１Ｄ１５

特許出願における発明者の識別
－ 二値分類による識別手法の性能評価 －

中山保夫 (NISTEP), ○細野光章 (NISTEP/東海国立大学機構), 富澤宏之 (NISTEP)
nakayama@nistep.go.jp

1. はじめに

産学連携の成果を分析するために、特許出願情報は極めて重要なデータソースである。特許出願情報を活用することで、産学連携ネットワーク内で中心的な役割を担う研究者を特定し、発明を通じた学界から産業界への技術移転とイノベーション促進のメカニズムを解明することが可能となる。これにより、社会への貢献度を評価し、研究の推進に必要な重要な情報を提供することができる。

このような研究プロセスの一環として、対象研究者と同姓同名の発明者を含む多数の特許出願の母集団から、対象研究者が関与した発明に基づく特許出願を正確に識別し、その情報を集約する必要がある。この目的を達成するために、特許出願メタデータ、自然言語処理、機械学習など、複数の識別手法を適用し、研究者が関与した特許出願の識別性能を検証した。

本稿では、これらの識別手法の性能を二値分類及び統計手法を用いて評価し、その結果と実用性について報告する。

2. 検証した発明者の識別手法

本稿にて、「特許出願における発明者の識別」とは、特許出願データベースから特定の人物が発明者として関与した特許出願を見つけ出すことである。表 1 に紹介する識別手法を用いて、その性能の検証を行っている。

表 1 検証した識別手法

番号	手法分類	手法名称	利用情報	概要
1-1	確定的二値分類	特許出願メタデータ	共同発明者	既知の対象者の特許出願から抽出した共同発明者リストを使い、合致者が存在する場合、対象者が発明した特許出願と識別する手法。単独発明には適用不可。
1-2			発明者住所	既知の対象者の特許出願から抽出した発明者住所を使い、識別対象特許出願の発明者住所と合致すれば、対象者が発明した特許出願と識別する手法。
1-3			出願人	識別する特許出願の出願人が対象発明者の所属機関と一致する場合、対象者が発明した特許出願と識別する手法。
1-4			住所情報の所属機関	識別する特許出願の発明者住所に含まれる機関名が対象者の所属機関と一致する場合、対象者が発明した特許出願と識別する手法。
2		引用情報ネットワーク	特許出願の後方引用	識別する特許出願を後方引用する特許出願の発明者に対象者が存在すれば対象者が発明した特許出願であると識別する手法。
3	確率的二値分類	技術分類コード	IPC8 技術ベクトルの類似性	既知の対象者の特許出願の IPC8 技術ベクトルと、識別する特許出願の IPC8 技術ベクトルのコサイン類似度を使い対象者が発明した特許出願を識別する手法。
4-1		文書の類似性	TF-IDF ベクトルを使った類似性	形態素解析・複合語処理と TF-IDF を用いて既知の特許出願と識別特許出願をベクトル化し、コサイン類似度で対象者の特許出願を識別する手法。
4-2			技術キーワード出現パターン	技術キーワードの出現頻度パターンを基に既知の特許出願と識別特許出願をベクトル化し、加重コサイン類似度で対象者の特許出願を識別する手法。
5-1		機械学習 (Random Forests)	TF-IDF による文書特徴化と RF 学習	陰陽両特許出願を形態素解析・複合語処理と TF-IDF を用いた特徴量により訓練したランダムフォレストモデルを使い、対象者の特許出願を識別する手法。
5-2			技術キーワードパターンの RF 学習	陰陽両特許出願の技術キーワードの出現頻度パターンにより訓練したランダムフォレストモデルを使い、対象者の特許出願を識別する手法。

3. 識別に使用する特許出願母集団と識別手順

識別に使用する特許出願母集団は、国立大学の研究者及びその同姓同名者が発明者として含まれる特許出願である。対象研究者は、識別手法の性能検証という目的から次の条件で 5 名を選定している。

(1) 同姓同名者が多い名前の研究者
(2) 数十件以上の特許出願実績を持つ研究者
(3) 識別の難度が高い要素を持つ研究者（同じ大学に特許出願する同姓同名者が存在、類似技術領域で特許出願する同姓同名者が存在するなど）

