Vol.2020-SE-204 No.19
2020/3/3

要求に応じた機械学習ソフトウェアの品質特性と測定方式の導出方法の提案と評価

仲道 耕二[1]　大橋 恭子[1]　難波 功[1]　山本 里枝子[1]　青山 幹雄[2]

概要：機械学習アルゴリズムの様々な分野への適用により，機械学習を組み込んだソフトウェア（MLS）の開発が急速に増加している．MLS の品質は，モデル学習時の学習データや運用時の入力データの量や分布傾向などに大きく依存するなど，従来のソフトウェアシステムとは異なる品質の考え方が必要である．これはエンタープライズ向け MLS 開発の品質保証における大きなリスクとなっている．この問題に対して，本稿では以下を提案する．(1)従来のソフトウェア品質を規定する ISO25010 の品質特性を MLS に特有な品質特性に拡張し，その測定方法を定義する．(2)開発対象システムにおいて重視する品質特性とその測定方法を導出するために，要求定義において決定すべき要求を特定する．これは開発対象システムの目的によって重視する品質特性とその測定方法が異なるためである．提案手法を評価するため，エンタープライズ向け MLS の機能正確性と成熟性に関する品質特性と測定方法の実証実験を行った．提案方法により導出した品質特性および測定方法と開発者が重視する品質特性および測定方法を比較し，提案方法の有効性を示す．

キーワード：機械学習ソフトウェア，品質，品質特性，品質尺度

Quality Characteristics and Measurement Method for Machine Learning Software Based on the Requirements and its Evaluation

KOJI NAKAMICHI[†1]　KYOKO OHASHI[†1]　ISAO NANBA[†1]
RIEKO YAMAMOTO[†1]　MIKIO AOYAMA[†2]

1. 研究の背景

機械学習アルゴリズムを組み込んだソフトウェアシステムの開発が様々な分野で広がっている．しかし，実際の開発においては様々な問題があることが指摘されている[3][14][15][26]．機械学習アルゴリズムを組み込んだソフトウェアの開発はまだ発展途上であり，ソフトウェア工学の視点からその開発の体系化して諸問題を解決の目指す機械学習ソフトウェア工学[2]や機械学習工学[15]の議論が開始されている．以降，本稿では機械学習アルゴリズムを組み込んだソフトウェアシステムを機械学習ソフトウェア(MLS: Machine Learning Software)と呼ぶこととする．

丸山は，統計的機械学習の限界として，「訓練データセットに現れない領域のデータでは十分な精度が出ない」，「本質的に確率的であるため，その出力はサンプリングバイヤスに影響される」等を指摘した．また，機械学習ソフトウェアの開発では，通常のソフトウェア開発で得られるソースコードの他に，訓練データセットと学習済モデルの２つの生成物があることも指摘した．これらの違いも背景に，従来のソフトウェア工学と異なる点として再利用と品質保証を議論した[15].

本稿では，企業情報システムへ MLS を導入する際の，MLS の特徴を捉えた品質問題に焦点をあてる．そのため，MLS が持つ従来のソフトウェアと異なる性質に着目し，品質モデルの定義とその具体的な品質特性を明らかにする．さらに，品質特性を評価するための尺度と測定方法を提案し，産業界で実行可能な実装を議論する．

本稿では 2 節で背景と研究課題，3 節で関連研究を述べ 4 節でアプローチを説明する．5 節と 6 節で MLS の品質モデルと従来のソフトウェアとは異なる品質特性及び品質評価方法を提案する．7 節で企業システム開発の上流工程で顧客と合意すべき要求仕様と MLS 品質特性の関連に着目した，品質測定方法とその評価を報告する．

2. 研究課題（Research Questions）

企業情報システムの開発においては，開発ベンダーがシステム開発を受託した場合に顧客と品質レベルを合意しながら開発・運用を行うことが実践されてきている．MLS を企業システムに導入する際，従来のソフトウェアと異なる特徴により，従来のソフトウェアの品質特性がそのままでは適用できないことが指摘されている[13]．また，Breck らは特定のモデルの実際の予測動作を事前に記述することが困難であるとした[5]．これまでこの困難に対して開発現場での開発プロセスを工夫して，PoC (Proof of Concept)開発

1 (株)富士通研究所
Fujitsu Laboratories Ltd.
2 南山大学
Nanzan University

でデータ収集やモデルの実現性を確認し，MLS で採用すべき技術やユーザシナリオを特定することが広く実施されている．机上で決められない多くの要求仕様は PoC を通じて具体化する[1][15]．その際にはソフトウェアとしての品質特性も併せて顧客と合意すべきであるが現状は品質の考え方を産業界で共通化する議論を進めている段階である[19].

このような背景のもと，著者らは企業情報システムへ MLS の導入する際の，MLS の特徴を捉えた品質問題に対して，開発するシステムの発注元である顧客との品質に関する合意形成を目的に，MLS の品質測定を可能とする方法を明確にする．具体的には，以下の三つを提案する．

(1)MLS で議論すべき要求仕様での考慮点
(2)MLS の特徴を捉えた品質測定を実現するための品質モデル，品質特性，及び，尺度
(3)顧客と合意すべき要求仕様での考慮点から，継続的に測定すべき品質特性と尺度の特定手法

このため，以下の研究設問を設定する．

RQ1：従来のソフトウェア開発と異なる特徴を捉えるための MLS の品質特性と尺度は何か？それらを顧客と合意することを目的として，要求仕様に取り込むべき考慮点は何か？

RQ2：提案する尺度と要求仕様は実際の MLS の開発に有効か？

3. 関連研究

3.1 MLS の品質問題

MLS は従来のソフトウェアとは本質的に異なる特性を持つことが指摘されている[15][21]．そのため，ソフトウェア工学における新たな課題を提起している[2][3][14].

