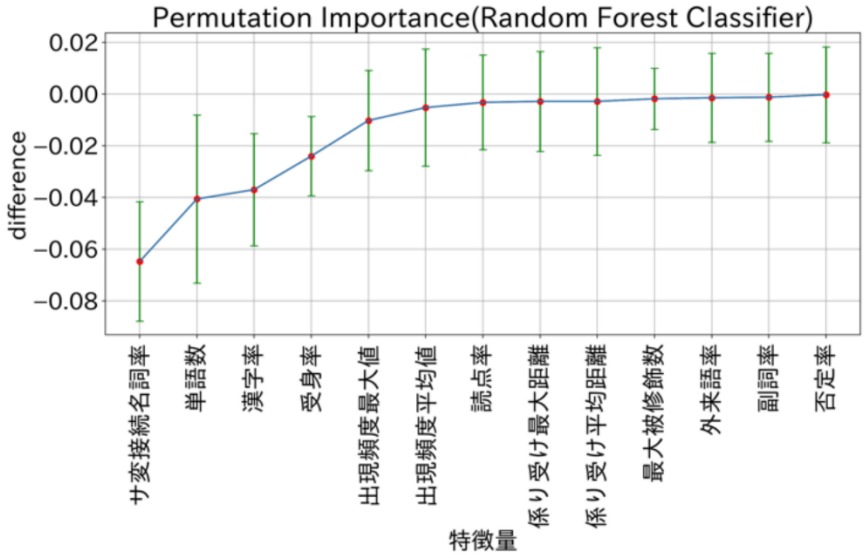
[1] https://www3.nhk.or.jp/news/easy/

[2] https://www3.nhk.or.jp/news/

表 4.4.2.（1）-1 補足_特徴量の作成方法

特徴量	作成方法
単語数	テキストに含まれる形態素の数
漢字率	テキストに含まれる漢字の個数を文字数で割った値
外来語率	全ての文字がカタカナである形態素の数を単語数で割った値
受身率	接尾語の「れる」「られる」の形態素の数を単語数で割った値
サ変接続名詞率	品詞がサ変接続名詞である形態素の数を単語数で割った値
副詞率	品詞が副詞である形態素の数を単語数で割った値
読点率	読点の個数を単語数で割った値
否定率	品詞が助動詞の「ない」「ぬ」「ん」の数を単語数で割った値
出現頻度最大値	品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の数の最大値
出現頻度平均値	品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の数の平均値
係り受け平均距離	1 文内で修飾-被修飾関係にある全ての文節間の係り受け距離の平均値
係り受け最大距離	1 文内で修飾-被修飾関係にある全ての文節間の係り受け距離の最大値
係り受け被修飾数	1 文内における全ての文節についての被修飾数の最大値

また、上記の特徴量を使用して、やさしい日本語と通常の日本語をランダムフォレストの手法で二値分類する AI モデルを構築し、どの特徴量が AI モデルの推定に寄与しているかを Permutation Importance にて測定している。ここで、Permutation Importance とは、AI モデルの推定に寄与する特徴量を評価する手法であり、各特徴量をランダムに並べ替えて確信度がどれだけ変化するかを評価することで重要度を算出する手法である。図 4.4.2.（1）-2 では、確信度の変化（difference）が大きい上位 5 項目として、「サ変接続名詞率」「単語数」「漢字率」「受身率」「出現頻度最大値」が示されている。



引用元：前川絵吏、村尾元(2021).「日本語の難易度に関する特徴分析」『言語処理学会 第 27 回年次大会 発表論文集』, c6-1

図 4.4.2.（1）-2 難易度判別に寄与する特徴量

(b) 本事業への適用

難易度別の基本統計量の違い及び難易度判別に寄与する特徴量の結果より、「サ変接続名詞率」「単語数」「漢字率」「出現頻度最大値」「係り受け最大距離」を特徴量として追加することを検証する。これらの特徴量を作成するためには、日本語文章に対する形態素解析及び係り受け解析が必要であり、それらの実施可否及びライセンス種別等について、表 4.4.2.（1）-2 にまとめている。本事業においては、単独で形態素解析及び係り受け解析が可能であり、かつ高度な自然言語処理が可能である GiNZA を採用する。

表 4.4.2.（1）-2 自然言語処理 Python ライブラリの比較

ライブラリ	ライセンス種別	形態素解析	係り受け解析	GitHub スター数	特徴
MeCab	GPL/LGPL/BSD	○	×	0.8k	C++で実装された高速な形態素解析エンジン。
CaboCha	LGPL/new BSD	○	○	0.1k	MeCab をベースにした係り受け解析エンジン。条件付き確率場を用いて係り受け関係を推定する。
KNP	Apache License 2.0	×	○	0.1k	JUMAN、Juman++の形態素解析結果を入力として、文節および基本句間の係り受け関係、格関係、照応関係を出力する。
GiNZA	MIT License	○	○	0.7k	spaCy をベースにした日本語処理ライブラリ。ニューラルネットワークを用いて形態素解析、固有表現抽出、依存構造解析、述語項構造解析等を行う。

(2) 事前学習済モデルによるテキスト特徴抽出/特徴ベクトルで検索者が経験した案件の傾向を表現

資料タイトル：特許庁における AI 技術の活用の現状と最新の取組
URL：https://www.jstage.jst.go.jp/article/jkg/73/7/73_256/_pdf/-char/ja

(a) 資料の概要

令和 4 年度「事前学習モデルに関する実証的研究事業」において、特許公報等で事前学習を行った事前学習モデル（以降、特許事前学習モデル）を構築し、特許関連タスクにおける精度調査が行われている。本資料では、BERT、ALBER、DeBERTa、Longformer の 4 手法において特許事前学習モデルが構築され、特許分類付与タスク及び類似文章ランキングタスクの精度比較がなされている。

図 4.4.2.（2）-1 及び図 4.4.2.（2）-2 では、特許分類付与タスクの microF 値、類似文章ランキングタスクの Recall@k はともに特許 DeBERTa が最もよい精度であることが示されている。

モデル	Precision		Recall		F 値	
	micro	macro	micro	macro	micro	macro
非特許 BERT	0.595	0.579	0.650	0.582	0.621	0.552
特許 BERT	0.503	0.531	0.728	0.679	0.595	0.562
特許 ALBERT	0.476	0.486	0.486	0.406	0.481	0.399
特許 DeBERTa	0.614	0.633	0.749	0.718	0.675	0.653
特許 Longformer	0.349	0.376	0.859	0.795	0.497	0.482

引用元：伊藤孝佑、久慈渉、後藤昌夫(2023).「特許庁における AI 技術の活用の現状と最新の取組」『情報の科学と技術』73（7）, pp.256-261

図 4.4.2.（2）-1 特許分類付与タスクにおける特許事前学習モデルの精度評価結果

モデル	Recall@k		
	10	100	1,000
BM25	0.019	0.091	0.258
非特許 BERT	0.022	0.108	0.372
特許 BERT	0.015	0.120	0.456
特許 ALBERT	0.001	0.019	0.160
特許 DeBERTa	0.028	0.163	0.474
特許 Longformer	0.024	0.135	0.397

引用元：伊藤孝佑、久慈渉、後藤昌夫(2023).「特許庁における AI 技術の活用の現状と最新の取組」『情報の科学と技術』73（7）, pp.256-261

図 4.4.2.（2）-2 類似文章ランキングにおける特許事前学習モデルの精度評価結果

(b) 本事業への適用

特許関連タスクにおいて特許 DeBERTa の精度優位性が示されていることから、特許 DeBERTa を第一候補として検証を実施する。本事業においては、AI モデル及びレコメンドアプローチにおいて、特許公報等のテキスト情報から埋め込みベクトルを抽出する際に特許 DeBERTa を利用する。(図 4.4.2. (2) -3)

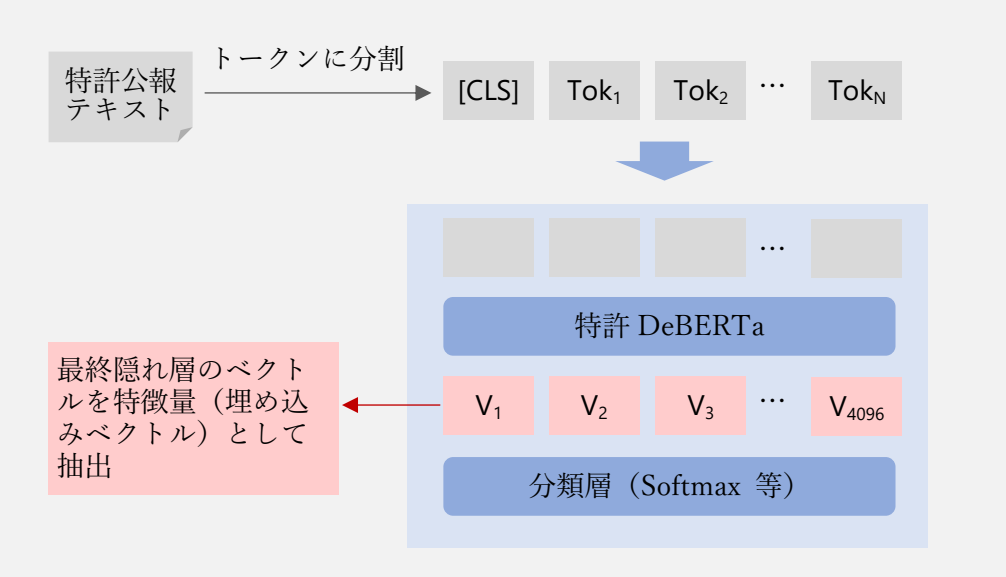


図 4.4.2. (2) -3 特許 DeBERTa の本事業への適用イメージ

(3) 高精度な分類アルゴリズム・ライブラリ

論文タイトル：TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED
URL：https://arxiv.org/pdf/2106.03253.pdf

(a) 論文の概要

表形式データの分類および回帰問題における、決定木ベースのモデルと深層学習モデルの比較が行われている。一般的に、これらの問題には決定木ベースのモデル（XGBoost 等）が用いられるが、近年、表形式データ用の深層学習モデルがいくつか提案されており、一部のユースケースでは XGBoost よりも優れたパフォーマンスを発揮すると主張されている。 本論文では、さまざまなデータセットで新しい深層学習モデルと XGBoost を比較し、これらの深層学習モデルが表形式データに本当に適しているかどうかを検討している。検証に用いたデータセットは、図 4.4.2.（3）-1 である。

Dataset	Features	Classes	Samples	Source	Paper
Gesture Phase	32	5	9.8k	OpenML	DNF-Net
Gas Concentrations	129	6	13.9k	OpenML	DNF-Net
Eye Movements	26	3	10.9k	OpenML	DNF-Net
Epsilon	2000	2	500k	PASCAL Challenge 2008	NODE
YearPrediction	90	1	515k	Million Song Dataset	NODE
Microsoft (MSLR)	136	5	964k	MSLR-WEB10K	NODE
Rossmann Store Sales	10	1	1018K	Kaggle	TabNet
Forest Cover Type	54	7	580k	Kaggle	TabNet
Higgs Boson	30	2	800k	Kaggle	TabNet
Shrutime	11	2	10k	Kaggle	New dataset
Blastchar	20	2	7k	Kaggle	New dataset

引用元：Shwartz-Ziv, R., and Armon, A. (2021). “Tabular Data: Deep Learning is Not All You Need.”

図 4.4.2.（3）-1 検討に用いたデータセット

上記データセットに対し、精度比較を行った結果が、図 4.4.2.(3)-2 である。

Model Name	Rossman	CoverType	Higgs	Gas	Eye	Gesture
XGBoost	490.18 ± 1.19	3.13 ± 0.09	21.62 ± 0.33	2.18 ± 0.20	**56.07**±0.65	80.64 ± 0.80
NODE	488.59 ± 1.24	4.15 ± 0.13	21.19 ± 0.69	2.17 ± 0.18	68.35 ± 0.66	92.12 ± 0.82
DNF-Net	503.83 ± 1.41	3.96 ± 0.11	23.68 ± 0.83	**1.44** ±0.09	68.38 ± 0.65	86.98 ± 0.74
TabNet	**485.12**±1.93	3.01 ± 0.08	**21.14**±0.20	1.92 ± 0.14	67.13 ± 0.69	96.42 ± 0.87
1D-CNN	493.81 ± 2.23	3.51 ± 0.13	22.33 ± 0.73	1.79 ± 0.19	67.9 ± 0.64	97.89 ± 0.82
Simple Ensemble	488.57 ± 2.14	3.19 ± 0.18	22.46 ± 0.38	2.36 ± 0.13	58.72 ± 0.67	89.45 ± 0.89
Deep Ensemble w/o XGBoost	489.94 ± 2.09	3.52 ± 0.10	22.41 ± 0.54	1.98 ± 0.13	69.28 ± 0.62	93.50 ± 0.75
Deep Ensemble w XGBoost	485.33 ± 1.29	**2.99** ±0.08	22.34 ± 0.81	1.69 ± 0.10	59.43 ± 0.60	**78.93** ±0.73
		TabNet			DNF-Net	

Model Name	YearPrediction	MSLR	Epsilon	Shrutime	Blastchar
XGBoost	77.98 ± 0.11	55.43±2e-2	11.12±3e-2	13.82 ± 0.19	20.39 ± 0.21
NODE	76.39 ± 0.13	55.72±3e-2	**10.39**±1e-2	14.61 ± 0.10	21.40 ± 0.25
DNF-Net	81.21 ± 0.18	56.83±3e-2	12.23±4e-2	16.8 ± 0.09	27.91 ± 0.17
TabNet	83.19 ± 0.19	56.04±1e-2	11.92±3e-2	14.94±, 0.13	23.72 ± 0.19
1D-CNN	78.94 ± 0.14	55.97±4e-2	11.08±6e-2	15.31 ± 0.16	24.68 ± 0.22
Simple Ensemble	78.01 ± 0.17	55.46±4e-2	11.07±4e-2	13.61±, 0.14	21.18 ± 0.17
Deep Ensemble w/o XGBoost	78.99 ± 0.11	55.59±3e-2	10.95±1e-2	14.69 ± 0.11	24.25 ± 0.22
Deep Ensemble w XGBoost	**76.19** ±0.21	**55.38**±1e-2	11.18±1e-2	**13.10**±0.15	**20.18**±0.16
		NODE		New datasets	

引用元：Shwartz-Ziv, R., and Armon, A. (2021). “Tabular Data: Deep Learning is Not All You Need.”