以上の条件で抽出した研究者の特許出願は表 2 に示す母集団を構成する。なお、文書間類似度及び機械

学習の評価には、データ準備の労力を考慮し、層別サンプリングを実施して特性を保持しつつ、特許出願数を減少させている。

また、特定の研究者が発明者として関与した特許出願を見つけ出すには、その人物を識別するための情報が必要である。このため、表 2 の識別を行う特許出願とは別に、同研究者が関与した既知の特許出願から共同発明者や発明者住所情報を抽出するとともに、文書の類似性評価のための特許出願全文などを準備する必要がある。この既知の特許出願データには、筆者らが過去に構築した『国立大学発特許出願データベース』を利用している[1]。

表 2 識別に使用する特許出願母集団

ケース番号	識別対象者	検証する特許出願母集団(出願数) 出願数合計	識別対象者の発明（陽性） 共同発明	識別対象者の発明（陽性） 単独発明	同名異人の発明（陰性） 共同発明	同名異人の発明（陰性） 単独発明
Ⅰ	A氏	331	253	2	75	1
Ⅱ	B氏	832	28	24	622	158
Ⅲ	C氏	108	54	0	32	22
Ⅳ	D氏	1,621	38	1	1,149	433
Ⅴ	E氏	910	86	2	664	158

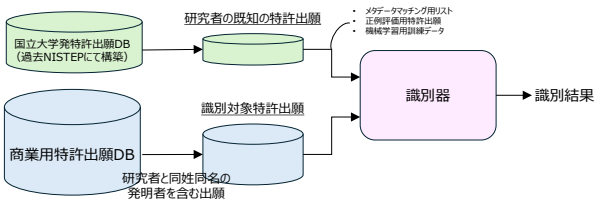


図 1 特許出願の識別手順イメージ

4. 識別手法の評価方法

識別とは、同姓同名の発明者を含む特許出願から、目的の研究者が関与したものを特定する作業を指す。ここでは、二値分類により識別した特許出願を実際の結果と比較し評価する[2][3]。

4. 1 二値分類による評価

4. 1. 1 混同行列

混同行列（Confusion Matrix）は、二値分類問題における識別結果と実際の関係を図 2 の表形式で可視化する手法である。データは 4 つのカテゴリに分類され、評価指標を算出するための基本データとして扱う。

		識別結果 陽性	識別結果 陰性
実際	陽性	(1)真陽性 (True Positive, TP)	(3)偽陰性 (False Negative, FN)
実際	陰性	(2)偽陽性 (False Positive, FP)	(4)真陰性 (True Negative, TN)

図 2 混同行列

(1)真陽性（True Positive, TP）:
対象研究者の特許出願を正しく識別した場合。

(2)偽陽性（False Positive, FP）:
同姓同名の別人の特許出願を誤って対象者の特許出願と識別した場合。

(3)真陰性（True Negative, TN）:
対象研究者の特許出願ではないと正しく識別した場合。

(4)偽陰性（False Negative, FN）:
対象研究者の特許出願を誤って別人の特許出願と識別した場合。

4. 1. 2 評価指標

識別手法の評価は、識別結果として得られる混同行列の各要素を基に、以下の指標を用いて行う。

(1)精度 (Accuracy):
対象研究者及び同姓同名の別人の特許出願を正しく識別できた割合。

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

(2)適合率 (Precision):
対象研究者の特許出願であると識別された中で、実際に正しかった特許出願の割合。

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

(3)再現率 (Recall):
実際に対象研究者が発明者である特許出願のうち、正しく識別された割合。

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

(4)マシューズ相関係数 (Matthews Correlation Coefficient, MCC):
陽性出願の不均衡に対しても強く、識別結果が実際のデータとどれだけ一致しているかを示す指標である。性能評価には、適合率と再現率のバランスを評価しつつ、陽性クラスの重視が求められる場合に F1 スコアがよく使用されるが、ここでは陰陽両クラスを含めた評価が可能な MCC を主に使用する。

$$\text{MCC}=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

4. 1. 3 確定的二値分類と確率的二値分類

二値分類には、確定的に分類される場合と、確率的な要素を含む場合がある。

確定的二値分類（Deterministic Binary Classification）は、特定のルールや基準、例えば、特許出願メタデータである共同発明者や発明者住所の一致に基づき、当該発明者が目的の人物と同一人物であるかどうかを確定的に二値に分類する手法である。この手法の識別性能は、共同発明者や発明者住所など、メタデータの充実度に依存する。

確率的二値分類（Probabilistic Binary Classification）は、発明者が目的の人物と同一人物であるかを、確信度や類似度の閾値に基づいて分類する手法である。この手法は確率的な要素を含み、閾値によって識別性能を調整できる。ROC 曲線や PR 曲線は、異なる閾値での識別性能を視覚化し、真陽性率（TPR）、偽陽性率（FPR）、適合率（Precision）、再現率（Recall）に対する影響を示す。これにより、最適な閾値を選択して識別性能を最適化できる。

