機械学習による特許分類モデルを「解釈」するための試み

How to Interpret a Patent Classification Model Trained by Machine Learning

株式会社 NTT データ数理システム　データマイニング部グループリーダー・主任研究員
岩本　圭介

1999 年株式会社数理システム（現：株式会社 NTT データ数理システム）入社。データマイニング・テキストマイニングに関わるツール・手法開発及び分析業務に従事。現職はデータマイニング部グループリーダー・主任研究員。

iwamoto@msi.co.jp

1　はじめに

文書を自動的に定められたカテゴリへと分類することは、自然言語処理技術と機械学習技術とが交差する応用分野として多大な関心が寄せられており、その実用性も極めて高いものである。文書分類の適用例としては

- Web から採取した個人の声をポジティブ・ネガティブに分類して、製品やサービスが世の中にどう受け止められているかを知る。
- 蓄積された不具合情報の内容から、問題が発生した部品・部位の種別や発生事象をカテゴライズし、それらの間の関連を把握する。
- コールセンターにおいて、オペレータが手動で付与していた問合せのカテゴリを見直し、分類付与を自動化することで業務効率化と客観性の確保を図る。

など幅広い業務において枚挙に暇がなく、特許文書を扱うに際してもこれは例外ではない。過去の Japio YEAR BOOK においても、機械学習による特許分類の付与については文献[1]，[2]で論じられており、また特許検索の効率化のため「正解文書」群を選り分けることについては 文献[3]で言及されている。

一方、機械学習による分類を業務に適用させようとした場合、そのブラックボックス性が問題となることがある。なぜ機械学習モデルがそのような判断を下したのか、その直接の理由を明示することは一般に難しい。もちろん機械学習モデルが下す判断はその背後に存在するアルゴリズムの所為であり、そういった意味で確たる判断の根拠は存在している。とはいえ、多くの場合それは数式上として表現される複雑な計算結果であって、人間にとって理解しやすい形式であるとはいえない。

しかし、機械学習のブラックボックス性は認めつつも、学習で得られたモデルがどのような特質を持つものか、モデルを「解釈」するための手法も開発・提案されてきており、「解釈可能な機械学習（Interpretable Machine Learning）」として近年研究の領域をなしている。文献[4]では、人工知能の適用における透明性を確保し、利用の上での説明責任を果たすためにこのような機械学習モデルの解釈性についての研究が注目されていることが述べられている。

本稿では、ある特許分類モデルを構築し、そこで生成されたモデルに対して「解釈可能な機械学習」として提唱されている一手法を更に適用させて、文書内でどういった箇所に着目して分類が行われたのか読み取ることを試みる。以降、2 章で手法の解説を行い、3 章で実験の結果を述べる。最後に、4 章で実装に用いたツールの解説とまとめを行う。

2　「解釈」可能な機械学習

2.1　モデルの解釈可能性

数理的な導出を経て判断を下す手法の全てにブラック

ボックス性が潜んでいるわけではなく、作成されたモデルがどのようなものであったか、という点の解釈を試みることできるものもある。一般に、定式化がそれほど複雑ではないものは解釈の余地があり、複雑になればなるほどその解釈は難しくなっていくであろう。

モデルに対する解釈の一例として、文献[5]で解釈可能（interpretable）な手法の一つとして挙げられているロジスティック回帰を挙げる。ロジスティック回帰では、説明変数群 $x_1, x_2, \cdots, x_k$ に対し、それらとある事象が発生する確率 p とを次の式で結びつける。

$$p = \{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)\}^{-1}$$

モデルの構築を行った結果、パラメータである $\alpha, \beta_1, \beta_2, \cdots, \beta_k$ の各量が求まる。$\beta_1, \beta_2, \cdots, \beta_k$ は $x_1, x_2, \cdots, x_k$ に乗ずる係数であり、ある x_i に対する β_i の絶対値が大きいほど x_i の変化に対し p の値が鋭敏に変化することが読み取れる。さらに、β_i の符号が正の場合は x_i の増大は p を上げる方向に寄与し、負の場合は逆であることもわかる。実際は、オッズ比や信頼区間・有意性等の概念を援用して更に厳密な議論を行うが、ロジスティック回帰のモデルパラメータやその結果から算出される統計量を読み解くことで、モデルがどの変数 x_i の影響を特に重視して扱っているのか、また特に「結果に対して効く」変数はどれなのか、といった点についての洞察を得ることができる。特に文書を題材とする場合は、各変数 x_i を、対応する単語の有無に相当する {0, 1} の値を取る変数として定式化することが多い（文書-単語ベクトル）。また、このとき文書内での単語の出現順は考慮せず、出現の有無のみを問題にする（Bag of Words）。