図 4.4.2.（3）-2 精度比較結果

結果、1 データセットにおいて深層学習モデルが決定木ベースのモデルの精度を上回ることはあるが、一貫して他のモデルを上回る深層学習モデルはないことを指摘している。また、深層学習モデルは、そのモデルが提案された論文に登場したデータセットでのみ、精度が優れていることを示している。総論として、決定木ベースのモデル（XGBoost 等）が、深層学習モデルを提案した論文で使用されたデータセットを含むデータセット全体で、深層学習モデルよりも優れたパフォーマンスを発揮することが示されている。

(b) 本事業への適用

本事業へ適用する決定木ベースのモデルの候補は、複数考えられる。選定のための観点として、精度に加え、本事業では扱うデータ数が多い(100 万件オーダー)ことから、学習・推論速度、メモリ使用量が挙げられる。また、検証の効率性の観点で、学習データの扱いやすさやハイパーパラメータの調整のしやすさ等も挙げられる。さらに、本事業ではモデル出力結果の説明性が求められることから、出力の解釈性も選定観点として重要である。

これらの観点をもとに、比較を行ったのが、表 4.4.2.（3）-1 である。選定の結果、LightGBM が最も適していると考え、本事業の検証で用いる。

表 4.4.2.（3）-1 アルゴリズム比較

アルゴリズム	精度	リソース使用量・学習速度	学習データの扱いやすさ	出力の解釈性
Decision Tree	×	○	○	○
Random Forest	△	×	○	△
XGBoost	○	×	×	△
LightGBM	○	△	○	△

精度の観点では、Decision Tree を改良した手法である XGBoost や LightGBM が、優れている。

リソース使用量・学習速度の観点では、Decision Tree が優れている。また LightGBM は、勾配ブースティング系の欠点であるリソース使用量や学習速度の増大に対処した手法であり、本観点で Random Forest や XGBoost と比較して優れている。

学習データの扱いやすさとは、各アルゴリズムがどれだけ容易にさまざまな種類のデータ（カテゴリカルデータ、テキストデータ、欠損値を含むデータ等）を取り扱えるか、また前処理が必要か否かという観点を意味している。特定の形式にデータを変換するための追加のステップが必要にならないという点で、Decision Tree、Random Forest、LightGBM が優れている。

解釈性の観点においては、素の実装の状態では、比較的シンプルなアルゴリズムである Decision Tree が優れている。一方 XGBosot や LightGBM に関しても、外部ライブラリを用いることで、説明性を向上させることが可能であることを確認済みである。

一方で、本事業では自然言語処理モデルを用いた文章の特徴ベクトルを用いることを考慮すると、成分間の関連などを学習するなど、深層学習モデルも一定の精度が期待できる。そこで、深層学習モデルについても学習・精度評価を行い、決定木ベースのモデルと並行して検証を行う。深層学習モデルを比較すると、まず各モデルが提案された論文で検証に用いられたデータセット以外のデータセットである、”New_datasets”に対する精度比較結果を確認すると、良い精度を示しているのは NODE と TabNet である。またこれらについて、NODE と TabNet の提案論文で検証に用いられたデータセットではないデータセットとして、DNF-Net の論文のデータセットに対する精度比較結果を確認すると、TabNet の精度がより優れていることが分かる。このことから、データセットに依らず精度が期待できる TabNet を、深層学習モデルとして本事業の検証で用いる。

論文タイトル：Benchmarking Multimodal AutoML for Tabular Data with Text Fields
URL：https://arxiv.org/pdf/2111.02705.pdf

(a) 論文の概要

本論文では、数値やカテゴリカルな列だけでなく、テキストフィールドも含まれたテーブルテーブルに対する、自動機械学習システム（以下 AutoML）の適用について考察が行われている。以下、数値やカテゴリカル変数に加え、テキスト情報など複数の形式を含むデータを、マルチモーダル表形式データと呼んでいる。論文内では、いくつかのテキストフィールドを含む、18 のマルチモーダル表形式データを用意している。

マルチモーダル表形式データのイメージを図 4.4.2.(3)-3 に示す。この例は、クラウドファンディングを行う Kickstarter という Web サイトにおいて、各プロジェクトが目標額を達成できたかに関するデータとなっている。具体的には、“name”列と“desc”列がテキスト、”goal”と“created_at”列が数値、”country”と”currency”がカテゴリカル形式となっており、”final_status”を予測するデータ

name	desc	goal	country	currency	created_at	final_status
The Secret Order - The Game that gives back Gl...	Can you trust your friends? Solve the puzzle? ...	5000.0	GB	GBP	1424101105	0
Booker Family Foods. Home made, the way food s...	Community based, home-made-foods producer, to ...	2500.0	US	USD	1404617242	0
J.A.E.S.A : Next Generation Artificial Intelli...	A true next generation AI with the ability to ...	30000.0	CA	CAD	1399078600	1

である。

引用元：X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola. “Benchmarking multimodal AutoML for tabular data with text fields.”

図 4.4.2.（3）-3 マルチモーダル表形式データの例

18 のマルチモーダル表形式データについて、図 4.4.2.（3）-4 に示す。#Cat.、#Num、#Text がそれぞれ、カテゴリカル、数値、テキスト形式の特徴量がいくつあるかを示している。

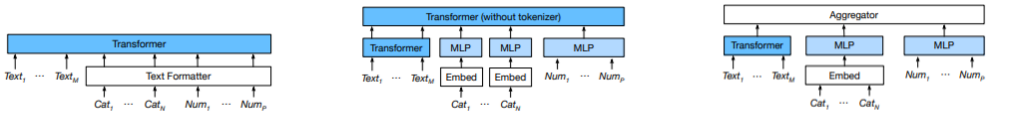
Dataset ID	#Train	#Test	#Cat.	#Num.	#Text	Task	Metric	Prediction Target
prod	5,091	1,273	1	0	1	multiclass	accuracy	sentiment associated with product review
salary	15,841	3,961	1	0	5	multiclass	accuracy	salary range in data scientist job listings
airbnb	18,316	4,579	37	24	28	multiclass	accuracy	price label of Airbnb listing
channel	20,284	5,071	1	15	1	multiclass	accuracy	news category to which article belongs
wine	84,123	21,031	0	2	3	multiclass	accuracy	which variety of wine (type of grape)
imdb	800	200	0	7	4	binary	roc-auc	whether film is a drama
fake	12,725	3,182	2	0	3	binary	roc-auc	whether job postings are fake
kick	86,502	21,626	3	3	3	binary	roc-auc	whether proposed Kickstarter project will achieve funding goal
jigsaw	100,000	25,000	2	27	1	binary	roc-auc	whether social media comments are toxic
qaa	4,863	1,216	1	0	3	regression	R^2	subjective type of answer (in relation to question)
qaq	4,863	1,216	1	0	3	regression	R^2	subjective type of question (in relation to answer)
book	4,989	1,248	1	2	5	regression	R^2	price of books
jc	10,860	2,715	0	2	3	regression	R^2	price of JC Penney products on their website
cloth	18,788	4,698	2	1	3	regression	R^2	customer review score for clothing item
ae	22,662	5,666	3	2	6	regression	R^2	price of American-Eagle inner-wear items on their website
pop	24,007	6,002	1	2	1	regression	R^2	online popularity of news article
house	37,951	9,488	1	18	20	regression	R^2	sale price of houses in California
mercari	100,000	25,000	3	0	6	regression	R^2	price of Mercari online marketplace products

引用元：X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola. “Benchmarking multimodal AutoML for tabular data with text fields.”

図 4.4.2.（3）-4 マルチモーダル表形式データの概要

これらのデータに対し、自然言語処理モデルを用いてテキストを特徴化し、その結果を用いてテーブル型データの AutoML を適用するという、標準的な 2 段階のアプローチを含む、様々なパイプラインを比較評価している。

マルチモーダル表形式データで分類/回帰を行う Multimodal-Net を図 4.4.2.（3）-5 に示す。図では省略されているが、各方法には最後段に 2 層の MLP が存在する。All-Text は、カテゴリカル、数値形式の特徴量を全てテキストとして扱う。Fuse-Early は、テキスト形式の特徴量を Transformer に、カテゴリカルと数値形式の特徴量を MLP に入力し、それぞれの出力を、後段の Transformer の入力とする。この際、カテゴリカル形式の特徴量については、各特徴量に埋め込みベクトルの取得と MLP からの特量ベクトル取得を行う。Fuse-Late は、Fuse-Early と反対に、カテゴリカル形式の特徴量をまとめて埋め込み層、MLP 層に入力する。また Fuse-Early では Transformer だった後段部分は、集約をするのみの Aggregator 層となる。

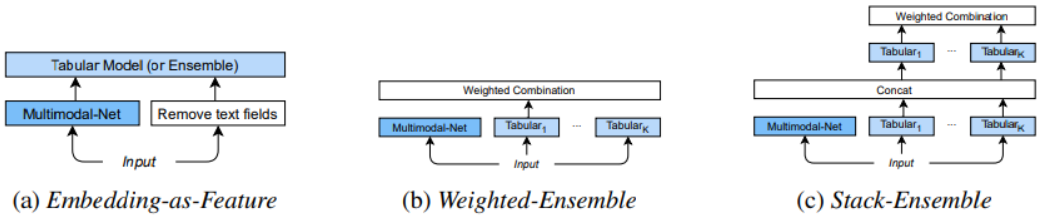


(a) *All-Text*. Convert numeric and categorical values into additional text tokens.
(b) *Fuse-Early*. Transformer operates on learned embeddings for each feature.
(c) *Fuse-Late*. Separate branches encode each modality, and aggregated via concatenation.

引用元：X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola. “Benchmarking multimodal AutoML for tabular data with text fields.”

図 4.4.2.（3）-5 マルチモーダル表形式データの分類・回帰を行うアルゴリズム詳細

上記の Multimodal-Net を特徴量抽出部として用いた、表形式データ分類モデルの全体アーキテクチャを図 4.4.2.（3）-6 に示す。Embedding-as-Feature は、Multimodal-Net から得た値を、そのまま後段の表形式データ用モデルが用いる特徴量とする。合わせて、表形式データ用モデルはテキストを直接扱えないため、テキスト形式の特徴量は削除したうえで、それ以外の特徴量も後段の表形式データ用モデルの特徴量とする。Weighted-Ensemble は、Multimodal-Net、及び複数の表形式データ用モデルを並列に用い、それぞれの出力の重みづけ平均をとる。Stack-Ensemble は、Weighted-Ensemble と同様に並列に出力を得た後、それらを結合する。そしてこの結合後の出力を特徴量として、さらに並列表形式データ用モデルから出力を得て、これら出力の重み付け平均をとる。



(a) *Embedding-as-Feature* (b) *Weighted-Ensemble* (c) *Stack-Ensemble*

引用元：X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola. “Benchmarking multimodal AutoML for tabular data with text fields.”

図 4.4.2.（3）-6 表形式データ分類モデルの全体アーキテクチャ

これら各手法について、精度比較を行った結果が、図 4.4.2.（3）-7 に示されている。Multimodal-Net の比較結果として、Fuse-Late が最も良い精度となった。また全体アーキテクチャの比較結果として、Stack-Ensemble が最も良い精度を示した。

Method	prod	qaq	qaa	cloth	airbnb	ae	mercari	jigsaw	imdb	fake	kick	jc	wine	pop	channel	salary	book	house	avg ↑	mrr ↑
Choosing Text-Net:	NLP Backbones and Fine-tuning Tricks (Section A.1)																			
RoBERTa	0.588	0.412	0.268	0.700	0.344	0.953	0.561	0.960	0.731	0.929	0.751	0.615	0.811	-0.000	0.301	0.396	0.151	0.821	0.572	0.07
ELECTRA	0.705	0.410	0.356	0.718	0.349	0.955	0.586	0.965	0.750	0.824	0.754	0.606	0.813	0.003	0.315	0.457	0.466	0.857	0.605	0.09
+ Exponential Decay $\tau = 0.8$	0.728	0.436	0.431	0.743	0.337	0.953	0.579	0.963	0.852	0.963	0.760	0.664	0.808	0.004	0.308	0.447	0.568	0.841	0.632	0.12
+ Average 3 ★	0.729	0.451	0.432	0.746	0.350	0.954	0.581	0.965	0.858	0.961	0.766	0.656	0.807	0.004	0.307	0.445	0.571	0.841	0.635	0.14
Choosing Multimodal-Net:	Fusion Strategy (Section A.2, Figure S2)																			
All-Text	0.907	0.454	0.419	0.746	0.366	0.957	0.599	**0.967**	0.840	0.967	**0.799**	0.645	0.810	0.013	0.480	0.465	0.585	0.892	0.662	0.28
Fuse-Early	**0.913**	0.441	0.418	0.745	0.377	0.953	0.596	**0.967**	0.843	0.960	0.770	0.653	0.806	0.013	0.474	0.458	0.548	0.901	0.658	0.21
Fuse-Late ★	0.907	0.449	**0.445**	0.747	0.395	0.958	0.603	0.966	0.857	0.961	0.773	0.639	0.812	0.015	0.481	0.468	0.571	0.907	0.664	0.22
Choosing Aggregation:	Multimodal Model Aggregation (Sections A.3 and A.4, Figure S3)																			
Pre-Embedding	0.895	0.216	0.247	0.642	0.449	0.972	0.433	0.586	0.871	0.926	0.743	0.491	0.680	0.012	0.526	0.460	0.581	0.939	0.593	0.11
Text-Embedding	0.867	0.446	0.432	0.748	0.430	0.972	0.434	0.587	0.855	0.962	0.790	0.658	0.830	0.008	0.502	0.438	0.594	0.932	0.638	0.17
Multimodal-Embedding	0.907	0.439	0.437	0.749	0.438	0.974	0.432	0.587	0.847	0.967	0.794	**0.683**	0.829	0.007	0.517	0.451	0.595	0.934	0.644	0.25
Weighted-Ensemble	0.907	0.439	0.429	0.744	0.453	0.976	0.597	0.957	0.876	0.923	0.787	0.641	0.814	0.018	0.554	0.483	0.620	0.941	0.676	0.25
Stack-Ensemble ★	0.909	**0.456**	0.438	**0.751**	0.459	0.977	**0.605**	**0.967**	**0.878**	0.964	0.797	0.624	0.836	**0.020**	**0.556**	0.496	**0.638**	**0.943**	**0.684**	**0.69**
	Tabular AutoML + Feature Engineering Baselines (Section A.3)																			
AG-Weighted	0.891	0.046	0.076	-0.002	0.426	0.841	0.098	0.587	0.845	0.686	0.668	0.004	0.173	0.016	0.549	0.226	0.222	0.934	0.405	0.08
AG-Stack	0.891	0.046	0.077	0.001	0.435	0.841	0.098	0.587	0.844	0.697	0.670	0.003	0.175	0.017	0.550	0.226	0.233	0.934	0.407	0.09
AG-Weighted+ N-Gram	0.892	0.426	0.382	0.610	0.450	0.978	0.526	0.909	0.842	0.966	0.772	0.357	0.829	0.019	0.546	0.484	0.591	0.941	0.640	0.17
AG-Stack+ N-Gram	0.895	0.414	0.383	0.654	**0.466**	**0.979**	0.569	0.915	0.850	**0.968**	0.775	0.612	**0.842**	**0.020**	0.548	0.494	0.600	0.943	0.663	0.43
H2O AutoML	0.869	0.247	0.159	0.163	0.329	0.976	0.430	0.531	0.813	0.756	0.669	0.411	0.478	0.014	0.530	0.525	0.444	0.939	0.516	0.11
H2O AutoML + Word2Vec	0.859	0.244	0.285	0.624	0.347	0.973	0.534	0.847	0.827	0.943	0.755	0.443	0.778	0.013	0.524	**0.528**	0.586	0.932	0.613	0.14
H2O AutoML + Pre-Embedding	0.846	0.227	0.312	0.644	0.367	0.969	0.282	0.572	0.874	0.893	0.738	0.549	0.571	0.007	0.483	0.483	0.523	0.933	0.572	0.09

引用元：X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola. “Benchmarking multimodal AutoML for tabular data with text fields.”

図 4.4.2.（3）-7 マルチモーダル表形式データセットにおける精度比較結果

また、データ分析コンペティションにおいて、同様の手法で高順位を残したことについても、論文で述べられている。このように、AutoML ツールを用いることで、複数モデル、または、それらの組み合わせを効率的に検証することが可能となり、精度の向上が期待できる。

(b) 本事業への適用

AutoML ツールは、どのツールも複数のモデルの学習やパラメータ最適化を実行することから、学習速度やリソース使用量には大きな差はないと想定される。

このような理由から、本事業への適用を考慮した AutoML ツールの選定においては、使用時の使いやすさや、対応しているモデル・アルゴリズムの多様さが重要な観点となる。具体的には、使用時の使いやすさとして実装の容易さ、また対応しているモデル・アルゴリズムの多様さとして、テキストデータへの対応や深層学習モデルの対応を観点とし、選定を行った。

候補としては、AutoML ツールとして長年使用されており実績のある auto-sklearn、TPOT、H2O、また歴史は浅いが大手 IT 企業が開発元であり、対応アルゴリズムの多様さなどで優位な ModelSearch、AutoGluon を挙げた。

選定の結果を表 4.4.2.（3）-2 に示す。

表 4.4.2.（3）-2 AutoML ツール比較

アルゴリズム	ライセンス	開発元	GitHub 公開日	実装の容易さ	深層学習モデル対応	テキストデータ対応
auto-sklearn	修正 BSD	フライブルク大学	2016 年 5 月	○	×	×
TPOT	LGPL-3.0 License	ペンシルベニア大学	2015 年 1 月	○	○	×
H2O	Apache License 2.0	H2O.ai	2014 年 3 月	×	○	○
ModelSearch	Apache License 2.0	Google	2021 年 2 月	○	○	×
AutoGluon	Apache License 2.0	AWS	2020 年 3 月	○	○	○

実装の容易さの観点で、H2O は Java で実装された GUI ベースのツールとなっており、コーディングを行うことなく複数モデルの学習を行うことが出来るのが特徴である一方、サーバを別途準備するなど環境構築の観点で難易度が高く、成果物の引継ぎも困難になることが想定される。この点において、Python のライブラリとして公開されている auto-sklearn、TPOT、ModelSearch、AutoGluon が優位である。

深層学習モデルの対応という観点では、auto-sklearn は Python の機械学習ライブラリである scikit-learn がベースとなっており、決定木や SVM、k 近傍法などのアルゴリズムが対応可能となっている一方、深層学習モデルについては未対応である。決定木などのモデル対応はどのツールも対応している中で、より多様なモデルが対応していることが望ましいという観点で、TPOT、H2O、ModelSearch、AutoGluon が適している。

テキストデータ対応の観点について、本事業の特性から、特許文献の本文から文章を抽出するなど、テキストデータを用いることが想定されるため、この観点が重要であると考える。H2O、AutoGluon がテキストデータに対応しており、これらが本事業の検証により適している。

以上の評価結果より、本事業では AutoGluon を検証に用いる。

(4) 説明可能な AI

論文タイトル：Explainable Artificial Intelligence for Tabular Data: A Survey
URL：https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9551946

