

## 企業情報システム向け機械学習システムの 要求駆動型品質尺度の選択方法の提案と評価

大橋 恭子<sup>1</sup> 仲道 耕二<sup>1</sup> 難波 功<sup>1</sup> 山本 里枝子<sup>1</sup> 青山 幹雄<sup>2</sup>

**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機械学習システム(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訓練データと実行時の入力データの量や分布に依存している点が従来のソフトウェアシステムとは大きく異なる。このことは企業情報システム向けの MLS の品質保証にとって大きな課題となっている。我々はこの課題を解決するため、MLS 向け要求仕様の作成観点である考慮点、従来のソフトウェア品質を定義している ISO/IEC 25010 の品質特性を MLS 向けに拡張した MLS 品質特性とその尺度、さらに考慮点の重要度に基づいて測定すべき MLS 品質特性や尺度を選択する方法を提案し、その有効性を確認した。一方、MLS の開発現場では QCD(品質、費用、納期)に基づいた要求や測定尺度等の取捨選択も行われている。我々は、この中で品質に注目し、品質評価尺度に対して MLS の品質向上に寄与する度合いを表す必須度を導入した。本稿では、MLS を利用した企業情報システムの要求定義で、(1)品質評価のための尺度の必須度、必須度の決定要因と算出方法の提案、ならびに、(2)尺度に対して必須度で優先順位付けを行うことで、品質尺度の選択を支援する方法を提案する。MLS 専門家に対するケーススタディの結果、尺度の必須度による優先順位と開発現場の MLS の専門家の尺度選択の判断とが整合していることを確認できた。このことから、提案方法は MLS 要求定義において MLS に対する高い専門性がない技術者に対して品質特性定義とその尺度選択の支援に効果が期待できる。

**キーワード:** 機械学習、機械学習システム、品質要求、ソフトウェア品質モデル、品質特性、尺度、品質保証

### 1. はじめに

機械学習アルゴリズムを組み込んだ機械学習システム(MLS)の開発が企業情報システムにおいても広がってきている[18]。しかし、実際の MLS の開発には様々な問題があることが指摘されている [1][14] [26]。そこで、機械学習や機械学習システムに対して、ソフトウェア工学の観点から MLS の開発を体系化することで、様々な問題のソリューションの発見を支援する機械学習のソフトウェア工学の研究が行われている[1][14]。

MLS の振る舞いは、従来のソフトウェアと同様にソースコードから影響を受ける。しかし、訓練データセットと訓練済モデルという二つのコンポーネントからも影響を受けるという新たな特性がある。この違いを踏まえ、従来のソフトウェア工学での再利用や品質保証の違いが議論されている[1][14]。

我々は、MLS が持つ従来のソフトウェアと異なる性質に着目し、ISO/IEC 25010 の品質モデルで定義されている品質特性を具体化し、MLS に求められる品質特性である MLS 品質特性を定義した[10][21]。また MLS 品質特性とその尺度を組み合わせた MLS 品質評価モデルを定義した。

一方、MLS の開発でも従来のソフトウェア開発と同様に機能要求だけでなく、品質要求の定義も必要である。定義された要求仕様の妥当性を判断するには、要求仕様为目标とする MLS の品質特性やその達成の度合いの定義が必要である。そのため、これまでの取り組みでは、MLS 向け要求仕様の作成観点である考慮点を導入し、考慮点と MLS 品質特性との対応関係を定義した。この定義により、開発対象の MLS における考慮点の重要度に応じた MLS 品質特性の抽出を可能にした。

しかし、社内の MLS 開発の専門家からは、企業情報システム向け MLS の場合には、適用される業務や利用環境の特徴、

あるいはステークホルダーから求められる MLS の品質の要求水準やコスト等様々な要因を考慮して、測定する尺度を選択していることが指摘されている。そこで我々は、尺度の絞り込みを支援するために、考慮点に基づいて抽出した尺度に対して優先順位付けをして MLS の品質要求を定義する方法が必要だと考えた。

本稿では、MLS を利用した企業情報システムの要求定義で、品質への要求尺度の優先順位を決定する要因を提案し、その要因に基づいて決定される優先順位の妥当性を議論する。なお、以降、品質に対する要求水準を品質レベルと呼ぶ。

### 2. 研究課題

企業情報システム開発において、開発ベンダは顧客との要求の合意に基づき、開発や運用を遂行する。Breck らは機械学習においては学習済モデルの実際の振る舞いを事前に記述できないことを示した[3]。機械学習を適用した企業の調査では、最も高いリスクは期待されない出力であるとの指摘をしている[16]。これらの問題を扱うため、PoC(Proof of Concept)で学習済モデルの実現性を確認し、MLS で採用すべき技術やユースナリオを特定することが広く実施されている[14]。これらは主にソフトウェアの機能要求を対象とする議論である。

一方、企業情報システムとしての MLS を開発する際には、開発対象の MLS に求められる性能や信頼性などの品質要求も顧客と合意する必要がある[15]。開発した MLS が品質要求を満たしているかは、測定によって判断される。しかし、MLS の品質要求定義とその具体的な定義方法に関する従来の研究は極めて限られている。

そこで我々は、品質要求の定義には、MLS が満たすべき品

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

質を表す品質特性、その品質特性をどの程度達成しているかを測定するための尺度の定義が必要と考えた。AI プロダクト品質保証ガイドライン[22]には 50 個のチェック項目が提示されている。しかし、これらのチェック項目によりどのような品質が向上するかは明示されていない。また、チェック項目には正解率のように尺度と呼べる程度に測定方法が明確なものもある一方で、測定方法の具体化が必要なものも含まれている。例えば、“発生しうる品質事故の致命度は許容できる程度に低く抑えられているか”というチェック項目がある。この中の“品質事故”や“致命度”、“許容できる程度”を具体化しておかないと、チェックした人によってチェック結果が異なる恐れがある。

これらから企業情報システム向けの MLS の品質保証を行うには、以下の実現が必要だと考えた。

- (1) MLS に求められる品質を表す品質特性と品質の達成度を測定する尺度を明確にすること
- (2) 1 節に述べたように MLS が適用される業務や利用環境の特徴が特定された場合に、その MLS の品質保証を行うために必要な品質特性や尺度を選択する方法

上記を実現するため、我々は図 1 に示す MLS の標準的な構成を想定し、以下の 3 つの研究設問を設定し研究をしてきた。

RQ1: 品質特性や尺度を特定するための、MLS 向けの要求仕様の作成観点である考慮点とは何か？

RQ2: MLS の品質モデルとその品質特性、品質特性の達成度を測定する尺度は何か？

RQ3: 考慮点に基づいて選択した品質特性、尺度は実際の MLS 開発において有効であるか？

その結果、MLS 向けの品質特性である MLS 品質特性と尺度を組み合わせた MLS 品質評価モデル、考慮点一覧、考慮点と MLS 品質特性間の対応関係定義、を定義した。また、これらを用いて、高重要度の考慮点から測定が必要な尺度を選択する方法を開発した。

しかし、前節に示した MLS の適用分野や求められる品質レベルに基づいた尺度は、RQ3 に示した考慮点と MLS 品質特性の関係では選択できない。そこで、新たに以下の 2 つの研究設問を追加した。本稿でこれらの研究設問について議論する。

