機械学習システムにおけるソフトウェアの品質評価の現状と課題

宇根正志（うねまさし）／清藤武暢（せいとうたけのぶ）

要　旨

機械学習を実装するシステム（機械学習システム）を活用する際には、その品質がビジネス要件を充足していることを予め確認しておく必要がある。ビジネス要件の充足度合いを評価するうえでは、機械学習システムにおいて判定や予測を実行するソフトウェア（判定・予測エンジン）の品質をどのように評価するかが重要であるが、従来のソフトウェアの品質評価の手法は、そのままでは適用できない場合が多い。本稿では、判定・予測エンジンの品質評価に焦点を当てて、その研究動向や課題を説明する。そのうえで、金融分野で活用される機械学習システムの事例を対象に、判定・予測エンジンの品質を評価する際の留意点や課題を考察する。

キーワード：　機械学習、ソフトウェア、品質評価

本稿の作成に当たっては、浦本直彦氏（三菱ケミカルホールディングス）から有益なコメントを頂いた。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者たち個人に属し、日本銀行あるいは有限責任監査法人トーマツの公式見解を示すものではない。また、ありうべき誤りはすべて筆者たち個人に属する。

宇根正志　日本銀行金融研究所企画役（E-mail: masashi.une@boj.or.jp）
清藤武暢　日本銀行金融研究所主査（現有限責任監査法人トーマツマネジャー、E-mail: takenobu.seito@tohmatsu.co.jp）

1. はじめに

近年、金融分野において、業務の効率化（コスト削減）や新たな金融サービスの創出等を実現する技術として、人工知能（artificial intelligence: AI）が注目を集めている。そうしたAIにおいては、各種データ（画像、音声やテキスト等）を認識し、判定や予測等の情報処理を行うための基礎技術として、機械学習（machine learning）が広く採用されている。金融機関が機械学習を実装したシステム（機械学習システム）の活用を検討する場合には、機械学習システムのビジネス要件を設定したうえで、導入を検討している機械学習システムがそれらを充足していることを予め確認しておく必要がある。

一般に、機械学習では、訓練データを一定のアルゴリズム（学習モデル）に適用し、新しいデータの判定や予測を実行するソフトウェア（判定・予測エンジン）を生成する。機械学習システムの主たる機能は判定・予測であり、判定・予測エンジンが重要な構成要素といえる。判定・予測エンジンは、通常、ソフトウェアとして実現されることから、機械学習システムのビジネス要件の充足度合いを評価するうえで、判定・予測エンジンのソフトウェアとしての品質をどう評価するかが重要となる。また、その生成に用いられた学習モデルと訓練データも、判定・予測エンジンの品質を左右する。

既存のソフトウェアについては、品質評価の方法論（ソフトウェア工学）が確立しているほか、関連する標準規格も規定されている（日本工業標準調査会［2013］、込山［2014］、東［2014, 2017］等）。入力と出力の関係を定式化したうえで、処理の流れをフローチャート等によって明確化し、それを実現するプログラムを生成するという手法（演繹的手法）が一般的である。ソフトウェアの品質は、（想定される）データを入力として与えたときに、出力されるデータが事前に定式化した入出力の関係とどの程度整合的かを確認することによって評価する。

一方、判定・予測エンジンは、入力と出力の関係を事前に定式化することなく、訓練データと学習モデルを用いて直接生成される（帰納的手法）。そのため、判定・予測エンジンにおける入出力の関係が明確でないケースでは、既存のソフトウェアで用いられる手法のみでは品質の評価が困難となる。機械学習システムを活用する際には、こうした点に留意する必要がある。

上記の問題意識を踏まえ、本稿では、機械学習システムにかかる品質のなかでも判定・予測エンジンの品質に焦点を当てて、それを評価する手法を概説する。さらに、金融分野で活用される機械学習システムの事例を取り上げ、想定されるビジネス要件を踏まえつつ、判定・予測エンジンの品質を評価する際の留意点や課題を考察する。

2. 機械学習システムとそのライフサイクル

(1) 機械学習システムの構成

本稿では、宇根［2019］に基づき、次の 4 つのエンティティから構成される機械学習システムを想定する。すなわち、①訓練データと学習モデルから判定・予測エンジンを生成する訓練実行者、②訓練実行者から判定・予測エンジンを受け取り、判定・予測を実行する判定・予測実行者、③判定・予測を依頼するシステム利用者、④訓練データを訓練実行者に提供する訓練データ提供者である[1]。判定・予測エンジンの訓練と判定・予測における処理の流れとしてそれぞれ以下を想定する（図表 1 を参照）。

【判定・予測エンジンの訓練】

（A） 訓練データ提供者は、訓練データの元となるデータを収集した後、システム利用者と連携しつつ、適宜加工するとともに、必要に応じてラベル（当該訓練データにかかる判定結果等を表すデータ）を付加して訓練実行者に提供する。

（B） 訓練実行者は、訓練データを学習モデルに入力し、判定・予測エンジンを生成する。

（C） 訓練実行者は判定・予測エンジンを判定・予測実行者に提供する。

【判定・予測の実行】

（D） システム利用者は、判定等を行うデータを判定・予測実行者に提示する。

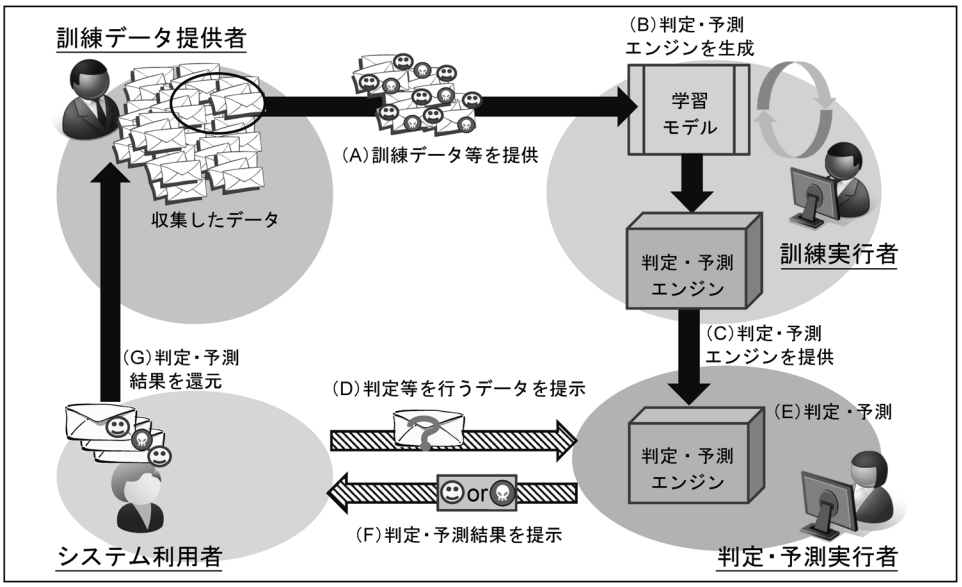
（E） 判定・予測実行者は、判定等を行うデータを判定・予測エンジンに適用し、判定・予測を実施する。

（F） 判定・予測実行者は判定・予測結果をシステム利用者に提示する。

（G） システム利用者は、上記（F）での判定・予測結果等を訓練データ提供者に還元する場合がある。

1 訓練データとして利用できるデータの量が少ない場合、他の訓練データによって生成された類似の判定・予測エンジンを入手し、それに自分の訓練データを適用することによって、自分のアプリケーションにより適したエンジンを生成するという手法（転移学習）が知られている。こうしたケースでは、訓練実行者は、訓練データを訓練データ提供者から受信するとともに、既存のエンジンを他のエンティティ（例えば、エンジン提供者）から受信し、それらを用いて新しいエンジンを生成することになり、機械学習システムを構成するエンティティが増える可能性がある。

図表 1　想定する機械学習システムの構成



(2)　機械学習システムのライフサイクル

情報システム一般のライフサイクルは、アセスメント、開発、運用の各フェーズから構成されるケースが多い。これらを機械学習システムに当てはめると、各フェーズでは以下の対応が考えられる（丸山［2017］、有賀・中山・西林［2018］、本橋［2019］等）。

