情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

機械学習応用システムの要件定義方法に関する考察

本橋 洋介[†1]　見上 紗和子[†1]　森本 麻代[†1]

概要：機械学習を活用したシステム（機械学習応用システム）では，評価方法や学習済みモデルの更新方法，精度と経済価値の関係の算出，異常値の対応など通常のシステムでは考慮しない要素が多くある．そのため通常のシステム開発の要件定義の方法論がそのまま適用できないという問題がある．本稿では，機械学習応用システムの要件を定義するために，実際のシステムの要件を調査し，要件定義の方法論確立に向けた考察を行う．

キーワード：機械学習，機械学習システム，機械学習工学

A survey of the method of requirement definition for systems using machine learning.

Yosuke Motohashi[†1]　Sawako Mikami[†1]　Mayo Morimoto[†1]

1. はじめに

近年，ビッグデータブームや Deep Learning の登場を経て，機械学習を業務で活用する事例が増加している．Deep Learning が画像認識分野を中心に精度改善をもたらしたことなどによって，適用の幅も広がっており，2018 年現在，第三次人工知能ブームと呼ばれるようになっている．しかし，機械学習に代表される人工知能を業務システムに取り入れ運用する事例はまだ多くないのが現状であり，各企業も「機械学習をどのように使っていけばよいか」「機械学習を使うようになるまでの手順」などのノウハウが不足している．特に，機械学習を用いたシステム（以下，機械学習応用システム）に関して，要件検討や開発上の留意点について，体系的な整理がない．これにより，システム開発時に，想定していなかった問題が発生して開発工程の遅延が起こるといった課題がある．これらの問題の早期発見や解決は，機械学習の経験が多くある一部の人に依存しており，機械学習応用システムの業務活用が拡がらない原因となっている．そこで，本発表では，機械学習応用システムの実例から，開発時の留意点や評価方法を洗い出し，システムの開発前に要件を定義するための基礎的な検討を行う．

2. 関連研究

機械学習を用いたシステムの開発方法や評価に関する研究やガイドラインは以下のようなものがある．

総務省 AI・ネットワーク社会推進会議においては AI を用いたシステムの開発ガイドラインを策定し公開している[1]．しかし，透明性や倫理性などの指針はあるものの，性能や利便性に関する評価の方法に対して言及していない．

機械学習を用いたシステムの運用方法・構築方法についての書籍[2][3]はあるが，限定的なユースケースにおける実例が記載されているのみで，複数のプロジェクトで共通する考え方なのかどうかがわからなくなっている．また，機械学習システムのテスト項目についてまとめた研究[4]もあるが，正確性や速度などを重視しており，利便性や有効性について評価する方法についての整理が不足している．

このように体系的な整理があまり存在しない中，「機械学習工学」を研究するコミュニティ[5]が立ち上げられ，体系的な整理をしたいという試みが始まっている．その中で，機械学習システムの評価方法について調査した結果がある[6]．しかし，要件定義において調査すべきことや，開発時に留意すべき点を予め決めることについての整理が不足している．

3. 調査方法

調査は，機械学習を用いたシステムの開発プロジェクトの関係者（主としてプロジェクトマネージャーや機械学習部分の設計者）に対してアンケート形式で実施した．

調査においては，まずシステムの概要としてシステムの目的・データの種類・業務フローを質問した．加えて，システムの評価方法・データの課題・開発上の留意点を質問した．各質問の選択項目は，事前に 10 プロジェクトの関係者からヒアリングを行って得た結果を基に作成した．

†1 (株)日本電気
NEC corporation

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

調査対象の PJ に関する情報を表１に示す．今回は，システムとして定常運用しているまたは開発中のものであり，教師あり学習タイプの機械学習を用いているものに限定して調査を行った．

表１　調査対象のプロジェクト

調査対象プロジェクト数	52
対象の業界	製造11，金融10，エネルギー6，流通5，交通4，自治体4，その他12
問題の種類	需要予測18，行動予測3，不正検知3，解約予測3，所要時間予測2，劣化予測2，顧客満足度予測2，その他19

質問項目の一覧を表２に示す．

質問は，システムの目的を除きすべて選択式で行った．以下に，各質問項目の内容および調査意図を説明する．

3.1 問題の種類に関する調査

教師あり学習を用いたシステムを対象としており，回帰・判別の２つを選択項目とした．

3.2 機械学習を用いたシステムの業務フローに関する調査

人と機械学習の役割分担によってパターンを作成し，以下の３通りのフローに分類し，選択項目とした．

パターン１．自動意思決定

機械学習結果のモデルを活用したシステムが自動的に実行するが，システムが判断の自信度を判定して自信がないときに人間に意思決定を委ねる流れである（図１）．
これは，自動でオペレーションして問題ない時によくあるフローである．全体として統計的に成功すればよく，個別のオペレーションの成功不成功がそこまで問題ではないケースとも言える．