本稿で用いた識別手法では、特許出願メタデータ及び引用情報ネットワークは確定的二値分類に、技術分類コード、文書の類似性及び機械学習は確率的二値分類に区分される。

4. 2 ROC 曲線と AUC

ROC 曲線(Receiver Operating Characteristic curve)は、識別手法の性能を評価するために使用され、真陽性率（TPR）と偽陽性率（FPR）の関係を示している（図 4 上図参照）。AUC（Area Under the Curve）は、この曲線の下の面積であり、値が 1 に近いほど性能が良いことを意味する。

4. 3 PR 曲線と AUC-PR

PR 曲線 (Precision-Recall curve)は、適合率と再現率の関係を示し、不均衡な出願母集団の性能評価に有用である（図 4 下図参照）。AUC-PR は、この曲線の下の面積であり、識別手法の性能を数値で評価する指標である。AUC と同じく、値が 1 に近いほど性能が良いことを意味する。

5. 識別結果と評価

5. 1 確定的二値分類

図 3 は、確定的二値分類の各識別手法について、表 2 に示した 5 つのケースの識別結果を評価指標別にボックスプロットで表示している。（見方は図 7 右下の説明を参照）

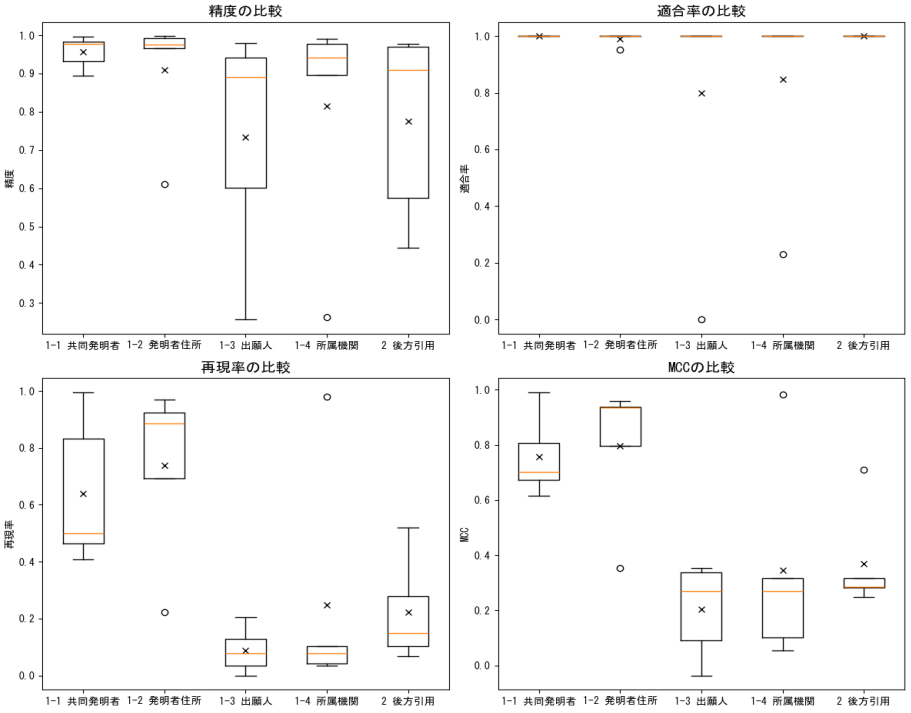


図 3 確定的二値分類による識別手法の性能評価

結論から言えば、特許出願メタデータとして、「1-1_共同発明者」及び「1-2_発明者住所」を用いた二つの識別手法が実用に足る性能を示し、その他の識別手法は識別性能が低いことが明らかとなった。