(a) 論文の概要

本論文では、表形式データに対する説明可能な AI（XAI）の技術について包括的な調査が行われている。XAI は、AI モデルの予測を人間が理解できるように説明する技術のことであり、AI モデルの信頼性や責任性を向上させるために必要である。

図 4.4.2.（4）-1 に示すように、XAI で解決すべき問題は下記の 3 つになる。

・モデル説明 … AI モデルの内部構造やパラメータを人間が理解できるように説明する問題である。例えば、ニューラルネットワーク等の複雑な AI モデルは、その動作や判定基準がブラックボックス化されており、人間にとって不透明である。そのため、モデル説明問題を解決することにより、AI モデルの信頼性や責任性を向上させることができる。

・モデル検査 … AI モデルの品質や性能を評価する問題である。例えば、AI モデルの精度、ロバスト性、公平性及び倫理性等の指標を測定し、その強みや弱みを分析する。そのため、モデル検査問題を解決することにより、AI モデルの改善や最適化を行うことができる。

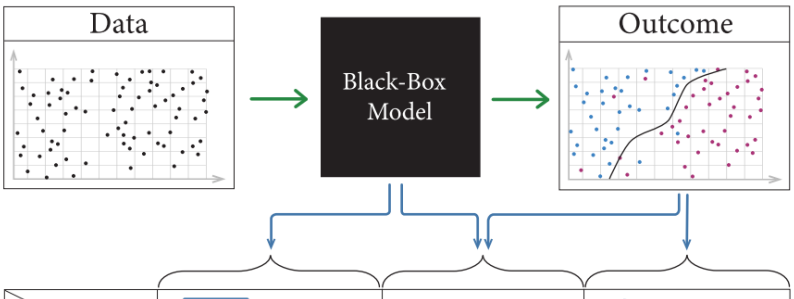
・予測説明 … AI モデルの予測結果を人間が理解できるように説明する問題である。例えば、AI モデルがある事例に対して行った予測結果について、その予測に影響した要因は何か、どの程度確信があるかを明らかにすることができる。そのため、予測結果説明問題を解決することにより、AI モデルの透明性や妥当性を確保することができる。

また、上記 3 つの問題に対して、2 つのアプローチによる解決がある。

・モデル特有の技術 … 特定の AI モデルに対応した XAI 技術であり、AI モデルの内部構造やパラメータを利用する。例えば、ニューラルネットワークにおいては、ニューロンや

重みの可視化、活性化マップの適用等の手法がある。この技術は、高い精度や効率性があるが、適用可能な AI モデルが限定されるため汎用性が低い。

・モデル非依存の技術 … 特定の AI モデルに限定せずに適用できる XAI 技術であり、AI モデルの入出力を分析して、その関係性や重要度を説明する。例えば、LIME や SHAP は、特徴量ごとに予測への寄与度を計算し、特徴量間の相互作用を考慮する。この技術は、高い汎用性や拡張性を持つが、計算コストが高くなる場合がある。



	Model Explanation	Model Inspection	Outcome Explanation
Model-Specific Techniques	Sections IV-B1 & IV-B2 [121-134], [135-148]	Section III-C [93-99, 108]	Sections II-A & II-B1 [37-42], [66-71]
Model-Agnostic Techniques	Section IV-A [24, 27, 34, 39], [110-120]	Sections III-A & III-B [83 - 86], [87-92]	Sections II-A & II-B2 [23-37], [76-82]

引用元：A. K. Singh, A. K. Singh, and A. K. Singh, “Explainable Artificial Intelligence for Tabular Data: A Survey,” IEEE Trans. Artif. Intell., vol. 2, no. 3, pp. 233-246, Sep. 2021

図 4.4.2.（4）-1 XAI の適用領域

（ア）予測説明のための XAI

予測説明のための XAI の技術は、グローバルな説明とローカルな説明に分けられる。グローバルな説明とは、AI モデルが全体的にどのような傾向やパターンを持っているかを示す説明であり、ローカルな説明とは、AI モデルが個別事例に対して、どのような判断や推論を行っているかを示す説明である。この分野における XAI 技術を図 4.4.2.（4）-2 に示す。この分野においては、LIME と SHAP が主流であり、その派生技術も多く登場している。

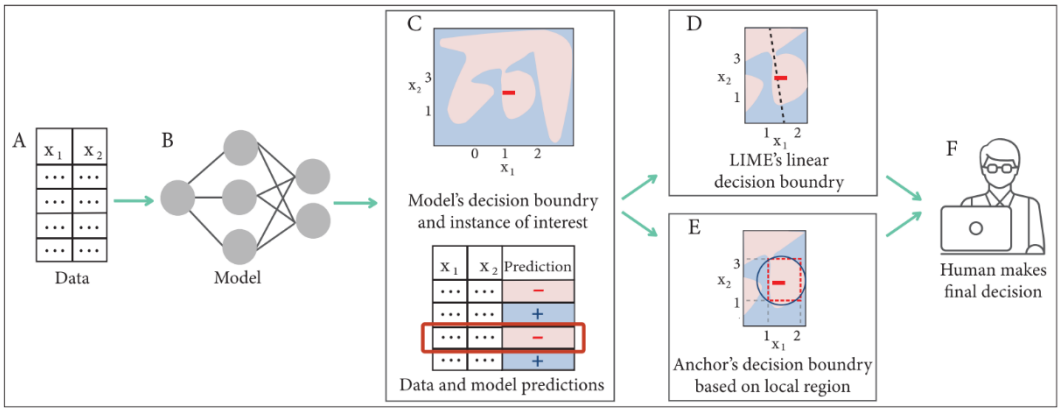
Name	Ref.	Year	Type	Underlying mechanism
ExplainD	[23]	2006	Agnostic	Graphical explanations based on additive models.
LIME	[24]	2016	Agnostic	Local linear models applied on random perturbations of the data and the predictions.
Anchors	[25]	2018	Agnostic	Reinforcement learning and graph search applied on random perturbations of the data and the predictions.
DLIME	[26]	2019	Agnostic	LIME combined with hierarchical clustering.
ILIME	[27]	2019	Agnostic	LIME, but with each perturbed instance weighed based on its influence on, and distance from, the instance to be explained.
ALIME	[28]	2019	Agnostic	LIME combined with an autoencoder.
Doctor XAI	[29]	2020	Agnostic	The idea behind LIME combined with a decision tree.
MAPLE	[30]	2018	Agnostic	Local linear modelling techniques combined with random forests.
LORE	[31]	2018	Agnostic	Genetic Algorithm.
LOCO	[32]	2018	Agnostic	Scoring of input features.
MES	[33]	2016	Agnostic	Monte Carlo algorithm.
MFI	[34]	2016	Agnostic	Feature ranking measure.
IME	[35]	2010	Agnostic	Coalitional game theory and the Shapley value.
-	[36]	2014	Agnostic	Coalitional game theory and sensitivity analysis.
Kernel SHAP	[37]	2017	Agnostic	LIME combined with the Shapley value.
DeepLIFT	[38]	2017	Specific	Assigning importance scores to inputs in a neural network.
Deep SHAP	[37]	2017	Specific	DeepLIFT combined with the Shapley value.
Tree SHAP	[39]	2020	Specific	SHAP, adopted for tree-based ML models.
Saabas	[40]	2014	Specific	Extracting explanations from a decision path in a decision tree.
-	[41]	2014	Specific	Transformation of backward models to forward models.
LionForest	[42]	2019	Specific	Explaining the decisions of random forests via unsupervised learning techniques and similarity metrics.

引用元：A. K. Singh, A. K. Singh, and A. K. Singh, “Explainable Artificial Intelligence for Tabular Data: A Survey,” IEEE Trans. Artif. Intell., vol. 2, no. 3, pp. 233-246, Sep. 2021

図 4.4.2.（4）-2 予測説明問題のための XAI 技術

（ア-1）LIME

Local Interpretable Model-Agnostic Explanations の略称であり、任意の AI モデルに対してローカルな説明を生成するために使われる XAI の技術である。LIME は、ある事例に対する予測への寄与度が高い特徴量を選択し、その特徴量を使って線形回帰モデル等の解釈可能なモデルを学習する。（図 4.4.2.（4）-3）そして、その解釈可能なモデルから予測への影響度や重み付けを算出し、グラフやテキストで表示する。LIME は、高い汎用性や拡張性があるが、計算コストが高く、精度や安定性が低い場合がある。LIME の派生 XAI 技術として、Anchors、DLIME、ILIME、ALIME、Doctor XAI 等がある。

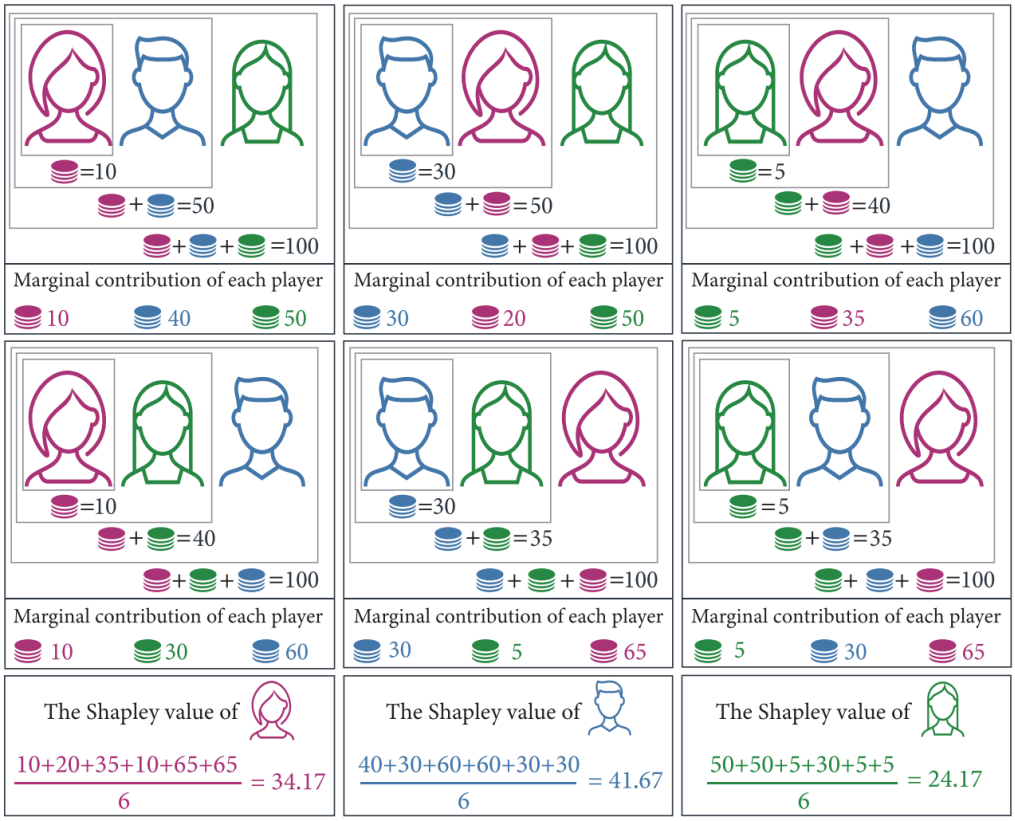


引用元：A. K. Singh, A. K. Singh, and A. K. Singh, "Explainable Artificial Intelligence for Tabular Data: A Survey," IEEE Trans. Artif. Intell., vol. 2, no. 3, pp. 233-246, Sep. 2021

図 4.4.2.（4）-3 LIME の概要

（ア-2）SHAP

SHapley Additive exPlanations（シャプレー加法的な説明）の略称であり、任意の AI モデルに対してローカルな説明を生成するために使われる XAI の技術である。SHAP は、ゲーム理論のシャプレー値という概念を用いて、ある事例に対する予測への寄与度が高い特徴量を選択し、その特徴量の組み合わせによる予測への影響度や重み付けを算出する。（図 4.4.2.（4）-4）そして、その影響度や重み付けをグラフやテキストで表示する。SHAP は、高い精度や一貫性があるが、計算コストが高く、特徴量間の相互作用を考慮しない場合がある。SHAP の派生 XAI 技術として、Kernel SHAP、Deep SHAP、Tree SHAP 等がある。



引用元：A. K. Singh, A. K. Singh, and A. K. Singh, “Explainable Artificial Intelligence for Tabular Data: A Survey,” IEEE Trans. Artif. Intell., vol. 2, no. 3, pp. 233-246, Sep. 2021

図 4.4.2.（4）-4 SHAP の概要

（イ）モデル検査のための XAI

モデル検査のための XAI の手法には、感度分析、部分依存プロット、モデル固有の手法がある。

（イ-1）感度分析

モデルの入力変数の値を少しずつ変化させ、出力にどのような影響があるかを調べる方法である。例えば、画像認識モデルに対して、画像の一部を隠したりぼかしたりして、どの部分が重要な特徴として認識されるかを見ることができる。

（イ-2）部分依存プロット

モデルの入力変数の一部を固定して、他の変数を変化させたときに、出力がどう変わるかをプロットする方法である。例えば、住宅価格予測モデルに対して、住宅の広さや立地などの変数を固定して、築年数や間取りなどの変

数を変化させて、価格にどのような影響があるかを見ることができる。

（イ-3）モデル固有の手法

特定の種類のモデルに対して適用できる分析手法である。例えば、決定木やニューラルネットワーク等のモデルは、その構造やパラメータを直接観察したり可視化したりすることができる。これにより、モデルがどのような判断基準や特徴量を学習しているかを理解することができる。

（ウ）モデル説明のための XAI

モデル説明のための XAI の手法には、モデル非依存の手法とモデル固有の手法がある。

（ウ-1）モデル非依存の手法

予測説明問題のための XAI 技術（LIME や SHAP 等）はローカルな説明だけではなく、AI モデル全体の傾向を説明可能なグローバルな説明も提供することができる。

（ウ-2）モデル固有の手法

特定の AI モデルに対して適用できる手法である。例えば、Attention 等の手法は、ニューラルネットワークの中間層や出力層の活性化を可視化することで、モデルがどの部分に注目して予測を行っているかを示すことができる。

(b) 本事業への適用

本事業においては、予測結果を出力する際にモデルの判断根拠を同時提供することにより、AI モデルの透明性や妥当性を確保する。予測説明のための XAI 技術が活用可能な Python ライブラリの比較表を表 4.4.2.（4）-1 に示す。説明性能や計算速度の面において、LIME より SHAP が優れていると評価されており、最近では SHAP が XAI ライブラリとして標準的になっており、本事業においても SHAP を適用する。

表 4.4.2.（4）-1 予測説明のための XAI Python ライブラリの比較

観点	LIME	SHAP
ライセンス種別	BSD 3-Clause License	MIT License
特徴	任意の機械学習モデルの個別の予測を局所的に近似する単純なモデルを生成し、そのモデルの重みから特徴量の寄与度を説明する。 近傍の定義や重み付け等に依存するため、安定性や一貫性に欠けるという欠点がある。	協力ゲーム理論のシャプレー値を機械学習に応用し、任意の機械学習モデルの出力を特徴量の寄与度の和として分解し、説明する。 LIME と違って一貫性や効率性などの性質を満たすことが理論的に保証されている。
GitHub スター （2023/10/04 時点）	10.9k	20.2k

(5) マルチラベル分類

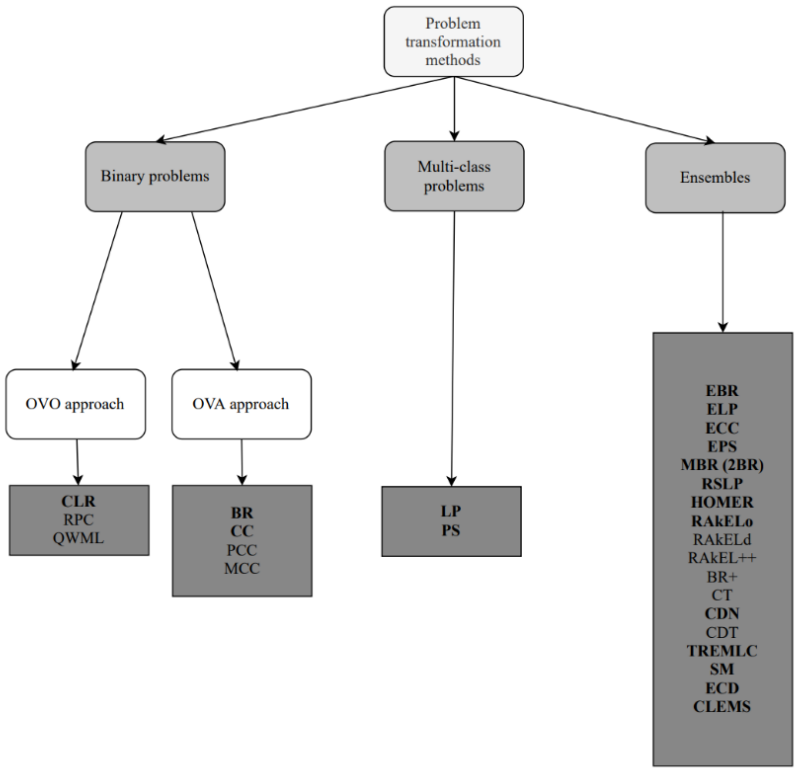
論文タイトル：Comprehensive Comparative Study of Multi-Label Classification Methods
URL：https://arxiv.org/pdf/2102.07113.pdf

(a) 論文の概要

1 つのデータに対して、複数の正解ラベルが同時に付与されうるデータをマルチラベルデータといい、マルチラベルデータの分類問題への対応方法として、問題変換手法とアルゴリズム適応手法の 2 種類が挙げられる。

（ア）問題変換手法

問題変換手法は、マルチラベル分類の問題をより単純なシングルラベルの問題に変換する手法であり、二値分類や多値分類に変換するというアイデアとそれらをアンサンブルするというアイデアの 3 つに大別される。（図 4.4.2.（5）-1）



引用元：J. Bogatinovski et al., “Comprehensive Comparative Study of Multi-Label Classification Methods,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 10, pp. 2345-2367, Oct. 2023.

