原著論文 Original Paper

確認修正コストに基づく機械学習評価手法

An Evaluation Method for Machine Learning Based on Verification and Correction Costs

木村 俊一 Shunichi Kimura 富士フイルムビジネスイノベーション株式会社 FUJIFILM Business Innovation Corp. shunichi.kimura.vk@fujifilm.com

久保田 聡 Satoshi Kubota （同　上） satoshi.kubota.sj@fujifilm.com

和田 直己 Naoki Wada （同　上） naoki.wada.kn@fujifilm.com

橋本 一成 Kazunari Hashimoto （同　上） kazunari.hashimoto.hq@fujifilm.com

Keywords: machine learning, OCR, model evaluation, performance measure, verification, reject model, reject option, selective classifier, classification with rejection, AUC.

Summary

We propose a new method to evaluate machine learning models, parameters, and algorithms. To design the method, we consider verification and correction costs of human labor, and we base it on machine learning systems with reject models. The models are proposed to handle problems where the machine learning systems have errors. In the processes of the reject models, a part of outputs of machine learning are rejected and we can get our desired accuracy, that means we can reduce errors, for the other results that are not rejected. On the other hand, we should verify and correct the rejected results. Therefore, the verification and correction costs should be considered to develop the evaluation method. In addition, the reject models are sensitive to thresholds that define decisions of rejection. Thus, the evaluation method should handle the varying thresholds. Hence, the method to evaluate the machine learning should have the following two features: (1) handling varying thresholds, (2) managing the verification and correction costs. Conventional methods such as ROC curve or PR curve can handle the varying thresholds. These conventional methods, however, cannot manage the verification and correction costs. In this paper, first, we define a performance measure to evaluate machine learning based on the verification and correction costs. Second, we propose ARAC curve(Acceptance Rate-Accuracy after Correction curve) and ARAC-AUC(Area Under Curve) in which the horizontal axis shows acceptance rate, and the vertical axis shows accuracy after correction, respectively. This ARAC curve can handle varying thresholds as well as the conventional methods. Third, we explain the relationship the performance measure and the ARAC curve. The horizontal axis is closely related to the verification costs, and the vertical axis is closely related to the correction costs. Accordingly, the ARAC curve can express verification and correction costs. Finally, we show experimental results where the proposed ARAC curve and ARAC-AUC can express the performance measure better than the conventional methods.

1. は　じ　め　に

機械学習システムの出力精度を 100%とすることは困難である．そのため，リジェクトオプション [De Stefano 00, Hendrickx 24]，リジェクトモデル [Kimura 17]，あるいは，selective classifier[Geifman 17, Huang 20] と呼ばれる方式が提案されている (以下，リジェクトモデルと総称する)．これらは，精度が低いと予想される一部の結果をリジェクトすることで，非リジェクト結果 (アクセプト結果) に対して所望の精度を実現する方式である [Chow 70].

リジェクトモデルは，リモートセンシング [Condessa 16, Giacinto 00]，医学生理学 [Hanczar 08, Quevedo 11, Zhang 12]，OCR[De Stefano 14, Navarro-Cerdan 15]，言語処理 [Fumera 03, Fumera 04]，画像分類 [Huber 05] 等，様々な領域で活用されている．

本論文では，このリジェクトモデルを利用した場合の機械学習評価手法の提案を行う．特に，人間が行う業務の代替を目的として機械学習を導入する場合を前提として考える．

リジェクトモデルを利用したシステムでは，非リジェクト結果 (アクセプト結果) に対しては人間の確認を必要としないが，リジェクトした結果に対しては，人間が確認修正を実施する前提をおく [De Stefano 14, Fumera 00, Tortorella 05]．予測器が出力した結果をリジェクトした場合，人間が予測結果を確認し，さらに，予測結果が

誤っている場合の修正を実施するため，精度向上が期待できる．その反面，リジェクトした場合の人間による予測結果の確認コストと，予測結果が誤っている場合の人間による予測結果の修正コストが必要となるデメリットがある．そのため，人間が行う業務の代替を目的とする場合，確認修正コストを基準とした評価手法が望ましい．

リジェクトモデルは，Separated Rejector，Dependent Rejector，Integrated Rejector の 3 種に分類される [Hendrickx 24]．この中で，本論文では特に Dependent Rejector 方式を対象とする．本方式は，予測器の確信度と閾値を比較し，確信度が閾値未満のときに予測結果をリジェクトするものである．確信度を出力する任意の予測器に適用可能な汎用性と，コストが大きい予測器の再学習が不要という利点を持つ [Hendrickx 24, Tang 14]．また，多くの文献 [De Stefano 00, Geifman 17, Pugnana 23, Sotgiu 20, Zhou 22] で Dependent Rejector 方式を前提とした検討が行われている．