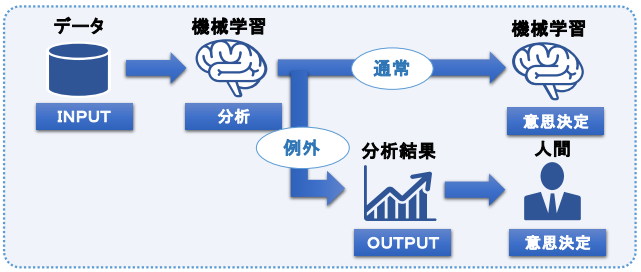

図１　業務フローパターン１．自動意思決定

パターン２．人による意思決定

機械が推奨する結果を参考に，実行に関する最終意思決定を人間が行う流れである（図２）．

これは，すでに行っている業務を機械がサポートするときに，よくあるフローである．小売業の発注自動化や，インフラや機器のメンテナンスのように，予測自体は人工知能が高性能に行うことができるが，そのあとのアクションにおいては複雑な要因が絡み合い最終判断は人間が行った方がよいようなケースはこのようなフローがよいと考えられる．

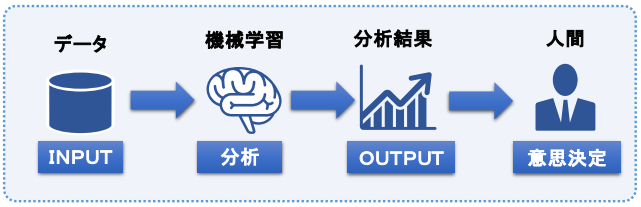

図２　業務フローパターン２．人による意思決定

パターン３．人によるルールの選択

機械学習が作成したルールを人が確認し，人が採用したルールに基づいて人工知能が自動的に実行するフローである．オペレーションに失敗したときの損害が大きいときや，オペレーションの論拠を正確に説明したり保証したりしなくてはならないケースである．

表２．質問項目

分類	質問項目	回答形式
目的	システムの目的	自由記述
問題の種類	回帰,判別	択一の選択式
業務フロー	自動意思決定,人による意思決定,人によるルールの選択	択一の選択式
データの種類	数値（センサ以外）,数値（センサ）,ラベル,自然言語,画像	複数選択式
データの量	1モデルあたりの学習データ量が多いか,説明変数の種類が多いか	各々6値（5,4,3,2,1,0）の選択式
データや対象の特性	目的変数の頻繁な変更があるか,モデル更新頻度が多いか,データが不正確かどうか	各々6値（5,4,3,2,1,0）の選択式
評価に用いた精度指標	MAE,RMSE,MAPE,平均誤差／平均実績値,最大誤差,一定以上の誤差値割合,上振れ誤差,下振れ誤差,適合率,再現率,特異度,F値,lift値,AUC	複数選択式
精度以外で評価に用いた項目	結果の解釈性,モデルの解釈性,意外性,安定性	複数選択式
開発上の問題・留意点	過学習しやすい,更新時のモデルの監視が重要,学習時の異常値処理が重要,推論時の異常値処理が重要,代替モデルや転移学習の必要性	6値（5,4,3,2,1,0）の選択式

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

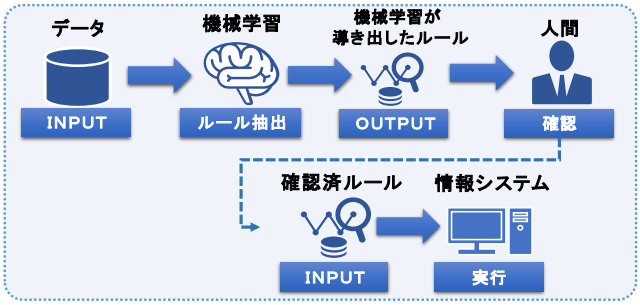

図３　業務フローパターン３．人によるルールの選択

このように，人と機械の役割分担を基に，業務フローをパターン化することとした．

3.3 データの種類に関する調査

数値（センサ以外），数値（センサ），ラベル，自然言語，画像の５通りを回答項目とした．数値データをセンサデータとセンサデータ以外に分けたのは，温度や振動などの物理量の観測結果と，売り上げや年齢などのトランザクションや人の入力によって登録されたデータでは，ノイズや欠損の混入が異なり，機械学習応用システムへの要件が異なることが想定されたからである．また，ラベルデータとは，「都道府県」・「エリアコード」などの非数値データ（カテゴリーデータ）である．