RQ4: 測定すべき尺度の絞り込みに必要な要因は何か？

RQ5: 要因に基づいて決定した尺度の優先順位は、尺度の選択において有効であるか？

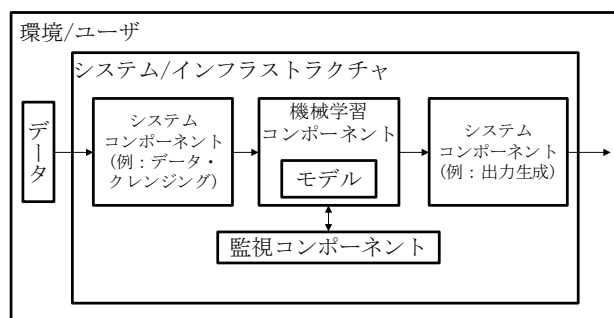


図 1：機械学習システム(MLS)の構成要素

### 3. 関連研究

#### 3.1 MLS の品質問題

MLS は、本質的に従来のソフトウェアシステムとは異なる特性を持つことが示されてきた[14]。これは、ソフトウェア工学に新しい問題を生じている[1]。

MLS としての品質、つまりソフトウェア全体の品質には、従来のソフトウェアの問題に加えてシステムの中に組み込まれた機械学習アルゴリズムの問題も生じている。例えば MLS の主要な応用として注目されている自動運転の分野では、安全性は最も重要な品質特性であるが、品質問題の研究はまだ初期段階である[15]。

MLS に特有な品質問題に性能ドリフト[21]がある。MLS の品質は、入力データと学習時期と運用時期の間のデータ変化に依存する。Chui らは 400 個以上の MLS アプリケーションの分析から 34%以上の学習済モデルが更新を必要とし、さらにその 77%は少なくとも月に 1 回の更新が必要であることを示した[5]。

TensorFlow を用いた MLS のバグの経験的な分析は、TensorFlow API の変更がバグの主な原因であるとしている[26]。これは、Web API でも同様の問題である[25]。フレームワークや Web API の利用方法は、MLS 開発において重要な品質上の課題である。

#### 3.2 MLS の品質モデルと品質保証

ISO25000 シリーズは、ソフトウェア製品の品質モデルとその尺度として広く知られている[10][11][12]。しかし、MLS は従来のソフトウェアシステムとは本質的に異なるので、MLS においては新しい品質モデルと尺度が求められている[13]。Masuda らの調査は、MLS の品質に様々な新しい課題を明らかにした[18]。

これらの背景に基づき MLS のガイドラインとその開発が提案されてきた。QA4AI のガイドラインでは一連の技術がカタログとして提示されている[22]。AI 開発ガイドラインのドラフト版が総務省から公開され、その中では幾つかの品質特性が提示されている[19]。Murphy らは、特定の MLS 品質保証フレームワークを議論しているが、その品質保証の範囲と議論は限定的である[20]。このように、MLS の品質モデルは個別的な定義に留まり、実践的な品質モデルの定義は確立されるに至っていないといえる。

#### 3.3 MLS の品質要求の特定と定義

MLS の要求定義について、Belani らは要求工学的な観点からの MLS 開発のための RE4AI 分類学の概要を提案した[2]。RE4AI 分類学は、RE の活動と AI 関連の 3 つの要素、データ、モデル、システム、の 2 つの軸で構成されている。しかしその分類学の中では品質要求は明示されていない。

Vogelsang らは、データサイエンティストにインタビューし、従来の要求とは異なる機械学習の重要な尺度である、機械学習性能、説明可能性、特定の法的要件を含む新しい品質要求を特定した[24]。彼らは従来の要求工学のプロセスに機械学習の要求を定義する活動を加えることを提案した。要求仕様の中に

において、頑健なニューラルネットワークの開発のための要求は、精度(検証精度)を優先して最適化し、他の性能尺度を追加すべきであることを Hyatt らが指摘した[8]。

## 4. アプローチ

本稿で示す測定すべき尺度の絞り込みに必要な要因、および尺度の選択方法のアプローチは我々の過去の取り組みを発展させたものである。4.1 節から 4.3 節(1)は、過去の取り組みの中で本稿に関わるものを簡潔に示す。本稿で新たに組み込んだ要因や選択方法のアプローチは 4.3 節(2)と 4.4 節に示す。

### 4.1 MLS の要求定義の流れ

本稿で想定する MLS の品質に関わる要求項目の獲得から要求仕様化までのプロセスを図 2 に示す。このプロセスでは、最初に MLS 要求定義の担当者が顧客の説明から MLS に対する要求項目を抽出する。次に、本稿で提案する考慮点一覧を参照し、要求項目に関わりの深い考慮点を抽出する。次に、MLS 品質評価モデル、および考慮点と MLS 品質特性間の対応関係定義(以後、対応関係定義と略す)を参照し、MLS 品質特性や尺度を選択する。最後に、選択した MLS 品質特性や尺度を用いて主要な非機能要求である品質要求の仕様を定義することを表している。

このように要求項目に応じた MLS 品質特性や尺度を用いて要求仕様化を行うことで、MLS に求められる品質の達成度の測定が可能になる。

これまでの我々の取り組みでは、図中の考慮点一覧や MLS 品質評価モデル、対応関係定義、これらを参照しながら行う選択方法評価し、その有効性を確かめた。しかし、2 節にも示したように MLS の適用分野や求められる品質レベルに基づいた尺度の選択には、図中の“選択”方法を新たに提案する必要がある。以降の節では、上記の図中に示された成果物や選択方法の概略を示す。また、選択方法の変更に伴って提案する MLS 品質要求定義方法にも変更を行ったので、その考え方も

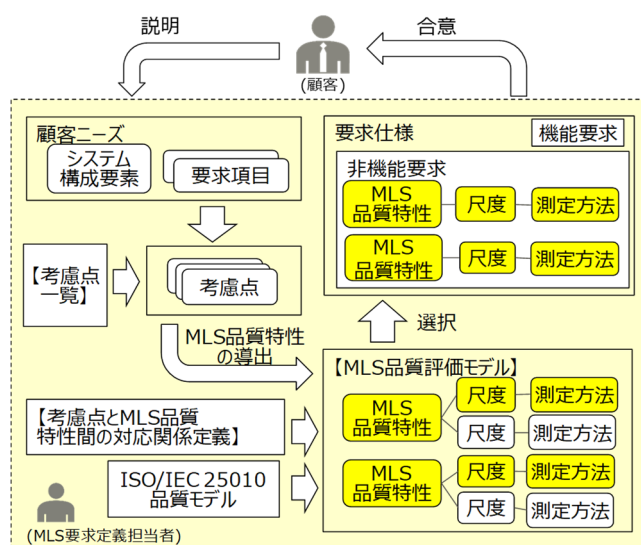


図 2: MLS 品質要求の獲得プロセス

示す。

### 4.2 考慮点

MLS の開発においても MLS の品質要求仕様の定義が必要である。我々は、MLS の要求定義担当者は、MLS の品質要求において着目すべき顧客の関心事を獲得しており、それに基づいて過不足なく要求仕様を定義していると考えた。そこで、本稿では、この MLS での品質要求の定義における顧客の関心事を“考慮点”と呼ぶ。

我々は MLS の専門家と共に 22 個の考慮点を特定した。その一部を表 1 に示す。考慮点は図 1 に示した MLS の概念的な典型的な構成要素である環境/ユーザ、システム/インフラストラクチャ、モデル、データで分類している[7]。考慮点の一つを例として説明する。要素“(a) 環境/ユーザ”に分類されている考慮点“(a-1) スコープへの追従”は、MLS の環境の変化をモニタリングに関わるもので、例えば、MLS を利用するタスクやユーザの変化である。