MLS の品質では組込まれる機械学習アルゴリズムに加えて，ソフトウェアシステム全体としての品質が課題となる．例えば，MLS の応用として注目する自動運転では安全性が最も重要な品質特性であるが，その品質問題の研究は萌芽的段階にある[13].

MLS 固有の品質特性として性能ドリフト(Performance Drift)の問題[20]がある．MLS の品質は学習時あるいは運用時の入力データとその変化に依存する．Chui らは 400 件以上の MLS の適用事例の分析から 34%で学習モデルの更新(refreshment)が必要であり，その中の 77%は少なくとも 1 ヶ月に 1 回の更新が必要であったことが報告されている[6]．田中らは MLS が利用環境の変化によってその精度が変化する問題を指摘している[23].

TensorFlow を利用した MLS のバグの実証分析からはバグの最大の原因が TensorFlow の API 変更であることが明らかになっている[26]．この問題は Web API でも問題となっている[25]．フレームワークや Web API を利用することが一般に行われている MLS 開発の品質問題として重要である．

3.2 MLS の品質モデルと品質保証

ソフトウェア製品の品質モデルとその尺度として ISO25000 シリーズが広く知られている[9][10][11]．しかし，MLS では，上述の特性から新たな品質モデルと尺度が求められる[12]．Masuda らのサーベイでも MLS の品質への多様な関心事を明らかにしている[16].

このような背景の元，MLS とその開発に関するガイドラインが提案されている．AI プロダクト品質保証ガイドラインでは技術カタログとして幾つかの技術が呈示されている[19]．総務省の AI 開発ガイドライン案でも幾つかの品質特性が指摘されている[22]．Murphy らは特定の MLS に対する品質保証のフレームワークについて議論しているが，その対象が限定されかつ品質保証の議論も限定されている[17]．このように，MLS の体系的な品質モデルの議論にまでは至っていない段階にあるといえる．

3.3 MLS の品質要求とその定義

MLS の要求定義に関して，例えば，Belani らは AI に関連したエンティティ(データ，モデル，システム)と要求工学のアクテビティ(要求獲得，分析，仕様化，検証，管理，文書化)を関連づけて，要求工学の視点からの MLS 開発の課題を整理した[4]．また，Vogelsang らは ML のエキスパートであるデータサイエンティストの作業をヒアリングし，従来の要求仕様からの変更点として，ML 性能の測定の重要性，説明可能性や特定の法的要求などの新しい品質要求も導入し，従来の要求工学プロセスに ML の要求定義のために追加するアクテビティを提案している[24]．要求仕様の項目に関連した研究では Hyatt らは堅牢なニューラルネットワーク開発の要求仕様として精度(Validation Accuracy)が優先して最適化されていることに対して，他の性能尺度も追加すべきであると指摘している[7].

4. アプローチ

4.1 MLS の品質要求仕様

一般的なシステム開発では、システムへの要求を明確した上で開発を進める．MLS の開発では，前述したようにモデルの実際の挙動予測を事前に記述することが困難である．そのため，顧客との合意は PoC 開発により，事前に挙動の確認を行うなどの工夫をしているが，PoC 開発では品質に関する考慮漏れが発生しやすい問題がある．このような問題に対して，MLS に対する要求に基づいた品質測定方法が決定できることが求められる．そのため本稿では，MLS の開発経験者による，MLS への要求仕様定義に関する考慮点を分類した上で，それらを継続的に測定するための方法を提供するアプローチを提案する．

4.2 MLS の特徴を捉えた品質特性の特定

本稿では，ソフトウェア製品の品質モデルを基礎として，MLS の特性とそれによって提起される品質に関する新たな問題を解決するために，品質モデルを拡張するアプロー

チをとる．

ソフトウェア製品の品質モデルは ISO/IEC25000 シリーズが広く認知され利用されている[9][10][11]．ISO25010 はソフトウェア製品の提供者とそのユーザ間での品質を規定する．本稿では，MLS の特性を捉えた品質特性を明確にするために品質モデルの基礎として ISO25010 を用いる．現在産業界で議論されている MLS の品質に関する発表資料[8][18][22]と社内の MLS 有識者による資料を参照し，ISO25010 のどの品質特性と品質副特性における考慮点を解析し，品質副特性で考慮すべき事項を特定した．考慮すべき事項が多い副特性を，MLS の品質で議論すべき副特性と判断した．その結果，本稿では以下の副特性を MLS の特徴を捉えるように拡張する．

(1)品質特性「機能適合性」の副特性の「機能正確性」
(2)品質特性「機能適合性」の副特性の「機能完全性」
(3)品質特性「信頼性」の副特性の「成熟性」

5. MLS の品質要求定義における考慮点

2 節で議論したように MLS は実行せずに要求仕様の詳細を決定することは難しい．実際の開発現場では開発プロセス上の工夫として，PoC 開発でデータ収集やモデルの実現性を確認することが広く行われている．