3.4 データの量に関する調査

１データあたりの学習データの量の多さと，機械学習に投入する説明変数の種類数を質問した．データの量や種類数は，定量的な数値であるが，他の回答項目と比較するために，以下の６段階評価とした．

学習データの量：0（100 以下），1 (101-500), 2 (501-1000), 3 (1001-5000), 4 (5001-10000), 5 (10001-100000), 6(100001 以上)

説明変数の種類数：0 (10 以下)，1 (11-50), 2 (51-100) ,3 (101-500), 4 (501-1000), 5 (1001-5000), 6 (5001 以上)

3.5 データや対象の特性に関する調査

目的変数及び説明変数の特殊な状況に関して，典型的な留意する観点として，目的変数の頻繁な変更があるか・モデル更新頻度が多いか・データが不正確かどうかの３つの質問を実施した（5 から 0 の 6 段階評価）．

3.6 評価に用いた精度指標に関する調査

筆者らが行った過去の調査を基に，実際のプロジェクトで評価指標に用いられたことがあるものを選択項目として，質問した．選択項目にした評価指標の一覧を以下に示す．

・機械学習が回帰問題を対象とする場合の精度指標

回帰問題とは，数値を推定する問題のことを指す．この場合は，評価対象データにおいて，正解の値（実績値）と機械が推定した値（予測値）の差（誤差）を基に評価するため以下の指標のいずれか１つ以上を回答させた．

・平均誤差（MAE）＝誤差の絶対値の平均値
・平均二乗誤差（RMSE）＝誤差の二乗の平均値の平方根
・誤差率（MAPE）＝（誤差の絶対値／実績値）の平均値
・平均誤差／平均実績値
・最大誤差値
・一定値以上の誤差値の割合
・上振れ誤差率
・下振れ誤差率

・機械学習が判別問題を対象とする場合の精度指標

判別問題とは，YES・NO などのラベルを推定する問題のことを指す．この場合は，図４にあるような混合行列に値を入れ精度を計算するのが通常である．

		実績値	
		P	N
予測値	P	TP	FP
	N	NP	TN

P：Positive（正例）
N：Negative（負例）

図４　混合行列の例

選択項目にした精度指標は以下のとおりである．（説明のため，図４の TP，FN，FP，TN の値を用いる．）

適合率(Precision)＝TP /　(TP＋FP)
再現率(Recall)・感度(sensitivity)＝TP /　(TP＋FN)
特異度(specificity)＝TN　/　（FP＋TN）
F 値=（適合率と再現率の調和平均）
Lift 値=「ランダムに推定したときの適合率」とテストデータの適合率の比率．

3.7 精度以外で評価に用いた項目に関する調査

精度以外の評価項目についても同様に，過去の調査を基に，用いられたことがあるものを選択項目として質問した．選択項目にした評価指標を以下に示す．

①　解釈性

結果を人が解釈しやすいか，人が理由（≒予測モデル）を理解して業務ができるかの指標を解釈性と呼ぶ．たとえば機械学習で異常検知する方法を作った時に，異常の検出

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

精度だけではなく,「どういう理由で異常と判断したか」という点を解釈できることは大切である．その理由を人が解釈することで，異常の原因を推定して修理に行くなどの行動ができるからである．
解釈性には，何を解釈するかによって以下の２つの種類があり，今回は以下の２つを選択項目とした．
・結果の解釈性
　機械学習の推定結果ここに対して，説明変数の何が影響してその結果になったのかを解釈できるかどうか
・モデルの解釈性
　学習結果のモデルが,説明変数の何を重視しているモデルなのかを解釈できるかどうか

②　意外性
機械学習が，人が従来持っていなかった知見を出せるかどうかを評価する指標を意外性と呼ぶ．定性的な評価になりやすいが，機械学習のプロジェクトでは，人が持っていない知見を得ることを記載されることがあり，その場合に用いるため選択項目とした．

③　安定性
機械学習の結果やモデルが，データが新しくなった時や追加された時に変わらないかどうかを評価する指標を安定性と呼ぶ．定常運用時に毎日実行して結果を人が解釈しながら用いるケースなど，モデルや結果の大幅な変化が業務に悪影響が出ることもあるため，選択項目とした．