顧客の説明に基づいて各考慮点の重要度を把握することは、要求定義担当者が MLS の品質要求を定義する際の支援になると考える。その理由は MLS 向けの品質要求の獲得の際に考慮漏れを防ぐことが可能になるからである。

以降本稿では品質要求への顧客の関心事である“考慮点”を用いて議論を進める。

### 4.3 MLS 品質評価モデルに対する必須度の導入

#### (1) MLS 品質評価モデル

ソフトウェア製品の品質モデルの国際標準として ISO/IEC

表 1: 考慮点一覧 (一部)

| 要素                          | 考慮点の名称               | 考慮すべきこと                                                                                            |
|-----------------------------|----------------------|----------------------------------------------------------------------------------------------------|
| (a) 環境 / ユーザ                | (a-1) スコープへの追従       | ・ 環境やその他条件がシステムが期待する範囲にあるか (スコープの監視)                                                               |
| (b) システム / インフラストラクチャ       | (b-1) 訓練効率性          | ・ モデル学習時の計算時間やメモリの使用率                                                                              |
|                             | (b-2) 実行性能の効率性       | ・ 訓練済モデル実行時の計算時間やメモリの使用率                                                                           |
|                             | (b-3) 出力監視アルゴリズムの効率性 | ・ 監視アルゴリズムによって、出力値の誤まり (偽陽性/偽陰性) をどの程度検出できるか                                                       |
|                             | ...                  | ...                                                                                                |
| (c) モデル                     | (c-1) モデル適切性         | ・ モデルの種類がタスク (分類等) や必要とされるデータ (型、範囲、量等) の適切かどうか                                                    |
|                             | (c-3) モデルの訓練時の性能     | ・ テストデータの性能 (精度、再現性など)                                                                             |
|                             | (c-4) 学習済モデルの実行時の性能  | ・ 実行時データの性能 (精度、再現性など)                                                                             |
|                             | ...                  | ...                                                                                                |
| (d) データ (訓練 / テスト / 実行時データ) | (d-1) 代表性            | ・ データが想定データを代表しているか<br>- 想定するデータセットの値域 (最小値/最大値) とデータセットの値域を比較<br>- 想定するデータセットと実際のデータセットの分布の類似性の検証 |
|                             | (d-2) 均衡性            | ・ データ内のカテゴリ間のバランス、カテゴリ毎のサンプルの数や、カテゴリ内のサンプルの確率分布の比較                                                 |
|                             | (d-3) 適時性            | ・ 訓練 / テスト用のデータ収集時刻と実行時データの時間の差<br>・ データの適時性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                         |
|                             | ...                  | ...                                                                                                |

表 2: MLS 品質評価モデル (一部)

| SQuaRE品質 |       | 新たに定義した品質                                                           | 尺度                                                                            |
|----------|-------|---------------------------------------------------------------------|-------------------------------------------------------------------------------|
| 品質特性     | 品質副特性 | MLS品質特性                                                             |                                                                               |
| 機能適合性    | 機能正確性 | ・学習済モデルの精度正確性<br>・MLSの精度正確性<br>・MLS出力データの適合性 …                      | ・モデル正解率の達成率<br>・MLS正解率の達成率<br>・MLS出力データ平均値の充足率 …                              |
|          | 機能完全性 | ・訓練データの適合性 …                                                        | ・モデル出力データ平均値の充足率 …                                                            |
| 信頼性      | 成熟性   | ・入力データの傾向変化に対するロバストネス<br>・入力データの品質維持手段の適切性<br>・MLS出力データ品質維持手段の適切性 … | ・実行時のモデル精度の適切度<br>・実行時のモデル正解率の達成率<br>・入力データ品質測定手段適切度<br>・MLS出力データ品質測定手段の適切度 … |
|          | 可用性   | ・稼働維持の適切性 …                                                         | ・リソース拡張の容易度 …                                                                 |
| 性能効率性    | 時間効率性 | ・時間性能の十分性 …                                                         | ・ターンアラウンドタイム …                                                                |
|          | 容量効率性 | ・リソース使用率の適切性…                                                       | ・CPU利用率 …                                                                     |

25010 が広く認知されている[11]. 産業界の MLS 専門家によって MLS の品質を議論した資料,あるいは MLS の品質保証の目的で作成され,様々な MLS に適用可能な汎用的なチェック項目等から MLS に必要な品質特性を抽出し,ISO25010 の品質特性の副特性に対応づけて具体化した提案がある[4][22][9][19].これらに基づき,MLS 品質特性を定義した.さらに,各 MLS 品質特性に対してその達成度を示す尺度を新たに定義した.これらの MLS 品質特性と尺度を合わせて,MLS 品質評価モデルと呼ぶ.このモデルの一部を表 2 に示す.

## (2) 尺度の属性必須度の導入

従来のソフトウェア開発や運用では,機能要求や非機能要求を定義する際に,開発に関する制約として QCD(品質,費用,納期)も決定し,これを遵守することが求められる. MLS で測定すべき尺度を選択する場合も同様に,要求定義の担当者が QCD を考慮しながら測定する尺度を選択していると考えた.

本稿では尺度の選択には QCD の中でも特に Q(品質)に注目する. MLS を組込んだ企業情報システムは従来の企業情報システム同様,品質レベルの達成が重要であるからである.

一般に,品質要求に対する要求レベルが低ければ,限定した少数の尺度で MLS の品質を測定する.逆に要求レベルが高ければ,様々な側面で MLS の品質を測定するので,測定する尺度が増えると考えられる.しかし,前述したように MLS 品質評価モデルの尺度は,様々な適用分野の事例や,汎用的なチェック項目を元に作成したものである.そのため,様々な品質レベルに対応するための尺度が混在している.

我々は企業情報システムとしての尺度選択方法を確立するため,尺度の測定が企業情報システムに求められる品質レベルの達成に必要である程度を表す“必須度”という属性を導入する.さらに,本稿では必須度を活用度と影響防止度という 2 つの属性に分けて評価する.活用度は尺度の適用範囲を意味し,影響防止度は MLS が被る可能性のある損害を防止する効果の尺度である.

必須度が高い尺度は,より多くの場面で使用できたり,MLS への損害を防止する効果が高いものである.逆に必須度が低い尺度は,限定された場面のみでしか使用できなかったり,損害防止効果が低いものである.

このような必須度の特性から,品質要求レベルが低い場合は必要な尺度が限定され,尺度の必須度は高くなると考えられる.逆に,品質要求レベルが高い場合は,様々なリスク低減をするために必須度が高い尺度だけでなく,必須度が低い尺度も使われると考えた.

MLS 品質評価モデルの定義のために,その構成要素や構成要素間の関係を示したメタモデルを定義した.メタモデルを図 3 に示す.このメタモデルは Web API の品質評価用のメタモデル[25]を元に,我々の取り組みで導入した考慮点と必須度を導入し,拡張したものである.この図に示すように,考慮点は,評価対象システムと品質特性間の関連クラスである.また,品質特性と尺度間にも関連がある.実際の品質特異性と尺度の関係は,表 2 に示した MLS 品質評価モデルで定義している.このメタモデルは,評価対象システムにとって重要度の高い考慮点を特定することで,考慮点に関わる品質特性や尺度を選択できることを示している.選択された尺度の属性“必須度”を用いることで,顧客が重視する品質要求と品質レベルに応じた尺度の選択が可能となることを示している.

次の 4.4 節で考慮点から尺度を選択するアプローチを示し,5 節で必須度の評価に用いる 2 要因と詳細な算出方法を示す.