イ．アセスメント

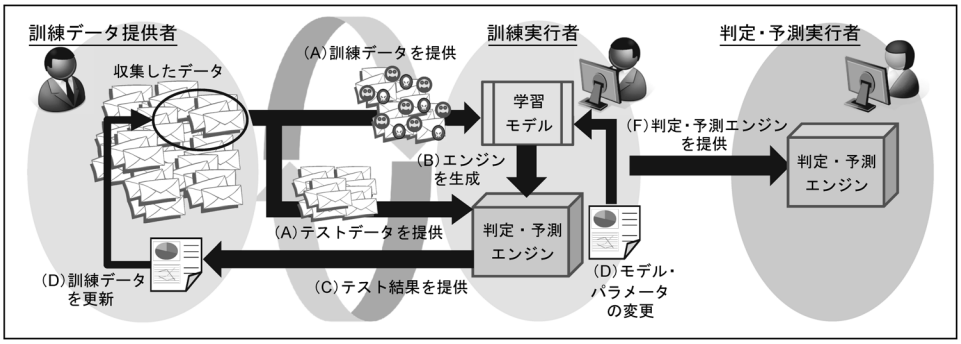
アセスメント・フェーズにおいては、金融サービスを提供する金融機関が、機械学習システムを活用する目的（新たな金融サービスの創出等）を明確にしたうえでビジネス要件を設定する。その後、ビジネス要件を充足する判定・予測エンジンを生成するために必要な学習モデルおよびデータの種類を設定する。

ロ．開発

開発フェーズにおいては、アセスメント・フェーズで設定した学習モデルやデータを用いて判定・予測エンジンを生成する。具体的な処理の流れは以下のとおりである（図表 2 を参照）。

（A）　訓練データ提供者は、アセスメント・フェーズで設定したデータを収集し、必要に応じて事前処理やラベル付加を行ったうえで、訓練データやテスト

図表 2　開発フェーズでの処理の流れ



データとして訓練実行者に提供する。

（B）　訓練実行者は、提示された訓練データを学習モデルに適用して判定・予測エンジンを生成する。

（C）　訓練実行者は、テストデータを用いて判定・予測エンジンの品質を評価（テスト）し、テスト結果を保管するとともに訓練データ提供者へ提供する。

（D）　訓練実行者と訓練データ提供者は、テスト結果を確認しつつ、訓練データの更新（追加や削除）や学習モデルの変更（種類やパラメータの変更等）を行う。

（E）　テスト結果がビジネス要件を充足するまで、上記（B）～（D）を繰り返す[2]。

（F）　訓練実行者は判定・予測エンジンを判定・予測実行者に提供する[3]。

ハ．運用

運用フェーズにおいては、判定・予測実行者が、判定・予測エンジンを用いたサービス等を提供するとともに、各エンティティが連携して機械学習システムのメンテナンス（判定・予測エンジンの更新等）を実施する。具体的な処理の流れは以下のとおりである（図表 3 を参照）。

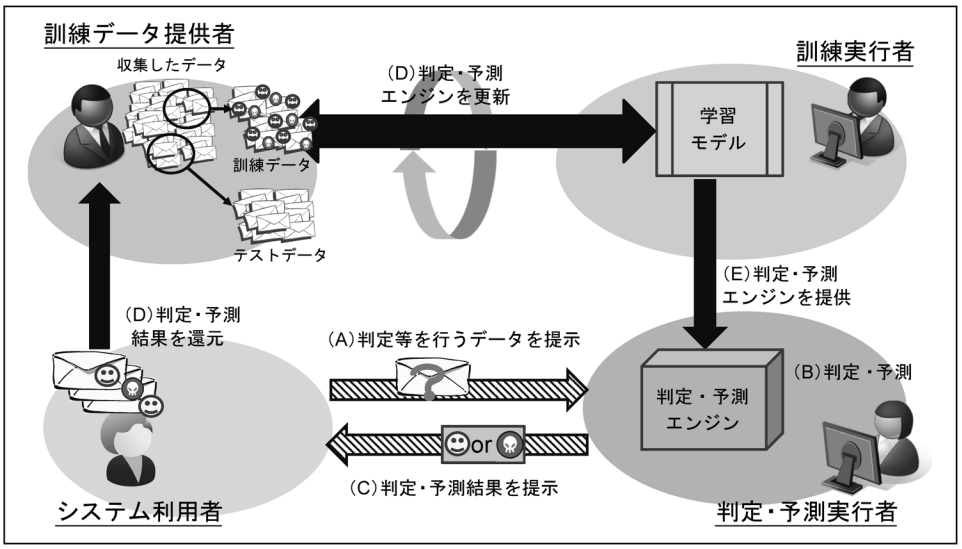
（A）　システム利用者は判定・予測を行うデータを判定・予測実行者に提示する。

（B）　判定・予測実行者は、上記（A）において受信したデータを判定・予測エンジンに適用し、判定・予測を行う。

2　これらの一連の処理は仮説検証（proof-of-concept）と呼ばれる。

3　訓練実行者は、入力データへの事前処理、ユーザー・インタフェースやエンジンの更新方法等の仕組みを判定・予測エンジンに組み合わせたうえで、判定・予測実行者に提供する場合がある。判定・予測エンジンを更新する主な手法として、バッチ手法とオンライン手法が挙げられる（本橋［2019］）。バッチ手法は、新しいデータと過去のデータをすべて訓練データとして用いるものであり、オンライン手法は、新しいデータのみを訓練データとして用いるものである。例えば、データの特性が変化する頻度が高い場合はオンライン手法、低い場合はバッチ手法が適している。

図表３　運用フェーズでの処理の流れ



（C）　判定・予測実行者は、判定・予測結果をシステム利用者に提示する。

（D）　判定・予測エンジンの品質が低下した場合や当該エンジンが一定の稼動期間を経過した場合、判定・予測エンジンを更新する[4]。例えば、システム利用者は、これまでの判定・予測結果を訓練データ提供者に還元し、訓練データ提供者は、それらの判定・予測結果の一部または全部を訓練データとして利用する。訓練実行者は、開発フェーズ（本節（2）ロ.）の（B）～（E）と同様の処理を行って判定・予測エンジンを更新する。

（E）　訓練実行者は、ビジネス要件を充足する（更新後の）判定・予測エンジンを判定・予測実行者に提供する。

(3)　ライフサイクルにおける留意点や課題

アセスメント・フェーズにおいては、機械学習システムの利用を前提とはせず、設定した目的（新たな金融サービスの創出等）の達成に機械学習システムの利用が必要か否かを十分に検討することが重要である。その結果、従来のソフトウェアでも目的が達成できる場合には、従来のソフトウェアの活用も十分に検討することが望ましい[5]。さらに、必要な訓練データを収集できるか否かも考慮し、その可能性

4　判定・予測エンジンの更新は、人手を介さず機械学習システムにおいて自動で実施される場合が多いが、本稿では、議論を簡単にするため各エンティティが連携するモデルを想定する。

5　機械学習システムは、「技術的負債の高利貸しクレジットカード（The high-interest credit card of

が低い場合には、開発フェーズを開始する時期を延期して訓練データの収集を行うことが有用である。

開発フェーズにおいては、訓練データを適切に管理することが重要である。判定・予測エンジンの品質は、生成に用いられた訓練データに依存する。訓練データと判定・予測エンジンの品質の関係が明らかになれば、ビジネス要件を充足する判定・予測エンジンを効率よく生成できる（すなわち、仮説検証の回数を削減できる）可能性が高くなる。そこで、仮説検証において、既存の訓練データを更新（訓練データの追加や削除等）した場合、それを新たな訓練データとしたうえで、テスト結果とともに別途保管しておくことが望ましい。また、訓練データの数とそれに要する処理時間の関係を検証し、開発に要するコストや期間の観点から最適な訓練データの数を決定することも重要となる[6]。判定・予測エンジンの更新については、判定・予測用のデータの特性（特徴が変化する頻度等）に応じて更新の方法や時期を検討することが有用である。