3.8 開発上留意すべき点に関する調査

開発中に発生する問題や，留意点について，過去のプロジェクトから典型的な問題の候補を挙げて質問した．
質問したのは，過学習しやすい・更新時のモデルの監視が重要・学習時の異常値処理が重要・推論時の異常値処理が重要・代替モデルや転移学習の必要性の５つであり，それぞれ５から０の６段階評価とした．なお，代替モデルとは，データが少ないなどの理由で機械学習が推定したい対象の学習結果を得ていないときに，他のデータで学習した結果のモデルを基に推定することである．

調査においては，3.1 から 3.8 で述べた項目について質問を行い，これらの回答間の関係性があるかなどを調査した．

4. 調査結果と考察

調査結果と考察を述べる．

4.1 問題の種類に関する調査結果

今回の調査対象は，回帰 35 件，判別 17 件であった．

4.2 システムの業務フローに関する調査結果

図５は，調査対象のプロジェクトの業務フローを分類した結果である．全体の 60%が，パターン２（人による意思決定）であったが，これは，調査対象の多くを占める需要予測プロジェクトにおいて，需要予測結果を基に，在庫管理や人員計画などを人が行うケースが多かったためである．

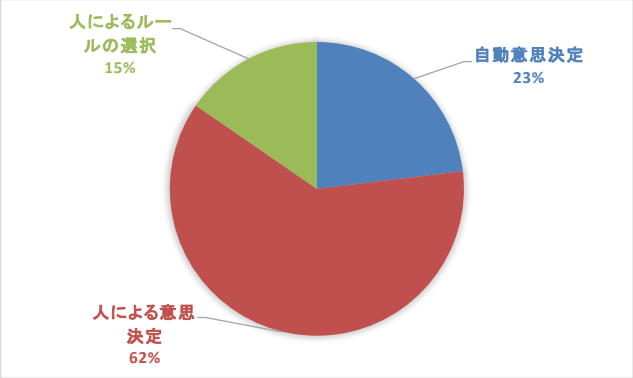

図５　調査対象の業務フローパターン

4.3 データの種類に関する調査結果

図 6 は，データの種類に関する質問の回答結果である．数値・ラベルデータが多く用いられている．

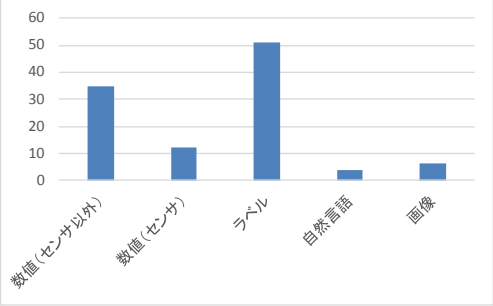

図６　調査対象のプロジェクトで用いられたデータ

4.4 データの量に関する調査結果

図７は，学習データの量・種類数に関する質問の回答結果である．一般に学習データ量が少ないことや，学習データに対して説明変数の種類数が多すぎることで過学習しやすくなる．図７に示すように学習データが 100 件以下（回答 0,1,2）のものが 22 プロジェクト，説明変数が 1001 種類以上（回答 5,4）のものが 15 プロジェクトあり，過学習の危険性があるプロジェクトが一定割合以上あることがわかる．

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

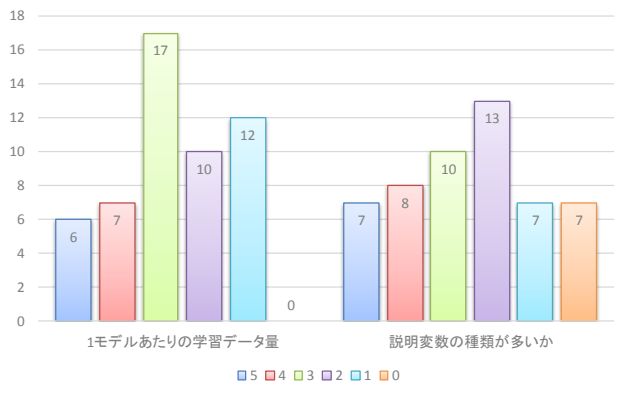

図 7　学習に用いたデータの量と種類数

4.5 データや対象の特性に関する調査結果

図 8 は，データや対象の特性に関する質問の回答結果である．目的変数の変更に関しては 7 割程度のプロジェクトが「0:　ない」と回答しているが，ごく一部頻繁な変更があるプロジェクトがあることがわかった．モデルの更新頻度は 29 プロジェクトが「1:　1 年に 1 回程度」との回答であったが，「5：　1 週間に 1 回以内」というものもあり，頻繁なモデル更新を必要とするプロジェクトもあることがわかった．データが不正確である度合いに関しては，「5:　非常に大きなノイズ混入や不正確な値が 2 割以上」というものは殆どなかったが，「3:　一部の変数に 5%以上の欠損がある，または，ノイズ処理が必要」が一定数あることから，一部のプロジェクトでは異常値の処理を行う必要があることがわかった．