## 4.4 考慮点に基づく品質特性や尺度の選択

MLS 品質特性や尺度の中から,MLS の適用分野や求められる品質レベルに基づいて品質特性や尺度の選択を可能とするために,本稿では 2 ステップの選択方法を提案する.

第 1 ステップは,個々の MLS での考慮点の重要度を元に尺度を抽出することが目的である.第 2 ステップは尺度の必須度

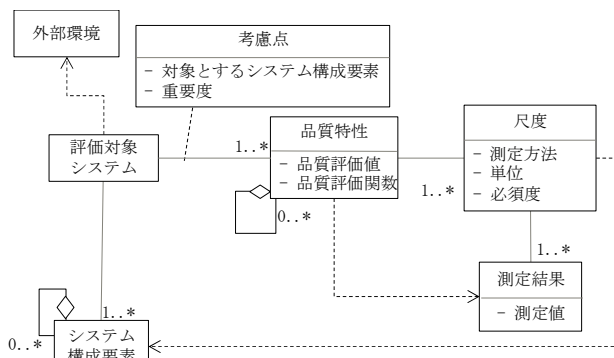


図 3: MLS 品質評価のメタモデル

表 3：考慮点と MLS 品質特性間の対応関係定義

| MLS品質特性                 | 考慮点       |     |                       |     |
|-------------------------|-----------|-----|-----------------------|-----|
|                         | データ       |     | 環境                    | ... |
| MLS品質特性                 | (d-1) 代表性 | ... | (a-1) ス<br>コープ追従<br>性 | ... |
| 学習済みモデルの精度正<br>確性       | *         |     |                       |     |
| MLSの精度正確性               |           |     | *                     |     |
| MLS出力データの適合性            |           |     | *                     |     |
| 入力データの変化に対す<br>るロバストネス  | *         |     |                       |     |
| 入力データの品質維持<br>手段の適切性    | *         |     |                       |     |
| MLS出力データの品質維<br>持手段の適切性 |           |     | *                     |     |

を利用して、開発対象の MLS に求められている品質レベルに  
 合致した尺度を抽出することが目的である。

各ステップを説明する。第 1 ステップでは、開発対象の MLS  
 において高重要度の考慮点と表 3 に示した対応関係の定義、  
 MLS 品質評価モデルを参照し、MLS 品質特性とそれに関連  
 する尺度を選択する。第 2 ステップでは、尺度の“必須度”を参  
 照して、要求定義の担当者が測定する尺度を決定する。

我々は MLS 開発の専門家への調査から考慮点と MLS 品  
 質特性の関係は事前に定義できると想定している。しかし、品  
 質レベルによって事前に選択することが困難な場合もあると想  
 定し、要求定義の担当者が選択を行うこととした。

## 5. 必須度の算出方法

### 5.1 必須度の考え方

本節では尺度の必須度の考え方を示す。前述したように必  
 須度は尺度の測定が企業情報システムに求められる品質レベ  
 ルの達成に必要である程度を表す。尺度の必須度を決定する  
 要因として、情報システムのリスクの深刻度評価(Severity  
 Rating)(ISO3100 ではリスクレベル(Levels of Risk)[6]と呼ぶ)の  
 概念をヒントとして、MLS の活用度とその影響防止度を定義し  
 た。

活用度は活用範囲の広さを意味し、測定結果を活用するア

クタとそのユースケースに基づいて算出する。これはリスク深刻  
 度評価でのリスクの発生頻度に相当する概念を適用したもので  
 ある。影響防止度は、技術負債を防止する度合いを意味し、運  
 用時に尺度を測定しないことによって生じる可能性のあるトラブ  
 ルが影響を及ぼすアクタとトラブルの対処に必要となる作業量  
 で算出する値である。これはリスク深刻度評価でのインパクト  
 に相当する概念を適用したものである。

本稿ではユースケースとして“開発”、“テスト”、“運用”の 3 つ  
 を想定する。また、アクタとして“顧客(利用部門)”、“顧客(情  
 シス部門)”、“AI 専門家”、“プロジェクトマネージャ”、“運用者”を  
 想定する。開発とテストに関わるアクタは、“顧客(利用部門)”、  
 “顧客(情シス部門)”、“AI 専門家”、“プロジェクトマネージャ”の  
 4 つ、運用に関わるアクタは“顧客(利用部門)”、“顧客(情シス  
 部門)”、“運用者”の 3 つとした。

図 4 は、ある尺度  $m$  の必須度、活用度、影響防止度の算出  
 方法の概要を図示したものである。以降の節でその詳細を示す。  
 なお、この図では全て同一の尺度  $m$  の必須度、活用度、影響  
 防止度を示しているの、図中では  $m$  の表記を省略している。

### 5.2 必須度の算出方法

まず必須度の算出に用いる変数や係数を示す。

$V(m)$ : 尺度  $m$  の必須度を算出する関数。

$A(m)$ : 尺度  $m$  の活用度を示す値。

$B(m)$ : 尺度  $m$  の影響防止度を示す値。

$w_a$ : 活用度  $A(m)$  の重み。

$w_b$ : 影響防止度  $B(m)$  の重み。

ここで、 $w_a + w_b = 1$  である。

必須度の算出は、図 4 に示すように、活用度と影響防止度  
 に重み  $w_a$  と  $w_b$  をつけて加算する。これを式(1)に示す。

$$V(m) = w_a A(m) + w_b B(m) \quad (1)$$

### 5.3 活用度の算出方法

活用度の算出に用いる関数と係数を示す。

$C(m, i, s_i)$ : 開発においてアクタ  $i$  による尺度  $m$  の使用状況  
 (使用する／しない)を示す引数  $s_i$  に応じて値が決まる関数。関  
 数の値はアクタ  $i$  が尺度  $m$  を使用しない場合は 0 である。使用  
 する場合は 0 より大きく 1 以下の値で、かつ全アクタが尺度  $m$   
 を利用した場合に、関数  $C()$  の値の総和が 1 となるよう  
 に設定する。

$D(m, j, s_j)$ : テストにおいてアクタ  $j$  による尺度  $m$  の使  
 用状況を示す引数  $s_j$  に応じて値を返す関数。関数の  
 値は、関数  $C()$  と同様である。

$E(m, k, s_k)$ : 運用においてアクタ  $k$  による尺度  $m$  の使  
 用状況を示す引数  $s_k$  に応じて値が決まる関数。関数  
 の値は、関数  $C()$  と同様である。

$w_c$ : 開発で尺度を利用することの重み。

$w_d$ : テストで尺度を利用することの重み。

$w_e$ : 運用で尺度を利用することの重み。

$N_c, N_d, N_e$ : 開発、テスト、運用に関わるアクタの数を  
 表す。

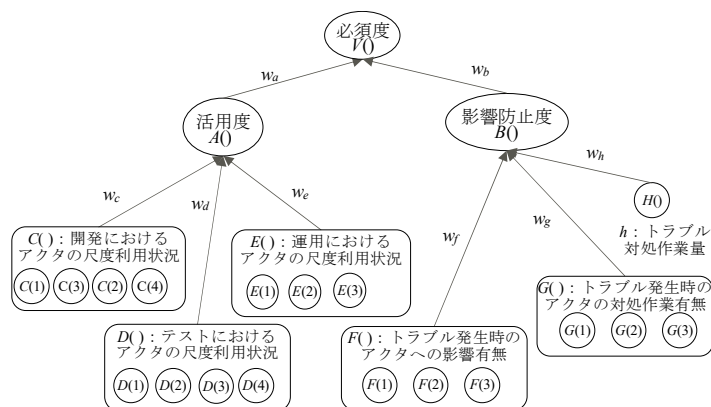


図 4: 必須度算出方法の概要

ここで  $w_c + w_d + w_e = 1$  である。アクタ  $i$  とアクタ  $j$  は、それぞれ“顧客(利用部門)”, “顧客(情シス部門)”, “AI 専門家”, “プロジェクトマネージャ”を示す識別子である。また、アクタ  $k$  は、それぞれ“顧客(利用部門)”, “顧客(情シス部門)”, “運用者”を示す識別子である。