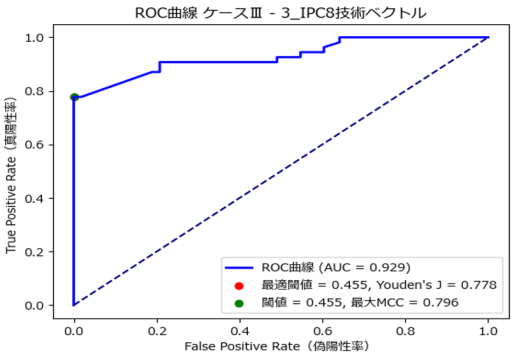
1-1_共同発明者を用いた手法では、適合率は 1.000 で、精度も高い（平均 0.957、中央値 0.977）。ただし、再現率に難があり、平均 0.640、中央値 0.500 と比較的低く、MCC は平均 0.757、中央値 0.701 で標準偏差や範囲も大きく、ケースごとの識別性能にばらつきがある（値は単独発明者の特許出願を除いた母集団による）。但し、これは手法の問題ではなく、マッチングに用いる共同発明者リストの網羅性に依存するものである。

1-2_発明者住所を用いた手法は、再現率が平均 0.738、中央値 0.885 で、MCC も 0.936 と高い水準を示し、実際の陽性出願を多く識別できており、共同発明者を用いた手法に比して良好な識別性能を示している。

これら二つの識別手法と対照的に、「1-3_出願人」以降の 3 つの識別手法は、図 3 で明らかなように、識別器の性能として不十分な値を示している。有意性の検定（Friedman 検定）においても、1-1、1-2 の手法とは適合率を除いた各指標で有意差があることを確認している。

5. 2 確率的二値分類

4.1.3 項にて説明したように、確率的二値分類では閾値によって識別性能の調整が可能である。一例として、図 4 に 3_IPC8 技術ベクトルの類似性を用いた手法のケースⅢに関する ROC 曲線と PR 曲線を示した。ここでは Cos 類似度の閾値を変えながら、ROC 曲線は真陽性率（TPR）と偽陽性率（FPR）の関係を、PR 曲線では適合率（Precision）と再現率（Recall）の関係を示している。ここから算出した MCC を最大化する（識別性能を最大化する）閾値は 0.455 であり、図 5 に見る通り適正な値が算出されている。図 5 は Cos 類似度の階級別に陰性・陽性両クラスの特許出願数をヒストグラムで表したものである。



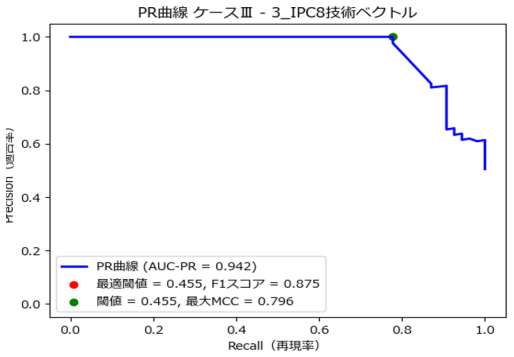


図 4 ROC 曲線と PR 曲線

図 7 は、確率的二値分類の各識別手法について、同じく表 2 に示した 5 つのケースの識別結果をもとに、評価指標別にボックスプロットで表示したものである。ここでは、識別方式全体の性能である AUC と AUC-PR 及び最適閾値における MCC を中心に評価する。

まず、3_IPC8 技術ベクトルの類似性を用いた手法は、AUC は中央値 0.986、平均値 0.974、標準偏差 0.024、範囲 0.067 とばらつきも少なく全体的な識別性能に優れている。一方、AUC-PR は中央値 0.937、平均値 0.889、標準偏差 0.089、範囲 0.234 で、最大・最小値に差のあるばらつきが見られ、陽性出願の識別性能がやや劣るケースが存在している。例えば、ケースⅠでは AUC が 0.986、AUC-PR も 0.996 と非常に高い一方、ケースⅣでは偽陽性（9 件）、偽陰性（10 件）の影響で適合率、再現率が低くなり AUC-PR は 0.762 と低くなり、陽性出願の識別性能に課題がある。

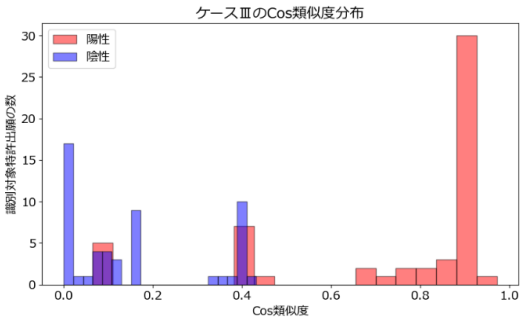


図 5 識別特許出願の Cos 類似度分布

次に、4-1_文書の類似性（TF-IDF ベクトルを使った類似性）を用いた手法では、AUC の中央値 0.971、平均値 0.974、標準偏差 0.014、範囲 0.043 と識別性能に優れ、3_IPC8 技術ベクトルとの違いは、AUC-PR も一 0.955 から 0.997 の範囲に収まっているが、MCC は 0.785 というケースもある。これは、図 6 の類似度ヒートマップに見るように、識別対象特許出願の中に対象者の既知特許出願との対比において、対象者の出願であっても類似性の低い出願や別人の出願であっても類似性の高い出願が混在しており、前者は偽陰性、後者は偽陽性の特許出願として誤識別された出願が存在するためである。