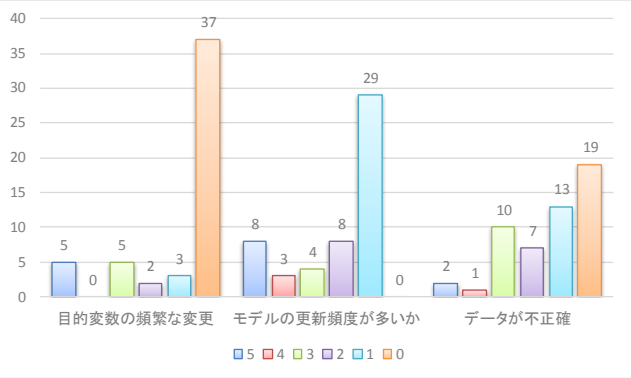

図 8　データや対象の特性

4.6 評価に用いた精度指標に関する調査結果

図 9 は，回帰における精度評価指標の採用率を示した図である．図 9 のように，MAE や平均誤差を平均実績値で割ったものが多く用いられていた．RMSE に比べて MAE が頻繁に用いられていたのは，直接経済価値に変換しやすい指標であるからだと推測される．実際に，調査対象のプロジェクトの中に，MAE をコストや利益に変換してプレゼンテーションを行った事例があった．

また，一定以上の誤差値割合や上振れ・下振れ誤差の度合を重視しているプロジェクトもあった．たとえば，需要予測結果において在庫管理を行う場合は，多めに予測する場合（＝在庫過多に繋がる）と少なめに予測する場合（＝欠品に繋がる）では運用者に与える被害の大きさが異なる．そのため，上振れ誤差と下振れ誤差を分けて評価する必要がある．

一般に機械学習は RMSE を小さくするように学習することが多いが，このような指標を重視すると必ずしも RMSE が最小のモデルが優秀とは限らず，運用に合わせた精度指標を設定して評価する必要があることがわかった．

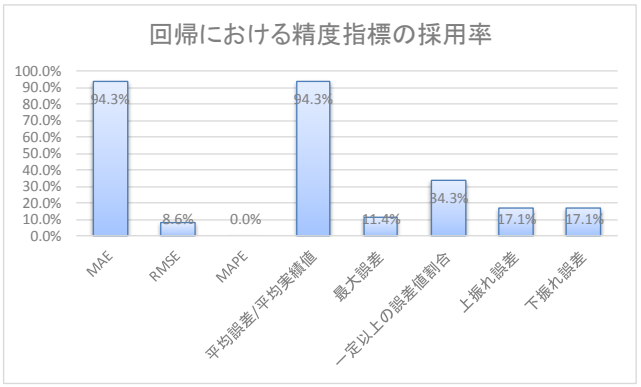

図 9　回帰における精度指標の採用率

図 10 は，判別における精度評価指標の採用率を示した図である．図 10 のように，F 値に比べて lift 値を用いることが多いのは，プロジェクトによっては判別問題における正例率が著しく低く（1%など），F 値を算出してもあまりに小さい値になり，十分価値がある精度か評価しづらいケースがあるからだとわかった．このような「解きたい問題の難しさ」を基に評価指標を基準化・正規化することが有効性の評価において重要であると考えられる．

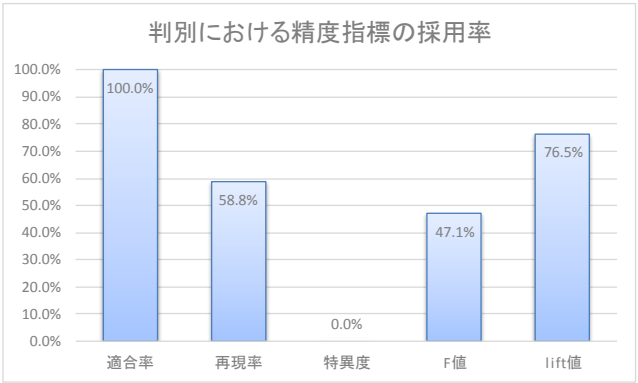

図 10　判別における精度評価指標の採用率

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

4.7 精度以外で評価に用いた項目に関する調査結果

図 11 は，精度以外の評価指標に関する調査結果である．図 11 のように，モデルの解釈性や結果の解釈性を多く用いていることがわかるが，意外性や安定性などの他の指標も用いていることが確認された．