次に活用度の算出式を式(2)に示す。

$$A(m) = w_c \sum_{i=1}^{N_c} C(m, i, s_i) + w_d \sum_{j=1}^{N_d} D(m, j, s_j) + w_e \sum_{k=1}^{N_e} E(m, k, s_k) \quad (2)$$

活用度の算出方法を図 4 を用いて説明する。図中、各関数の尺度  $m$  以外にも各アクタによる使用状況を省略した。使用状況が異なっても、アクタ数分の関数値が計算するので、図の簡略化のために省略した。

$C()$  の枠内に示す円は、 $C(m, i, s_i)$  の値を示し、この枠で式(2)右辺の第一項の総和の部分を表している。枠内の円で示した  $C(i, m)$  の値の総和をとることを表している。図中の  $D()$   $E()$  の枠も同様に右辺第二項、第三項の総和を表している。これらの値にそれぞれ重み  $w_c, w_d, w_e$  を掛けて加算したものが活用度  $A(m)$  である。我々は開発に関わるアクタは(1)節に示したように 4 種類だと想定している。そこで、関数  $C()$  の枠内にはアクタ数と同数の円を配置した。関数  $D()$  と  $E()$  の枠内に配置している円の数も、関数  $C()$  と同様にテストや運用時に想定しているアクタ数である。

#### 5.4 影響防止度の算出方法

影響防止度の算出に用いる関数、値と係数を示す。

$F(m, i, s_i)$ : 尺度  $m$  を測定しなかったことで発生したトラブルによって、アクタ  $i$  に影響が及ぶか否かを示す引数  $s_i$  に応じて値が決まる関数。関数の値は、影響が及ばない場合は 0 である。影響が及ぶ場合は、0 より大きく 1 以下の値で、かつ全アクタに影響が及ぶ場合に関数  $F()$  の値の総和は 1 となるように設定する。

$G(m, j, s_j)$ : 尺度  $m$  を測定しなかったことで発生したトラブルによって、アクタ  $j$  がトラブルの対処作業を実施するか否かを示す引数  $s_j$  に応じて値が決まる関数。関数の値は、対処作業を実施する場合は 1、しない場合は 0。対処作業を実施する場合は 0 より大きく 1 以下の値で、かつ全アクタが対処作業を実施した場合に関数  $G()$  の値の総和は 1 となるように設定する。

$H(m)$ : 尺度  $m$  を測定しないことによって発生したトラブルへの対処作業量を示す値で 0 から 1 の実数値をとる。作業量が多いほど 1 に近い値、作業量が小さいほど 0 に近い値をと

る。

$w_f$ : 尺度を測定しないことで発生したトラブルの影響に対する重み。

$w_g$ : トラブル対処作業に対する重み。

$w_h$ : トラブルへの対処作業量に対する重み。

ここで  $w_f + w_g + w_h = 1$  である。また、アクタ  $i$  とアクタ  $j$  の値 1 から 3 は、それぞれ“顧客(利用部門)”, “顧客(情シス部門)”, “運用者”を示す。

次に影響防止度  $b$  の算出式を式(3)に示す。

$$B(m) = w_f \sum_{i=1}^{N_f} F(m, i, s_i) + w_g \sum_{j=1}^{N_g} G(m, j, s_j) + w_h H(m) \quad (3)$$

影響防止度の算出方法を図 4 を用いて説明する。図中、各関数の尺度  $m$  以外にも各アクタへの影響や対処作業の有無を示す引数を省略した。それらが異なっても、アクタ数分の関数値が計算するので、図の簡略化のために省略した。

$F()$  と  $G()$   $H()$  の枠内に示す円は  $F(m, i, s_i)$  と  $G(m, j, s_j)$  値を示し、この枠で式(3)右辺の第一項と第二項との総和を表している。計算方法は活用度と同様である。これらと  $H(m)$  の値にそれぞれ重み  $w_c, w_g, w_h$  を掛けて加算したものが影響防止度  $B(m)$  である。我々は影響防止度に関わるアクタは 3 種類を想定している。そこで、図 4 の  $F()$  と  $G()$  の枠内には、想定するアクタ数と同数の円を配置している。

#### 5.5 必須度の算出例

本節では、必須度、活用度、影響防止度の算出例を表 4 を用いて示す。図中のピンク色で示した列は活用度、青い列は影響防止度の算出に用いる。右端の列は算出された必須度の値を示す。評価値の行は、その上の行のアクタが尺度を利用したり、トラブルの影響を受ける場合等の関数  $C()$  から  $H()$  の値を示している。表の左列の評価列から下の各行は尺度を示している。表中の行と列の交点には尺度を使用したりトラブルの影響を受ける場合等に“\*”を設定する。一方、尺度を使用しない場合やトラブルの影響を受けない場合には交点に“-”を設定する。このような表を用いると、尺度”学習中のモデル正解率の達成率”の場合、関数  $C()$  (学習中のモデル正解率の達成率、顧客(利用部門)、使用する)の値は 0.33 であることが分かる。一方、 $E()$  (学習中のモデル正解率の達成率、顧客(利用部門)、使用しない)の値は、0 であることが分かる。これらの関数値と以下の重みを用いて、活用度、影響防止度、必須度を算出した。

$$w_d = w_b = 0.5, w_c = w_d = w_e = 1/3, w_f = w_g = w_h = 1/3$$

表 4: 必須度の算出例

| 評価値<br>尺度名      | 設計時の尺度利用者    |               |           |     | テスト時の尺度利用者   |               |           |     | 運用時の尺度利用者    |               |      | 活用度  | トラブル影響範囲     |               |      | トラブル<br>対処作業<br>の大きさ<br>大:1<br>中:0.8<br>小:0.4 | 影響防止度 | 必須度  |
|-----------------|--------------|---------------|-----------|-----|--------------|---------------|-----------|-----|--------------|---------------|------|------|--------------|---------------|------|-----------------------------------------------|-------|------|
|                 | 顧客<br>(利用部門) | 顧客<br>(情シス部門) | AI<br>専門家 | PM  | 顧客<br>(利用部門) | 顧客<br>(情シス部門) | AI<br>専門家 | PM  | 顧客<br>(利用部門) | 顧客<br>(情シス部門) | 運用者  |      | 顧客<br>(利用部門) | 顧客<br>(情シス部門) | 運用者  |                                               |       |      |
|                 | 0.33         | 0.27          | 0.2       | 0.2 | 0.33         | 0.27          | 0.2       | 0.2 | 0.33         | 0.27          | 0.18 |      | 0.45         | 0.36          | 0.18 |                                               |       |      |
| 学習中のモデル正解率の達成率  | *            | *             | *         | *   | *            | *             | *         | *   | -            | -             | -    | 0.67 | *            | -             | -    | 大                                             | 0.82  | 0.74 |
| MLS出力データ平均値の充足率 | *            | *             | *         | *   | *            | *             | *         | *   | -            | *             | *    | 0.85 | -            | *             | *    | 大                                             | 0.70  | 0.77 |