著者らは MLS 開発を実施してきた技術者と，MLS に固有な開発の考慮点を特定した．PoC を通じた要求仕様の獲得においてもこれらの考慮点が顧客との確認事項の候補となり得る．図 1 は MLS の構成の中で品質に影響するコンポーネントの構成を示す．要求仕様の考慮点を，環境／ユーザ，システム／インフラ，モデル，データに大別した．表 1 に環境／ユーザ，表 2 にデータに関する考慮点の一部を示す．環境／ユーザは 6 項目，データは 15 項目を特定した．モデルに関しては，「モデルの型」や「訓練済モデル」に関して，訓練時の性能，実行時の性能，リソースを含む 13 項目の考慮点を抽出した．システムに関しては 12 項目を特定した．これらを要求仕様で考慮すべき項目として，MLS システムの品質モニタリングの対象とモニタリング観点の入力とすることを提案する．7 節で，著者らが整理した顧客からのパースペクティブでモニタリングすべき品質特性を議論する．

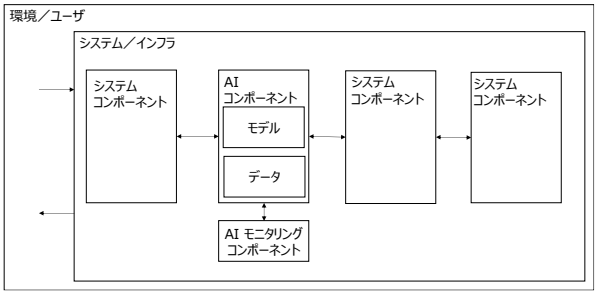

図 1 MLS の構成要素の概要

表 1　考慮点（環境／ユーザ）

構成子	分類	確認すべき考慮点
環境／ユーザ	環境の効果	(a)サプライチェーンの環境への親和性 (b)開発時／利用時のリソースの使い方
	社会的効果	(a)人が介入する可能性の提供 (b)AIと利用者が会話することを利用者がわかっているか (c)AIシステムによる社会的関係への影響
	スコープの追従	(a)スコープの変化の境界を監視するか (b)特異点検出を使って未知の境界を検出するか

表 2　考慮点（データ）

構成子	分類		確認すべき考慮点
データ	訓練時／テスト時／実行時のデータ	代表性	(a)ターゲットアプリの値域（最小値／最大値）とデータセットの値域の比較する， (b)ギャップを検出するためのデータの密度 (c)ターゲットアプリの想定分布とデータセットの分布の類似性， (d)データサンプリング方法の正確さ
		均衡性	(a)各クラスのサンプルの数， (b)クラスにあるサンプルの確率分布の比較
		適時性	(a)訓練／テスト用のデータ収集と実行時データ間の時間， (b)データに影響を及ぼす可能性のあるシステムまたは環境の変更があったか
		完全性 ・・	(a)欠損値の数，(b)・・・

6. 品質評価モデルと品質評価方法

本節では，4 節で示した 3 つの品質副特性のうち，後述の評価に関連する機能正確性と成熟性を詳細に述べる．

6.1.1 品質評価メタモデル

MLS に求められる品質評価への認識の共通化のため，MLS の品質評価メタモデルを定義する(図 2)．

本メタモデルは 3 つのパートから構成した．MLS 品質モデルパートでは，品質特性を階層的に定義できるようにした．MLS 品質測定パートでは，最下層の品質特性を測定するための尺度，その測定結果，品質評価対象のシステムやその構成要素を関連づけた．これらは，一般的なソフトウェアと同様である．

本メタモデルでは，尺度とそのサブクラスとして定量的尺度を定義した．MLS の代表的尺度として学習済みモデルの精度がある．これは定量的尺度である．測定された精度は，それが特定の MLS 開発プロジェクトにおいて十分か否かは測定値だけでは判断できない．モデル精度に対して適切な目標値(基準)が定義されており，実測したモデル精度と目標値を比較し，モデルの精度が十分かの判断を行う必要がある．これらは，定量的には測定できないため，定量評価以外に定性評価が必要だと考えた．

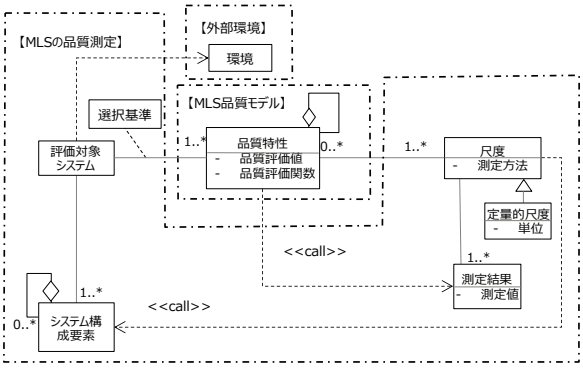

図 2　MLS 向け品質評価のメタモデル

定量的評価と定性的評価の達成度を判定するための測定方法が必要である．一方，定量的評価では測定値の単位の定義も必要である．そこで，定性的評価をスーパークラス「尺度」の属性「測定方法」を，定量的評価をサブクラスとして属性「単位」を定義した．

次に図 2 の中で MLS として特徴的なパートである外部環境の要素「環境」を説明する．この要素は，MLS の品質の良さに影響を及ぼす MLS の外部の要素で，MLS の運用中に変化する可能性がある．例えば想定ユーザ，MLS に対する要求， MLS への入力となるデータ，あるいは MLS を利用した業務で参照している法律や基準である．