運用フェーズにおいては、システム利用者や判定・予測実行者等が判定・予測エンジンの品質を定期的に検証し、低下を検知した場合に、その原因（判定等を行うデータの特性が当初の想定よりも早く変化したなど）を明確にしたうえで、更新の処理に反映させることが重要となる。もっとも、データの特性の変化を適切に反映させる更新方法は研究途上であり、今後の研究動向をフォローすることが望ましい。また、こうした品質の観点からの判定・予測エンジンの定期的な検証は人手によって対応することになるため、運用にかかるコストが従来のシステムよりも大きくなる可能性にも留意する必要がある。

3. 機械学習システムの品質保証

(1) 従来のソフトウェアにおける品質特性

ソフトウェアやそれを中心とするシステム、特に、情報通信技術を活用したソフトウェアやシステムの品質については、これまでソフトウェア工学をはじめとする各分野において研究や標準化が行われてきた（東［2014］）。ソフトウェアの

technical debt）」と呼ばれることもあり、従来のソフトウェアよりも実装上の不備の問題が蓄積しやすいデメリットを有することが指摘されている（Sculley *et al.* [2014]）。

6 一般に、判定・予測エンジンの生成に用いる訓練データの数に比例して、その品質は向上する一方、その数に比例して処理時間も増加する。もっとも、訓練データに関しては、判定・予測エンジンが訓練データに対して過度に最適化されたもの（過学習）とならないように、属性等に偏りがないように選択するなどの留意が必要である。

品質モデルを規定する国内標準 JIS X 25010「システム及びソフトウェア製品の品質要求及び評価（Systems and software engineering—Systems and software Quality Requirements and Evaluation: SQuaRE）」では、ソフトウェアの品質を、「明示された状況下で使用されたとき、明示的ニーズ及び暗黙のニーズをソフトウェア製品が満足させる度合い」と定義している（日本工業標準調査会［2013］）。これに基づくと、ソフトウェアの品質を検討する際には、まず、「明示的ニーズ及び暗黙のニーズ」（品質特性）を明確にしたうえで、ソフトウェア製品やデータがそれらのニーズを満足させる度合い（品質特性の評価尺度）を決定する必要がある。JIS X 25010では、ソフトウェア製品の品質モデルにおいて、品質特性として、①機能適合性、②性能効率性、③互換性、④使用性、⑤信頼性、⑥セキュリティ、⑦保守性、⑧移植性が規定されている（図表 4 を参照）。上記①から⑧の品質特性を、機械学習システムの判定・予測エンジンに適用することが考えられる。その際、機能適合性、セキュリティ、保守性については、従来のソフトウェアとは異なる考慮や対応が求められる。

(2) 機械学習システムにおける品質

イ. 機能適合性

機能適合性は、「明示的ニーズ及び暗黙のニーズを満足させる機能を、製品又はシステムが提供する度合い」と定義されており、機械学習システムの主たる目的が判定・予測の実行である点に着目すれば、特に重要な品質特性であるといえる。機能適合性は、より細かく、①機能正確性（正確さの必要な程度での正しい結果を提供する度合い）、②機能完全性（機能が作業や目的を網羅する度合い）、③機能適切性（作業と目的の達成を機能が促進する度合い）に分けられる。これらのうち、機能正確性と機能完全性に関して、従来のソフトウェアとは異なる対応が必要になると考えられる。

（イ） 機能正確性

機能正確性の評価では、テストデータをソフトウェアに入力し、期待される出力が得られる度合い（正確性）を確認するケースが多い。機械学習システムにおける判定・予測エンジンにおいても、同様の方法で評価することがまず考えられる。その場合には、従来のソフトウェアにおけるテストの評価指標を用いる。例えば、判定を実行するエンジンの評価指標として、適合率（precision）、再現率（recall）、正解率（accuracy）が用いられるケースが多いほか、予測を行うエンジンの場合には、

図表 4　ソフトウェアの品質特性（JIS X 25010）

項番	品質特性	定義
1	機能適合性（functional suitability）	明示された状況下で使用するとき、明示的ニーズ及び暗黙のニーズを満足させる機能を、製品又はシステムが提供する度合い。 ・機能正確性：正確さの必要な程度での正しい結果を、製品又はシステムが提供する度合い。 ・機能完全性：機能の集合が明示された作業及び利用者の目的の全てを網羅する度合い。 ・機能適切性：明示された作業及び目的の達成を、機能が促進する度合い。
2	性能効率性（performance efficiency）	明記された状態（条件）で使用する資源の量に関係する性能の度合い。
3	互換性（compatibility）	同じハードウェア環境又はソフトウェア環境を共有する間、製品、システム又は構成要素が他の製品、システム又は構成要素の情報を交換することができる度合い、及び／又はその要求された機能を実行することができる度合い。
4	使用性（usability）	明示された利用状況において、有効性、効率性及び満足性をもって明示された目標を達成するために、明示された利用者が製品又はシステムを利用することができる度合い。
5	信頼性（reliability）	明示された時間帯で、明示された条件下に、システム、製品又は構成要素が明示された機能を実行する度合い。
6	セキュリティ（security）	人間又は他の製品若しくはシステムが、認められた権限の種類及び水準に応じたデータアクセスの度合いを持てるように、製品又はシステムが情報及びデータを保護する度合い。
7	保守性（maintainability）	意図した保守者によって、製品又はシステムが修正することができる有効性及び効率性の度合い。
8	移植性（portability）	1 つのハードウェア、ソフトウェア又は他の運用環境若しくは利用環境からその他の環境に、システム、製品又は構成要素を移すことができる有効性及び効率性の度合い。

資料：日本工業標準調査会［2013］

一般に平均 2 乗誤差が用いられる[7]。

もっとも、判定・予測エンジンは帰納的に生成されることから、入力と出力の関係を事前に定式化できない場合には、従来のソフトウェアとは異なる方法で判定・予測の正確性を評価しなければならない。こうした課題を解決するために、判定・予測エンジンの特性を考慮した品質評価の手法がいくつか提案されている。主な手法として、メタモルフィック（metamorphic）手法、サーチ・ベースド（search based）

7 例えば、入力を 2 つのカテゴリー（A と B）に分類するエンジンを想定する。適合率は「A と判定された入力のうち、実際に A に分類されるものの割合」、再現率は「A に分類される入力全体のうち、正しく A と判定したものの割合」、正解率は「入力全体のうち、正しく判定したものの割合」を示す。いずれも、値が大きいほど、期待される出力が得られる度合いが高まる。また、平均 2 乗誤差については、予測によって得られた値と実際の値の差分（の 2 乗）をそれぞれ計算して平均し、その平方根を計算して求める。この値が小さいほど、期待される出力が得られる度合いが高まる。

手法、ランダム・テスティング（random testing）手法、形式検証手法が挙げられる（科学技術振興機構［2018］、石川・徳本［2019］）[8]。ただし、いずれの手法も、現時点では技術的な課題（一部の判定・予想エンジンには適用困難であるなど）が残されており、特定の手法のみで十分なテストを実施できるとはいえない。

（ロ） 機能完全性

機能完全性は、ソフトウェアの機能が作業や目的を網羅する度合いと定義され、それらがソフトウェアの機能によってカバーされているか否かを評価することになる。したがって、具体的な評価項目は、ソフトウェアの種類やその応用分野に依存して変化しうる。

例えば、判定・予測エンジンの出力の理由・根拠を知る必要があるアプリケーションでは、人間が解釈可能な出力を実現する機能の有無が評価項目の 1 つとなる。判定・予測エンジンのなかには、判定・予測の決定過程を条件分岐の階層構造によって表現可能な決定木を用いた学習モデル等、判定・予測結果を人間が解釈しやすいものもある。もっとも、そうした学習モデルは、深層学習等に比べて判定・予測の精度が相対的に低い場合が少なくない[9]。他方、深層学習は、判定・予測の精度が相対的に高いものの、判定・予測を出力する際に多数の重み（パラメータ）が用いられ、それらの意味や効果を人間が理解困難なケースがある。このように、アプリケーションに求められる精度（機能正確性）を達成しつつ、判定・予測の理由・根拠を人間が解釈できるようにする手法の開発が重要な研究課題と位置付けられており、研究が活発化している（Adadi and Berrada [2018]、科学技術振興機構［2018］）[10]。