## 6. 尺度の選択方法

### 6.1 選択方法

本節では、開発対象の MLS で測定すべき尺度を考慮点に基づいて、尺度を選択する方法を述べる。この選択方法は 4.4 節に述べたように 2 ステップで行う。2 ステップの概要を図 5 に示す。図 5 中、考慮点は五角形、品質特性を円、尺度を正方形で示している。各図形や図形間の線に設定された添え字付きの  $w$  は、算出に用いる点数を表している。以降では各ステップを説明する。

#### (1) 第 1 ステップ

本ステップでは、開発対象の MLS において高重要度の考慮点に基づき、関連する尺度をすべて抽出する。具体的には以下のステップで行う。

1-1) 高重要度の考慮点の抽出: 表 1 から開発対象の MLS で高重要度の考慮点を抽出する。

1-2) 考慮点への点数づけ: 前ステップで抽出した高重要度の考慮点それぞれに 0 より大きい値の点数  $w_{si}$ 、ただし  $i$  は  $0 \leq i \leq N$  の整数を与え、低重要度の考慮点には 0 を設定する。

1-3) MLS 品質特性の抽出: 高重要度の考慮点に対し、表 3 を用いて関連する MLS 品質特性を抽出する。

1-4) MLS 品質特性の点数づけ: 抽出された各 MLS 品質特性に対し、関連付けられた全考慮点の点数の合計値  $w_{qj}$ 、ただし  $j$  は  $0 \leq j \leq M$  の整数、を算出する。

1-5) MLS 品質特性の点数によるフィルタリング: MLS 品質特性の点数  $w_{qj}$  が閾値以上の MLS 品質特性を抽出する。

1-6) 尺度の抽出: 前ステップ 1-5) で抽出された各 MLS 品質特性に対して表 2 を利用して関連する尺度を抽出し、出力とする。

#### (2) 第 2 ステップ

要求される品質レベルに基づいた尺度の選択を支援するために尺度の“必須度”を用いる。第 1 ステップで出力された尺度を必須度の降順に並べたものを第 2 ステップの出力とする。

### 6.2 想定する選択手順

本稿では、尺度の選択を実施するのは、機械学習の知識を持った要求定義の担当者だと想定する。その担当者が尺度を選択するプロセスを示す。

#### 1) 準備

1-1) 考慮点のチェックリスト作成: 考慮点一覧の考慮すべきこと欄の内容を顧客に分かりやすく表現し、高重要度の考慮点を抽出するためのチェックリストを作成する。

1-2) 品質レベルの想定: 開発対象の MLS に要求される品質レベルを想定する。

1-3) 開発/運用費用の把握: 開発対象の MLS で想定される開発環境で尺度を測定するためのプログラムの開発費用と測定作業や測定値を元に判断する作業に要する費用を見積る。

#### 2) 顧客へのヒアリング

2-1) チェックリストの確認: 各考慮点の重要度を顧客からヒアリングし、高重要度の考慮点を決定する。

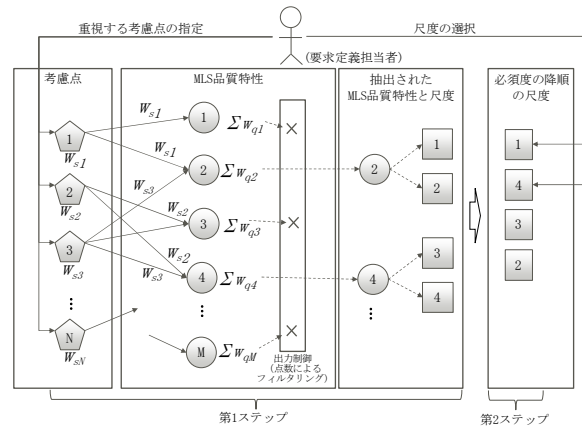


図 5: 尺度選択方法の概要

2-2) 品質レベルの確認: 1-2) で想定した品質レベルが妥当かを顧客に確認する。必要に応じて変更する。

2-3) 開発/運用予算の確認: 顧客に MLS の開発費用や運用中に測定に要する費用の見積りを示し、合意する。

#### 3) 尺度の選択

3-1) 尺度リストの作成: ヒアリングで把握した高重要度の考慮点を用い、6.1 節の第 1 ステップの方法で尺度を抽出する。さらに、第 2 ステップの方法で抽出した尺度を必須度の降順に並べ、尺度リストを作成する。

3-2) 尺度リストからの選択: 尺度のリストから、開発対象の MLS に求められる品質レベルや尺度の開発や運用のコストを考慮して、測定する尺度を選択する。

3-3) 顧客との合意: 顧客に対して選択した尺度の説明や目標値の合意を行う。

## 7. 評価

本論文に示した尺度選択方法のベースである“必須度”の有効性を評価するためにケーススタディを実行した。

### 7.1 評価項目

必須度が有効であるとは、以下の 2 点が示せることであると考える。

- (1) 評価項目 1: 必須度の高い尺度が、必須度の低い尺度よりもより多く測定が実施されていること。
- (2) 評価項目 2: 各尺度の活用度と影響防止の値が正しいこと。

本稿では必須度が高い尺度の方がより多く測定が実施されると想定している。そのためには、1 点目に示すように、必須度の高低と測定に実施/未実施が一致していることを確認すれば良いと考えた。2 点目は必須度の有効性を補強するためのもので、必須度算出に用いた活用度と影響防止の値の妥当性を示すものである。

この評価に用いる尺度に、技術的な理由で測定が困難な尺度、あるいは MLS の品質評価に効果がない尺度が含まれていると、その尺度は要求定義担当者には選ばれることはないはずである。従って評価の前提として、これらの理由で選ばれない尺度がないことの確認が必要である。

表 5：評価用の尺度リスト

| No. | 名称                | 概要                                                                     | 必須度  | 活用度  | 影響防止度 |
|-----|-------------------|------------------------------------------------------------------------|------|------|-------|
| 1   | 訓練データ数            | 訓練データの数                                                                | 0.83 | 0.67 | 1.00  |
| 2   | 訓練データ標準偏差の充足率     | 訓練データの各次元の目標とする標準偏差に対する実際の訓練データの各次元の標準偏差の割合                            | 0.83 | 0.67 | 1.00  |
| 3   | 訓練データ値域の充足率       | 訓練データの各次元の目標とする最大値と最小値に対する実際の訓練データの各次元の最大値と最小値の割合                      | 0.83 | 0.67 | 1.00  |
| 4   | 運用中のモデル正解率の達成率    | 学習済みモデルの目標とする正解率に対する運用環境での実測の正解率の割合                                    | 0.81 | 0.80 | 0.82  |
| 5   | モデルF値の達成率         | モデルのF値の目標値に対する、実測のF値の割合                                                | 0.78 | 0.80 | 0.79  |
| 6   | モデルAUCの達成率        | モデルのAUCの目標値に対する、実測のAUCの割合                                              | 0.78 | 0.80 | 0.79  |
| 7   | 訓練データプライバシー法適用項目数 | プライバシー法に則った扱いが必要なデータの項目数                                               | 0.76 | 0.67 | 0.85  |
| 8   | 入力データ平均値の充足率      | 入力データの各次元の目標とする平均値に対する実際の入力データの各次元の平均値の割合                              | 0.75 | 0.89 | 0.62  |
| 9   | 入力データ標準偏差の充足率     | 入力データの各次元の目標とする標準偏差に対する実際の入力データの各次元の標準偏差の割合                            | 0.75 | 0.89 | 0.62  |
| 10  | 入力データ値域の充足率       | 入力データの各次元の目標とする最大値と最小値に対する実際の入力データの各次元の最大値と最小値の割合                      | 0.75 | 0.89 | 0.62  |
| 11  | 学習中のモデル正解率の達成率    | 学習済みモデルの目標とする正解率に対する開発/テスト環境での実測の正解率の割合                                | 0.74 | 0.67 | 0.82  |
| 12  | 訓練誤差の充足率          | 出力データの各次元の損失の目標値に対する訓練データを用いた場合の出力データの各次元の損失値の割合。損失関数として、平均二乗誤差を使用した場合 | 0.73 | 0.67 | 0.79  |
| 13  | 訓練データの説明変数の割合     | 訓練データのデータ項目数                                                           | 0.73 | 0.67 | 0.79  |