以降の節では，機能正確性，成熟性に注目し，前節に示したメタモデルの品質特性と尺度，定量的尺度を具体化する．具体化にあたっては，AI に関連したエンティティ（データ，モデル，システム）の中のどの要素の品質を評価するのかを明確にすること，その要素の評価すべき性質を明確にする．その理由は，特に定性的な尺度を評価では，回答者の主観による回答の個人差を可能な限り低減するためである．なお，MLS で扱うデータは， MLS の開発や運用で扱うデータは，訓練データやテストデータ等があり，それぞれ異なる役割がある．その役割応じて評価すべき性質が異なると考え，訓練データ，テストデータ，入力データ，出力データ，MLS 出力データに詳細化した．MLS の挙動は本質的に確率的であり，異常値が出力される可能性がある．そのような場合であっても MLS としての出力を保証するため，学習済みモデルの出力データに対して MLS が処理を行うことがあるので両者を分けた．

6.2 機能正確性の品質評価モデル

6.2.1 機能正確性の評価モデル

本節では，機能正確性に注目し，前節に示したメタモデルの品質特性を具体化する．MLS の機能としての正確性は，その出力データの正しさで判断すると考えられる．そこで，MLS の出力結果に関わるエンティティとして MLS の出力データ，テストデータ，学習済みモデルを特定した．これらのエンティティに機能正確性の面から評価すべき性質を定めた．以下にその性質を列挙する

(1)傾向の一致：データに対する性質である．データセットとして想定される統計量に対して実際のデータセットの統計的な特徴が一致している程度．

(2)独立性：テストデータに対する性質である．テストデータが訓練データと独立している程度．

(3)精度：MLS や学習済みモデルの出力データの性質で，正しさの程度．

機能正確性の品質特性は，前述の機能正確性で注目するエンティティと性質を組合せて具体化した．その構造を図 3 に示す．図中で取り上げた品質特性の概要を以下に示す．

a)学習済みモデルの精度正確性：学習済みモデルから得られた結果が正しい程度

b)MLS の精度正確性：MLS の出力データが正しい程度．

c)出力データの適合性：学習済みモデルの出力データ想定している出力データと類似している程度．

d)テストデータセットの独立性：テストデータが訓練データと独立している程度．

6.2.2 機能正確性の尺度と測定方法

学習済みモデルの正確さは，通常，精度という定量的尺度で評価されることが多い．しかし，この尺度だけではモデルに求められる正確さが得られたかは判断できない．一般に，何らかの基準が設定されている．また，その基準が適切でない場合は比較をしても意味がない．さらに，正確さを実際に測定したり，環境の変化などにより基準を変更する必要がある可能性もある．そこで，定性的な尺度の達成度の順位として以下の 7 項目を設定した．上側ほど達成度が低く，下にいくほど達成度が高いとする．

(1) 基準となるような目標値の有無
(2) その目標値の妥当性
(3) 目標値対してテスト環境や運用環境で得られた実測値の十分さ
(4) 目標値の利用手順の有無
(5) 目標値の更新手順の有無
(6) 運用環境での継続的な測定の実施有無

この順位に基づき，前節に示した 4 つの品質特性のうち，最初の 3 品質特性に対して尺度と測定方法の定義を示す．

(1) モデル精度: 学習済みモデルの正確さの程度を表す定性的尺度である．正確さの達成度は本節の冒頭に述べた順位に基づいて測定する．

(2) モデル正解率の達成率(A_{model}): 上述のモデル精度の判断のために測定する．学習済みモデルが目標とする正解率に対する，実測した学習済みモデルの正解率の達成度を表す．その定義を式(1)に示す．値は 0 以上である．実測した正解率が目標値を下回っている場合は 1 未満，目標値と等

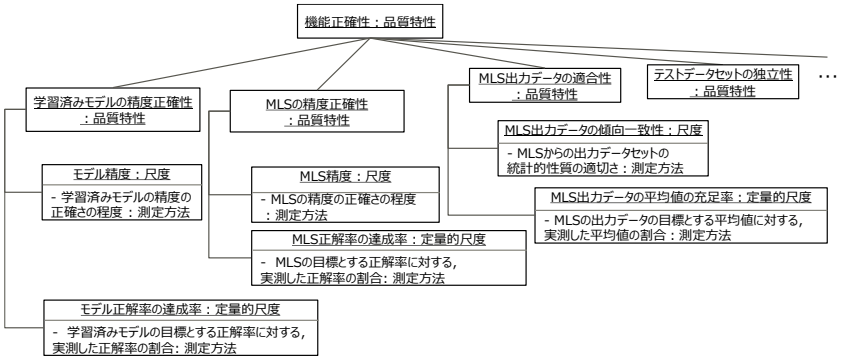

図 3　MLS 向け品質評価モデルの一部(機能正確性)

しければ 1，目標を上回ると 1 より大きくなる．

$E_{model\_accuracy}$：目標とする正解率

$R_{model\_accuracy}$：実測された正解率

$$A_{model} = \frac{R_{model\_accuracy}}{E_{model\_accuracy}} \quad \cdots(1)$$

(3) MLS の精度: MLS 運用時の出力の正確さを表す定性的尺度である．モデル精度と同様に，正確さの達成度は本節の冒頭に述べた考え方に基づいて測定する．

(4) MLS 正解率の達成率(A_{mls}): 上述の MLS の精度の判断のための尺度である．モデル正解率の達成率と同様に MLS が目標とする正解率に対する，実測した MLS の正解率の達成度を表す．その定義を式(2)に示す．値域はモデル正解率の達成度と同一である．

$E_{mls\_accuracy}$：目標とする正解率

$R_{mls\_accuracy}$：実測された正解率

$$A_{mls} = \frac{R_{mls\_accuracy}}{E_{mls\_accuracy}} \quad \cdots(2)$$

(4) MLS 出力データの傾向一致性: MLS の出力データセットの統計的な傾向が，想定している傾向と一致する程度を表す定性的尺度である．モデル精度と同様に，一致の程度は本節の冒頭に述べた順位に基づいて評価する．