図 4.4.2.（5）-1 マルチラベル問題における問題変換手法の体系的整理

（ア-1）二値問題への変換

二値問題への変換ではマルチラベル分類問題をラベル個数分の二値分類問題に変換し、各ラベルに対して 1 つの二値分類器を学習する。この方法の利点は、学習データに登場しないラベルの組み合わせに対しても予測することができる点である。欠点は、ラベル間の依存関係を考慮できないことや、ラベル数が多い場合に学習に時間がかかることである。この方法の代表的な手法には下記がある。

・BR … Binary Relevance。ラベルごとに 1 つの二値分類器を学習する。この方法は単純で汎用性が高いが、ラベル間の依存関係を考慮しないため、予測精度が低下する可能性がある。

・CLR … Calibrated Label Ranking。ラベルのペアごとに二値分類器を構成し、それぞれのラベルに対して重み付きマージンロスを最小化するよう学習する。この方法は、各ラベルのランキングと二値分類を同時に生成できるが、ラベル数が多い場合は計算コストがかかる。

・CC … Classifier Chains。BR と類似しているが、ラベルごとに二値分類器を学習する際に、前のラベルの予測結果も特徴量として追加する。この方法は、ラベル間の依存関係をある程度反映できるが、ラベルの順序に影響されやすい。

（ア-2）多値問題への変換

多値問題への変換ではマルチラベル分類問題を 1 つあるいは複数個の多値分類問題に変換する方法である。この方法では、各サンプルに対応するラベルの組み合わせを別々のクラスとみなし、多値分類器を学習する。この方法の利点は、ラベル間の依存関係を考慮できることである。欠点は、学習データに存在しないラベルの組み合わせを予測できないことや、ラベルの組み合わせの数が多い場合に学習が困難になることである。この方法の代表的な手法には下記がある。

・LP … Label Powerset。マルチラベル問題を直接多値分類問題に変換し、任意の多値分類器を適用する。この方法は単純でラベル間の依存関係を考慮できるが、見たことのないラベル組合せを予測できないため、汎化性能が低い可能性がある。

・PSt … Pruned Sets。LP の欠点を緩和するために提案された方法であ

り、出現頻度の低いラベル組合せを学習データから除外し、その代わりに頻度の高いラベル組合せから部分集合をサンプリングして再導入する。これにより、学習データに存在しないラベル組合せも予測できるようになる。

（ア-3）問題変換手法のアンサンブル

複数の問題変換手法を組み合わせてマルチラベル問題に対処するアンサンブルな手法として、代表的な手法には下記がある。

・MBR … Meta Binary Relevance。スタッキングと呼ばれる手法を用いて、Binary Relevance（BR）を 2 段階に適用する方法である。まず、BR と同様に、各ラベルに対して二値分類器を学習し、その出力を元の特徴量に追加して、再び各ラベルに対する二値分類器を学習する。この結果、MBR は BR が無視していたラベル間の依存関係を考慮することができ、BR の欠点を緩和しているが、ラベル数が多い場合に計算コストがかかるという欠点は引き継ぐ。

・RakEL … RAndom k-labELsets。Label Powerset（LP）の欠点を緩和するために提案された方法である。この方法は、ラベル集合をランダムに k 個の部分集合に分割し、各部分集合ごとに LP を適用して多値分類器を学習する。次に、各分類器の出力を多数決で統合する。この方法は、ラベル間の依存関係を部分的に考慮できるが、部分集合のサイズや数に影響される。

・ TREMLC … Tree-based Ensemble Method for Multi-label Classification。Classifier Chains（CC）の欠点を緩和するために提案された方法である。この方法は、CC のアンサンブルであり、各分類器の学習時にラベルの順序をランダムに変えることで、順序への依存性を低減する。また、各分類器の予測結果を重み付き多数決で統合することで、予測精度を向上させる。

・SM … Stacking Method。問題変換法とアルゴリズム適応法を組み合わせた方法である。この方法は、まず問題変換法でラベルごとに二値分類器を学習し、その予測結果を新しい特徴として追加する。次に、アルゴリズム適応法で多値分類器を学習する。この方法は、二段階の学習でラベル間の依存関係を考慮できるが、学習時間が長くなる可能性がある。

（イ）アルゴリズム適応手法

アルゴリズム適応手法は、シングルラベルの分類アルゴリズムをマルチラベルに対応させるために修正する方法である。例えば、木構造、ニューラルネットワーク及び SVM などの分類器がある。この方法の利点は、ラベル間の依存関係を考慮できることである。欠点は、特定のアルゴリズムに依存することや汎化性能が低いことである。代表的な手法には下記がある。(図 4.4.2. (5) -2)

・ANN … Artificial Neural Network。ニューラルネットワークをマルチラベル分類に適応するため、出力層で各ラベルに対してシグモイド関数やソフトマックス関数等の活性化関数を適用し、各ラベルの確率を出力する。例えば、BPNN や ML-RBF 等が該当する。

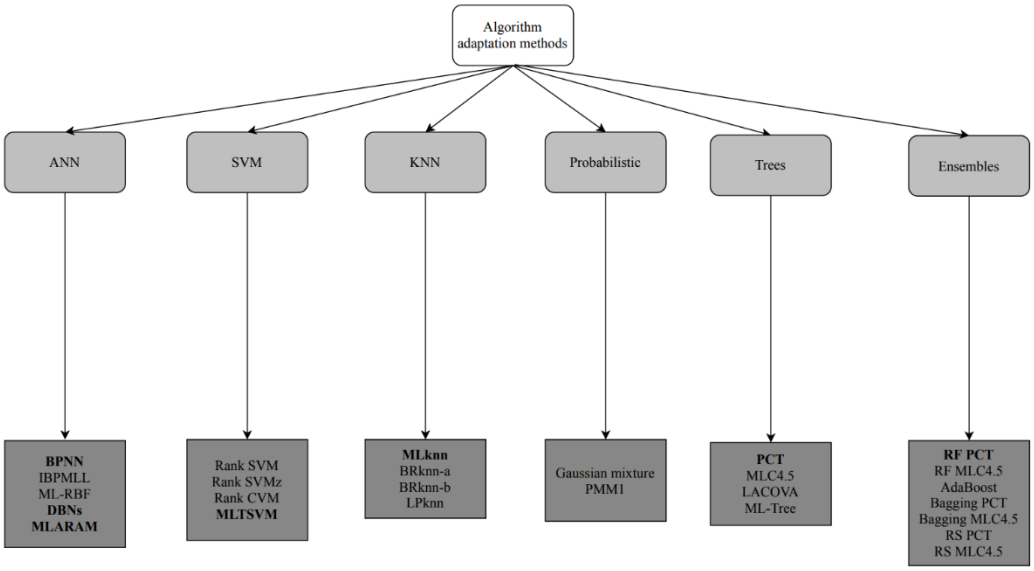
・SVM … Support Vector Machine。マージン最大化とカーネルトリックという 2 つの原理に基づく分類器であり、マルチラベル分類に適応するため、ラベルごとに二値分類器を作り、各ラベルに対して SVM を学習する。また、予測時には閾値関数を導入することで、新しいラベル組み合わせを生成できるようにする。例えば、Rank-SVM や MLTSVM 等が該当する。

・kNN … k-Nearest Neighbor。インスタンスベースの分類器であり、学習データをそのまま保存し、予測時には新しいサンプルと最も近い k 個のサンプルを探してそのラベルを参考にする。マルチラベル分類に適応するため、距離関数やラベル集合の統合方法を工夫する必要がある。例えば、ML-kNN や BR-kNN 等が該当する。

・Probabilistic … 確率的なモデルを用いてラベルの分布や依存関係を表現する方法である。例えば、ガウス混合モデルや PMM1 等が該当する。

・Trees … 木構造を用いて分類問題を解く方法であり、各ノードで特徴やラベルに基づいて分割条件を決める。マルチラベル分類に適応するため、各ノードで複数のラベルを予測し、分割条件を決める際にラベル間の依存関係を考慮する。例えば、PCT や ML-Tree 等が該当する。

・Ensembles …複数の分類器を組み合わせて予測精度や安定性を向上させる方法である。例えば、RF-PCT や Bagging-PCT 等が該当する。



引用元：J. Bogatinovski et al., “Comprehensive Comparative Study of Multi-Label Classification Methods,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 10, pp. 2345-2367, Oct. 2023.

図 4.4.2.（5）-2 マルチラベル問題におけるアルゴリズム適応手法の体系的整理

マルチラベル分類においては、ラベルの依存性と高次元のラベル空間という 2 つの特性に対処する必要がある。それぞれの特性への対処方法として、下記がある。

（ウ）ラベルの依存性

マルチラベル分類では、ラベル間に依存関係がある場合がある。例えば、1 つの画像に「犬」と「猫」の両方のラベルが付与される可能性は低いが、「犬」と「首輪」のラベルが同時に付与される可能性は高い。このようなラベルの依存性を考慮することにより、予測精度を向上させることができる。ラベルの依存性を考慮する方法として、下記がある。

・ラベルセットを 1 つのクラスとして扱う方法（LP、PSt 等）
・ラベルセットを分割して複数のクラスにする方法（RakEL 等）
・ラベルセットを埋め込み空間に写像する方法
・ラベルごとに個別に学習後、出力結果を組み合わせる方法（MBR、CC 等）

（エ）高次元のラベル空間

マルチラベル分類では、ラベル数が多くなると、学習や予測に必要な計算コストが増大する。また、各ラベルに対応するデータが少なくなり、過学習や未

学習のリスクが高まる。高次元のラベル空間に対処する方法として、下記がある。

・ラベル空間を削減する方法（PSt 等）
・ラベル空間を分割する方法（RAkEL 等）
・ラベル空間を埋め込む方法
・ラベル空間をサンプリングする方法（TREMLC 等）

(b) 本事業への適用

本事業のデータセットにおいては、検索オプションの付与がマルチラベル分類の対象となる。検索オプションの各ラベルの依存関係はどの程度であるのかは、「貸与データの分析」にて詳細に検証する必要があるものの、プレアンケートの結果から一部のラベルにのみ依存関係は限定されると想定する。また、ラベルの次数については、全 11 検索オプションであることから、ラベル空間は高次元ではない。これより、マルチラベル問題への対処方法として、複雑な手法を適用するのではなく、シンプルな手法で効率的にモデル構築を実施することを優先し、問題変換手法のうちの BR 法、アルゴリズム適応手法のうちの Trees（ランダムフォレスト 等）を本事業に適用する。

(6) 検索者のクラスタリング

論文タイトル：A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects

URL：https://bibbase.org/service/mendeley/bfbbf840-4c42-3914-a463-19024f50b30c/file/3bf7d292-fec6-2dc7-20f4-c81abae91b5e/1_s20_S095219762200046X_main.pdf.pdf

(a) 論文の概要

本論文では、クラスタリング手法を体系的に整理している。(図 4.4.2.(6)-1) まず、クラスタリングは階層的クラスタリングと分割的クラスタリングに大別される。

(ア) 階層的クラスタリング

階層的クラスタリングはデータを階層的な構造に分割する手法であり、データを個別から集合へ統合する手法 (凝集法) と集合から個別へ分割する手法 (分割法) に分けられる。本手法は大規模なデータセットに対して計算コストが大きくなる課題やノイズや外れ値への感受性の課題がある。

(ア-1) 凝集法

凝集法は、各オブジェクトを単一クラスタとして始め、最も類似したクラスタを反復的に統合していき、単一のクラスタまたは停止基準に達するまでクラスタリングを行う、ボトムアップのアプローチである。統合プロセスにおいて、クラスタ間の類似度を計測する尺度として、最短距離法、最長距離法、平均距離法が用いられる。凝集法は任意の類似度尺度や属性タイプに対応できるが、計算コストが大きく、ノイズや外れ値に敏感であるという欠点がある。凝集法のアルゴリズム例としては、SLINK, CLINK, CURE, Ward's method 等がある。

(ア-2) 分割法

分割法は、凝集法とは逆に、データをトップダウンで分割していく手法である。最初に全てのデータを 1 つのクラスタとし、必要な数のクラスタになるまで、クラスタを小さく分けていく。分割プロセスでは、クラスタ内のデータ間の類似度や距離を基準にする。分割法は、単一特徴法と多重特徴法の 2 種類がある。単一特徴法は、1 つの変数に基づいてクラスタを分割し、多

重特徴法は、複数の変数に基づいてクラスタを分割する。分割法は大規模なデータや高次元データに対応できるが、計算コストが大きくなるという課題がある。分割法のアルゴリズム例としては、bisecting K-Means, DIVFRP, IDPSO, HGCUDF 等がある。

（イ）分割的クラスタリング

分割的クラスタリングは、データを互いに排他的なクラスタに分割する方法であり、データ点が 1 つのクラスタのみに属するハードクラスタリングとデータ点が複数のクラスタに属することを許容するファジークラスタリング、データを複数の確率分布の混合としてモデル化する混合モデルに分けられる。分割的クラスタリングは、事前にクラスタ数を指定する必要があることや、初期化や距離尺度に依存する課題がある。

（イ-1）ハードクラスタリング

ハードクラスタリングは、データ点が 1 つのクラスタのみに属することを前提とする分割的クラスタリングの手法であり、下記の 6 つの主要なカテゴリが存在する。

・検索ベース … 最適化手法やメタヒューリスティック手法を用いて、クラスタ中心やクラスタ割り当てを探索する。アルゴリズム例としては、K-Means、K-Medoids、K-Modes 等がある。

・グラフ理論ベース … データ点を頂点、類似度を辺としたグラフを構築し、グラフの分割や最小カット等の操作でクラスタを形成する。アルゴリズム例としては、Spectral Clustering や CLICK 等がある。

・密度ベース … データ空間における密度の高い領域をクラスタとして抽出し、密度の低い領域や離れた点を外れ値として除外する。アルゴリズム例としては、DBSCAN や OPTICS 等がある。

・モデルベース … データ点がある確率分布や関数から生成されたと仮定し、尤度や情報量などの基準で最適なクラスタモデルを選択する。アルゴリズム例としては、Gaussian Mixture Model や Self-Organizing Map 等がある。