## 7.2 評価方法

本評価では、社内で開発された3つのMLSを用い、この3つのMLSに関わった社内のMLSの専門家を被験者として実施した。

評価に用いた尺度リストの一部を表5に示す。このリストは、3つのMLSで高重要度であった考慮点を用いて、26個の尺度を選択し、必須度の降順に並べたものである。表5には、必須度の降順に並べた尺度、必須度の算出に用いた活用度、影響防止度を示している。この尺度リストは、6.2節のステップ3-1)までを実施した場合のアウトプットである。その次のステップ3-2)は要求定義担当者による尺度の選択である。顧客から要求された品質レベルに基づいて尺度を選択する作業の支援は、尺度リストにおいて必須度による尺度の優先順位付けが正しく行われていることで可能になると考える。そこで、本評価では、6.2節のステップ3-1)のアウトプットである尺度リストを用いて、前節に示した2つの評価項目と評価の前提の確認を実施する。

評価の前提である、MLS品質評価モデルの妥当性を確認するため、尺度リスト中の各尺度に対して、過去の3つのMLSでの測定実施の有無と、もし利用していない場合はその理由を質問した。理由は選択肢式で問い、選択肢は、1)技術面で困難、2)コスト面で困難、3)対象のMLSは該当しない、4)効果がない、5)その他とした。選択肢1)と4)が選ばれた場合、MLS品質評価モデルには実現が困難な尺度や役立たないと判断された尺度が含まれていることを表す。従って、この2つの選択肢が選ばれた場合は、評価の前提となるMLS品質評価モデルの妥当性が主張できない。

この2つ以外の選択肢を設けた理由を説明する。我々はQCDのQ(品質)に注目して必須度を決定した。しかし実際の開発現場では、C(コスト)やD(納期)も鑑みて実施する／しないが判断されるはずである。これらの理由で実施されなかったことが分かるように設けたのが選択肢2)である。選択肢3)は尺度選択

方法のステップ1の妥当性を確認するために設けた。

評価項目1を確認するため、尺度リストに含まれる尺度を上位13個と下位13個の2グループに分けた。上位グループの尺度の方が下位グループの尺度よりもより多く測定が実施されていれば、必須度が有効だと判断できる。評価に用いた尺度数が26個のみで、かつ、被験者数や評価に用いたMLS数が少ないことから、尺度を細かいグループに分けて分析することに意味はないと考え、上位と下位の2グループのみに分けた。

評価項目2を確認するため、尺度リスト中の26個の中で、実際のMLSで使用された尺度に対し、2つの要因、活用度と影響防止度のどちらを重視して測定の実施を決定したかを被験者に質問する。2つの要因のうち

値の大きい要因と、被験者の回答が一致すれば、活用度と影響防止の値は妥当であると判断できると考えた。本来、値の妥当性を示すには値の大小だけでなく、値の差も議論すべきである。しかし、今回の評価では被験者数や評価に用いたMLSの数が少ないことから、詳細な分析は困難であると考え、値の大小のみで妥当性を判断する。

## 7.3 結果

被験者の回答を表6、図6、図7に示す。

表6は尺度の測定の実施／未実施の割合、および未実施の理由別の割合を示したもの、図6は必須度の上位グループと下位グループ間の尺度の実施率を比較したものである。また、図7は必須度の値の算出に用いた2要因の大小関係と被験者が尺度の測定実施の判断をする際に重視した要因の一致割合を示したものである。

表6からは全体としては実施と未実施がおおよそ半々であった。また、この表には未実施理由の各選択肢の割合も記載している。選択肢“1)技術的に困難”と“4)効果がない”の回答は0%であった。これらの回答が0%であることは、MLS品質評価モデルには測定実施が困難な尺度や有用でない尺度は含まれていないことを表している。従って、今回のケーススタディに関わる範囲ではあるが、MLS品質評価モデルで定義した品質特性やその尺度は適切であったと判断する。選択肢“3)対象のMLSでは該当しない”が7件と最も割合が高い。この選択肢が選ばれた尺度は、例えば個人情報の取り扱いの適切さに関わる尺度であった。しかし、本ケーススタディで扱ったMLSでは個人情報に関わるデータを扱っていなかったため、この選択肢

表 6: 尺度の実施/未実施状況

| 実施  | 未実施       |            |                 |          |        |
|-----|-----------|------------|-----------------|----------|--------|
| 67% | 33%       |            |                 |          |        |
|     | 1) 技術面で困難 | 2) コスト面で困難 | 3) 対象のMLSは該当しない | 4) 効果がない | 5) その他 |
|     | 0%        | 8%         | 42%             | 0%       | 42%    |
|     |           |            |                 |          | 回答無し   |
|     |           |            |                 |          | 8%     |

3)が選ばれていた。これは、MLS で扱うデータの特徴に関わる考慮点による選択である第 1 ステップで除外すべきものと考ええる。従って、考慮点の内容を見直しが必要と判断した。この例以外にも類似の理由で選択肢 3)が選ばれていた。

図 6 からは必須度が高い尺度の方が低い尺度と比べあ倍多く実施されていることが分かる。このことから、必須度の高低が、実際の MLS で尺度測定の実施有無の判断と整合する傾向はあり、必須度を用いて測定すべき尺度を提案することはある程度は有効だと判断した。

図 7 では、被験者が尺度の測定を実施する／しないの判断に用いた要因が算出した各要因の値の大きい方と一致している場合が不一致の場合の 9.5 倍であることが分かる。このことから、必須度の算出に用いた活用度と影響防止の値は妥当であると判断した。

活用度と影響防止度の値と尺度の測定を実施する／しないの判断とが一致するにも関わらず、必須度の上位グループと下位グループの実施割合の差が小さい理由を考察する。尺度には例えば学習済みモデルの精度を表す典型的な尺度は正解率である。その他、AUC や F 値などがある。我々が提案した MLS 品質評価モデルには、同じ目的の複数の尺度を含んでいる。それらは同一目的であるので、必須度は同程度としていた。そのため、同一目的の複数の尺度が同じグループに含まれている。しかし被験者が利用した尺度は、同じ目的の中の一部のみであった。これは開発コストや運用コストを考えると妥当な結果である。今後、MLS に要求品質レベルが低い場合は同一目的の尺度の中から一部を選ぶ等の工夫が必要だと思われる。

各要因の重視する要因として選択された割合を図 8 に示す。この図からは、影響防止度選択されている割合が高いことが分かる。このことから MLS のカットオーバー後の問題防止の観点で多くの尺度を測定していることが推察される。一方で影響防止度として考慮する範囲が広すぎる可能性もある。影響防止という観点でどのような要因が必要であるかを検討する必要があると考える。