(5) MLS 出力データ平均値の充足率(V_{mls}): 上述の MLS 出力データの傾向の判断のための尺度である．MLS 出力データ値の平均値の目標に対する，実測した MLS の出力データの平均値の達成度を表す．その定義を式(3)に示す．値域はモデル正解率の達成度と同一である．

$E_{mls\_average}$：目標とする平均値

$R_{mls\_average}$：実測された平均値

$$V_{mls} = \frac{R_{mls\_average}}{E_{mls\_average}} \quad \cdots(3)$$

出力データが多次元の場合は，次元毎に計算する．

MLS 出力データの傾向一致を判断する際に平均値を利用したが，標準偏差や AUC(Area Under the Curve)を用いたり組み合わせたりすることもある．

6.3 成熟性の品質評価モデル

6.3.1 成熟性の評価モデル

本節では，成熟性に注目し，前述したメタモデルの品質特性を具体化する．MLS の成熟性とは，定められた機能正確性を維持する能力と考える．

成熟性に関わるエンティティとして MLS が持つ，入出力データを監視し異常の予兆がある場合や発生した場合に対応するための各種手段，入力データに注目した．これらのエンティティに成熟性の面から評価すべき性質として，手段の適切さと入力データの変化に対するロバストネスを定めた．

成熟性の品質特性は，注目するエンティティと性質を組合せて具体化した．その構造の一部を図 4 に示す．図中で取り上げた品質特性を以下に示す．

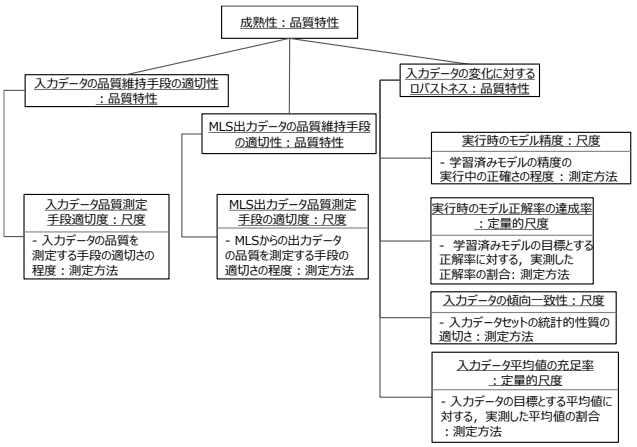

図 4　MLS 向け品質評価モデルの一部(成熟性)

(1)入力データ品質維持手段の適切性: 入力データの傾向をモニタリングしたり，外れ値や欠損値を除去する等の力データの品質を維持するための手段が適切である程度を示す．

(2)MLS 出力データ品質維持手段の適切性: ML の出力データの傾向をモニタリングしたり，ML の出力異常時に代替系の値を採用するなど，MLS としての品質を維持するための手段が適切である程度を示す．

(3)入力データの変化に対するロバストネス: 実行中に入力される入力データが学習時と変化しても学習済みモデルの精度が変化しない程度を示す．

6.3.2 成熟性の尺度と測定方法

前節に示した 2 つの品質特性の尺度と測定方法の定義を示す．

(1) 入力データ品質測定手段の適切度: 入力データ品質測定手段の適切さを表す定性的な尺度である．適切さの程度は，モデル精度の考え方を応用した．6.2.2 節に示した「基準となるような目標値の有無」は「手段の有無」,「その目標値の妥当性」は「手段の適切さ」として測定する．

(2) MLS 出力データ品質測定手段の適切度: MLS 出力データ品質測定手段の適切さを表す定性的尺度である．これも(1)と同様に測定する．

なお，ここに挙げた 2 つの尺度は定量的尺度ではない．

(3) 実行時のモデル精度: 学習済みモデルの実行中の正確さの程度を表す定性的尺度である．精度を実行中に測定すること以外は, 6.2.2 の(1)に示したモデル精度と同様である．

(4) 実行時のモデル正解率の達成率: 上述の実行時のモデル精度を判断するための尺度である．学習済みモデルを用いた MLS の実行の正解率である．精度測定を実行中に行うこと以外は, 6.2.2 の(2)に示したモデル正解率の達成率と同様である．

(5) 入力データの傾向一致性: 入力データの統計的な傾向が，想定している傾向と一致する程度を表す定性的尺度

である．測定の対象が入力データであること以外は，6.2.2 の(4)に示した MLS 出力データの傾向一致性と同様である．

(6) 入力データ平均値の充足率: 上述の入力データの傾向一致の判断のための尺度である．入力データの平均値の目標に対する，実測したデータの平均値の達成度を表す．測定の対象が入力データであること以外は，6.2.2 の(5 に示した MLS 出力データ平均値の充足率と同様である．

7. 考慮点に基づく品質特性と尺度の選択方法

5 節で述べたように，MLS 開発時の考慮点は要求仕様の確認項目になる．しかし，この確認項目は品質評価のための具体的な尺度と関係づけられていない．そこで本節では，考慮点を元に 6 節の品質評価モデルから，対象とする MLS で重視すべき品質特性と尺度を選択する方法を提案する．以下ではまず考慮点と品質モデルの関連を示し，次に関連に基づいて品質特性と尺度を選択する方法を示す．