また、判定・予測エンジンの出力に偏りがない（公平性を有する）ことも、アプリケーションによっては求められる場合があり、そのための機能の有無が評価項目の 1 つとなりうる。偏りを排除する方法として、例えば、訓練フェーズにおいて出力に偏りを発生させる特徴量を訓練データに使用しないことが考えられる。もっ

8 これらの手法については補論を参照されたい。

9 深層学習は、広く用いられている学習モデルであり、人間の脳神経系回路を模した数学モデル（ディープ・ニューラル・ネットワーク）を用いて、訓練データから判定・予測エンジンを生成するものである。

10 こうした研究開発の代表的なプロジェクトとして、米国国防高等研究計画局（Defense Advanced Research Projects Agency）の説明可能な AI（eXplainable Artificial Intelligence）が知られている。機械学習の処理性能（精度等）を維持しつつ、人間が出力の意味や根拠を理解できるようにするために、2017 年に開始された（Gunning [2017]）。情報科学に加えて心理学の観点も取り入れつつ、既存の学習モデルを改善する手法や、出力を人間が理解しやすいように表示するインタフェースの研究が進められている。プロジェクトは 2021 年までとなっており、改良モデルやインタフェースを含むツールキット・ライブラリを主な成果物とすることが予定されている。このほか、学習モデルを解釈可能にするさまざまな手法が学会に提案されるようになってきている。例えば、任意の学習モデルの出力を、人間が理解可能なモデルによって局所的に近似する手法（local interpretable model-agnostic explanations）が挙げられる（Ribeiro, Singh, and Guestrin [2016]）。

とも、望ましくない特徴量と相関が高いデータが訓練データに含まれていた場合、最終的に出力に偏りが生じる可能性が残る。最近では、判定・予測エンジンの出力の偏りを検知・解消する技術分野として「フェアネス・アウェア・データマイニング（fairness-aware data mining）」が知られている（Kamishima [2019]）。出力の偏りを誘発しうる訓練データのタイプを定義したうえで、それらを効率的に探索する手法（unfairness discovery）、偏りが発生しうる出力を別のエンジンによって偏りのないものに変換する手法（post-process unfairness prevention）等、さまざまな視点からの研究が活発に進められている。

ロ. セキュリティと保守性

セキュリティは、「製品又はシステムが情報及びデータを保護する度合い」と定義されている。機械学習システムにおいて特に留意すべきは、判定・予測エンジンから訓練データが推定されるなど、機械学習に特有の脆弱性が知られていることである（宇根［2019］）。それらに対応するために、従来のソフトウェアとは異なる対策が必要となる場面が想定される（井上・宇根［2019, 2020］）。

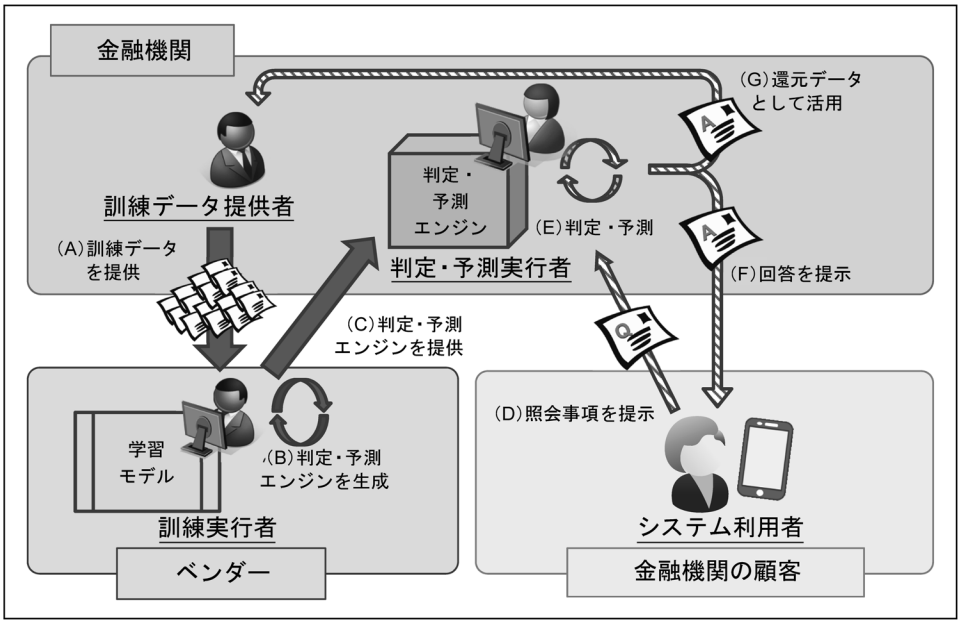
また、保守性は、「製品又はシステムが修正することができる有効性及び効率性の度合い」と定義される。機械学習システムでは、判定・予測結果等を考慮して、判定・予測エンジンを自動的に更新する場合がある。そうした場合には、既存のソフトウェアに比べて効率的に判定・予測エンジンを更新できる可能性がある。もっとも、判定・予測エンジンの更新に人間が必ず介在しなければならないアプリケーションでは、保守性は、従来のソフトウェアと同程度、あるいは、逆に低下する可能性もある。

4. 金融分野で活用される機械学習システムの品質保証上の留意点

本節では、金融分野で活用される主な機械学習システムの事例として、井上・宇根［2019, 2020］で検討されたもののうち、①チャットボットを用いた顧客応答のシステム、②個人向けローンの信用度評価のシステム、③金融市場等の異常検知のシステムを取り上げ、ビジネス要件や品質を検討する際の留意点や課題を考察する[11]。

11 セキュリティについての考察は井上・宇根［2019, 2020］を参照されたい。

図表５　チャットボットを用いた顧客応答のシステム（概念図）



備考：井上・宇根［2020］図表５に基づき作成。

(1)　チャットボットを用いた顧客応答のシステム

このシステムでは、金融機関がチャットボットを活用して顧客からの照会に自動で応答したり顧客の状況に合わせて金融商品を提案したりする。そうしたモデルの１つとして、ベンダーが提供する汎用的な学習モデルを活用して判定・予測エンジンを生成するケースを取り上げる。具体的な処理の流れとして以下が想定される（図表５を参照）。

（A）　金融機関は、照会対応にかかる過去の事例や標準的な応答のシナリオ、顧客の属性とそれに合致する金融商品のデータ等を、訓練データとしてベンダーに提供する[12]。
（B）　ベンダーは、チャットボット用の学習モデルに訓練データを適用し、判定・予測エンジンを生成する。
（C）　ベンダーは判定・予測エンジンを金融機関に提供する。

12　訓練データとして利用されるものに顧客の個人情報が含まれる場合には、照会にかかる応答の利用につき顧客から事前の了承を得ておく、あるいは、それらに対して匿名加工を実施して個人を識別・特定困難にしておくなどの対応が求められる。

（D） 金融機関の顧客は、スマートフォン・アプリや SNS を用いて照会事項を金融機関に提示する。
（E） 金融機関は、照会事項を判定・予測エンジンに適用し、判定・予測を行う。
（F） 金融機関は判定・予測結果を回答として顧客に提示する。
（G） 金融機関は必要に応じて判定・予測結果を還元データとして活用する。

このシステムの主たる機能は、音声やテキストによる顧客からの照会事項に対して適切な情報をタイムリーに回答することである。したがって、主なビジネス要件として、①顧客の属性等を勘案しつつ照会事項に対して正確な回答（判定・予測結果）を提示することと、②照会事項の受信から回答の提示までの時間を一定の値以下とすることが考えられる。また、金融商品を提案するなど、顧客の属性に応じて回答が変化しうる場面においては、上記①の要件から派生して、③回答の根拠にかかる照会に対応できることもビジネス要件となりうる。さらに、出力を還元データとして利用し、判定・予測エンジンを適宜更新することから、④回答の正確さを維持することもビジネス要件となる。上記①、③に対応する品質特性はそれぞれ機能正確性と機能完全性である。上記②に対応する品質特性は使用性であり、上記④に対応する品質特性は保守性である。