この Dependent Rejector 方式では閾値が変化すると，それに伴い精度が変化する．そのため，単純な評価が困難であり，変化する閾値を前提とした評価手法が必要となる．従来，変化する閾値に対応する評価曲線を描画する機械学習評価手法が種々用いられている [Bradley 97, Davis 06, Eban 17, Fumera 02, Hogan 23, Nadeem 09, Pietraszek 07]．これらの従来手法では，確認修正コストを基準とした評価を実施できない点が課題である．

一方，コスト評価が可能，かつ，変化する閾値を前提とした評価手法としてコスト曲線 [Abbas 19, Hanczar 19] が提案されている．しかしながら，コスト曲線を用いるためには，予めコスト定数を定めておく必要がある点が課題である．

そこで本論文では，リジェクトモデルおよび Dependent Rejector 方式を前提とし，コスト定数に影響されない，かつ，確認修正コストを反映可能な，機械学習評価曲線の考察および提案を実施する．以下，2 章に従来手法，3 章に提案手法，4 章に実験結果を示す．5 章でまとめを行う．

2. 従 来 手 法

前提となるリジェクトモデルの説明，および，従来手法としての ROC 曲線，PR 曲線，Risk-Coverage 曲線，および，コスト曲線の説明を行う．

2·1 リジェクトモデル (Dependent Rejector)

本論文における評価手法の適用対象であるリジェクトモデル (Dependent Rejector) を示す．

図 1 に示すように，ベクトル $\boldsymbol{u}$(Input $\boldsymbol{u}$) を入力しベクトル $\boldsymbol{v}$(Output $\boldsymbol{v}$) を出力する機械学習システムを考える．まず入力ベクトル $\boldsymbol{u}$ は予測器 (Predictor) において分類され $\boldsymbol{v}$ を出力する．同時に予測器は，確信度 t(Confidence t) を出力する．典型的には，確信度 t は入力 $\boldsymbol{u}$ に対し $\boldsymbol{v}$ が出力されるときの条件付き確率 $P(\boldsymbol{v}|\boldsymbol{u})$ である．ただし，t は必ずしも条件付き確率 $P(\boldsymbol{v}|\boldsymbol{u})$ に関連している必要はなく，単に分類の確信度を示す数値であればよい．

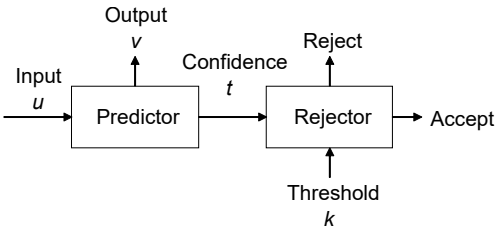


図 1: リジェクトモデル (Dependent Rejector)．予測器 (Predictor) はベクトル $\boldsymbol{u}$(Input $\boldsymbol{u}$) を入力し，出力ベクトル $\boldsymbol{v}$(Output $\boldsymbol{v}$) と確信度 t(Confidence t) を出力する．リジェクト器 (Rejector) は確信度 t と閾値 k(Threshold k) を比較し，予測器の出力ベクトル $\boldsymbol{v}$ をアクセプトするかリジェクトするかを判断する．

次に，確信度 t は，2 クラス分類器であるリジェクト器 (Rejector) において分類され，アクセプト (Accept)，あるいは，リジェクト (Reject) のいずれかのクラスを出力する．リジェクト器の出力がアクセプトの場合 (予測器の結果が正しいと推定した場合)，予測器の出力ベクトル $\boldsymbol{v}$ はそのまま機械学習システムの出力となる．一方，リジェクト器の出力がリジェクトの場合 (予測器の結果が誤りと推定した場合)，予測器の出力ベクトル $\boldsymbol{v}$ はリジェクトされ，入力ベクトル $\boldsymbol{u}$ の分類は別の手法 (例えば人間による分類) に委ねられる．

文字認識システムで例示する．入力 $\boldsymbol{u}$ は文字画像であり，出力 $\boldsymbol{v}$ は文字画像内に記載されている文字の文字コードとなる．予測器は OCR であり，文字コードと認識確信度を出力する．リジェクト器は，確信度を用いて出力文字コードをそのまま利用するかどうかを判定する．出力文字コードをそのまま利用しない場合 (リジェクトする場合) は，人間が出力文字コード $\boldsymbol{v}$ をチェックする．さらに，出力文字コード $\boldsymbol{v}$ が間違っている場合は人間がキーパンチで文字コードを修正する．