7.1 考慮点と品質評価モデルの関連

考慮点と品質評価モデルの関連を表 3 に基づき説明する．縦軸は 6 節で定義した品質評価モデルの品質特性と尺度である．ここでは「機能正確性」「機能完全性」「成熟性」に対して具体化された MLS の品質特性の内の一部を示している．横軸は 5 節で抽出した考慮点を示している．考慮点と品質特性／尺度と考慮点の交点は，考慮点に対して品質特性や尺度が関連する場合に〇で示してある．

例として 2 つの考慮点に対する品質特性と尺度の関連付けを説明する．考慮点「データ／代表性／（c）ターゲットアプリの想定分布とデータセットの分布の類似性」は，例えば入力データに対するデータ分布の当初想定分布からの変化，つまりコンセプトドリフトに関わるものとみなせる．そこでコンセプトドリフトに関連する品質特性と尺度の組として，実行中のモデルの精度に関わる「学習済みモデル精度の正確性／モデル精度」、入力データの分布に関わる「入力データに対するロバストネス／入力データ傾向一致」、入力データの監視手段に関わる「入力データの品質維持手段の適切性／入力データ統計量測定手段の十分さ」をそれぞれ関連付けた．

もう 1 つの考慮点「環境／スコープへの追従／（a）スコープの変化の境界の監視」は，環境の変化によって同じ入力に対して分類タスクの分類先が変化する状況に関連している．分類先が変化することで，入力データに対するモデルからの出力結果が変わらないにも関わらず，ある時点でその出力に対する正解／不正解のラベルが変わることになり、その結果システムとしての精度が低下する．そこで分類先の変化に関連する品質特性と尺度の組として，分類先変化後の精度変化に関する「MLS の精度正確性／MLS の精度」，分類先変化前後での出力データの変化状態に関わる「MLS の精度正確性／出力データ傾向一致」，MLS 出力データの監視手段に関わる「MLS 出力データの品質維持手段の適切性／MLS 出力データ統計量測定手段の十分さ」をそれぞれ関連付けた．また上記以外の考慮点に関しても同様の関連付けが可能である．

表 3　考慮点と品質評価モデルの関連

品質評価モデルにおける品質特性と尺度			考慮点				
			データ		環境		…
品質特性／副特性	MLSの品質特性	尺度	代表性	…	スコープへの追従	…	
			(c)	…	(a)	…	
機能適合性／機能正確性	学習済みモデルの精度正確性	モデル精度	〇				
	MLSの精度正確性	MLSの精度			〇		
	MLS出力データの適合性	出力データ傾向一致			〇		
	…						
機能適合性／機能完全性	訓練データの適合性	訓練データ傾向一致					
	…						
信頼性／成熟性	入力データの変化に対するロバストネス	入力データ傾向一致	〇				
	入力データの品質維持手段の適切性	入力データの統計量測定手段の十分さ	〇				
	MLS出力データの品質維持手段の適切性	MLS出力データの統計量測定手段の十分さ			〇		
	…						

7.2 品質特性と尺度の選択方法

次に表 3 の考慮点と品質モデルとの関連付けに基づいて，重視する品質を選択する．図 6 に品質特性と尺度の選択方法の概要を示す．

まず考慮点毎に，その考慮の必要性を問う質問を実施する．例えば上記「データ／代表性／（c）ターゲットアプリの想定分布とデータセットの分布の類似性」に対する質問は，「実行時の入力データセットの分布が想定からの変化する可能性はあるか？」のようなコンセプトドリフトの発生可能性を問うものとなっている．同様に「環境／スコープへの追従／（a）スコープの変化の境界の監視」に対する質問は，例えば「同一の入力に対する予測や分類の適用先が変わる可能性があるか？」のように環境の変化に起因した分類先変化の可能性を問うものになっている．

各考慮点に対して品質特性と尺度が関連付けられていることから，品質特性と尺度はそれぞれの考慮点から重複して関連付けられる．そこで考慮点に対して重み値 w_n を与え，品質特性と尺度に対して関連付けられた重み値の合計値 Σw_m を算出する．さらに合計値が閾値以上のものを出力

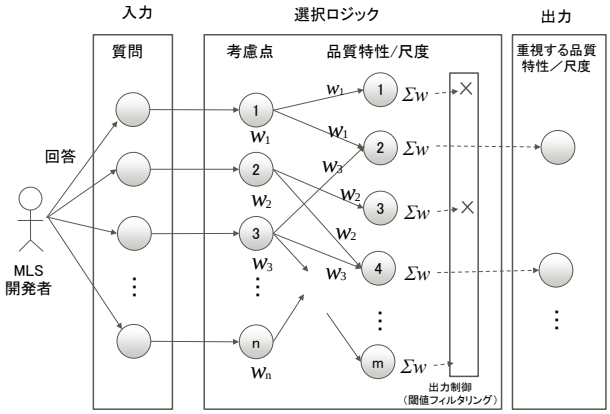

図 6　品質特性と尺度の選択方法の概要

表 4　利用状況に対する質問紙

質問	選択肢
品質特性「学習済みモデルの精度正確性」に関連する尺度の利用状況は？	1. 実際に使用している
	2. 使いたい／試したい
	3. 使いたいが難しい（技術面）
	4. 使いたいが難しい（コスト面）
	5. 必要性を感じない
	6. 定性／定量評価の意図が不明

表 5　有効性に対する質問紙

質問	選択肢
「(c)ターゲットアプリの想定分布とデータセットの分布の類似性」がある場合，提示された品質特性や尺度は有効か？	0. 無効
	1. 有効

表 6　回答集計結果

利用状況						有効性	
1. 実際に使用	2. 使いたい/試したい	3. 使いたいが難しそう（技術面）	4. 使いたいが難しそう（コスト面）	5. 必要性を感じない	6. 定性／定量評価の意図が不明	0. 無効	1. 有効
38.9%	22.2%	16.7%	0.0%	0.0%	11.1%	83.3%	16.7%
77.8%							