・部分空間ベース … 高次元データに対応するために、各クラスタが異な

る部分空間に存在すると仮定し、有効な特徴集合や射影を探索する。アルゴリズム例としては、CLIQUE や PROCLUS 等がある。

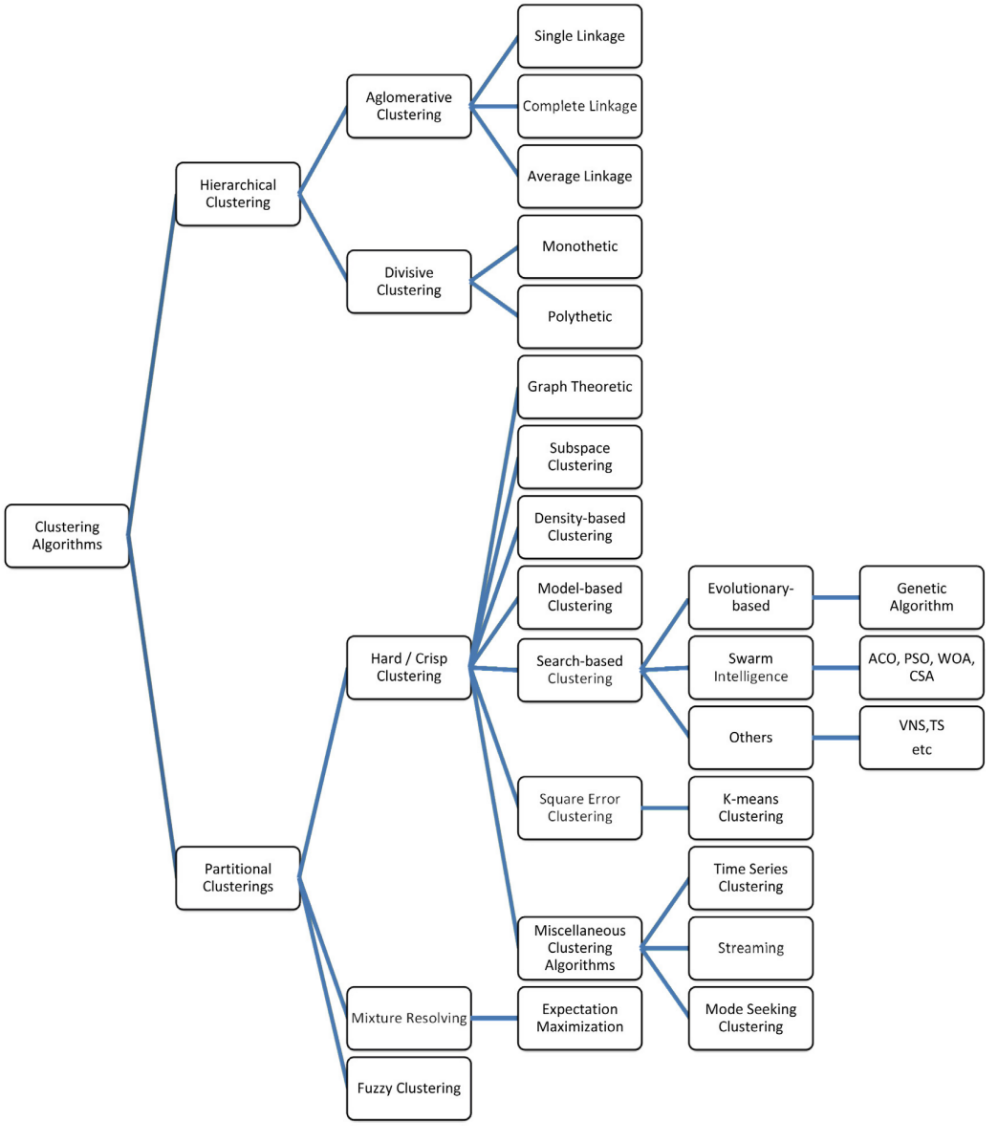
・その他 … 上記のカテゴリに分類されないアルゴリズム例として、Grid-Based Clustering や Co-Clustering 等がある。

（イ-2）混合モデル

混合モデルは、データを複数の確率分布の混合としてモデル化する方法であり、混合モデルのパラメータを推定するために、最尤推定法やベイズ推定法などの統計的手法が用いられる。本手法は、データの潜在的な構造やクラスタ数を自動的に発見できるという利点がある。また、混合モデルには、ガウス混合モデルやディリクレ過程混合モデルなど、さまざまな種類のモデルを用いることができる。

（イ-3）ファジークラスタリング

ファジークラスタリングは、データ点が複数のクラスタに属する可能性があることを許容する分割的クラスタリングの一種である。ファジークラスタリングは、データ点が各クラスタに対して持つ所属度を計算し、所属度の高いクラスタを割り当てる。所属度は、距離や密度等の基準に基づいて決定される。ファジークラスタリングは、データの曖昧さや不確実さに対応できるという利点があり、さらに、データの潜在的な関係や構造をより柔軟に表現できる。アルゴリズム例としては、Fuzzy C-Means や Fuzzy C-Medoids 等がある。



引用元：Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2020a). “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects.” Engineering Applications of Artificial Intelligence, 110, 104743.

図 4.4.2.（6）-1 クラスタリング手法の体系的整理

(b) 本事業への適用

登録調査機関に属する約 3,800 者の検索者を数十~数百程度のグループにクラスタリングし、検索者への案件割り振りの判定ロジックを簡易化する。クラスタリング手法としては、作成されたクラスタの解釈性を重視し、分割法による階層的クラスタリングの考え方を適用する。本事業への適用イメージを図 4.4.2.（6）-2 に示す。

検索者 ID	機関	区分	経験年数	…
XXX1	A 社	1	3 年	…
XXX2	A 社	1	5 年	…
XXX3	A 社	1	1 年未満	…
XXX4	A 社	2	1 年未満	…
XXX5	B 社	1	3 年	…
XXX6	B 社	2	3 年	

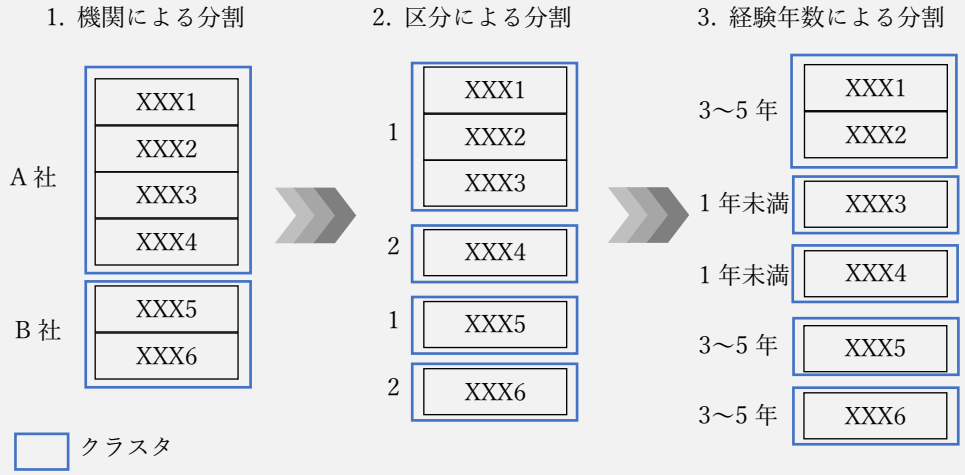


図 4.4.2.（6）-2 検索者に対する分割法による階層的クラスタリングの適用イメージ

(7) コンテンツベースフィルタリング/コールドスタート問題の対応/人材育成を考慮したレコメンド

論文タイトル：Recent Developments in Recommender Systems: A Survey
URL：https://arxiv.org/pdf/2306.12680.pdf

(a) 論文の概要

レコメンドシステムの手法として、協調フィルタリングベース、コンテンツベース、知識ベース及びそれらのハイブリッドのアプローチについて言及されている。

（ア）協調フィルタリング（CF）ベースレコメンドシステム

レコメンドシステムでよく使われる手法であり、多数のユーザの集合知を活用することにより、ユーザの嗜好や意見を予測する。本手法は、同じような嗜好や行動を共有するユーザは、同じような意見を持つという考え方に基づいている。CF の手法は、メモリベース CF とモデルベース CF に大別されるが、近年ではコンテキストを考慮した CF も注目されている。

（ア-1）メモリベース CF

メモリベース CF はユーザやアイテム間の類似性を利用して推薦を行う手法であり、下記の 2 つのステップで構成される。

1. ユーザ又はアイテム間の類似度を最近傍法に基づいた手法で計算
2. ソートされた類似度を用いて、ユーザへの推薦を生成

メモリベース CF はシンプルなアプローチではあるが、ユーザやアイテムが大規模になった場合、これらの相互作用行列の計算コストが課題となる。相互作用行列は高次元の疎行列であるため、埋め込み技術を使用して、低次元の密行列へ変換を行い、計算コストの問題を解決する手法が提案されている。

（ア-2）モデルベース CF

モデルベース CF はユーザとアイテムの関係からユーザの嗜好を予測する手法であり、近年ではレコメンドシステムにニューラルネットワークを応用した手法が提案されている。モデルベース CF の中でも、ニューラルベース CF では、ユーザとアイテムの関係をニューラルネットワークでマッピング

し、出力を因数分解法で精緻化する。また、因数分解機（FM）は、ユーザとアイテムの相互作用を効果的にモデル化することができ、多くのレコメンドシステムにおいて不可欠な要素となっている。

（ア-3）コンテキストを考慮した CF

コンテキストを考慮した CF は、時間や場所、気分、天気等のコンテキスト情報を考慮して、より正確な推薦を提供する手法である。コンテキスト情報を利用する方法としては、事前フィルタリング法とコンテキストモデリング法がある。

事前フィルタリング法は、推薦に関係のないデータを除外してからコンテキストをレコメンドシステムに入力する手法であり、コンテキストモデリング法は、深層学習やグラフニューラルネットワーク等の技術を使用して、コンテキストをレコメンドシステムに直接統合する手法である。

（イ）コンテンツベースレコメンドシステム

コンテンツベースレコメンドシステムは、アイテムの特徴に基づいて分類器を学習することにより、推薦をユーザ固有の分類問題として扱い、より個人的な推薦を提供する手法である。本手法の利点は、アイテムに対する評価は必要なく、アイテムに含まれるコンテンツ情報のみに基づいて推薦を行うことができる点である。また、本手法においてはアイテムの説明文等のテキストデータが重要な情報源であり、アイテムの特徴抽出のため、単語埋め込み技術が使用される。

（ウ）知識ベースレコメンドシステム

知識ベースレコメンドシステムは、ユーザとアイテムの相互作用の履歴に依存せず、ドメイン知識を利用して推薦を行う手法である。本手法にとって最も重要な情報はドメイン知識であるが、ユーザやアイテムに関する情報、またはユーザとアイテムの相互作用を追加で利用することもできる。

（エ）ハイブリッドレコメンドシステム

ハイブリッド推薦システムとは、協調フィルタリングやコンテンツベース等の異なる推薦手法を組み合わせて、それぞれの手法の長所を生かし、短所を補うことにより、より高品質な推薦を提供する手法である。組合せの手法としては、下記が挙げられる。

・ 重み付け型 … 複数手法の推薦スコアを重み付けして統合する方法

・ 切り替え型 … 推薦に適した手法を状況に応じて切り替える方法
例えば、コールドスタートの場合はコンテンツベース、
それ以外は協調フィルタリングの手法を使用する 等

・ 混合型 … 複数手法の推薦リストを 1 つにまとめる方法

・ 特徴拡張型 … ある手法で得られた特徴量を別手法の入力として使う方法
例えば、協調フィルタリングで得られたユーザの潜在特徴
量をコンテンツベースの手法に入力する 等

・ カスケード型 … ある手法で得られた推薦リストを別手法で再ランキング
する方法

・ メタレベル型 … ある手法で得られたモデルを別手法の入力とする方法
例えば、協調フィルタリングで得られた相関行列をコン
テンンツベースの手法に入力する 等

(b) 本事業への適用

本事業においては、ユーザ=検索者、アイテム=外注案件とみなし、外注案件リストの 1 件ごとに、ランキング形式で複数の検索者を推薦するアプローチを検証する。業務の特性上、1 つの案件に対して複数の検索者が割り当てられることがないことから、協調フィルタリングベースのアプローチではなく、コンテンツベースのアプローチを適用する。データ準備フェーズにおける適用イメージを図 4.4.2.（7）-1 に示す。

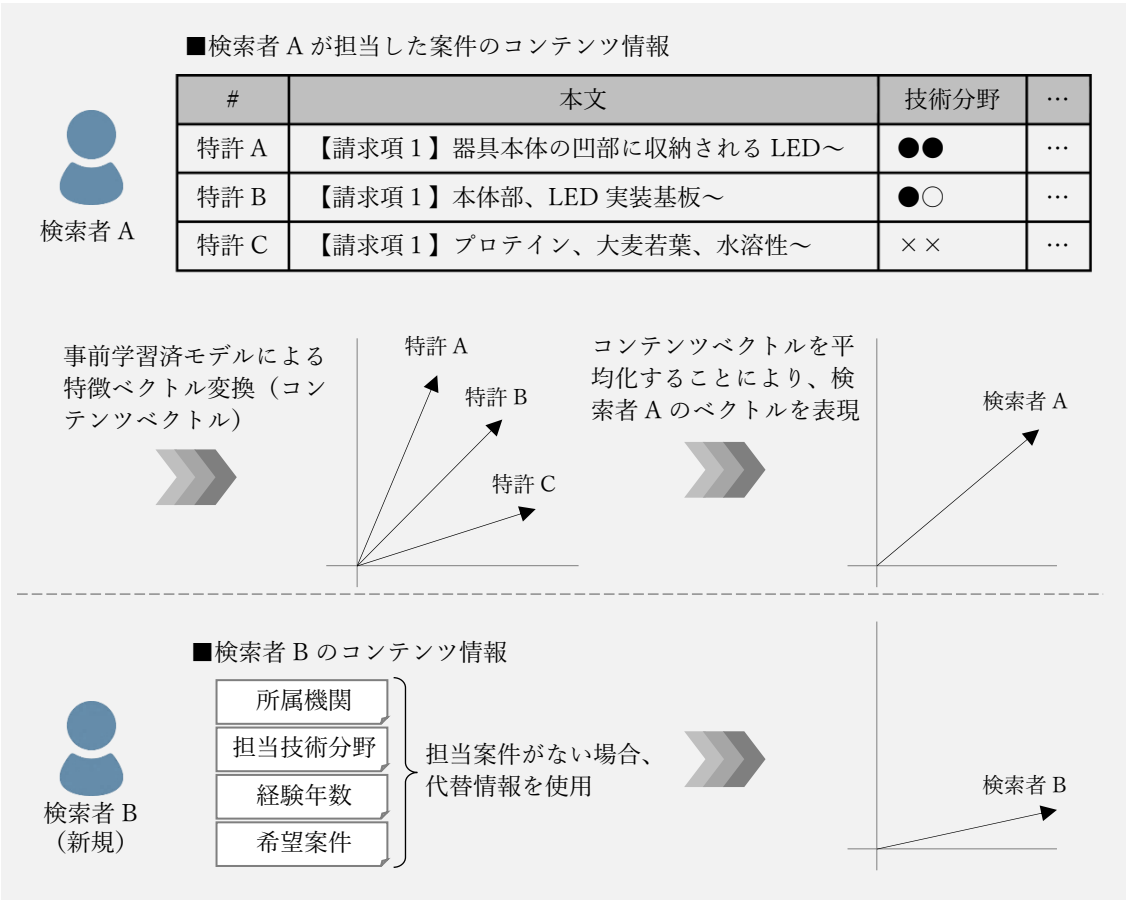


図 4.4.2.（7）-1 コンテンツベース手法の適用イメージ（データ準備フェーズ）

データ準備フェーズで作成した検索者ベクトルをレコメンド対象リストとして、外注案件 1 件ごとにベクトル化した案件ベクトルとの類似度を計算し、類似度順にソートした結果を検索者の候補リストとして推薦する。また、人材育成観点を考慮するため、知識ベースレコメンドシステムの要素を取り入れ、難易度の情報などの制約をドメイン知識として加味して推薦を行うことも検証する。（図 4.4.2.（7）-2）

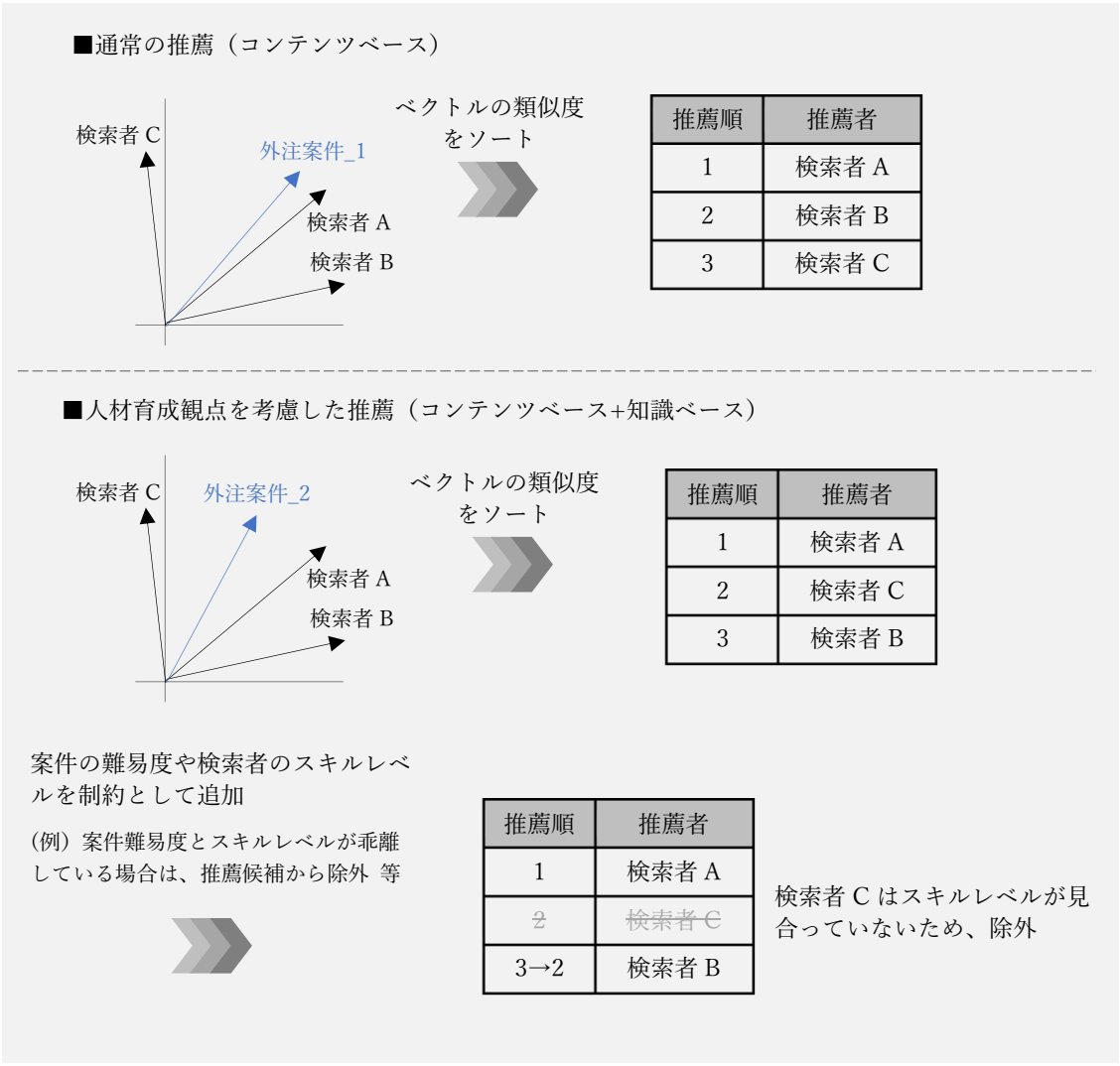


図 4.4.2.(7)-2 コンテンツベース及び知識ベース手法の適用イメージ(推薦フェーズ)