上記①（機能正確性）の要件に関しては、まず、どのような回答を正解と判断するかを明確にするとともに、正確性を評価する指標を定義する必要がある。そのうえで、指標の値が一定のレベル（金融機関が設定）以上であるか否かをテストで確認する。例えば、照会事項等のサンプルとそれに対する正解（「正確な」回答）をテストデータとして準備し、サンプルを判定・予測エンジンに入力して得られる判定・予測結果が正解と一致する確率（正解率）を指標として計測することが考えられる。照会事項等のサンプルは、対応する回答を予測可能な（メタモルフィック性を有する）場合が多いと考えられるため、メタモルフィック手法を用いてテストデータを準備することが想定される。計測された正解率が予め設定された値以上である場合、上記①が充足されていると判断することができる。

また、顧客の照会事項等、判定・予測エンジンへの入力が微小に変化した際に、不正確な出力となる可能性にも留意する必要がある。例えば、照会事項が自然言語のテキストによって提示される場合、同一の意味であっても表現の変化が回答の内容に影響しうる。そうした影響をテスト（ランダム・テスティング手法）で評価することが望ましい。もっとも、判定・予測エンジンへの入力（顧客の照会事項）のバリエーションが非常に大きく、十分なテストを実施困難なケースでは、不適切な回答が発生した場合の影響と運用面での対応を検討することが求められる。例えば、顧客に不適切な金融商品を提案・推奨する回答が生じ、そうした事象を無視できないと判断するならば、顧客に回答を提示する前に金融機関の職員が判定・予測結果を確認するなどの対応が考えられる。

上記②（使用性）の要件に関しては、例えば、金融機関が照会事項を受信して顧客に回答を提示するまでの時間を測定し、それが一定の値に収まるならば、要件が充足されていると判断することができる。上記の時間にかかる上限値は、個々のシステムのサービス内容や通信環境等を勘案しつつ、金融機関が決定することになる。また、金融機関の職員が、顧客に回答を提示する前に、判定・予測エンジンの出力を確認するケースでは、判定・予測エンジンの処理時間に加えて、金融機関の職員による確認にかかる時間も含めて測定する必要がある。

上記③（機能完全性）の要件に関しては、例えば、回答の理由・根拠について人間が解釈しやすい学習モデルを採用するとともに、顧客とのやり取りや属性等が回答とどのように関係しているかについて、顧客が理解できる情報として提示する機能を照会応答のシステムに持たせることが考えられる。これが容易ではないケースでは、運用による対応が必要となる。例えば、金融商品の提案・推奨等に対する照会では、その根拠にかかる問合せも想定されるため、職員への直接の照会・相談を勧めることが考えられる。

上記④（保守性）の要件に関しては、金融機関の職員が還元データを確認し、不正確な判定・予測結果が増加していないか、あるいは、機能正確性の指標の値が悪化していないかなどを随時モニタリングすることが考えられる（運用による対応）。その際、上記①（機能正確性）のテストと同様に、更新された判定・予測エンジン等を対象にテストを実施しつつ指標の値を測定・確認することとなる。また、指標の値が一定レベルを下回らないように、判定・予測エンジンを適宜修正することが求められる[13]。

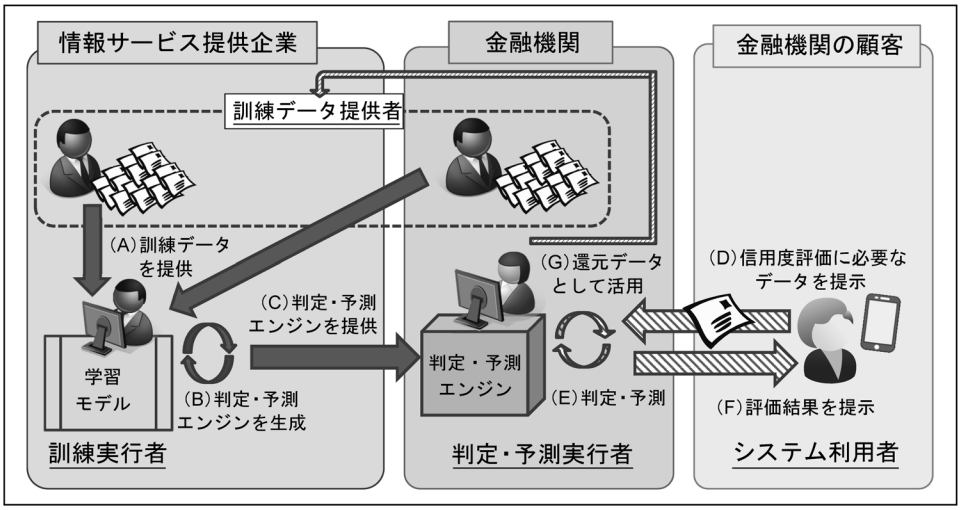
なお、各要件を充足する方法を検討する際には、それらの対応にかかるコストや作業負担等を勘案することも必要となる。

(2) 顧客の信用度を評価するシステム

融資審査における顧客の信用度評価は、金融機関における判断・予測の支援を目的とした機械学習システムの代表的な適用事例の 1 つである。こうしたシステムが実現できれば、1 件あたりの信用度評価にかかる手間やコストを低減させることが可能となり、その結果として、一定の時間においてより多くの案件を審査できるようになると期待される。最近では、金融機関が顧客に関して既に有するデータに加えて、顧客の行動やソーシャル・メディア上のデータ等も活用する動きがみられており、金融機関と取引関係が薄い顧客に対しても信用度評価が可能になるとの見方

13 ただし、判定・予測エンジンの出力の根拠を把握困難な場合、どのように判定・予測エンジンを修正すればよいかが明確とならない可能性がある点に留意が必要である。

図表 6　顧客の信用度を評価するシステム（概念図）



備考：井上・宇根［2020］図表 6 に基づき作成。

もある。

信用度評価システムの事例の 1 つとして、顧客のデータを保有するとともに、機械学習システムを構築するためのノウハウを有する企業（情報サービス提供企業）と金融機関が連携するモデルが想定される。例えば、双方が有する顧客のデータを訓練データとして用いる場合には、次のように処理が実行されることが想定される（図表 6 を参照）。

（A）　金融機関は、顧客の属性や金融取引にかかるデータ等を訓練データとして情報サービス提供企業に提供する。
（B）　情報サービス提供企業は、上記（A）の訓練データに加えて、自社が有する顧客にかかるデータを用いて、判定・予測エンジンを生成する。
（C）　情報サービス提供企業は、判定・予測エンジンを金融機関に提供する。
（D）　金融機関の顧客は、スマートフォン・アプリや SNS を用いて、信用度評価に必要なデータを金融機関に提示する。
（E）　金融機関は、上記（D）のデータを判定・予測エンジンに適用し、判定・予測結果を得る。
（F）　金融機関は、判定・予測結果に基づき、信用度の評価結果を顧客に提示する。
（G）　金融機関と情報サービス提供企業は、必要に応じて、判定・予測結果等を還元データとして活用する。

このシステムの主たる機能は、顧客の信用度を適切かつタイムリーに評価することである。ビジネス要件としては、①顧客から受信したデータから信用度を正確に評価すること、②顧客からのデータの受信から評価結果の提示までの時間を一定の値以下とすることである。また、本節（1）の顧客応答のシステムと同様に、上記①から派生して、③評価結果の根拠を顧客に説明できることも要件として設定されうるほか、還元データ等を用いて判定・予測エンジンを適宜更新することから、④評価結果の適切性を維持することが要件となりうる。さらに、⑤信用度の評価結果が公平性を有していることもビジネス要件となりうる。例えば、金融サービスの提供における公平性を維持する観点から、顧客の属性（人種、性別等）によって評価結果が偏らないようにすることが求められる。上記①に対応するのが機能正確性であって、上記③、⑤は機能完全性である。また、上記②に対応するのは使用性であり、上記④に対応するのは保守性である。