文献[5]で挙げられている解釈可能な手法の一部について、モデルに対しどういった「解釈」が可能であるかを表 1 にまとめる。また、これらに対し、利用される場面の多い Neural Network（Deep Learning も含む）や Support Vector Machine といった手法は、ブラックボックス性の高いものであると見做されている。

表 1　モデル構築手法とそれに対する解釈の例

手法	解釈
線形回帰 ロジスティック回帰	係数や統計量から変数毎の影響度合い・有意性を読み取る
Decision Tree	モデルそのものが等式や不等式で表される階層的なルールであり、それがそのままモデルが判断する際のロジックである
K-NN 法	モデル全体の性質を簡単に語ることはできないが、ある予測結果がどのようにして得られたか、という根拠は「そのデータに対してこのような K 個の近接点が現に存在していたから」というもので明白である

2.2　Surrogate モデル

2.1 章 冒頭の繰り返しになるが、モデルの定式化が複雑になるほど解釈は難しくなる傾向にある。しかし、単純なモデルでは必要とされる精度が確保できないおそれも多分にあり、実用上の要請とモデルの解釈可能性は両立しない面がある。こういった点の解消を試みるべく、Surrogate Models という手法が提案されている[6]。これは、最初から解釈可能性の高い手法でモデルを作ろうとするのではなく、まずは一般の手法で実用モデルを作成しておき、後からそのモデルの挙動を近似するような代理（surrogate）モデルを解釈可能な手法で作成してそこから情報を読み取ろうとするものである。

表 1 で示した例には、実は解釈を試みる観点として次の 2 つが存在している。

・モデルが、全体としてどの変数のどういった値に特に着目しているのか読み解く
・モデルがある予測結果をはじき出したとき、その 1 点の予測結果がどこに着目して得られたものか読み解く

前者はモデル全体に対する解釈であり、例えばロジスティック回帰の変数から読み取れることは、そのモデル自体が全体としてどの変数を重要視しているか、ということである。後者は、ある 1 点のデータ点の予測に対する解釈であり、K-NN 法の例がこれにあたる。

代理モデルを作成するにあたっても、実用モデル全体を代理モデルで表現しようとするか、実用モデルに対する特定の入力の予測結果を表現しようとするかで大域的（Global Surrogate Models）・局所的（Local Surrogate Models）の２つのアプローチが存在する。

ここでは、文献 6）で示されている 局所的アプローチ LIME（Local Interpretable Model-Agnostic Explanations）を取り上げる。まず、実用モデルにおいて、図 1a)のような分類の状況が得られたとする。実際に文書を題材とする場合、x_i は各々の単語に対応する非常に多次元の空間となるが、ここでは（x_1, x_2）の２変数 の入力に対して○, ×どちらかの分類を予測するという問題設定とする。この分類の境界線は複雑であり、x_1 の値がどのような場合に○である、x_2 の値がどのような場合に×である、という大域的な傾向を論じることは一般に難しいであろう。