3.5. 適用可能な人工知能技術等の選定

4.4.2 の調査結果から本事業に適用可能と選定した人工知能技術等を表 4.5-1 に示す。

表 4.5-1 適用可能な人工知能技術等の選定結果

#	技術要素	選定した人工知能技術等
1	可読性を考慮した難易度の定量化	GiNZA（構文解析、形態素解析）
2	事前学習済モデルによるテキスト特徴抽出/特徴ベクトルで検索者が経験した案件の傾向を表現	特許 DeBERTa
3	高精度な分類アルゴリズム・ライブラリ	LightGBM
		TabNet
		AutoGluon
4	説明可能な AI	SHAP
5	マルチラベル分類	Binary Relevance（BR）、アルゴリズム適応（Randomforest）
6	検索者のクラスタリング	分割法による階層的クラスタリング
7	コンテンツベースフィルタリング/コールドスタート問題の対応/人材育成を考慮したレコメンド	コンテンツベース+知識ベースレコメンドシステム

4. 貸与データ等の分析、検証用データの作成

4.1. 貸与データ等の分析

4.1.1. 分析対象

下記を分析対象データとする。

- 特許公報データ（H25 年度-R4 年度）
- 発注リスト（H30 年度-R4 年度）
- 非発注リスト（H30 年度-R4 年度）
- 評価票（H30 年度-R4 年度）
- 検索者データ（H30 年度-）

4.1.2. 分析手法

特許公報データをメインデータとして、発注リスト、評価票及び検索者データを紐づけた表形式のデータを作成する。なお、特許公報データを除くデータは H30 年度以降のみ存在するため、特許公報データについても H30 年度以降を利用する。

作成した表形式のデータに対して、下記の観点で分析を実施する。

- 各課題の目的変数（外注有無、報告形態、検索オプション、検索者）の分布
- 目的変数と相関が高いと考えられる項目（特許分類、請求項の数）の分析

4.2. 検証用データ（学習、検証、評価データ）の作成

4.2.1. 実施内容

特許公報データや発注/非発注リストに含まれる項目を変数として活用して、検証用データを作成する。

4.2.2. 検証用データ（学習、検証、評価データ）

分析対象データから各アプローチで使用するデータを作成した。

(1) 課題① 外注対象案件の自動選定

(ア) ルールベースの検証用データ

(イ) AI モデルの検証用データ

(2) 課題② 選定案件の報告形態（対話型・補充型）

(ア) ルールベースの検証用データ

(イ) AI モデルの検証用データ

(3) 課題③ 選定案件への検索オプションの自動付与

(ア) ルールベースの検証用データ

(イ) AI モデルの検証用データ

(4) 課題④ 検索者への案件割り振りの高度化・効率化

(ア) ルールベースの検証用データ

(イ) AI モデルの検証用データ

(ウ) レコメンドの検証用データ

5. 人工知能技術等モデルの構築及び精度評価

5.1. 実施内容

選定した人工知能等技術等の各技術要素を用いて、各業務課題を解消するための AI モデルの構築及び評価を行う。AI モデルの学習と精度評価・分析は複数回実施し、1 回目終了後、改善検討をふまえた上で 2 回目以降を実施する。

5.2. 人工知能技術等モデル構築

5.2.1. 構築環境

(1) システム環境

本事業においては、大規模な学習データを用いたモデル構築を実施するため、大容量メモリを搭載したオンプレミスサーバ環境を使用する。

(2) ソフトウェア環境

本事業において利用するソフトウェア環境を表 6.2.1.（2）-1 に示す。

表 6.2.1.（2） -1 使用する主なソフトウェア

#	種別	ソフトウェア	バージョン	ライセンス	商用利用
1	自然言語処理ライブラリ	ginza	5.1.3	MIT License	可
2	自然言語処理ライブラリ	ja-ginza	1.1.3	MIT License	可
3	自然言語処理ライブラリ	spacy	3.6.1, 3.7.2	MIT License	可
4	配列処理ライブラリ	numpy	1.3.5	BSD License	可
5	データ解析ライブラリ	pandas	2.8.0	BSD License	可
6	データ解析ライブラリ	scipy	1.7.3	BSD License	可
7	データ解析ライブラリ	statsmodels	0.13.5	BSD License	可
8	データ可視化ライブラリ	matplotlib	1.30.1	PSF License	可
9	データ可視化ライブラリ	japanize-matplotlib	3.3.5	MIT License	可
10	データ可視化ライブラリ	seaborn	0.12.2	BSD License	可
11	機械学習ライブラリ	lightgbm	3.5.3	MIT License	可
12	機械学習ライブラリ	scikit-learn	1.0.2	BSD License	可
13	深層学習ライブラリ	mxnet	1.21.2, 1.21.6	Apache License 2.0	可
14	深層学習ライブラリ	pytorch-	4.1.0	MIT License	可

#	種別	ソフトウェア	バージョン	ライセンス	商用利用
		tabnet			
15	深層学習ライブラリ	tabnet	0.1.6	MIT License	可
16	深層学習ライブラリ	torch	1.10.0	BSD License	可
17	深層学習ライブラリ	transformers	4.23.1	Apache License 2.0	可
18	自動機械学習ライブラリ	autogluon	2.7.12	Apache License 2.0	可
19	実験管理ライブラリ	mlflow	1.9.1	Apache License 2.0	可
20	数理最適化ライブラリ	PuLP	0.6.2	BSD License	可
21	AI モデル解釈ライブラリ	shap	0.42.1	MIT License	可
22	GPU 処理ライブラリ	cuda	12.2	NVIDIA EULA	可
23	言語・標準ライブラリ	Python	3.7.11	PSF License	可
24	OS	Ubuntu	18.04.05	GPL	可

5.2.2. 構築手順及びモデル構成

(1) 課題① 外注対象案件の自動選定

(ア) ルールベース

令和 4 年 4 月~5 月に実施された特許庁審査室外注推進担当官を対象としたプレアンケートで挙げられた意見のうち、外注あり/外注なしの基準として有効活用できる意見からルールを設定する。6 つのルールを適用し、1 つでも該当すれば、外注不適と判定するルールベースを構築する。

(イ) AI モデル

カテゴリ属性の説明変数に対してはエンコーディング処理を実施し、テキスト属性の説明変数に対しては特許 DeBERTa で 1536 次元にベクトル化する。前処理後の各属性の説明変数を結合した加工データセットに対して、3 つの AI モデル（LightGBM、TabNet、AutoGluon）を適用し、推論結果を得る。

(2) 課題② 選定案件の報告形態（対話型・補充型）

(ア) ルールベース

令和 4 年 4 月~5 月に実施された特許庁審査室外注推進担当官を対象としたプレアンケートで挙げられた意見のうち、対話型/補充型の基準として有効活用できる意見からルールを設定する。3 つのルールを適用し、1 つでも該当すれば、補充型と判定するルールベースを構築する。

(イ) AI モデル

AI モデルパイプラインは課題①と同様。

(3) 課題③ 選定案件への検索オプションの自動付与

(ア) ルールベース

令和 4 年 4 月~5 月に実施された特許庁審査室外注推進担当官を対象としたプレアンケートで挙げられた意見のうち、検索オプション付与の基準として有効活用できる意見からルールを設定する。各検索オプションにおいてルールを適用し、1 つでも該当すれば、オプション付与と判定するルールベースを構築する。

(イ) AI モデル

AI モデルパイプラインは課題①と同様。なお、課題③においては検索オプション数分の AI モデルを構築するため、3 つのアルゴリズムのうち処理が高速な LightGBM を一律適用する。

(4) 課題④ 検索者への案件割り振りの高度化・効率化

(ア) ルールベース

令和 4 年 4 月~5 月に実施された特許庁審査室外注推進担当官を対象としたプレアンケートで挙げられた意見のうち、機関/検索者割り振りの基準として有効活用できる意見からルールを設定する。調査機関の割り振りと熟練度の割り振りを組み合わせて、27 グループに対してスコアを計算し、上位 10 件をランキングする。

(イ) AI モデル

AI モデルパイプラインは課題①と同様。なお、課題①において AutoGluon が最も精度が高いことが確認できたため、課題④においては AutoGluon のみを適用する。

(ウ) レコメンド

レコメンドパイプラインを図 6.2.2.（4）.(ウ)-1 に示す。5.2.2. で示したテキスト属性の説明変数に対しては特許 DeBERTa で 1536 次元にベクトル化し、案件のベクトルデータベース（案件ベクトル DB）を作成する。評価用データに対して案件ベクトル DB の中から類似度が高い上位 10 件をリストアップし、推論結果を得る。

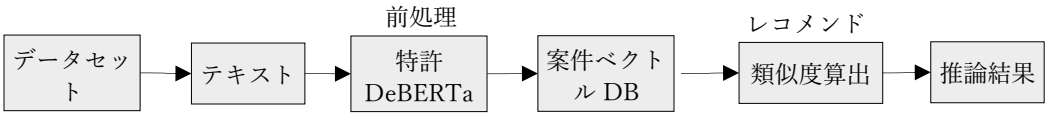
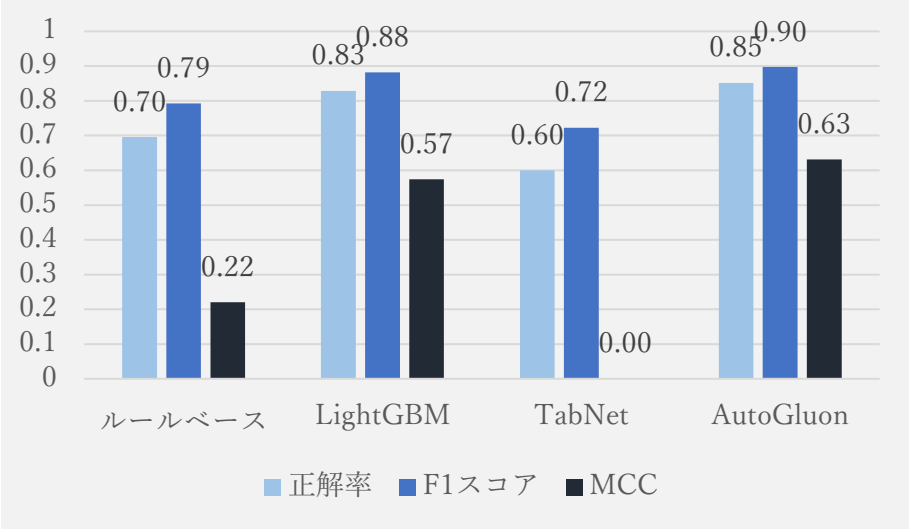


図 6.2.2.（4）.(ウ)-1 レコメンドパイプライン

5.3. 精度評価

5.3.1. 課題① 外注対象案件の自動選定の精度評価結果

AutoGluon の精度が最も高く、LightGBM が 2 番目の精度である。AutoGluon 及び LightGBM では正解率は 83%～85%、F1 スコアは 0.88～0.90、MCC は 0.57～0.63 と高い精度で外注あり/外注なしの判定が可能である。ルールベースは上記 2 つの AI モデルと比較すると、MCC 0.4 ポイント程度劣っており、TabNet は MCC が 0 付近であり、学習が上手く出来ていないことを示す。

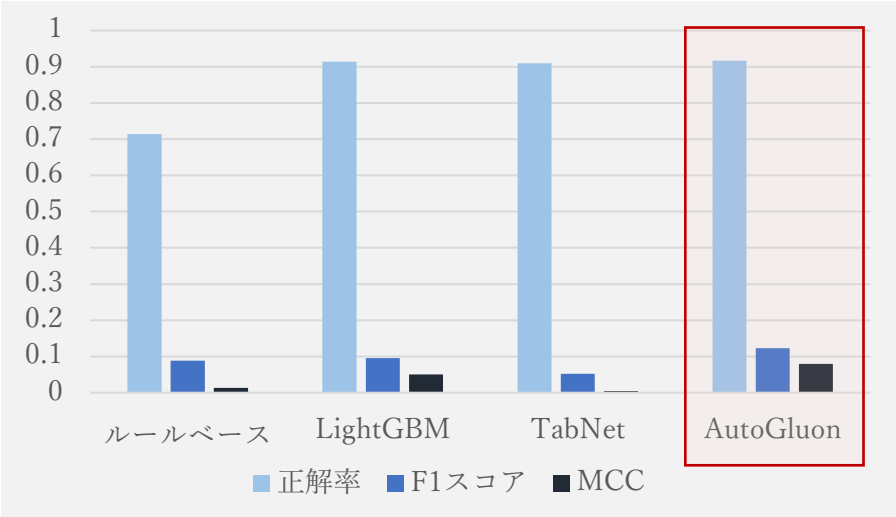


※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：外注ありの推論結果のうち、実際に外注ありである割合（0~1）
再現率　　：実際に外注ありのうち、外注ありと推論した割合（0~1）
MCC　　：外注ありと外注なしの両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.3.1-1 課題①の精度評価（評価指標※）

5.3.2. 課題② 選定案件の報告形態（対話型・補充型）の精度評価結果

4 つのアプローチの中では AutoGluon が最も精度が高いが、F1 スコアは 0.12、MCC は 0.08 と精度は低い。なお、補充型が全体の 4.76%であり、分布に偏りがあるため、正解率は 92%と高くなる。



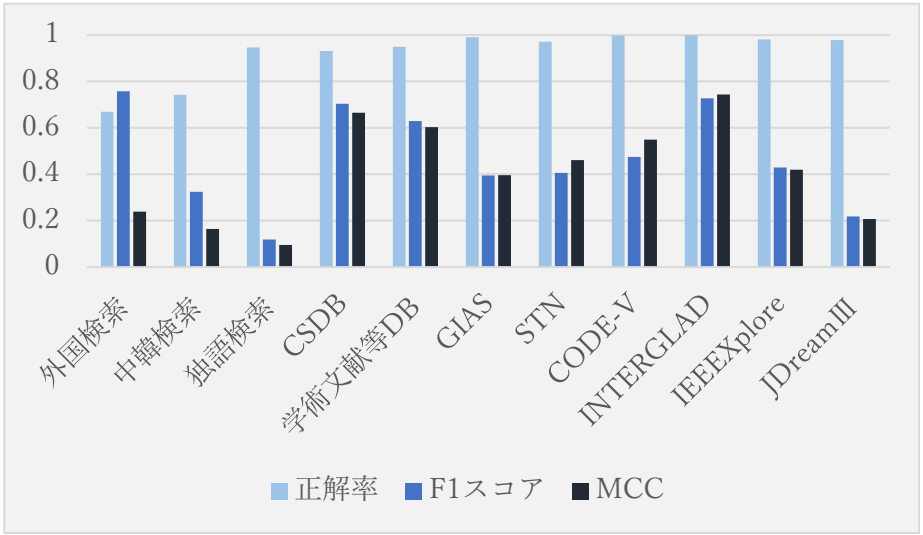
※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：補充型の推論結果のうち、実際に補充型である割合（0~1）
再現率　　：実際に補充型のうち、補充型と推論した割合（0~1）
MCC　　：補充型と対話型の両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.3.2-1 課題②の精度評価（評価指標※）

