D-003

# 特許埋め込みを用いた企業ポートフォリオ情報による提携企業候補選定手法の基礎検討

久保 春喜 [1] 伊藤 由佳 [2.] 西本 恵太 [1.] 浅谷 公威 [1.] 坂田 一郎 [1.]
Harunobu Kubo Yuka Ito Keita Nishimoto Kimitaka Asatani Ichiro Sakata

東京大学 [1] ダイキン工業株式会社 [2]

## 1. 概要

企業の持続的成長には外部組織との連携が不可欠であり, 有望な提携候補の探索は重要な経営課題である. 本研究は, 特許情報を活用した新たな提携候補選定手法を提案する. 特許文書を大規模言語モデルでベクトル化し企業技術ポートフォリオを生成, さらに異なる技術分野間の潜在的補完性（シナジー）を行列として定量化した. このシナジー情報と機械学習モデルを組み合わせ, 企業間の技術的親和性に基づく提携候補選定の精度を検証した. M&A 事例を代理データとした評価の結果, 提案手法は IPC 分類に基づく従来手法と比較し, Best F1 Score で顕著な性能向上（0.22→0.51）を示した. 特に技術同士の関連性に注目したシナジー指標が精度向上に大きく貢献し, 本手法がデータ駆動型の高精度な提携戦略支援に寄与する.

キーワード: 特許分析, 提携候補選定, 機械学習, 技術シナジー, 埋め込み表現

## 2. 諸言

企業の競争力維持・強化において戦略的提携は重要であり, 特に技術主導型産業では有望な提携候補の効率的探索が喫緊の課題だが, 技術情報の複雑性から実現は容易ではない.

従来, 特許情報を活用する試みとして, 「技術ポートフォリオが類似する企業は提携しやすい」との仮説に基づき, 国際特許分類（IPC）を技術分類軸とするアプローチが存在する. これらは主に IPC ベースのポートフォリオベクトル間の類似度で企業間の技術的近接性を評価し, 提携候補を予測していた[1][2]. しかし, IPC 分類に基づく手法には限界が指摘される. IPC は固定的な分類で技術進化への追従が困難な上, 特許の「意味内容」のニュアンスや IPC レベルで捉えきれない特許間の微細な近接性, 異なる IPC 分類間の意味的「距離」の定量的評価も難しかった.

本研究はこれらの課題に対し, 特許文書の**埋め込み表現**の活用を提案する. 特許抄録を大規模言語モデルでベクトル化しクラスタリングすることで, 従来 IPC では困難だった詳細粒度での技術分類と, 技術分類間の意味的関係性の解釈を可能にする[3]. さらに, この新たな技術表現を基盤とし, 企業間の技術的相互作用や補完性を捉える新たな「**技術シナジー指標**」（**Personalized PageRank (PPR)[4]**や**Pointwise Mutual Information (PMI)**等）を設計・導入した. これらの特徴量と機械学習モデルを用い, 過去の提携成立事例（M&A 事例を代理データとして使用）から提携成立確率を予測し, 有望な提携候補を高精度に選定することを目指し, その有効性を検証する.

## 3. 提案手法

本研究では, 企業間の将来的な提携成立を予測する機械学習モデルの入力特徴量として, 特許の埋め込み表現を核に, 企業間の技術的関連性や潜在的シナジーを多角的に捉える指標群を設計した. 主要な特徴量群と, その設計における仮説・期待効果は以下の通りである.

・**基本ポートフォリオ特徴量:**技術ポートフォリオの類似性が提携可能性を示唆するとの仮説に基づき設計した.各特許の Abstract（抄録）の埋め込みベクトル群に対し, K-means クラスタリング（クラスタ数：1652）を行い, 技術分類軸を定義する.この技術分類に基づき, 各企業が各技術分類に保有する特許数を要素とし, 企業全体の特許数で L1 正規化することで, 企業の**技術ポートフォリオベクトル**を構築する. このベクトル間の類似性を内積 (dot_A_T_pre)やコサイン類似度 (portfolio_cos_1652dim)で評価する.

・**企業 Centroid 特徴量:**企業全体の技術焦点の近接性やその時間変化が提携の兆候となるとの仮説に基づき, 全特許埋め込みの平均ベクトル（企業 Centroid）を利用. 静的類似度 (cos_c 等) 及び, 累積 Centroid 間類似度の時系列変化とその傾き (cum_cos_v3_y-*等) を動的特徴量として導入.