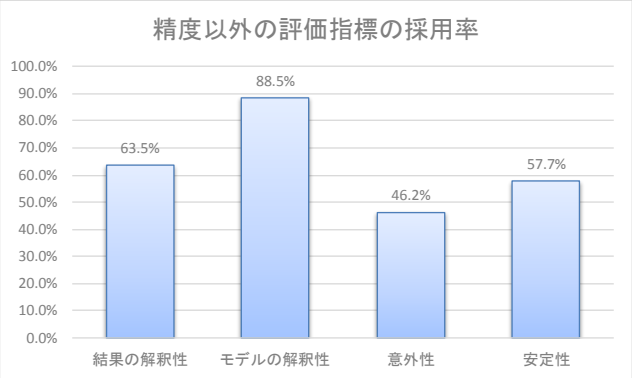

図 11　精度以外の評価指標の調査結果

表 3 は，精度以外の評価指標と，プロジェクトの特徴に関する他の調査結果の関係を分析した結果である．表 3 内の値は，評価指標の回答結果と，業務フロー・データの種類・データの量・データの特性の回答結果について，スピアマンの順位相関係数を算出した結果である．表 3 のように，用いられる評価指標は業務フローと関係が深いことがわかる．業務フローが自動意思決定の場合は安定性，人による意思決定の場合はモデルの解釈性，人によるルールの決定の場合は安定性と結果の解釈性が重要であることがわかる．また，モデルの更新頻度が多い時や，目的変数の頻繁な変更があるときは，モデルの解釈性や安定性が重要になりやすいことがわかった．

4.8 開発上留意すべき点に関する調査結果

図 12 は，開発上留意すべき点の調査結果である．図 12 のとおり，それぞれ 10 プロジェクト以上で重要である（回答が 3 以上）という回答となっており，これら 5 つの留意点は開発上典型的な留意点であることがわかる．

表 4 は，開発上留意すべき点と，プロジェクトの特徴に関する他の調査結果の関係を分析した結果である．表 4 内の値は，開発上留意すべき点と，業務フロー・データの種類・データの量・データの特性の回答結果について，スピアマンの順位相関係数を算出した結果である．
表 4 を参照しながら，それぞれの留意点が重要になりやすいプロジェクト特性をまとめる．

表 3．精度以外の評価指標と，プロジェクトの特徴との関係

	結果の解釈性	モデルの解釈性	意外性	安定性
【業務フロー】 自動意思決定	−0.44	−0.52	−0.51	0.28
【業務フロー】 人による意思決定	0.14	0.33	0.18	0.04
【業務フロー】 人によるルールの選択	0.32	0.15	0.35	−0.39
【データ種類】 数値（センサ以外）	0.15	0.26	0.15	0.07
【データ種類】 数値（センサデータ）	0.04	0.20	0.04	−0.18
【データ種類】 ラベル	0.18	0.39	0.13	−0.12
【データ種類】 自然言語	−0.08	0.10	0.17	0.25
【データ種類】 画像	−0.35	−0.81	−0.33	0.19
【データの量】 1モデルあたりの学習データが多い	0.06	−0.39	0.16	0.24
【データの量】 説明変数の種類が多い	0.23	0.37	0.26	0.27
【データ特性】 モデル更新頻度が多い	−0.04	0.31	−0.07	0.24
【データ特性】 データが不正確	−0.13	0.03	−0.03	−0.04
【データ特性】 目的変数の頻繁な変更	0.25	0.23	0.02	0.31

