特許分類における機械学習の課題と解決策となる特許自動分類ツール

― PatentNoiseFilter ―

A Patent Auto-Classification Tool (PatentNoiseFilter) as a Solution to the Issues of Machine Learning in Patent Classification

IRD 国際特許事務所　所長・弁理士／株式会社アイ・アール・ディー
谷川　英和

1986 年神戸大学工学部システム工学科卒業。同年、松下電器産業（株）［現パナソニック］に入社し、中央研究所等において、データベース管理システム等の研究開発に従事。1997 年同社知的財産権部門に異動。1999 年弁理士試験合格。2002 年 1 月、IRD 国際特許事務所を開設。所長、弁理士。2003 ～ 2007 年 3 月京都大学 COE 研究員、2007 年 4 月～京都大学非常勤講師、2011 年 4 月～大阪大学非常勤講師（現招聘教授）2019 年 4 月～関西学院大学非常勤講師、博士（情報学）。弁理士会、日本知財学会、情報処理学会各会員。2007 年度から特許版産業日本語委員会委員。

① はじめに

特許を自動分類する技術に対するニーズが大きく、機械学習を用いた特許自動分類のツールは、多数、存在している。機械学習を用いた特許自動分類のツールは、通常、ユーザが分類した教師データをＡＩに学習させ、学習器を取得する学習モジュール（図１参照）と、学習器を用いて特許データの分類予測を出力する予測モジュール（図２参照）とを有する。

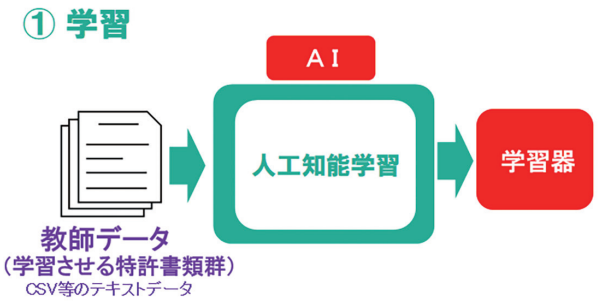

図１　学習モジュールの概念図

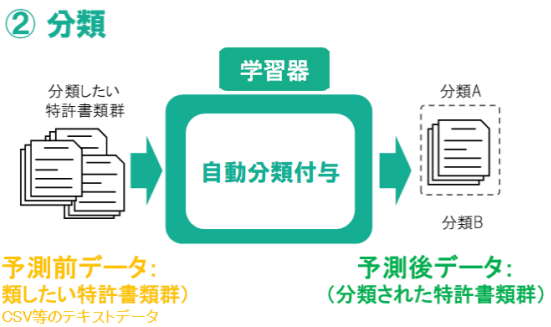

図２　予測モジュールの概念図

② 機械学習を用いた特許自動分類の特性と課題

機械学習を用いた特許自動分類には、以下の４つの特性があり、各特性に起因する課題がある。

（１）アルゴリズム依存性

特許分類に利用可能な機械学習には、深層学習、ランダムフォレスト、決定木、サポートベクターマシン（SVM）等の複数のアルゴリズムがあり、各アルゴリズムにおいて、多数のモジュールやツールが開発されている。

一方、ユーザが与える教師データの内容、数、分類の候補の数等により、使用するべきアルゴリズムは異なり、基本的に、実際に分類を行い、精度を確かめないと、どのアルゴリズムが適切かは分からない。

一方、既存の特許自動分類ツールの多くは、固定の一つのモジュールを用いており、精度の高い学習器が構成される場合もあるが、精度が上がらない学習器が構成される場合もある、と考えられる。

（２）データ依存性

一つの特許データには、特許請求の範囲、明細書、要約書、特許分類コード（例えば、IPC、FI、F ターム）等の情報が含まれる。また、明細書には、【発明が解決しようとする課題】【発明の効果】【発明を実施するための形態】における種々の説明文等の情報が含まれる。これらの情報のうち、どの情報を使用すれば、精度の高い特許自動分類が可能な学習器（学習モデル等と言っても良い）が構築できるかをユーザが判断することは極めて困難である。

特許データのうちのどの情報を用いるべきかも、基本的に、実際に分類を行い、精度を確かめないと判断できない。

一方、既存の特許自動分類ツールにおいて、ユーザが与えた特許データのうち特許自動分類ツールが予定している特許データを固定的に使用する、またはユーザが与えた特許データをすべて使用して学習器を構築することが一般的である。