・**技術間シナジー特徴量:** 直接的なポートフォリオ類似性に加え, 異なる技術群が相互に作用し補完しあうことで生まれる潜在的な価値が, 提携の可能性を強く示唆するという仮説に基づき設計した.

  ・**PPR シナジー (ppr_synergy):** 各技術分類の代表埋め込みベクトル間類似度を重みとするグラフ上で, 各企業の技術ポートフォリオを初期影響力とし PPR[4]を計算. 影響が「拡散」した新ポートフォリオ表現ベクトル間の内積を計算し, 潜在的技術近接性や影響可能性を評価する.

  ・**PMI シナジー (synergy_pmi_Ddim):** 企業群の特許データから, 二技術分類の「同時保有」頻度が偶然の共起を統計的に超える度合いを PMI で定量化. この技術分類間 PMI 値を要素とする行列(1652*1652)を介したポートフォリオベクトル間の内積を計算することで, 両社が統計的に関連・補完し合う技術ペア群のカバー度合いを反映.

  基本 1652 次元及び階層 300,600,1000 次元分類で評価.

これら特徴量を機械学習モデルの入力とする.

## 4. 実験

### 4.1 実験設定

本研究の予測モデル構築と評価に用いたデータ及び設定の概要は以下の通りである.

**特許データ**は, 企業・特許データベースより 2010 年～2020 年の「Semiconductor」関連特許約 200 万件を抽出し,

OpenAI 社製 Embedding モデル (text-embedding-3-small) で抄録の埋め込みベクトルを生成した.

**M&A データ**は, 2015 年〜2020 年の完了済買収案件のうち, Acquiror/Target 双方が上記期間に特許を保有する 458 件を提携成立のポジティブサンプルとした. 各ポジティブサンプルに対しネガティブサンプルを 1:20 で作成した[2].

**評価**は, Best F1 Score（最適閾値での F1）, PR-AUC, AUC-ROC で行う.

比較実験は次の 3 系列で構成する.

1. IPC ベース閾値法[1][2]：IPC 区分で作ったポートフォリオベクトルの内積とコサイン類似度 を計算し, F1 が最大となるカットオフ値を求めて二値判定する.
2. 埋め込みベース閾値法：上記と同手順だが, IPC の代わりに　LLM 埋め込みに基づく技術分類を利用.
3. 機械学習モデル（提案手法）：作成した全特徴量を投入し, LightGBM と Random Forest で学習.

この 3 系列を同一データ分割で比較し, 提案特徴量の効果を検証した.

## 4.2 主要結果と考察

予測結果を表 1 に示す. 先行研究で用いられる IPC ベース閾値法（IPC-cos: Best F1 0.218）に対し,埋め込みベース閾値法（Emb-cos_port）は Best F1 を 0.378 へと大幅に改善した. この結果は, 特許埋め込みが IPC 分類と比較して技術の「意味内容」をより精密に捉え, 基本的な技術的近接性評価の精度を向上させることを示唆している.

さらに, 本研究で設計した全特徴量（基本ポートフォリオ特徴量, 企業 Centroid 特徴量, 技術間シナジー指標）を LightGBM モデル（LGBM-Full）に適用した場合, Best F1 Score は 0.511 (PR-AUC 0.483) に達し, 比較した全てのモデルを大きく上回る性能を示した（表 1）. この性能向上は, 単に技術表現の高度化だけでなく, 企業間の多様な技術的関係性や潜在的シナジーを捉える特徴量群と, それらを統合的に学習する機械学習モデルの組み合わせが極めて有効であることを示している.

表 1. モデルの予測性能比較.

| モデル／特徴量セット | ROC-AUC | PR-AUC | Best F1 |
|---|---|---|---|
| IPC-cos (Thresh.) (先行研究) | 0.692 | 0.147 | 0.218 |
| IPC-dot (Thresh.) (先行研究) | 0.689 | 0.131 | 0.219 |
| Emb-cos_port (Thresh.) | 0.799 | 0.290 | 0.378 |
| Emb-dot_port (Thresh.) | 0.708 | 0.179 | 0.301 |
| **LGBM (Full Feats.) (提案)** | **0.878** | **0.483** | **0.511** |
| RF (Full Feats.) (提案) | 0.893 | 0.411 | 0.487 |