上記①（機能正確性）の要件に関しては、指標を定義したうえで、開発したシステムにおいて測定した指標の値が一定のレベル以上であるか否かをテスト（メタモルフィック手法やサーチ・ベースド手法）で確認することとなる。どのような指標を用いるかは判定・予測結果の表現方法等に依存する。例えば、判定・予測結果を整数等の数値で表す場合、3節で示したように、その値と「正解」との差分（平均2乗誤差）を指標とすることが考えられる。判定・予測結果として数値をそのまま用いる代わりに、ランクA、B、C等のカテゴリーに変換して示す場合には、得られたカテゴリーが「正解」と一致する確率（正解率）を指標とすることが考えられる[14]。また、より厳密な機能正確性を評価することが求められる場合には、適合率や再現率を指標として加えることも考えられる。

こうした確率が、顧客によって提示されたデータの微小な変化（摂動）によって著しく変化する（誤判定を誘発する）可能性があることから、ランダム・テスティング手法を利用して摂動の影響を評価することが考えられる。もっとも、顧客によって提示されるデータが複数の項目から構成され、各項目のデータが（顧客が誤って入力する以外は）変動しづらいケースや、チェックシート形式でデータを入力するケースでは、こうした可能性を考慮する必要がない場合もありうる。

上記②（使用性）の要件に関しては、例えば、顧客からの信用度評価用のデータを受信してから評価結果を提示するまでの時間を測定し、それが一定の値以下となるか否かを確認することとなる。

上記③（機能完全性）の要件に関しては、照会応答システムと同様に、評価結果

14 セキュリティの観点からは、判定・予測エンジンの入出力から当該エンジンや訓練データにかかる情報が漏洩するという脆弱性が知られている（井上・宇根［2020］）。そうした脆弱性をなるべく悪用されないようにする方法の1つとして、数値としての出力をそのまま利用するのでなく、それを何らかのかたちで変換して用いることが考えられる。カテゴリーに変換することもそうした方法の1つと考えることができる。

の理由・根拠を提示する機能を追加することが考えられる[15]。ただし、それが容易でない場合には、評価結果の根拠の提示にかかる技術的な限界を顧客に予め説明し、提示が困難な場合があることを了承してもらうなどの対応が考えられる。

上記④（保守性）の要件に関しては、還元データを用いて、不正確な判定・予測結果の事例が増加していないか、あるいは、その結果として機能正確性の指標の値が変化していないかなどを随時チェックすることが考えられる。また、指標の値が（金融機関が設定した）一定のレベルを下回らないように、判定・予測エンジンを適宜修正するという運用も求められる。

上記⑤（機能完全性）の要件に関しては、属性によって評価結果が偏らないようにするために、そうしたデータを訓練データとして使用しないという対応が挙げられる。もっとも、望ましくない属性と相関が高い属性が訓練データに含まれていた場合、最終的に評価結果に偏りが生じる可能性があることから、出力の偏りを検知・解消する手法が適用可能か否かを検討することが望ましい。こうした手法を適用できない場合には、顧客に提示する前に、評価結果に偏りがないか否かを別途確認するなどの運用による対応が考えられる。

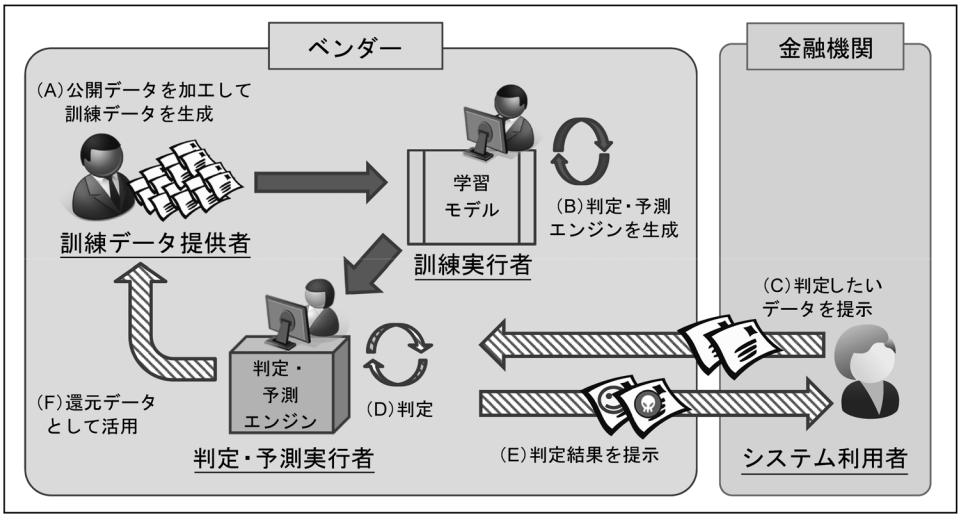
なお、上記①～⑤の要件に関連して、ここで取り上げるシステムでは、複数のエンティティ（金融機関と情報サービス提供企業）がそれぞれ訓練データを提供する点に留意する必要がある。すなわち、複数の先から訓練データを収集して活用する際には、それぞれのエンティティにおいて、データの種類や形式、外れ値や欠損値の処理方法、誤りを含むデータの修正方法等、訓練データの取扱いがエンティティ間で整合的であることが前提となる。したがって、金融機関と情報サービス提供企業との間で訓練データをどのように取り扱うかを、事前に調整・決定しておくことが求められる。

(3) 異常や不正取引を検知するシステム

異常検知も、機械学習の主要なアプリケーションの 1 つであり、従来から盛んに研究開発が進められてきた。金融分野では、金融市場の異常な振舞いを検知したりクレジットカード取引等における不正取引を検知したりするツールとして注目されている。金融市場の異常検知では、過去の注文、市場流動性、価格変動等を用いて金融市場の正常な状態を学習し、それに基づいて異常を検知する。また、クレジットカード取引等の不正取引検知では、過去の不正取引のパターンを学習し、類似のパターンを有する取引を不正取引として検知する。ここでは、金融市場の公開デー

15 最近では、エクィファックス（Equifax）社が、出力結果を解釈しやすくする機能を有する（機械学習を用いた）クレジット・スコアリングのシステム「NeuroDecision® Technology」の導入を発表している（Equifax [2018]）。

図表７　異常や不正取引を検知するシステム（概念図）



備考：井上・宇根［2020］図表９に基づき作成。

タや取引データを訓練データとする機械学習システムをベンダーが構築し、金融機関がそれを利用するという形態を取り上げる。そうした場合、以下のような処理の流れが想定される（図表７を参照）。

（A）　ベンダーは公開データを加工して訓練データを生成する。
（B）　ベンダーは、訓練データを用いて、判定・予測エンジンを生成する。
（C）　金融機関は金融市場のデータや取引データを「判定したいデータ」としてベンダーに提示する。
（D）　ベンダーはデータを判定・予測エンジンに適用し判定結果を得る。
（E）　ベンダーは判定結果を金融機関に提示する。
（F）　ベンダーは、必要に応じて、判定結果を還元データとして活用する。

このシステムの主な機能は、金融機関によって提示されたデータが異常（あるいは不正）か否かを正確かつタイムリーに判定することである。したがって、ビジネス要件として、①提示されたデータを正確に判定することと、②データの判定を一定時間内で実行することが挙げられる。還元データを用いて判定・予測エンジンを適宜更新することから、③判定の正確さを維持することが要件となる。また、判定結果に基づいて金融取引を実施する場合等では、④判定結果の根拠を金融機関に対して説明できることも要件となりうる。