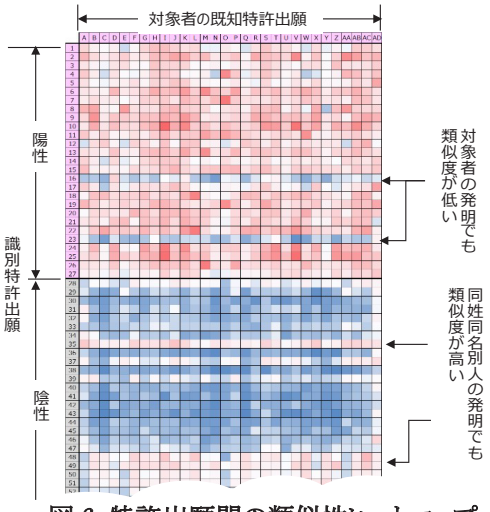


図 6 特許出願間の類似性ヒートマップ

4-2_文書の類似性（技術キーワード出現パターン）を用いた手法では、同じ文書の類似性でも 4-1 の手法に比して AUC 及び AUC-PR の値は小さくケースによるばらつきが大きく見られる。適合率や再現率もケースにより変動し、MCC も 0.626 から 0.928 と大きな差が見られた。これは、ケースごとに異なるデータの特性や、識別対象の特許出願の技術内容の違いが影響しているためである。

機械学習（Random Forests, RF）を用いた手法の中で、5-1_TF-IDF による文書特徴化と RF 学習を用いた手法は、AUC-PR が高水準で安定しており、0.960 から 1.000 の範囲であった。適合率や再現率も非常に高く、MCC も全体的に高く、特にケースⅢでは 1.000 を記録するなど、非常に高い性能を発揮することが確認された。

機械学習の 5-2_技術キーワードパターンの RF 学習を用いた手法も、AUC-PR が 0.951 から 1.000 と高水準で安定していた。適合率や再現率も高く、特にケースⅢで 1.000 を記録するなど、非常に高い性能を示している。MCC もケースによって 0.859 から 1.000 と高い数値を示しており、全体的に高い性能を発揮する手法である。

総合的に見ると、機械学習を用いた手法が最も高い識別性能を示しており、5-1 の手法が、MCC を含めたすべての評価指標において高スコアを得ており、識別対象者に依存しない、ばらつきの少ない識別性能を発揮している。

5-2 の手法も紙一重の性能差であるが高スコアを得ている。この手法は、対象者の研究実績に基づく技術キーワードの出現パターンを学習する手法であり、発明の技術的な特徴が研究実績の中で創出されている場合に強みを発揮する。ただ、「KAKEN: 科学研究費助成事業データベース」の研究者情報に掲載されていない対象者も考えられ、汎用性に限界がある。他の識別手法、特に「3_IPC8 技術ベクトルの類似性」や「4-1_TF-IDF 法を使ったコサイン類似度」は、全体的に高い AUC-PR を示しているものの、最大 MCC の値では 5-1 及び 5-2 の手法に劣るケースが多い。例えば、4-1_TF-IDF 法を使ったコサイン類似度による手法は、AUC-PR が各ケースで 0.95 以上の値を示す一方で、最大 MCC ではやや低い値を示し、特定の閾値における識別性能が不安定であることが見て取れる。