注: *Thresh.* モデルの *ROC-AUC, PR-AUC* はスコアベース。*Best F1* は各モデルの最適閾値での値。

LGBM-Full モデルにおける特徴量重要度（LightGBM の”gain”に基づく. 図 1）を分析すると, PMI シナジー指標（synergy_pmi_300dim, synergy_pmi_1652dim 等）, 企業の特許規模（pat_count_a）, ポートフォリオ間コサイン類似度（portfolio_cos_1652dim）, そして PPR シナジー（ppr_synergy）などが上位を占めた. この結果は, 特に本研究で提案したシナジー関連指標群が高い予測貢献度を持つことを明確に示している.

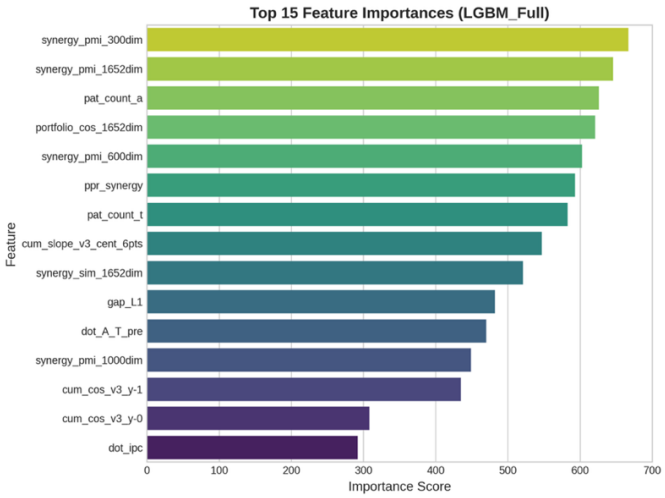


図 1 . LightGBM 重要特徴量.

注目すべきことに, 異なる粒度での PMI シナジー指標群がトップグループを形成した. これは, 提案手法で意図した通り, 企業が統計的に有意に「共に保有する」技術クラスタペアのパターンが, 過去の成功技術の組み合わせや補完的技術セットを示唆し, 提携予測において強力なシグナルとなることを裏付けている. 詳細レベル（1652 次元）と粗視化レベル（300 次元等）の PMI の双方が有効であったことは, ミクロな専門技術とマクロな技術分野双方の連携の重要性を示唆する.

PPR シナジーも同様に高い重要度を示した. これは, 技術クラスタグラフ上での影響の拡散を捉える PPR が, 直接的な技術重複の少ない企業間の「隠れた技術的経路」や潜在的影響力を効果的に数値化し, 従来の類似度指標では見過ごされがちな提携可能性の検出に貢献した可能性を示している.

基本的な埋め込み類似度や特許規模も依然として重要であり, これらの基礎情報と, PMI や PPR といった高度で多角的なシナジー指標の組み合わせが, 最終的な高い予測精度達成の鍵であると言える.

## 5. 結論

本研究は, 特許埋め込みと多様な技術シナジー指標（特に PPR および PMI）を組み合わせた機械学習モデルが, M&A 提携候補選定において高い予測精度を達成できることを示した. 提案手法は, Best F1 Score で先行研究ベースの閾値法を 0.22 から 0.51 へと大幅に改善し, 特に PPR シナジー等の新規特徴量の有効性が確認された[1][2].

### 参考文献

[1] Albora Giambattista, Straccamore Matteo, Zaccaria Andrea, “Machine Learning-Based Similarity Measure to Forecast M&A from Patent Data”, *arXiv preprint arXiv:2404.07179* (2024).

[2] Arsini Lorenzo, Straccamore Matteo, Zaccaria Andrea, “Prediction and Visualization of Mergers and Acquisitions Using Economic Complexity”, PLOS ONE, Vol. 18, No. 4, e0283217 (2023).

[3] Whalen Ryan, Lungeanu Alina, DeChurch Leslie, Contractor Noshir, “Patent Similarity Data and Innovation Metrics”, Journal of Empirical Legal Studies, Vol. 17, No. 3, pp. 615–639 (2020).

[4] Jeh G., Widom J., “Scaling Personalized Web Search”, Proceedings of the 12th International World Wide Web Conference (WWW 2003), pp. 271–279 (2003).