## 8. 議論

### 8.1 RQ4：測定すべき尺度の絞り込みに必要な要因は何か？

5 節に示したように尺度の絞り込みを容易にするため、二つの提案を行った。一つめは尺度の活用度と影響防止度から算出される必須度である。もう一つは算出された必須度を用いて、顧客が求める品質レベルに適した尺度を選択する方法である。この必須度の算出が可能であること、重視する考慮点に尺度の選択方法を用いた尺度の選択が可能であることを示した。

### 8.2 RQ5：要因に基づいて決定した尺度の優先順位は、尺度の選択において有効であるか？

7 節に示したケーススタディでは、提案した活用度と影響防止度という 2 つの要因に基づいて算出した必須度で優先順位付けをした尺度を用いて、必須度の高さと尺度の測定状況の

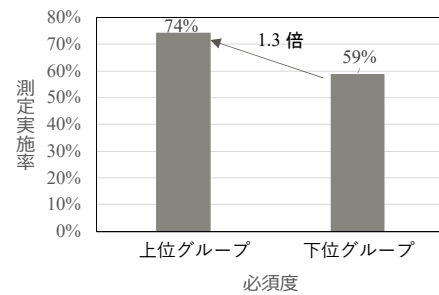


図 6：グループ別の測定実施状況

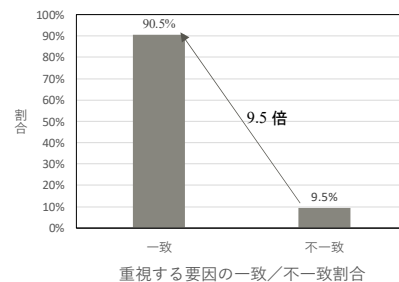


図 7：重視する要因の一致／不一致割合

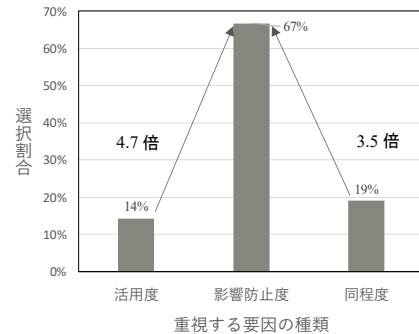


図 8：重視する要因の選択割合

整合性を評価した。その結果、優先順位付けの上位の尺度は、下位の尺度と比較してより多く測定される傾向があった。従って、RQ5 で提案した要因とそれに基づく尺度の優先順位付けは概ね有効であると判断する。但し、特に影響防止度に関してはより詳細化する必要があると考える。

### 8.3 関連研究との比較

MLS の企業情報システムへの導入が広がると共に、その品質定義は実務において重要な課題となっている[5][17][18]。しかし、現在の MLS の品質に関する研究は品質モデルのフレームワークに焦点が当たっており、実務において適用できる方法の提案までには至っていない。本研究では、MLS 品質要求モデルの提案[21]を基礎とし、MLS 開発者からのフィードバックを元に、実務での適用を目的に MLS 品質要求とその評価尺度の具体的な導出方法を提案し、ケーススタディで有効性を確認した点で意義があると言える。

## 9. まとめと今後の課題

本論文では、企業情報システム向け機械学習システムの品質を測定する尺度の選択方法を提案し、MLS の開発時や運用時に発生しうる品質問題の回避方法を述べた。

これまでの我々の取り組みでは、MLS の経験豊富な開発者との議論に基づき、MLS で検討すべき要求仕様に関連する 22 の考慮点と MLS の品質特性や尺度を定義していた。これらの取り組みを元に、本稿では、尺度には尺度の活用範囲の広さやトラブルを未然に防ぐ効果に基づいて算出する必須度という尺度の新たな属性を提案した。さらに考慮点や尺度、尺度の必須度を用いて、要求定義の担当者がMLSの品質確保・維持のために測定すべき尺度の選択を支援する方法を提案した。

本研究では、必須度の有効性を示すケーススタディを実施し、MLS の専門家からは有効であるとの結果を得た。しかし、このケーススタディは限定された考慮点に基づいたものであり、被験者数や対象とした MLS が限定されている。今後、さらに多くの考慮点を用いて選択方法の評価を実施する必要がある。

## 参考文献

- [1] A. Arpteg et al., Software Engineering Challenges of Deep Learning, Proc. SEAA 2018, IEEE, Aug. 2018, pp. 50-59.
- [2] H. Belani et al., Requirement Engineering Challenges in Building AI-Based Complex Systems, Proc. REW (RE Workshop) 2019, IEEE, Sep. 2019, pp. 252-255.
- [3] E. Breck et al., The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction, Proc. IEEE Big Data 2017, IEEE, Dec. 2017, pp. 1123-1132.
- [4] 内閣官房, 人間中心の AI 社会原則, Mar. 2019, <https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/aigensoku.pdf>.
- [5] M. Chui et al., Notes from the AI Frontier: Insights from Hundreds of Use Cases, Discussion Paper, McKinsey & Company, Apr. 2018, pp. 1-32.
- [6] D. Cooper, et al., Project Risk Management Guidelines: Managing Risk with ISO 31000 and IEC 62198, 2<sup>nd</sup> ed., Wiley, 2014.
- [7] DIN-SPEC 92001-1, Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 1: Quality Meta Model, <https://www.en-standard.eu/din-spec-92001-1-artificial-intelligence-life-cycle-processes-and-quality-requirements-part-1-quality-meta-model-text-in-english/>.
- [8] 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>.
- [9] IPA SEC, AI 社会実装推進調査報告書, Jun. 2018, <https://www.ipa.go.jp/files/000067229.pdf>.
- [10] ISO/IEC 25010:2011, Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - System and Software Quality Models, 2011.
- [11] ISO/IEC 25020:2007, Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Measurement Reference Model and Guide, 2007.
- [12] ISO/IEC 25030:2007, Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Quality Requirements, 2007.
- [13] 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>.
- [14] F. Kohn, et al., Software Engineering for Machine-Learning Applications: The Road Ahead, IEEE Software, Vol. 35, No. 5, Sep./Oct. 2018, pp. 81-84.
- [15] 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.
- [16] B. Lorica et al., AI Adoption in the Enterprise, O'Reilly, 2019, <https://www.oreilly.com/data/free/ai-adoption-in-the-enterprise.csp>.
- [17] R. Magoulas, et al., AI Adoption in the Enterprise 2020, O'Reilly, Mar. 2020, <https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2020/>.
- [18] S. Masuda et al., A Survey of Software Quality for Machine Learning Applications, Proc. of ICSTW 2018, IEEE, Apr. 2018, pp. 279-284.
- [19] 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/000499625.pdf](https://www.soumu.go.jp/main_content/000499625.pdf).
- [20] C. Murphy et al., A Framework for Quality Assurance of Machine Learning Applications, Computer Science Technical Report, CUCS-034-06, Columbia University, Apr. 2011.
- [21] K. Nakamichi et al., Requirements-Driven Method to Determine Quality Characteristics and Measurements for Machine Learning Software and Its Evaluation, Proc. of RE 2020, IEEE Computer Society, Sep. 2020 [採録決定].
- [22] 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/>.
- [23] 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.
- [24] A. Vogelsang et al., Requirements Engineering for Machine Learning: Perspectives from Data Scientists, Proc. REW (RE Workshop) 2019, 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. ISSA 2018, ACM, Jul. 2018, pp. 129-140.