なお、図３は、我々が用意する４つのアルゴリズム（ディープラーニング１，ランダムフォレスト，ディープラーニング２，Few-shot）と分類に使用する情報とを変更した場合の精度、適合度、再現率、およびＦ値の４つの各指標の実験結果を図３に示す。分類に使用する情報は、ここでは、「要約書と分類コード」を用いた場合と、「要約書と特許請求の範囲と分類コード」を用いた場合の２通りである。また、図３において、比較的、簡単な二値分類（○または×）である。図３において、４つのアルゴリズムの判定結果の多数決、AND 演算、およびＯＲ演算についての４つの各指標の実験結果が含まれる。また、図３において、各指標の最大値（最良値）が１であり、最小値（最悪値）が 0.5 である。そして、図３において、赤字・太字は、各指標において最良値であり、赤字・細字は、各指標において最悪値である。

実験結果によれば、用いるアルゴリズムと用いる情報とにより、精度では「0.092（9.2%）」、適合率では「0.243（24.3%）」、再現率では「0.326（32.6%）」、およびＦ値では「0.084（8.4%）」の差が生じたことを示す。

（３）ユーザ指向性

特許分類が必要な局面として、研究プロジェクトの発足時の大規模特許調査、製品開発の開始時の大規模特許調査、特許出願前の先行特許調査、SDI 特許調査、他社特許権を無効にするための先行特許調査等の種々の局面がある。

そして、特許分類の局面により、「再現率」を上げたい場合、「適合率」を上げたい場合、「Ｆ値」を上げたい場合、「精度（正解率）」を上げたい場合等、上げたい指標が異なる。なお、再現率とは「検索された正しい特許数／本来検索されるべき全ての特許数」であり、見逃さない確率である、と言える。また、適合率とは「検索された正しい特許数／検索された特許数」であり、誤検出しない確率である、と言える。また、Ｆ値は「（再現率＋適合率）／２」である。

一方、既存の特許自動分類ツールにおいて、ユーザが上げたい指標を指定できる機能を有しないことが多い。

（４）ブラックボックス性

機械学習の予測処理を行った結果の精度が見えないブラックボックス性の課題がある。

つまり、教師データの数が不足しているため、分類に与える特許データが不適切であるため、アルゴリズムが分類対象の特許データとミスマッチであるため等の原因により、精度の高い学習器が構築できていない場合も多い。

一方、特許自動分類ツールの学習処理により構築された学習器の精度が分からずに、学習器を使用せざるを得ない場合には、ユーザは不安を持ちながら、特許自動分類ツールが使用している。

アルゴリズム	使用する情報	精度	適合率	再現率	F値
アルゴリズム1（ディープラーニング1）	要約書，分類コード	0.776	0.757	0.772	0.762
アルゴリズム2（ランダムフォレスト）	要約書，分類コード	0.822	0.783	0.863	**0.820**
アルゴリズム3（ディープラーニング2）	要約書，分類コード	0.787	0.777	0.772	0.769
アルゴリズム4（Few-shot）	要約書，分類コード	0.833	0.845	0.803	**0.820**
4つのアルゴリズムの多数決	要約書，分類コード	0.812	0.784	0.825	0.803
4つのアルゴリズムのAND	要約書，分類コード	0.805	0.926	0.636	0.751
4つのアルゴリズムのOR	要約書，分類コード	0.769	0.687	**0.947**	0.795
アルゴリズム1（ディープラーニング1）	要約書，特許請求の範囲，分類コード	0.748	0.724	0.750	0.736
アルゴリズム2（ランダムフォレスト）	要約書，特許請求の範囲，分類コード	0.812	0.783	0.833	0.806
アルゴリズム3（ディープラーニング2）	要約書，特許請求の範囲，分類コード	0.805	0.793	0.803	0.796
アルゴリズム4（Few-shot）	要約書，特許請求の範囲，分類コード	**0.840**	0.865	0.779	0.819
4つのアルゴリズムの多数決	要約書，特許請求の範囲，分類コード	0.826	0.809	0.826	0.817
4つのアルゴリズムのAND	要約書，特許請求の範囲，分類コード	0.801	**0.929**	0.621	0.743
4つのアルゴリズムのOR	要約書，特許請求の範囲，分類コード	0.766	0.686	0.931	0.789
差		**0.092**	**0.243**	**0.326**	**0.084**

図３　アルゴリズムと使用情報を変更した場合の精度等の実験結果

③ 最適学習器構築技術を搭載したPNF（PatentNoiseFilter）

ユーザが安心して、精度の高い特許自動分類を行うために、最適学習器構築技術を搭載した特許自動分類システム（PNF）を開発した。

最適学習器構築技術の特徴は、以下の（１）から（３）の３点である。

（１）複数の各モジュールと内容の異なる複数の各特許データとの多数の組を用いた学習器の評価

第一には、４つの各アルゴリズムに対応する４つの各モジュールに、種々の組み合わせの特許データ（例えば、「要約書」または「特許請求の範囲」）を与え、モジュールと特許データとの組を、多数、構成し、組ごとに、ユーザから指定された指標における学習器の精度を測定する点である[1]。その結果、特定の指標の値が最大となるモジュールと特許データとを用いた学習器が自動構築できる。