することで，重視すべき品質特性と尺度の組を選択する．ここで考慮点に対する重み値 w_n は考慮点に対する重要性などに基づいて決定される．

このようにして考慮点を元に MLS 品質評価モデルの中から重視すべき品質特性や尺度を選択できる．

8. 評価と考察

考慮点に対する確認事項に基づき品質評価モデルから対象とする MLS において重視する品質特性と尺度を選択する方法の妥当性を検証するために，7 節で述べた 2 つの考慮点に対する品質特性と尺度の選択結果を開発者に提示し，提示内容の妥当性を回答してもらう実験を行った．以下 8.1 節で評価手順を説明し，8.2 節で結果を示し，8.3 節は RQ に対する考察を述べる．

8.1 評価手順

(1) 考慮点に対する重要な品質特性と尺度の選択

7 節で説明した 2 つの考慮点に対し，その重みを 1，選択の閾値を 1 以上として，重要な 6 個の品質特性と尺度を選択した．

(2) 選択された重要な品質特性と尺度の提示

選択された 6 つの品質特性と尺度に対して，複数の MLS 開発者に対して以下の項目を提示した．

a) 品質特性名
　ISO25010 の品質特性，副特性，MLS の品質特性
b) 定性評価尺度
　i) 定性評価の名称
　ii) 定性評価尺度の達成度レベル．達成度のレベルとして「品質向上手段(機能)の有無」「目標や手段の validation」「目標や手段の verification」「目標や手段の利用手順や見直し手順，手順の実施有無」を提示．
c) 定量評価尺度
　定量評価尺度名，定量評価尺度の概要，定量評価尺度の計算式

(3) 開発者に提示した品質特性と尺度に対する調査

開発者に提示した 6 つの品質特性と尺度毎に質問紙を用いて「尺度（定性／定量評価）の利用状況」と「実施条件に合致している場合の尺度の有効性」の調査を実施した．表 4，表 5 に質問の例を示す．

(4) 調査結果の集計

表 6 に回答結果を示す．提示した 6 個の MLS 品質特性と尺度に対する回答結果を合算し，選択肢毎の回答率を示す．

8.2 考察

8.2.1 RQ1：**従来のソフトウェア開発と異なる特徴を捉えるための MLS の品質特性と尺度は何か？それらを顧客と合意することを目的として，要求仕様に取り込むべき考慮点は何か？**

著者らは，MLS の特性を捉えた品質特性を明確にするために品質モデルに ISO25010 を用い，MLS の品質に関する産業界の発表資料を元に，MLS において重要な以下の副特性を特定した．

(1)品質特性「機能適合性」の副特性の「機能正確性」
(2)品質特性「機能適合性」の副特性の「機能完全性」
(3)品質特性「信頼性」の副特性の「成熟性」

さらに MLS の特徴を表す品質特性に細分化し，それぞれに定量／定性尺度を規定した品質評価モデルを定義した．

また MLS の開発時の考慮点を MLS 開発に関わる有識者との議論に基づいて定義した．これらは環境／ユーザ，システム／インフラ，モデル，データの観点で分類され，MLS の品質評価モデルと関連付けることで，考慮点に対するに質問に基づき重視する品質特性や尺度の選択が可能となる．これにより MLS 品質に関する顧客との合意が可能となる．

8.2.2 RQ2：**提案する尺度と要求仕様は実際の MLS の開発に有効か？**

品質特性と尺度の利用状況については，表 6 より，開発者に提示された品質特性と尺度の実施率（回答 1）は 38.9%であった．それほど高いとは言えないが，「使いたい／試したい」（回答 2）、「使いたいが難しい」（回答 3, 4）を含めると 77.8%が提示された尺度の利用に肯定的であると考えられる．

ただし，今回の評価は 2 つの考慮点に対応した 6 つの品質特性と尺度についてのみの評価である．今後，より多くの品質特性や尺度に関して同様の結果が得られるかどうかを確認する必要がある．

「使いたいが難しいそう（技術面）」（16.7%）における回答では、定量尺度であるモデル精度に関して，実行時にモデル出力値に対する正解／不正解ラベルを人が付与するため，正解率が出せないケースがある等 MLS の適用に応じてモデル精度の適用の容易性に差が出ることが分かった．

一方、品質特性と尺度の有効性に関しては，提示された尺度に対して有効であるとの回答が 83.3%であることから，提示された尺度の有効性はある程度確認できたといえる．ただしこれも限られた数の尺度に対する評価結果であり，今後，より多くの品質特性や尺度に関して評価することが課題である．

9. まとめ

本稿では，企業情報システムへ導入する MLS の特徴を捉えた品質問題に焦点を当て，開発するシステムの顧客と開発者間で品質に関する合意形成を目的に，MLS の品質測定を可能とする方法を提案した．

まず MLS で議論すべき要求仕様での考慮点について、MLS 開発時の考慮点として MLS 開発者との議論を元に定義した．これらは環境／ユーザ，システム／インフラ，モデル，データの観点で分類した．

さらに，MLS の特徴を捉えた品質測定を実現するための品質評価モデルとして，ISO25010 の品質モデルを元に MLS において重要な副特性を特定し，MLS の品質特性として具体化し，その尺度を定義した．