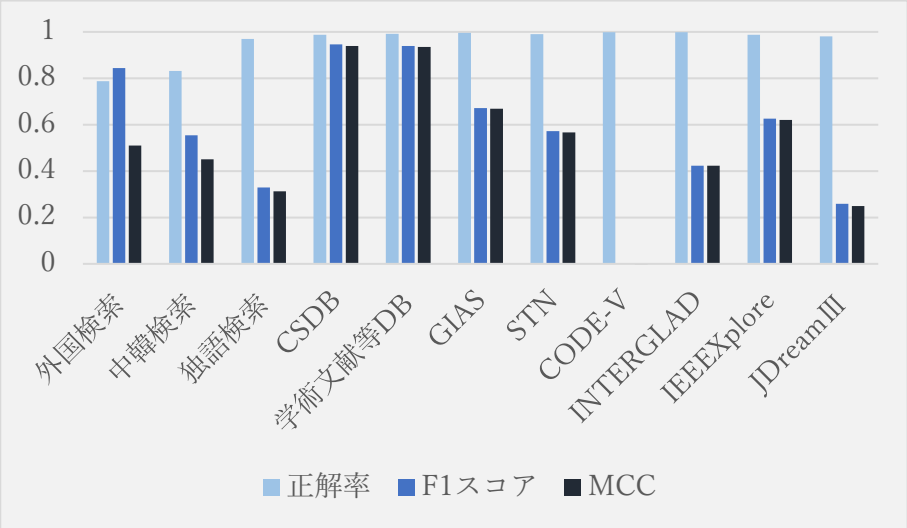
5.3.3. 課題③ 選定案件への検索オプションの自動付与の精度評価結果

課題③においては検索オプション数分の AI モデルを構築するため、3 つのアルゴリズムのうち処理が高速な LightGBM を一律適用する。

2 つのアプローチの中では AI モデル（LightGBM）の方が全体的に高い精度を示すが、CODE-V、INTERGLAD はルールベースの方が高い精度を示す。



(a) ルールベース



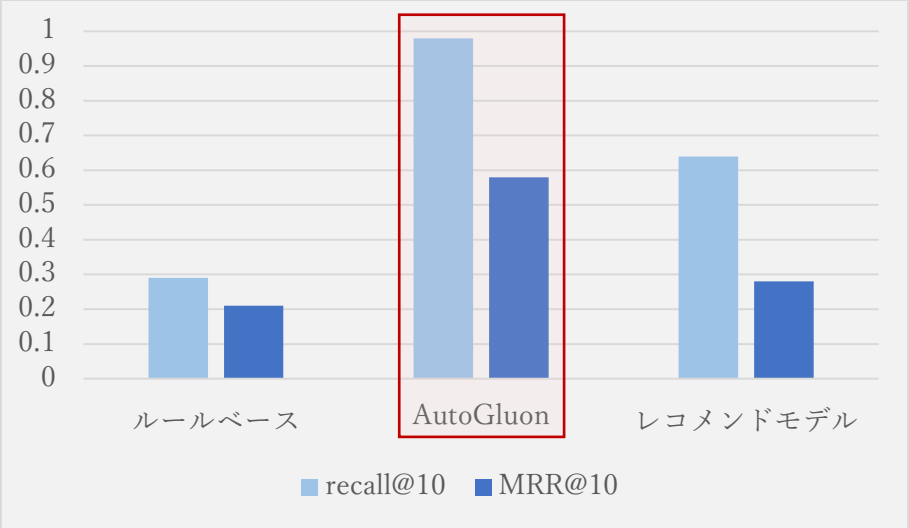
（b）AI モデル（LightGBM）

※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：オプション付与の推論結果のうち、実際にオプション付与である割合（0~1）
再現率　　：実際にオプション付与のうち、オプション付与と推論した割合（0~1）
MCC　　：オプション付与とオプションなしの両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.3.3-1 課題③の精度評価（評価指標※）

5.3.4. 課題④ 検索者への案件割り振りの高度化・効率化の精度評価結果

課題①において AI モデルの中では AutoGluon が最も精度が高いことが確認できたため、課題④の AI モデルアプローチでは AutoGluon のみを適用する。3 つのアプローチの中では AutoGluon が最も精度を示し、recall@10 が 0.98、MRR@10 が 0.58 である。この値は、推論結果の上位 10 件に正解である検索者グループが含まれる確率が 98%であり、さらに平均して上位 1.72 番目（=1/0.58）に正解である検索者グループが登場することを意味する。



※ recall@10：正解を上位 10 位内として推薦できているのかを評価する
MRR@10：正解をどれだけ上位で推薦できているかを評価する

図 6.3.4-1 課題④の精度評価（評価指標※）

5.4. モデル改善

5.4.1. 課題① 外注対象案件の自動選定のモデル改善

課題①においては、最も精度の高いアプローチである AI モデル(AutoGluon)のチューニングによるさらなる精度改善を実施するとともに、現在の外注選定の運用に最も近いルールベースの改善を実施する。また、それぞれのアプローチにおける推論結果出力イメージを作成する。

(1) AI モデルの改善

改善前の AI モデルにおいて、どの特徴量が推論に寄与しているかを確認するため、特徴量の重要度を算出する。

LightGBM の特徴量重要度では FI 主分類が最も重要度が高い。FI 主分類とは、特許分類の一つであり、FI セクション ＞FI クラス ＞FI サブクラス ＞FI メイングループ ＞ FI 分類 ＞ … といった階層構造をもつ項目をひとまとめに表す特徴量である。特許分類を 1 項目で表せる便利な項目であるが、それゆえにユニーク数が大きくなり、それぞれの値に過剰に適応してしまう過学習が生じやすい。そのため、FI 主分類を階層構造に分解して、FI セクションや FI クラスといった、より上位のグループにまとめることにより、モデルがより汎化されたパターンを学習できるよう特徴量を追加する。(観点①)

AutoGluon の特徴量重要度では請求項の数が最も有用性が高い。請求項の数が極端に少ないあるいは多い場合に外注不適として、外注対象から除外されている運用を AI モデルが学習していると考えられる。同様の考え方により、請求項に関する数値データを特許公報データより取得し、特徴量に追加する。(観点②)

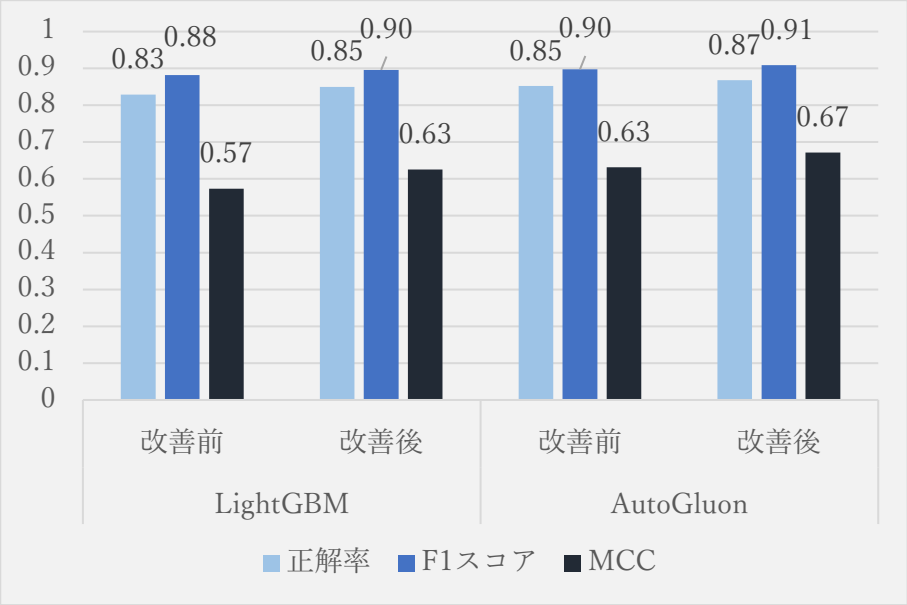
後述する課題②の改善タスクに基づき、アンケート結果をふまえて有効と考えられる特徴量を追加する。(観点③)

後述する課題③の改善タスクに基づき、検索オプション付与における重要単語の登場数に関する特徴量を追加する。(観点④)

以上より、4 つの観点から追加した特徴量を表 6.4.1. (1) -2 に示す。なお、課題②③においても同一の特徴量で AI モデルの改善を実施する。

特徴量の追加により、LightGBM の正解率は 83%→85%、F1 スコアは 0.88→0.90、MCC は 0.57→0.63 に向上し、AutoGluon の正解率は 85%→87%、F1 スコアは 0.90→0.91、MCC は 0.63→0.67 に向上している。

また、追加した特徴量(表内の網掛け項目)が特徴量重要度の上位に登場し、精度向上へ寄与したことが確認できる。



※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：外注ありの推論結果のうち、実際に外注ありである割合（0~1）
再現率　　：実際に外注ありのうち、外注ありと推論した割合（0~1）
MCC　　：外注ありと外注なしの両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.4.1.（1）-1 課題①AI モデルの改善前後の精度評価（評価指標※）

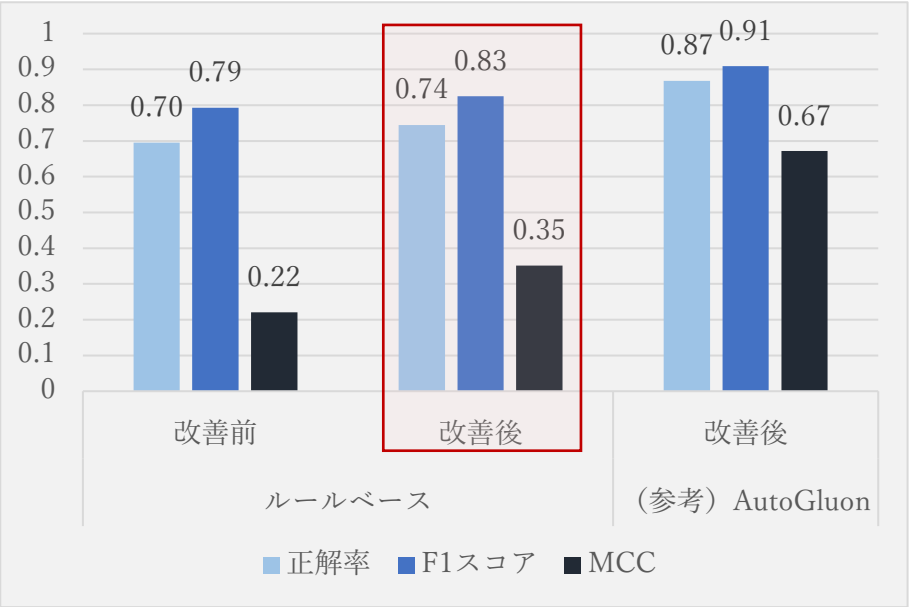
(2) ルールベースの改善

ルールベースは精度の面では AI モデルに劣るものの、判定根拠が明確であり、運用ルールに基づきルールを調整することができるため、現状の審査室の外注選定業務に近い運用が可能である。業務分析の結果をふまえて、改善後のルールを検討した。

改善後のルールベースでは総合スコアを用いて判定することにより、任意の件数を外注不適と判定するよう改善する。

ルールの改善及び総合スコアの導入により、ルールベースの正解率は 70%→74%、F1 スコアは 0.79→0.83、MCC は 0.22→0.35 に向上し、AutoGluon の正解率は 85%→87%、F1 スコアは 0.90→0.91、MCC は 0.63→0.67 に向上した。

本事業におけるルールベースはプレアンケートや審査室へのヒアリング結果を反映したものであるが、すべての審査室のルールを網羅してはいないため、審査室ごとに精度の差があると考えられる。



※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：外注ありの推論結果のうち、実際に外注ありである割合（0~1）
再現率　　：実際に外注ありのうち、外注ありと推論した割合（0~1）
MCC　　：外注ありと外注なしの両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.4.1.（2）-3 課題①ルールベースの改善前後の精度評価（評価指標※）

(3) 推論結果出力イメージの作成

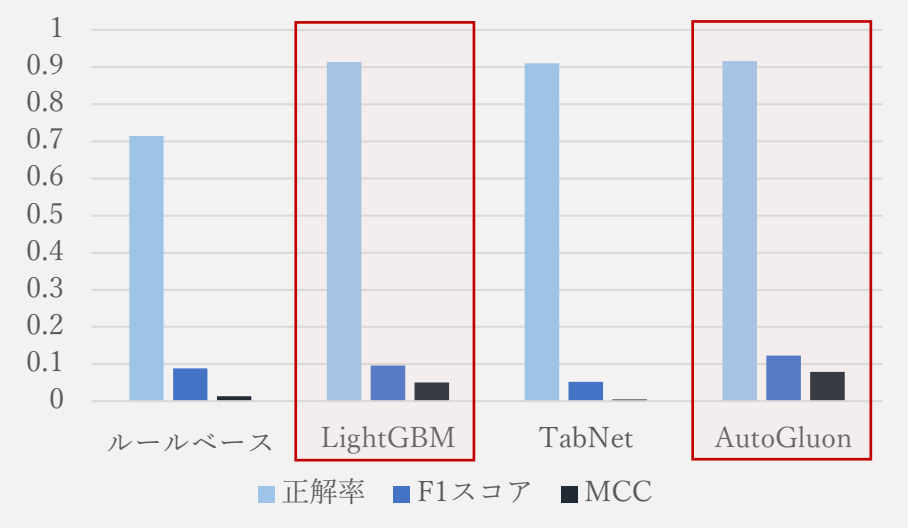
AI モデルの出力イメージでは、外注不適の推論確率とその判断に寄与した項目を上位 3 件出力する。寄与した項目は 4.4.2.（4）の説明可能な AI の技術（SHAP 値）により算出する。寄与した項目は実運用にて利用者が参考にしていない項目が上位に出現し、納得感を得られない等の否定的な意見も想定されるため、出力有無や出力方法については検討が必要である。

ルールベースの出力イメージでは、総合スコアと各ルールに該当したか否かを出力する。出力イメージにあるとおり、ルールベースは判定根拠が明確である点が特長である。

5.4.2. 課題② 選定案件の報告形態（対話型・補充型）のモデル改善

課題②においては、最も精度の高いアプローチは AI モデル（AutoGluon）であるが、MCC は 0.08 であり、ランダムな推論と同程度である。精度が低い要因の一つとして、補充型の割合（4.8%）がデータセット全体に対して小さく、不均衡なデータであり、補充型の特徴を学習できていないと考えられる。

そのため、データセットの多数を占める対話型のデータをダウンサンプリングし、不均衡を緩和することによる精度向上を試みる。また、これまでは全課題共通して同一対象期間（H30 年度～R4 年度）のデータセットを使用していたが、補充型の報告形態の運用が開始された R3 年度以降のデータに限定することにより課題②のデータセットの質の向上を図る。さらに、課題①と同様に特徴量を追加し、精度向上を図る。



※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：補充型の推論結果のうち、実際に補充型である割合（0~1）
再現率　　：実際に補充型のうち、補充型と推論した割合（0~1）
MCC　　：補充型と対話型の両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.4.2-1 課題②の精度評価（評価指標※）【再掲】

(1) AI モデルの改善（ダウンサンプリングによる不均衡の緩和）

AI モデルの学習時における不均衡データの対策として、下記が考えられる。

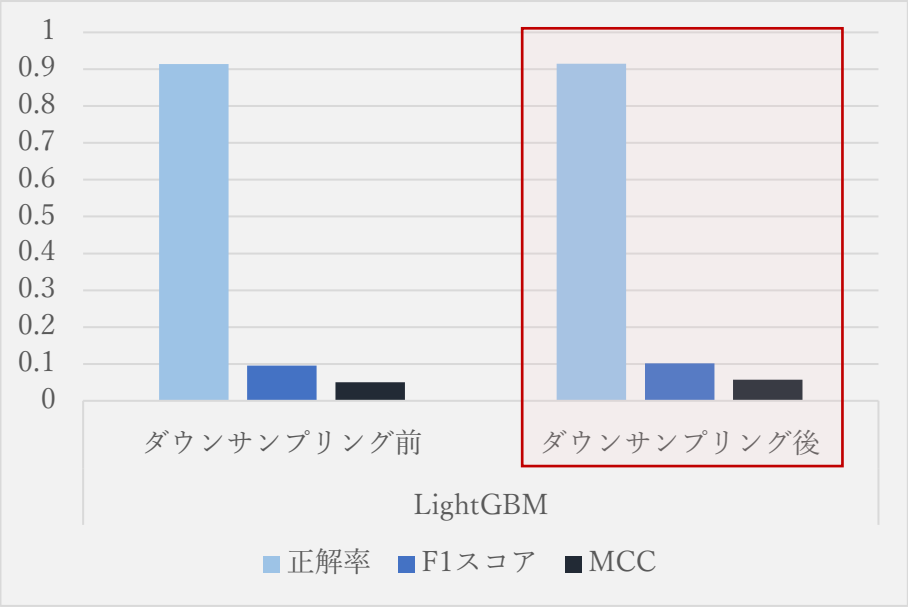
- 多数クラスのダウンサンプリング、少数クラスのアップサンプリング
- AI モデルパラメータによるクラスの重み付け
- 異常検知アルゴリズムの適用