（２）ユーザ指定の指標の評価値を最大にする学習器

第二は、再現率、適合率、Ｆ値、精度（正解率）の４つの指標のうち、ユーザが指定した指標の値を測定し、その値が最大となる学習器が自動構築できる点である。

（３）複数の各モジュールと内容の異なる
各特許データとの組ごとの論理演算結果の評価

第三は、使用する特許データの内容ごとに、複数のモジュールの予測結果の論理演算（ＡＮＤ、ＯＲ、多数決）の結果についても、ユーザから指定された指標における学習器の精度を測定する点である。その結果、複数のモジュールの予測結果の論理演算の結果も含めて、ユーザから指定された指標の値が最大となるモジュールと特許データとを用いた学習器が自動構築できる。

（４）最適学習器構築技術の成果物

最適学習器構築技術において、以上の第一から第三の点を実行し、ユーザから指定された指標の値が最大となるモジュールと特許データとの組または論理演算を決定し、当該情報（学習器仕様情報と言う）と対に、最適学習器を蓄積する。

なお、学習器仕様情報とは、ユーザが重視する指標の値を最大にする学習器の仕様であり、予測処理に使用するモジュールおよび特許データを特定する情報である。

予測処理において、学習器と対に格納されている学習器仕様情報を用いて、使用する特許データと使用するモジュールが決定され、決定した特許データとモジュールと学習器とを用いて、特許を分類する。

④ SDI 調査における PNF の利用

4.1　SDI における学習器管理技術

SDI（Selective Dissemination of Information）とは、ユーザが指定した検索式に合致する特許データを、定期的に自動提供する情報配信サービスである。ここで、検索式に合致した特許データをユーザに提供するだけではなく、特許自動分類ツールにより自動分類された結果をユーザに提供することが望まれている。

そこで、我々は、検索式情報と学習器（分類器）とを対にして管理し（図４参照）、ユーザが指定した検索式に合致する特許情報を取得し、かつ検索式と対なる学習器を用いて、機械学習の予測処理を行い、分類結果を含む特許情報を構成し、ユーザに配信するユニークな技術を開発し、提供している[2]。このような、１または２以上の各検索式情報と学習器とを対にして管理し、利用する我々の特許技術を学習器管理技術と呼んでいる（図４参照）。

ID	検索式情報	分類器
1	公開日:2010/03/01-2010/03/31	分類器A
2	IPC:A or IPC:B or IPC:G or IPC:H	分類器B
3	1 and 2	分類器C
：		

図４　SDI における学習器管理技術

4.2　最適学習器を用いた SDI 調査

SDI 調査における検索式情報ごとに、ユーザによる過去の分類結果（教師データ）を PNF に与えて、学習処理を実行することにより、上述したように、ユーザ条件に対応する最適な学習器を得ることができる。なお、ユーザによる過去の分類結果は、例えば、関連特許（○）と非関連特許（×）とに分類されている。

そして、SDI 調査を支援する SDI システムにおいて、検索式情報に対応付いて、学習器が管理される。

次に、SDI システムにおいて、発行される公開特許公報または登録公報に対して、ユーザが指定した検索式に基づいて、多数の公開特許公報等を自動抽出する。

次に、PNF の予測モジュールに、検索式情報と対になる学習器と自動抽出した多数の公開特許公報等とを与え、PNF の予測モジュールを実行する。すると、PNF の予測モジュールは、各公開特許公報等に対する分類結果（「○」または「×」）とスコアとを対応付けた CSV ファイルを出力する。

次に、SDI システムにおいて、分類結果およびスコアをキーとして降順にソートした CVS ファイルを構成

し、蓄積する。

以上の処理により、毎週発行される公開特許公報等をチェックする必要がある研究者、技術者、または知財担当者は、AI が重要であると判断した特許情報から順番に内容を確認でき、また、スコアの高い（信頼性の高い）特許情報の確認が不要となり、SDI 調査の効率が大幅に向上する。

5 まとめ

機械学習を用いた特許自動分類の４つの特性と課題とを詳述した。そして、４つの課題をすべて解決できる独自技術である、最適学習器構築技術を搭載した特許自動分類ツール（PNF）について紹介した。

また、独自技術であり、SDI 調査における検索条件ごとの学習器の学習器管理技術を紹介し、かかる学習器管理技術と PNF とを組み合わせた SDI 調査について提案した。

今後、PNF を使用した特許分類の精度をさらに上げるために、機械学習の各モジュールの改善、第５のモジュールの導入等を行っていきたい。

参考文献

[1] 特許第 7620866 号公報
[2] 特許第 6973733 号公報