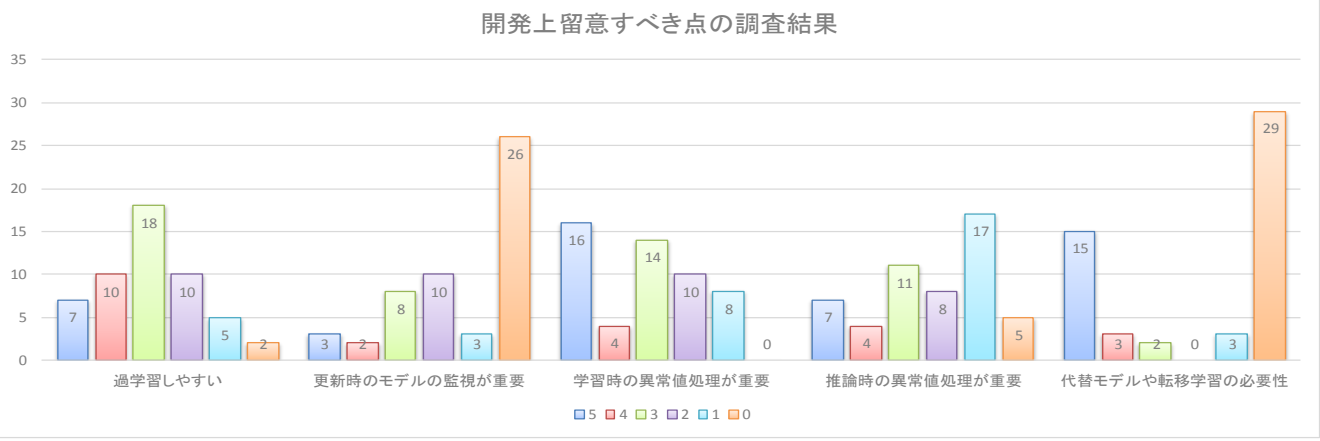

図 12．開発上留意すべき点の調査結果

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

表 4. 開発上留意すべき点と，プロジェクトの特徴との関係

	過学習しやすい	更新時のモデルの監視が重要	学習時の異常値処理が重要	推論時の異常値処理が重要	代替モデルや転移学習の必要性
【業務フロー】 自動意思決定	−0.11	−0.23	0.05	−0.10	−0.26
【業務フロー】 人による意思決定	0.23	0.37	−0.25	0.06	0.32
【業務フロー】 人によるルールの選択	−0.18	−0.24	0.27	0.04	−0.13
【データ種類】 数値（センサ以外）	0.49	0.29	−0.19	−0.02	0.09
【データ種類】 数値（センサデータ）	−0.33	−0.11	0.19	0.29	−0.01
【データ種類】 ラベル	0.13	0.13	−0.17	−0.22	0.12
【データ種類】 自然言語	0.40	−0.16	0.21	−0.09	−0.25
【データ種類】 画像	−0.34	−0.34	0.11	−0.26	−0.17
【データの量】 1モデルあたりの学習データが多い	−0.17	0.11	0.07	0.23	−0.32
【データの量】 説明変数の種類が多い	0.58	0.20	0.14	0.19	−0.04
【データ特性】 モデル更新頻度が多い	0.14	0.79	−0.33	0.12	0.46
【データ特性】 データが不正確	0.05	−0.11	0.34	0.25	−0.04
【データ特性】 目的変数の頻繁な変更	0.02	0.73	−0.18	0.22	0.51

・過学習しやすい…説明変数の種類が多いとき，1 モデルあたりの学習データが少ない時，自然言語・数値データの時に重要となりやすい．これは，過学習が発生しやすい条件の一般的な知見と一致する．

・更新時のモデルの監視が重要…業務フローが「人による意思決定」の時，モデル更新頻度が多い時，目的変数の頻繁な変更が多い時に重要となりやすい．人による意思決定の場合，人にとって直感的ではないモデルの変更が行われることで人が機械学習の結果を利用しづらくなることが現れていると考えられる．

・学習時の異常値処理が重要…業務フローが「人によるルールの選択」の時，データの種類がセンサデータや自然言語データの時，データが不正確な時に重要となりやすい．これは，センサデータや自然言語データはノイズや表記ゆれの問題が起こりやすいことが関係していると考えられる．なお，「人によるルールの選択」は，センサデータを用いていることとの相関が高いことから結果的に相関が高くなっていると考えられる．

・推論時の異常値処理が重要…データの種類がセンサデータの時，1 モデルあたりの学習データが多い時，データが不正確な時に重要となりやすい．学習時の異常値処理と大きな違いはないが，モデル更新頻度が多く目的変数の頻繁な変更があるようなケースでは，推定対象のデータが不安定で，推論時にも異常値が混入しやすい可能性を示していると考えられる．

・代替モデルや転移学習の必要性…学習データが少ない時，モデル更新頻度が高い時，目的変数の頻繁な変更がある時に重要となりやすい．目的変数が変更されるようなケースでは，学習データ量が少ない対象が発生しやすく，モデルが不安定になりやすいことを示していると考えられる．

4.9 要件定義時点での開発上留意すべき点の推定

これまでの調査結果によると，評価指標や，開発上の留意点など，要件定義時に決めるべき情報は，プロジェクトの特徴との関係が深いことがわかった．そこで，プロジェクトの情報を基に開発上の留意点を推定することができるかの分析を行った．