上記①（機能正確性）の要件に関しては、判定結果の正確さを表す指標を定義したうえで、テスト（メタモルフィック手法やサーチ・ベースド手法）において得ら

れた判定結果から指標の値を計算し、ベンダーが予め設定した値と比較して要件の充足度合いを評価することとなる。通常、異常なデータはなるべく高い確率で「異常」と判定するとともに、正常なデータはなるべく高い確率で「正常」と判定することが望ましい。例えば、評価指標として再現率を採用する場合、再現率を高めるようにパラメータ（判定のしきい値等）を変更すると、一般に、正常なデータを異常と判定する事象が増加する傾向がある（逆も成り立つ）[16]。金融市場の異常検知や不正取引検知においては、「正常なデータの誤判定をある程度許容しても、異常あるいは不正なデータをなるべく見逃さないようにしたい」というニーズが想定される。そのようなケースであれば、なるべく高い再現率を実現するようにパラメータを設定することとなる。

上記②（使用性）の要件に関しては、例えば、ベンダーが金融機関からデータを受信した時点から判定結果を送信する時点までの時間を測定し、それが一定の値以下となるか否かを確認することとなる。判定までの時間の設定に関しては、ベンダーが金融機関のニーズを考慮して決めることが想定される。

上記③（保守性）の要件に関しては、誤った判定結果にかかる還元データを用いて、誤判定の事象が増加していないか、または、その結果として再現率が低下していないかなどを随時チェックすることが考えられる。また、再現率が目標値を下回らないように、判定・予測エンジンを適宜修正するという運用も求められる。

上記④（機能完全性）の要件に関しては、判定の対象となるデータの種類によっては、判定の根拠となりうる情報を示すことができる。例えば、金融市場における異常検知において、判定の対象となるデータが時系列データの場合には、実際にそのデータの時系列グラフを表示させ、「異常」と判断されたデータとその他の（「正常」とみなされた）データの関係（当該データが他のデータの系列から大きく外れた値になっているなど）を視覚的に示すという方法が考えられる。もっとも、判定の根拠を説明することが困難なケースでは、金融機関に対して、根拠の説明にかかる技術的な限界を予め説明し了承を得たうえで、ベンダーの職員が当該データを確認するなどの対応が考えられる。

5. おわりに

機械学習システムのソフトウェア（判定・予測エンジン）の品質評価について

16 ここでの再現率は、（判定・予測エンジンへの）入力の集合に含まれる異常なデータ（の全集合）のうち、実際に「異常」と判定されたデータの割合を指す。入力の集合に含まれる正常なデータ（の全集合）のうち、実際に「正常」と判定されたデータの割合を示す指標としては正答率（detection rate）が知られている。

は、従来のソフトウェアにおける評価手法だけでは対応できない場合が多い。そのため、機械学習システムにおけるソフトウェアの品質特性の定義やその評価手法について、近年、研究が活発化している。また、機械学習システムを実際に開発・運用する際のガイドラインを策定する動きもみられている（AI プロダクト品質保証コンソーシアム［2018］等）。現時点では、品質評価の手法が確立しているとは言い難いものの、今後、研究やガイドラインの策定等の取組みが進展し、品質を評価するための方法論が体系化されることが期待される。

今後、金融機関が機械学習システムの活用を検討する際には、品質評価にかかる最新の研究動向等をフォローしていくことが望ましい。品質評価の技術的手法の効果や限界を理解しつつ、技術的な対応が十分でないと考えられるケースにおいては運用面での対応を適切に検討することが求められる。

参考文献

東　基衞、「ICT 応用システムおよびソフトウェア（S&S）の品質向上のための課題と取り組み」、『情報処理』Vol. 55 No. 1、情報処理学会、2014 年、4〜9 頁

————、「ICT 応用 S&S 製品の品質不良のリスクと SQuaRE シリーズ国際標準（及び翻訳 JIS）：その歴史と概要」、ソフトウェアテストシンポジウム 2017 講演資料、2017 年

有賀康顕・中山心太・西林　孝、『仕事ではじめる機械学習』、オライリー・ジャパン、2018 年

石川冬樹・徳本　晋、「機械学習応用システムのテストと検証」、『情報処理』Vol. 60 No. 1、情報処理学会、2019 年、25〜33 頁

井上紫織・宇根正志、「金融分野における機械学習システムの活用とセキュリティ対策」、日銀レビュー 2019-J-2、日本銀行、2019 年

————・————、「金融分野で活用される機械学習システムのセキュリティ分析」、『金融研究』第 39 巻第 1 号、日本銀行金融研究所、2020 年、17〜48 頁（本号所収）

宇根正志、「機械学習システムのセキュリティに関する研究動向と課題」、『金融研究』第 38 巻第 1 号、日本銀行金融研究所、2019 年、97〜123 頁

科学技術振興機構、「AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」、戦略プロポーザル CRDS-FY2018-SP-03、科学技術振興機構、2018 年

込山俊博、「システムおよびソフトウェアの品質基準の体系化」、『情報処理』Vol. 55 No. 1、情報処理学会、2014 年、10〜16 頁

中島　震、「データセット多様性のソフトウェア・テスティング」、『コンピュータソフトウェア』第 35 巻第 2 号、日本ソフトウェア科学会、2018 年、2_26〜2_32 頁

日本工業標準調査会、『JIS X 25010：システム及びソフトウェア製品の品質要求及び評価（SQuaRE）―システム及び品質モデル』、日本規格協会、2013 年

丸山　宏、「機械学習工学に向けて」、日本ソフトウェア科学会第 34 回大会講演論文集、2017 年（https://jssst.or.jp/files/user/taikai/2017/GENERAL/general6-1.pdf、2019 年 9 月 10 日）

本橋洋介、「機械学習応用システムのプロジェクト管理と組織」、『情報処理』Vol. 60 No. 1、情報処理学会、2019 年、48〜55 頁

AI プロダクト品質保証コンソーシアム、「『AI プロダクト品質保証コンソーシアム』の設立について」、AI プロダクト品質保証コンソーシアム、2018 年（http://www.qa4ai.jp/QA4AI-pr-20180326.pdf、2019 年 9 月 10 日）

Adadi, Amina, and Mohammed Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, 6, IEEE, 2018, pp. 52138–52160.

Equifax, "Equifax Launches NeuroDecision® Technology," Equifax, 2018 (available at: https://investor.equifax.com/news-and-events/news/2018/03-26-2018-143044126、2019 年 9 月 10 日).

Gunning, David, "Explainable Artificial Intelligence (XAI)," 2017 (available at: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf、2019 年 10 月 29 日).

Huang, Xiaowei, Marta Kwaitkowska, Sen Wang, and Min Wu, "Safety Verification of Deep Neural Networks," *Proceedings of International Conference on Computer Aided Verification (CAV) 2017*, *Lecture Notes in Computer Science*, 10426, Springer-Verlag, 2017, pp. 3–29.

Jarman, Darryl C., Zhi Quan Zhou, and Tsong Yueh Chen, "Metamorphic Testing for Adobe Data Analytics Software," *Proceedings of International Workshop on Metamorphic Testing (MET) 2017*, IEEE, 2017, pp. 21–27.

Kamishima, Toshihiro, "Fairness-Aware Data Mining," 2019 (available at: http://www.kamishima.net/archive/fadm.pdf、2019 年 9 月 10 日).

Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana, "DeepXplore: Automated White-box Testing of Deep Learning Systems," *Proceedings of Symposium on Operating Systems Principles (SOSP) 2017*, Association for Computing Machinery, 2017, pp. 1–18.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD) 2016*, Association for Computing Machinery, 2016, pp. 1135–1144.

Sculley, David, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young, "Machine Learning: The High-Interest Credit Card of Technical Debt," talk at Software Engineering for Machine Learning, Tufts University, 2014 (available at: https://www.eecs.tufts.edu/~dsculley/papers/technical-debt.pdf、2019 年 9 月 10 日).

Segura, Sergio, Gordon Fraser, Ana B. Sanchez, and Antonio Ruiz-Cortés, "A Survey on Metamorphic Testing," *IEEE Transactions on Software Engineering*, 42(9), IEEE, 2016, pp. 805–824.

Xie, Xiaoyuan, Joshua W. K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen, "Testing and Validating Machine Learning Classifiers by Metamorphic Testing," *Journal of Systems and Software*, 84(4), Elsevier Science, 2011, pp. 544–558.