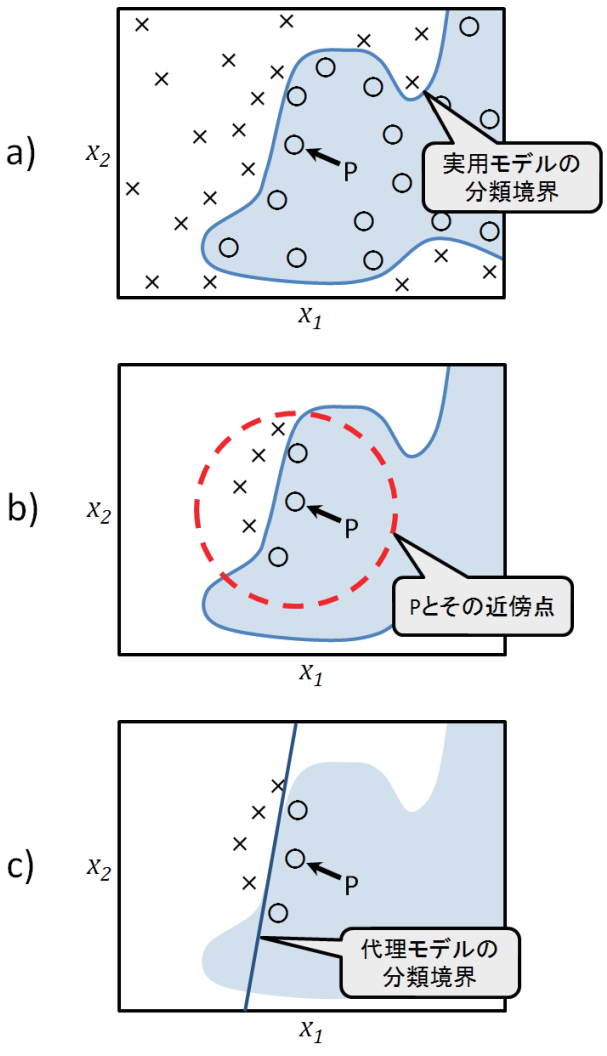

図１　実用モデルと代理モデル

ここで観点を変え、図 1a)内の点 P は「○」に分類されているが、これは x_1 が重視された結果の「○」であるのか、それとも x_2 が重視された結果なのか、ということを理解することを試みる。ここで、点 P の近傍の点のみを用いて、解釈可能な手法で代理モデルを構築するのが LIME の考え方である。この代理モデルは、点 P の付近の情報はある程度正しく反映されたものであるが、点 P から離れた点に対しては全く実用モデルとは異なる外れた結果を示すであろう。しかし、点 P における状況を理解するにはこれで十分なのである。

点 P の近傍のみを用いて例えば線形回帰を行った結果図 1c)のような境界線が得られたとすると、これは x_2 よりも x_1 を２分割するような形になっているので、「点 P 付近では x_1 の挙動がクリティカルである、特に効いている」と結論することができる。

この手法は、実用モデルがどういった手法で構築されたものであるかを問わない（Model-Agnostic である）点も特色であり、きわめて一般的な場面に適用できるものである。

3　特許文書分類モデルへの応用

本章では、特許文書の分類モデルに対し 2.2 章の手法を適用させ、分類に特に影響している変数（単語）を可視化することを試みる。データ諸元や分類モデルなどの問題設定を表２に示す。

表２　問題設定

対象データ	
検索条件 データ条件	「パンケーキ」で全文検索した 国内特許公報の要約部分
件数	1371 件
期間	出願日が 1996 年～2016 年
実用モデル	
説明変数	形態素解析ののち、出現した公報数が多い上位 500 単語(500 次元)を用いて {0, 1} の 文書-単語ベクトル を作成
目的変数	特定のテーマコードを含む/含まない の 2 値分類
手法	Support Vector Machine 全データのうち 80%を学習に使用