考慮点に基づいて測定すべき品質特性と尺度の特定する方法として，考慮点と品質評価モデルの品質特性と尺度関係づけることで，考慮点に対する質問に基づいて，考慮点に対応した重視すべき品質特性と尺度を選択する方法を提案した．

提案手法を評価するために，一部の考慮点に基づいて事前に選択した品質特性と尺度を開発者に提示し，それらに対する実際の利用状況や有効性を回答する実験を実施した．その結果，提示された品質特性や尺度の利用には肯定的であり，尺度の有効性にも高い回答が得られたことから，提案方法の有効性を一定の水準で確認した．しかし，本実験では限定された考慮点に基づいて選択された品質特性と尺度のみの評価であることから，今後さらに評価が必要である．

参考文献

[1] アビームコンサルティング P&T Digital ビジネスユニット Advanced Intelligence セクター, AI システム構築実践ノウハウ, 日経 BP, 2019.
[2] 青山幹雄, ソフトウェア工学基礎から機械学習ソフトウェア工学基礎への考察, FOSE2019 論文集, 日本ソフトウェア科学会/近代科学社, Nov. 2019, pp139-144.
[3] A. Arpteg, et al., Software Engineering Challenges of Deep Learning, Proc. SEAA ‘18, IEEE, Aug. 2018, pp50-59.
[4] H. Belani, et al,Requirement Engineering Challenges in Building AI-Based Complex Systems, Proc. RE ‘19 Workshops, IEEE, Sep. 2019, pp. 252-255.
[5] E. Breck, et al., The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction, Proc. IEEE Big Data ‘17, IEEE, Dec. 2017, pp.1123-1132.
[6] M. Chui, et al., Notes from the AI Frontier: Insights from Hundreds of Use Cases, Discussion Paper, McKinsey & Company, Apr. 2018, pp. 1-32.
[7] J. S. Hyatt et al., Requirements for Developing Robust Neural Networks, Proc. AAAI FSS-19, AAAI, Nov 2019, pp. 1-4, https://arxiv.org/abs/1910.02125.
[8] IPA SEC: AI 社会実装推進調査報告書, June 2018, available from https://www.ipa.go.jp/files/000067229.pdf.
[9] ISO/IEC 25010:2011, Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - System and Software Quality Models, 2011.
[10] ISO/IEC 25020:2007, Software Engineering- Software Product Quality Requirements and Evaluation (SQuaRE) – Measurement Reference Model and Guide, 2007.
[11] ISO/IEC 25030:2007, Software Engineering- Software Product Quality Requirements and Evaluation (SQuaRE) - Quality Requirements, 2007.
[12] 科学技術振興機構研究開発戦略センター, 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, CRDS-FY2018-Sp-03,Dec.2018, https://www.jst.go.jp/crds/pdf/2018/SP/CRDS-FY2018-SP-03.pdf.
[13] H. Kuwajima, et al., Open Problems in Engineering and Quality Assurance of Safety Critical Machine Learning Systems, arXiv:1812.03057, Dec. 2018, pp. 1-7.
[14] L. E. Lwakatare, et al., A Taxonomy of Software Engineering Challenges for Machine Learning Systems, Proc. XP 2019, LNBIP Vol.355, Springer, May 2019, pp227-243.
[15] 丸山 宏, 城戸 隆, 機械学習工学へのいざない, 人工知能, Vol. 33, No. 3. Mar. 2018, pp. 124-131.
[16] S. Masuda, et al., A Survey of Software Quality for Machine Learning Applications, Proc. of ICSTW 2018, IEEE, Apr. 2018, pp. 279-284.
[17] C. Murphy, et al., A Framework for Quality Assurance of Machine Learning Applications, Computer Science Technical Report, CUCS-034-06, Columbia University, Apr. 2011.
[18] 日本マイクロソフト: 保証できる AI の実現のために, July 2018, available from https://www8.cao.go.jp/cstp/tyousakai/humanai/4kai/siryo2-1.pdf
[19] QA4AI コンソーシアム, AI プロダクト品質保証ガイドライン, 2019.05 版, May 2019, http://www.qa4ai.jp/.
[20] P. Santhanam, et al., Engineering Reliable Deep Learning Systems, Proc. of AAAI FSS-19: AI in Government and Public Sector, Nov. 2019, pp. 1-8, arXiv:1910.12582.
[21] D. Sculley, et al., Machine Learning: The High Interest Credit Card of Technical Debt, Proc. of SE4ML ’14 (NIPS Workshop), Dec. 2014.
[22] 総務省 AI ネットワーク社会推進会議: 国際的な議論のための AI 開発ガイドライン案, Jul. 2017, https://www.soumu.go.jp/main_content/000490299.pdf.
[23] 田中 優之 他, 機械学習ソフトウェアシステムの環境変化適応の課題とアプローチ: スマートフォンのナビゲーションアプリケーションを例として, 第 2 回機械学習工学研究会論文集, 日本ソフトウェア科学会, Jul. 2019, pp.49-54.
[24] A. Vogelsang et al., Requirements Engineering for Machine Learning: Perspectives form Data Scientists, 2019 International Requirement Engineering Conf. Workshops, IEEE, Sep. 2019, pp.245-251.
[25] 山本 里枝子 他, Web API の習得容易性と相互運用性, およびその定量評価方法の提案と適用評価, 情報処理学会論文誌, Vol. 60, No. 10, Oct. 2019, pp. 1896-1914.
[26] Y. Zhang et al., An Empirical Study on TensorFlow Program Bugs, Proc. ISSTA’18, ACM, Jul 2018, pp129-140.