補論. 判定・予測エンジンにおけるテスト手法

判定・予測エンジンの機能正確性にかかる主なテスト手法として、メタモルフィック手法、サーチ・ベースド手法、ランダム・テスティング手法、形式検証手法について説明する。

(1) メタモルフィック手法

メタモルフィック手法では、判定・予測エンジンを用いることなく出力を予測できるという特性（メタモルフィック性）を利用してテストデータを生成する（Xie *et al*. [2011]、Jarman, Zhou, and Chen [2017]、中島［2018］等）。

例として、あるショッピングサイトにおける購買履歴が入力されたときに、商品の推薦ランキングを出力する判定・予測エンジンを想定し、メタモルフィック手法を用いてテストデータを生成する流れを示す（図表 A-1 を参照）。

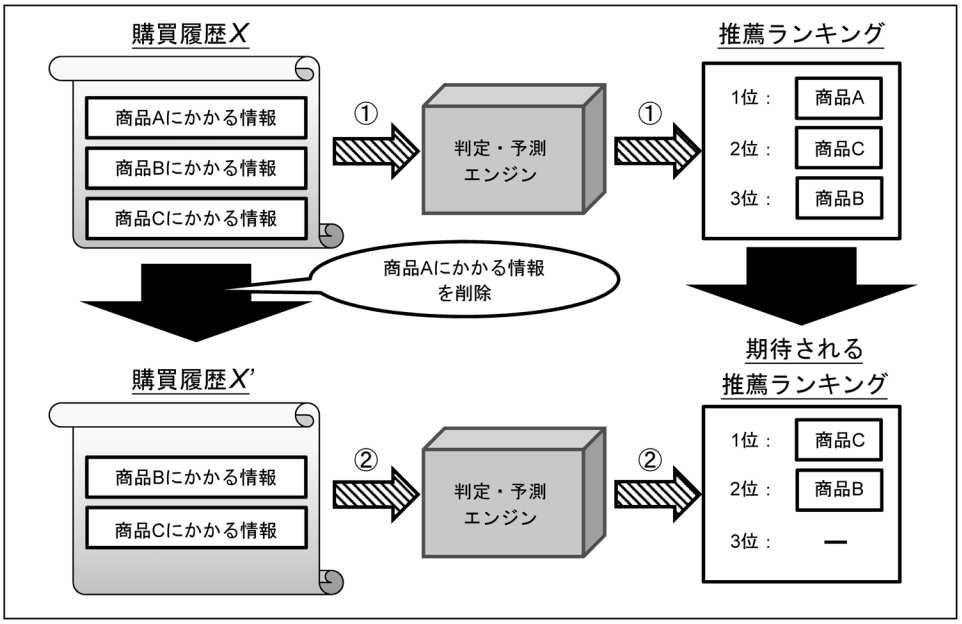
① 実際の利用者の購買履歴（X）を判定・予測エンジンに入力し、推薦ランキングを生成する。
② 次に、推薦ランキングで 1 位となっている商品にかかる情報を除外した購買履歴（X'）を新たに生成する。X' を判定・予測エンジンに入力したとき、上記①の推薦ランキングでは 2 位と 3 位であった商品がそれぞれ 1 位と 2 位となる推薦ランキングが出力されることが期待される（メタモルフィック性）。
③ 上記②と同様の考え方に基づき、さまざまな順位の商品データを除外した購買履歴を複数生成し、これらをテストデータとする。

機能正確性の評価では、メタモルフィック手法を用いて生成したテストデータを判定・予測エンジンに入力し、出力された判定・予測結果が期待されるものに合致する度合いを指標に基づき評価する。

このように、1 つのテストデータから多数の異なるテストデータを生成できることや、さまざまな学習モデルによって生成された判定・予測エンジンに対して適用できることがメリットとして挙げられる。一方で、メタモルフィック性を有するテストデータが存在しない判定・予測エンジンに対しては原理的に適用できないことにも留意が必要である。

また、現時点では、適切な評価のためにどのようなメタモルフィック性を有するテストデータを生成すればよいかが明確になっておらず、テストを実行するエンティティ（訓練データ提供者や訓練実行者）の経験則に基づいて選択されているの

図表 A-1　メタモルフィック手法のアイデア（概念図）



が実情である。そのため、単一のメタモルフィック性を有するテストデータのみを用いた場合には、厳密に機能正確性を評価できない可能性がある（Segura *et al.* [2016] 等）。対策としては、複数のメタモルフィック性を有するテストデータを組み合わせて評価を行うことが考えられる。

(2)　サーチ・ベースド手法

サーチ・ベースド手法は、判定・予測エンジンの機能正確性を満たすうえで大きな影響を与えうる入力（テストデータ）を探索する手法である（Pei *et al.* [2017]）。例として、自動運転技術において自動車の周辺画像が入力されたときに、運転制御データを出力する判定・予測エンジンを想定し、サーチ・ベースド手法を利用してテストデータを生成する流れを示す。

① 実際の周辺画像に対して、さまざまな画像操作（ノイズ追加、画像の一部欠損等）を行い、それらを入力としたときの運転制御データ（自動車に加速、停止等を指示するもの）を生成する。
② これらの運転制御データによって、危険な誤動作（急停車や車線変更等）が引き起こされる確率を求める。

③ 上記②の処理を複数回実施し、危険な誤動作を引き起こす確率が最も高い画像操作を施した周辺画像を特定する。
④ 特定された周辺画像を訓練データとして用いて判定・予測エンジンを更新することにより、同様の周辺画像が入力されても誤動作を引き起こさないようになることが期待される。

サーチ・ベースド手法のメリットとしては、入力の探索にかかる一連の処理を自動化できることや、既存の探索手法（遺伝的アルゴリズム等）の活用により、全数探索よりも効率よく探索できることなどが挙げられる（Pei *et al.* [2017]）[17]。一方、現時点では、一部の学習モデルにより生成された判定・予測エンジンに対してのみ適用可能であることや、探索した入力よりも影響の度合いが高いものが存在しうることなどが課題として挙げられる。

(3) ランダム・テスティング手法

ランダム・テスティング手法は、主に入力に摂動（画像の場合には、ゆがみや欠損等）が加えられたときに判定・予測エンジンの機能正確性に与える影響を評価するための手法である（石川・徳本［2019］等）。具体的には、テストデータに対して、ランダムに選択したさまざまな摂動を付加したテストデータを生成する。これらを入力としたときの出力結果を指標に基づき比較することにより、摂動の影響を評価する。

この手法は、テストデータの生成にかかる処理がシンプルであり実施しやすい。一方、機能正確性に大きな影響を与えるテストデータを得られることが保証されておらず、この手法のみで厳密に機能正確性を評価することは難しい。

(4) 形式検証手法

形式検証手法は、判定・予測エンジンが機能正確性を満たすために必要な条件を満たしていることを数学的に検証する手法である（Huang *et al.* [2017] 等）。例として、サーチ・ベースド手法と同様に、自動運転技術の事例を用いて、形式検証手法による評価の流れを説明する。

17 遺伝的アルゴリズムとは、多数のデータからある条件を満たすものを探索するアルゴリズムであり、探索処理とその結果に基づき、探索範囲等を絞り込む処理を繰り返すことにより、全数探索よりも効率よく条件を満たすデータを探索できる。

① 判定・予測エンジンの処理を、数式（論理式等）を用いて抽象化したモデル（抽象モデル）に変換する。
② 判定・予測エンジンが満たすべき要件（例えば、判定・予測エンジンは周辺画像に付加された摂動の影響を受けないこと）を、数式（論理式等）を用いて定義する。
③ 抽象モデルにおいて上記②の要件が成立することを数学的に証明する。

形式検証手法については、判定・予測エンジンの正確性を網羅的に評価することができる反面、判定・予測エンジンの処理が複雑な場合には、抽象モデルへの変換や証明に要する時間が膨大になることが課題として挙げられる。そのため、例えば、開発フェーズの仮説検証における（試作段階の）判定・予測エンジンの機能正確性の評価に利用することが考えられる。