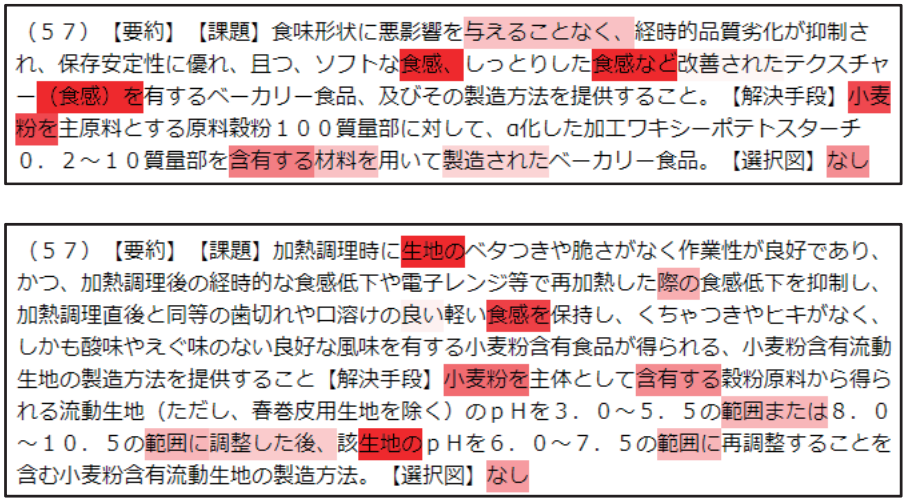

図２　テーマコード“4B032”への分類例

実用モデル構築後、次の手順で代理モデルを構築する。

1. 学習に用いなかったデータ（残り 20%）から、正しく予測できた例を 1 件選択する。
2. その 1 件を「注目データ」とし、注目データのベクトルの 1 の箇所をランダムに 0 に置き換えたデータを「近傍データ」として 10000 件作成する。ランダムの確率は 20% とした（単語毎に独立で、20% の確率で出現しなかったことになる）。
3. 「近傍データ」10000 件を実用モデルに投入し、予測値を得る。
4. 説明変数を「近傍データ」10000 件、目的変数をそれらに対する実用モデルの予測値、として ロジスティック回帰による代理モデル を作成する。
5. 代理モデルの係数を観察し、特に効いている単語の情報を得る。

実用モデル学習の観点は特定のテーマコードを含む／含まないというものであるため、代理モデルの解釈においてそのテーマコードであることを特徴付ける単語が分類の根拠として抽出されることを期待する。

原文に対して、代理モデルにおける変数への影響度合いを図示したものを図２に示す。テーマコード“4B032”、「ベイカリー製品及びその製造方法」に正しく分類された公報文章の中で、「食感」「生地」「小麦粉」といった関連が深いと推測されるキーワードがハイライトされていることが読み取れる。モデルの判断根拠として、確かに特定のテーマコードに関連が深いとみられる単語が抽出されていることがわかる。

4　まとめ

本稿では、一般にブラックボックスであると見做されがちな機械学習のモデルに対して解釈を試みるための手法について解説を行い、特許分類のモデルに対してその手法を適用させた例を示した。

本稿の分析は、当社 NTT データ数理システムで開発･販売を行っている分析ツール **Text Mining Studio**, **Visual Mining Studio**, **Visual R Platform** を組み合わせて実現した。主として日本語の言語処理解析部分をテキストマイニングツール **Text Mining Studio** が、データ整形や実用モデルである Support Vector Machine によるモデル構築をデータマイニングツール **Visual Mining Studio** が、ロジスティック回帰による代理モデル構築と評価の部分を統計解析ツール **Visual R Platform** が担当している。これらの分析ツール群は共通の基盤プラットフォーム上で提供され、それらの間をシームレスに連係させて利用することができる。この様子を図３に示した。これら分析ツールの組合せによって、多様化するデータに対しての様々な要請に応えることができると当社では考えている。

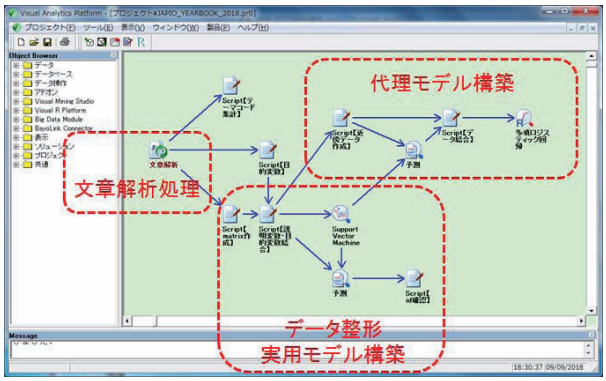

図３　利用イメージ

参考文献

[1] 富永泰規：“外国特許文献へ分類付与に関する機械学習活用可能性調査について”Japio YEAR BOOK 2017 pp. 212-217 (2017)

[2] 近藤浩之：“特許分類の自動推定への取り組み”Japio YEAR BOOK 2016 pp. 188-191 (2016)

[3] 安藤俊幸：“機械学習を用いた効率的な特許調査法”Japio YEAR BOOK 2016 pp. 150-161 (2016)

[4] 原聡：“機械学習における解釈性”人工知能 33 巻 3 号 pp. 366-369 (2018)

[5] Christoph Molnar：“Interpretable Machine Learning – A Guide for Making Black Box Models Explainable.”https://christophm.github.io/interpretable-ml-book/ (2018)

[6] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin：“Why Should I Trust You?: Explaining the Predictions of Any Classifier” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 1135-1144 (2016)