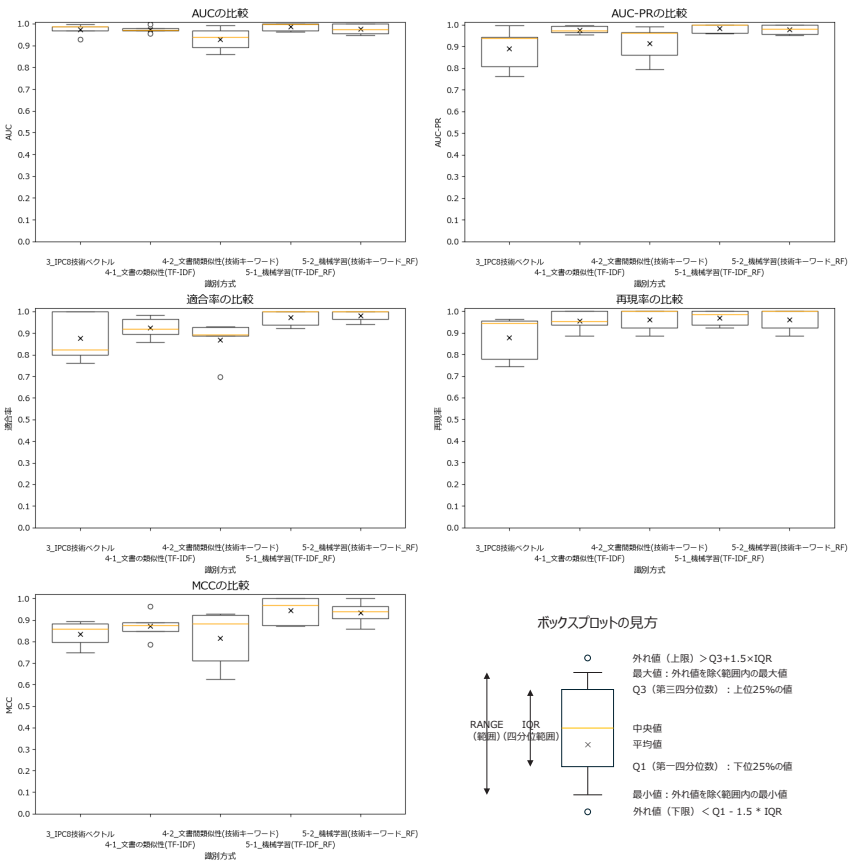


図 7 確率的二値分類による識別手法の性能評価

6. まとめと改善策

各識別手法の中では、一部の特許出願メタデータを使用した手法のみが実用性に欠けており、それ以外の手法は、優劣はあるものの、実用性に欠けるものはなかった。しかし、確率的二値分類に属する識別手法では、手法の違いにかかわらず、既知の特許出願との技術的な類似性を基に識別を行う際に、誤識別が特定の特許出願に集中する傾向が見られた。

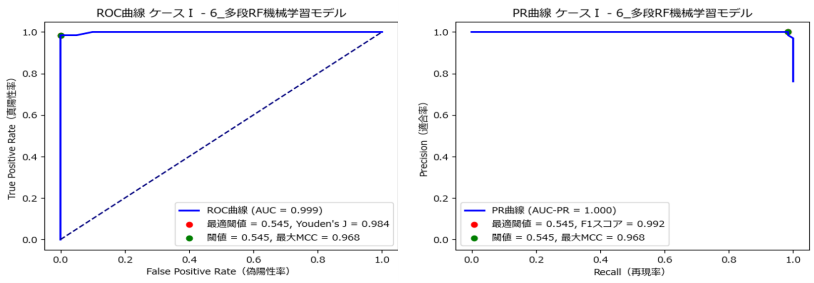


図 8 多段 RF 機械学習モデルによる識別結果

この結果から、技術領域が類似する発明を技術的類似性のみで陽性・陰性クラスに識別することには限界があることが明らかになった。そこで、確定的二値分類に分類される識別手法の中で、成績が良好であった 1-1_共同発明者と 1-2_発明者住所、さらに 5-1_機械学習（TF-IDF による文書特徴化と RF 学習）を組み合わせた「多段 RF 機械学習モデル」を考案し、その適用を試みた。この手法は、特許出願全文を用いた RF 機械学習による識別に加え、RF を共同発明者や発明者住所のバイナリマッチングにも適用することで、識別精度を向上させることを目指したものである。

この手法の識別結果は図 8 に示す通り、本稿で使用した 5 つのケースにおいて、全特許出願の中で誤識別（偽陰性）はケースⅠの 1 件のみという非常に高い性能を示した。

【参考文献】
[1] 科学技術・学術政策研究所. (2017).『国立大学の研究者の発明に基づいた特許出願の網羅的調査』. 調査資料-266, 科学技術・学術政策研究所.
[2] G-K. (2024). "ROC 曲線と PR 曲線-分類性能の評価方法を理解する", Qiita, https://qiita.com/g-k/items/b47b9b0ee2015a3b0b94 (参照 2024-8-10)
[3] チームカルポ. (2023).『Python 統計分析＆機械学習マスタリングハンドブック』. 秀和システム.