分析の概要を表 5 に示す．開発上の留意点（5 種類）は 0 から 6 までの値を取る数値であり，その数値を回帰木で推定したときの精度を検証した．学習時の説明変数を表 5 のように 3 通りの説明変数の組で行うことで，どの情報があることで正しく推定できるかを調べた．

表 5　開発上の留意点の推定に関する分析概要

学習方法		回帰木で学習して推定
学習・評価データ		学習34データ・評価18データ（3分割交差検証）
目的変数		開発上の留意点（5通り）
説明変数	変数セット1	問題の種類・データの種類・業務フロー
	変数セット2	問題の種類・データの種類・業務フロー・データの量
	変数セット3	問題の種類・データの種類・業務フロー・データの量・データと対象の特性

分析結果を図 13 に示す．図 13 における精度は MAE である．図の通り，変数セット 3 の時に最も良い推定精度であることがわかる．また，変数セット 1 と変数セット 2 の間では，「代替モデルや転移学習の必要性」の推定精度が上昇している．これは，データの量が変数に加わることで，代替モデルの必要性がわかることを示していると考えられる．さらに，変数セット 2 と変数セット 3 の間では，「更新時のモデル監視が重要」についての推定精度が上昇している．これは，目的変数の頻繁な変更があるどうかなどのデータと対象の属性が変数に加わることで，モデル監視の重要性がわかることを示していると考えられる．

Vol.2019-GN-106 No.18
Vol.2019-CDS-24 No.18
Vol.2019-DCC-21 No.18
2019/1/24

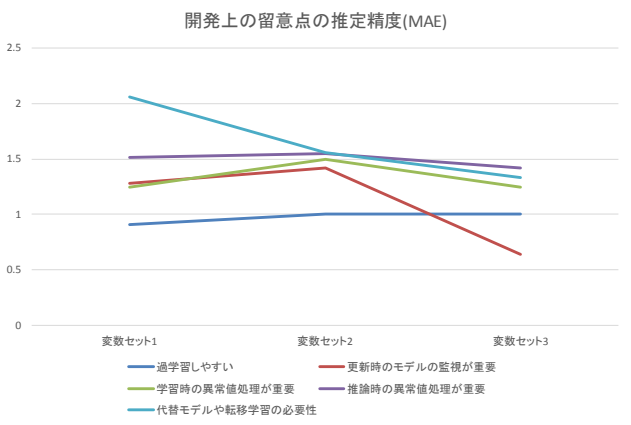

図 13　開発上の留意点の推定精度

一方，他の 3 つに関しては，変数セット 1,2,3 の間に大きな精度の差が見られない．これは，現在の調査では，これらを推定するために必要な情報が不足していることが考えられ，今後別の項目を加えた調査を行うことが必要であることが考えられる．

5. まとめと今後の展望

本論文では，機械学習応用システムの実例から，開発時の留意点や評価方法を洗い出し，システムの開発前に要件を定義するための基礎的な検討を行った．調査の結果，精度指標にも運用に合わせた評価指標を用いる必要性や，精度以外の指標も併せて評価する必要があることがわかった．また，精度以外の指標については，機械学習を用いたシステムと人の役割のパターン(＝業務フローのパターン)や，データの特性によって求められる指標が異なることがわかった．さらに，過学習のしやすさ，モデルの監視，異常値の処理，代替モデルの用意といった開発上の留意点についても，プロジェクト特性によって重要となる場合に違いがあることが確認された．

さらに，開発上の留意点を，プロジェクトの特徴から推定することができるかの分析を行い，データの特性などプロジェクトの情報がわかることで，開発上の留意点について推定できる可能性があることを確認した．

今後，さらに調査対象を拡げると共に，要件定義を自動的に行う方法の検討を行い，多くの人が機械学習応用システムの要件定義を円滑に行えることの支援を行うことを目指していきたい．

参考文献

[1] 総務省 AI ネットワーク推進会議 2017 報告書 http://www.soumu.go.jp/menu_news/s-news/01iicp01_02000067.html.
[2] 有賀康顕他. 仕事で始める機械学習，オライリージャパン社，2018.
[3] 本橋洋介, 人工知能システムのプロジェクトがわかる本，翔泳社，2018
[4] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, D. Sculley. What's your ML test score? A rubric for ML production systems. NIPS 2016 Workshop (2016)
[5] 日本ソフトウェア学会 機械学習工学研究会 https://sites.google.com/view/sig-mlse/
[6] 本橋洋介,機械学習を用いた業務システムの機能と評価に関する考察，情報処理学会第 105 回研究会研究報告,2018