異常検知アルゴリズムは多数クラスのデータが均質である場合（工場ラインでの生産品 等）に異常値を検出する際に有効なアプローチであるが、本事業におけるデータセットの場合、多数クラスの中でもバリエーションが多数存在するため、異常検知アルゴリズムは適さないと考える。1 点目と 2 点目の対策はともに少数クラスの傾向を多数クラスの傾向に埋もれないように学習

させるアプローチであり、本事業においては 1 点目の多数クラスのダウンサンプリングを採用する。

学習及び検証データの対話型をダウンサンプリングすることにより、補充型の割合が約 5%→約 17%に向上し、不均衡が緩和される。

ダウンサンプリング前後の精度評価を図 6.4.2.(1)-2 に示す。ダウンサンプリング後の正解率は 91.39%→91.45%、F1 スコアは 0.0958→0.1017、MCC は 0.0506→0.0569 の変化であり、精度向上の結果はあまり見られない。また、課題③の GIAS 検索や STN 検索はオプション付与の割合が約 1%であり、課題②より不均衡なデータであるが、MCC は 0.5 以上であり、精度が高い。以上より、データの不均衡が課題②の精度が低い主な原因ではないと考えられる。



※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：補充型の推論結果のうち、実際に補充型である割合（0~1）
再現率　　：実際に補充型のうち、補充型と推論した割合（0~1）
MCC　　：補充型と対話型の両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

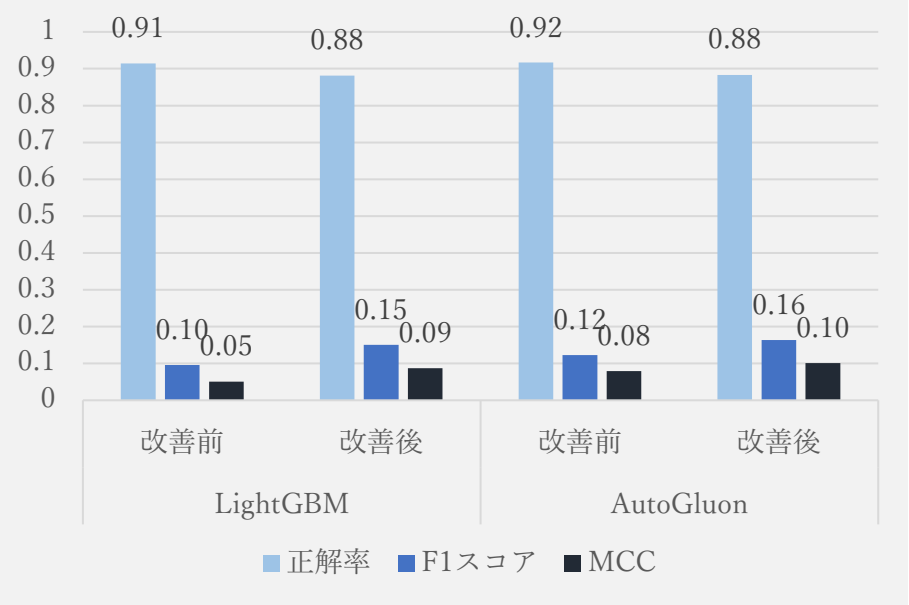
図 6.4.2.(1)-2 課題②AI モデルのダウンサンプリング前後の精度評価（評価指標※）

(2) AI モデルの改善（データセットの対象期間変更及び特徴量の追加）

補充型の報告形態は R3 年度以降に登場するため、R3 年度以降のデータに限定することにより課題②のデータセットの質の向上を図る。

併せて、課題①と同様に特徴量を追加し、精度向上を図る。登録調査機関ヒアリングでは課題②に関する判断基準の情報は多くは得られなかったが、本願発明が理解し易い、ボリュームが小さい文献などの意見があった。そこで、特許公報データより得られる全頁数や参照ファイルの数等の数値データを追加する。また、選定判断の時期によって対応件数を決める等の運用も見られたため、審査請求月等の時期的な特徴量も追加する。

データセットの対象期間変更及び特徴量の追加により、LightGBM の正解率は 91%→88%に減少するものの、F1 スコアは 0.10→0.15、MCC は 0.05→0.09 に向上する。AutoGluon の正解率は 92%→88%に減少するものの、F1 スコアは 0.12→0.16、MCC は 0.08→0.10 に向上する。データセットの対象期間変さらにより、多数クラスの対話型の件数が相対的に減少しているため、正解率は減少しているが、モデル全体の精度を示す MCC は向上していることより、改善の効果が見られる。しかしながら、最も精度の高い改善後の AutoGluon であっても、MCC は 0.10 程度であり、精度は低い。



※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：補充型の推論結果のうち、実際に補充型である割合（0~1）
再現率　　：実際に補充型のうち、補充型と推論した割合（0~1）
MCC　　：補充型と対話型の両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.4.2.（2）-1 課題②AI モデルの改善後の精度評価（評価指標※）

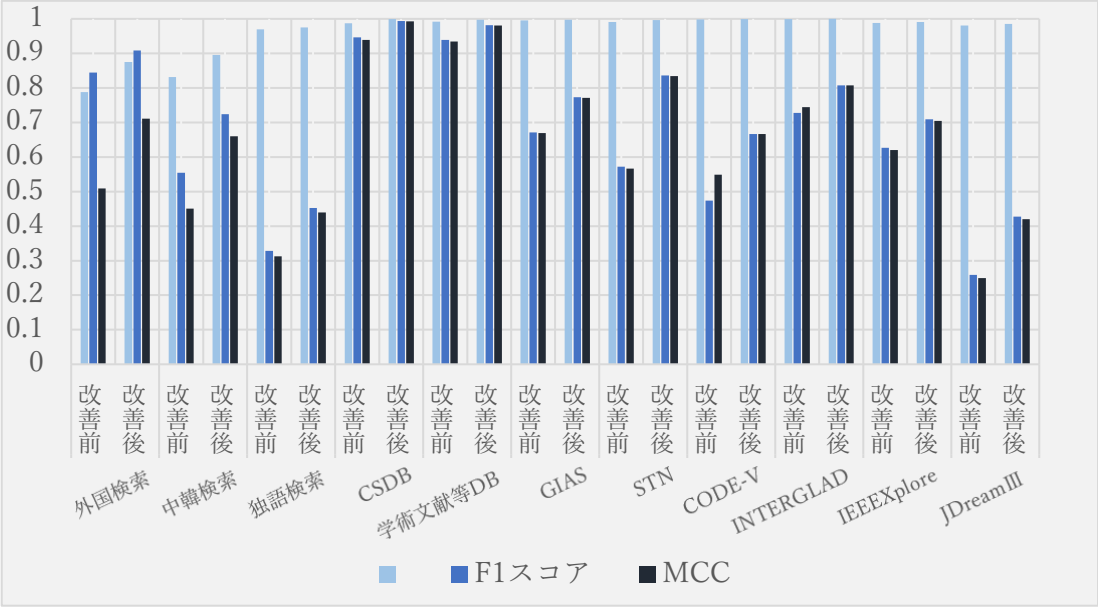
5.4.3. 課題③ 選定案件への検索オプションの自動付与のモデル改善

正解データの件数が少ない、CODE-V 及び INTERGLAD を除き、AI モデル（LightGBM）が高い精度を示すため、課題①と同様に特徴量を追加し、精度向上を図る。

(1) AI モデルの改善

課題③のデータセット分析の結果、各検索オプションに対応する特徴的な単語が存在することが確認できたため、重要単語リストを作成し、テキスト項目に含まれる重要単語の登場回数をカウントした特徴量を追加する。

上記で作成した重要単語の登場回数に関する特徴量を含めて、特徴量を追加し、改善した AI モデルによる精度評価を図 6.4.3.（1）-1 に示す。



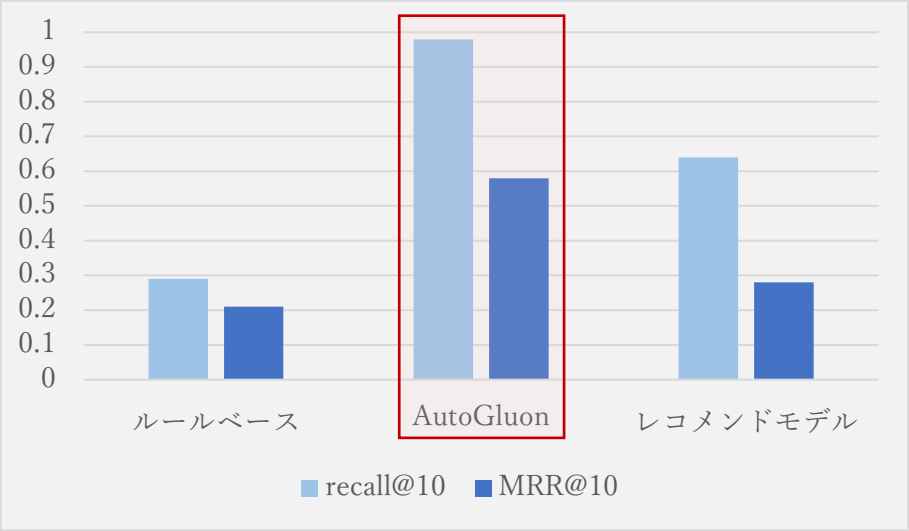
※ 正解率　　：正解と推論結果が一致した件数を全体の件数で除算した値（0~1）
F1 スコア　：適合率と再現率の調和平均（0~1）
適合率　　：オプション付与の推論結果のうち、実際にオプション付与である割合（0~1）
再現率　　：実際にオプション付与のうち、オプション付与と推論した割合（0~1）
MCC　　：オプション付与とオプションなしの両方の精度を評価する指標（-1～+1）
+1 は完璧、0 はランダム、-1 は正反対な推論をそれぞれ示す。

図 6.4.3.（1）-1 課題③AI モデルの改善後の精度評価（評価指標※）

5.4.4. 課題④ 検索者への案件割り振りの高度化・効率化のモデル改善

AI モデル（AutoGluon）が高い精度を示すため、過去の案件割り振り傾向を模倣する観点では AI モデルが最も優れている。一方、実運用を想定した場合、案件割り振りの前提として調査機関及び区分ごとに外注計画件数が定め

られており、この計画件数内での割り振りが必要となる。AI モデルの推論結果は外注計画件数を考慮していないため、外注計画件数を制約条件とし、効果的に案件割り振りが可能となるよう数理最適化を適用することにより、AI モデルを改善する。また、検索者の検索結果に対する特許庁審査官の評価点を分析することにより、高い評価点を得るためのパターンを明らかにする。



※　recall@10：正解を上位 10 位内として推薦できているのかを評価する
　　MRR@10：正解をどれだけ上位で推薦できているかを評価する

図 6.4.4-1 課題④の精度評価（評価指標※）【再掲】

(1) AI モデルの改善

課題④の案件割り振りの結果と外注計画について、区分 1 であれば、A 社に 58%、E 社に 22%、F 社に 21%の外注計画の割合であるが、AI モデルの推論では A 社に 81%を割り振っており、計画を大きく上回る。実運用を想定した場合、外注計画から大きく乖離することは望ましくないため、AI モデルの推論結果を外注計画の割合を制約条件としつつ、AI モデルの推論確率は最大化するよう数理最適化処理を実施する。

数理最適化処理を実行することにより、最適化前は外注計画の割合を超過していた調査機関及び区分であっても、外注計画の割合に合うよう最適化が行われる。

表 6.4.4.（1）-3 に最適化前後の精度評価を示す。recall@1 は 0.57→0.53 に低下しているものの、実運用に則した推論となることが確認できた。

※ recall@1：正解を上位 1 位に推薦できているのかを評価する

(2) 検索者の評価点の分析

課題④に関するこれまでのアプローチは、過去の案件割り振り傾向を模倣するものであるが、登録調査機関ヒアリングによれば、ランダムに割り振るなどの運用が行われているため、最良な判断の結果ではないことが考えられる。そのため、検索者の検索結果に対して、特許庁審査官が点数付けした評価点を分析することにより、評価点が高くなるような案件割り振りのパターンを分析する。

つづいて、検索者と案件の相性による評価点を分析する。分析にあたり、検索者は似通った傾向をもつ者同士でグルーピングし、検索者クラスタを作成する。また、案件についてはテーマ群でグルーピングする。そして、テーマ群単位に平均評価点が高い検索者クラスタをランキング形式で出力する。

クラスタリング結果を解釈すると、検索者 1 人当たりの平均案件数が少ないクラスタ群は初心者クラスタ、検索者 1 人当たりの平均テーマ群数が 1 であり、検索者 1 人当たりの平均案件数が多いクラスタ群はテーマ群専属クラスタ等と意味づけることができる。主要テーマ群の件数や割合を確認することにより、どのようにクラスタリングされたかを解釈することができる。

6. リソース、学習時間等の調査

6.1. 実施内容

本事業を継続的に進めていくためには、AI モデルの構築等に必要となるマシンリソースやソフトウェアを準備し、追加の学習や精度検証を実施する必要がある。そこで、本事業で使用したリソースや学習に要した時間を実績ベースで記載する。

6.2. 調査結果

本事業で使用したハードウェア構成は図 7.2-1 のとおり。なお、リソース測定時には、CPU に対して全体の半数の 20 コアを占有する設定としている。

図 7.2-1 構築環境

6.3. 調査の考察

本事業におけるデータセットに対しては、LightGBM は少ないマシンリソースである程度高い精度が期待でき、AutoGluon はマシンリソースを大量に使用するが、LightGBM より高い精度が期待できる。以上をふまえて、ベースラインモデルの構築や特徴量選択の試行錯誤のフェーズにおいては、LightGBM を用いて高速に AI モデルを構築し、精度追及のフェーズにおいては、AutoGluon を用いることが望ましい。

7. 総合分析

7.1. 実施内容

総合分析は、前章までの結果をもとに、各観点の実用化に向けた方向性を特許庁審査推進室担当者と合意したうえで、本事業の検証資材等から実現可能な本事業の成果、実用化に向けた要件整理、次年度以降にて対応すべき事項としての課題、ロードマップを整理することとした。

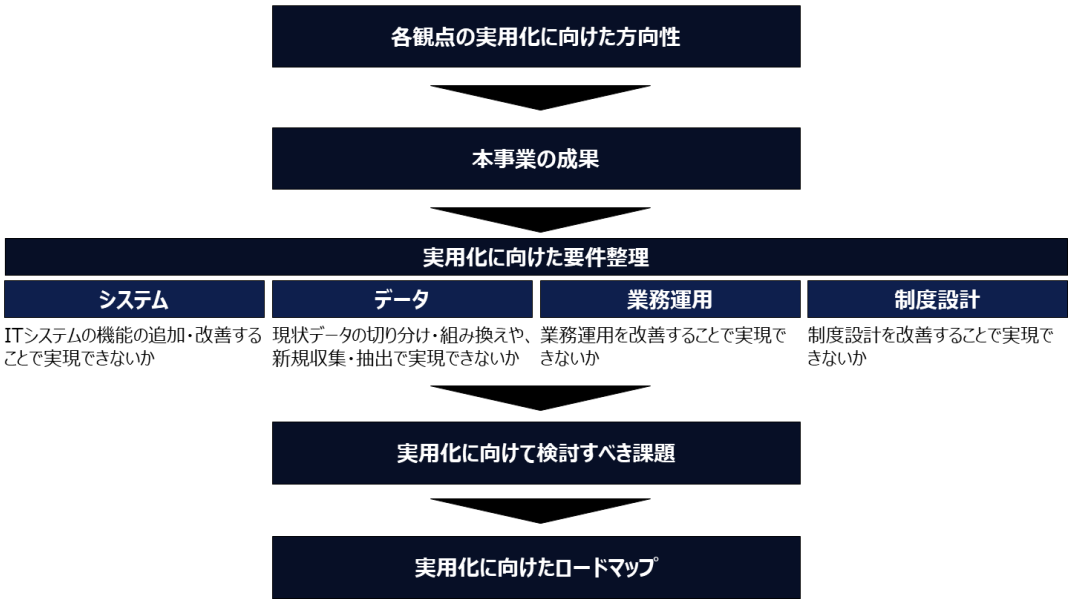


図 7.1-1 業務分析の実施内容

※以降の整理内容は最終報告会にて特許庁へ報告し、承認・合意済みではあるが、公開情報としての提示は省略することとする。

令和５年度
人工知能技術等を活用した先行技術文献調査事業の案件選定業務
における高度化・効率化に関する実証的研究事業

報告書

発　行　　令和６年３月 22 日

株式会社 NTT データ