さて，ここで対象としているリジェクトモデルは，Dependent Rejector であり，リジェクト器における 2 クラス分類として，閾値 k(Threshold k) を用いて「$t \geq k$ ならば，アクセプト」「$t < k$ ならば，リジェクト」という手法を採用する．また，以下では，説明および計算を単純化するため，t の範囲は $t \in [0,1]$ とした議論を行う．t の定義域がこれと異なる場合は，この範囲となるように何らかの単調変換を行えば良いため，上記としても一般性を失うことはない．

2·2 混 同 行 列

以下の説明の準備としてリジェクト器における 2 クラス分類の混同行列の内容を示す．

機械学習の性能を表現するため，混同行列が利用される

表 1: 混同行列．各行はリジェクト器 (Rejector) がアクセプト (Accept) したかリジェクト (Reject) したかを示す．各列は予測器 (Predictor) の予測が正しい (Correct) か誤っている (Incorrect) かを示す．リジェクト器の結果と予測器の結果の組み合わせで，T_A，F_A，F_R，T_R の 4 種の領域に分割される．

		Predictor	
		Correct	Incorrect
Rejector	Accept	$T_A(k)$ True Accept	$F_A(k)$ False Accept
	Reject	$F_R(k)$ False Reject	$T_R(k)$ True Reject

ことがある．Dependent Rejector では，表 1 に示される 4 象限の表が用いられる [Hendrickx 24]．本表において，行はリジェクト器 (Rejector) がアクセプト (Accept) したか，リジェクト (Reject) したかを示す．列は予測器 (Predictor) の予測が正しい (Correct) か，誤っている (Incorrect) かを示す．

これらの分類の結果，表 1 に示すように $T_A(k)$，$F_A(k)$，$F_R(k)$，$T_R(k)$ の 4 種の指標を得る．各指標は，それぞれの事象の数とみなすことができる．あるいは，確率として考えても良い．予測器の正誤とリジェクト器の出力結果により，入力ベクトル $\boldsymbol{u}$ がこれら 4 種のいずれかに分類される．また，各指標はリジェクト器における閾値 k の影響で変わるため，4 種の指標は閾値 k の関数となる．この混同行列に，本論文で必要な機械学習システムの評価情報が含まれている．

ここで，下記 3 種の尺度を示す．

- 真陽性率: $R_{TP}(k) = T_A(k)/[T_A(k) + F_R(k)]$.
- 偽陽性率: $R_{FP}(k) = F_A(k)/[T_R(k) + F_A(k)]$.
- 適合率: $R_P(k) = T_A(k)/[T_A(k) + F_A(k)]$.

真陽性率 (TPR: True Positive Rate)$R_{TP}(k)$ は予測器で正しく予測された中で，リジェクト器で正しくアクセプトされた割合である．一般的には再現率 (Recall) と称する．2·3 節で説明する ROC 曲線では歴史的に「真陽性率」の用語を用いているため，本論文では ROC 曲線に関連する場合は「真陽性率」，その他の場合は「再現率」を用いる．

偽陽性率 (FPR: False Positive Rate) $R_{FP}(k)$ は予測器で誤って予測された中で，リジェクト器で誤ってアクセプトされた割合である．

適合率 (Precision) $R_P(k)$ はリジェクト器でアクセプトされた中で，予測器の結果が正しい割合である．なお，適合率は精度とも呼ばれることがあるが，本論文中では「精度」は明確な定義を実施しない一般的な用語として用いる．

これらの尺度に関して次のように言うことができる．真陽性率は高いほうが良く，偽陽性率は低いほうが良い．しかしながら，真陽性率と偽陽性率はトレードオフの関係がある．閾値 k を大きくすると，確信度が大きい結果のみアクセプトされるため偽陽性率は低下するが，同時に正解数が少なくなるため，真陽性率も低下する．閾値 k を小さくするとその逆の結果となる．

また，適合率は高いほうが良く，再現率も高いほうが良い．しかしながら，適合率と再現率はトレードオフの関係がある．閾値 k を大きくすると，確信度が大きい結果のみアクセプトされるため適合率が向上するが，同時に正解数が少なくなるため，再現率は低下する．閾値 k を小さくするとその逆の結果となる．

上記のようなトレードオフの関係があるため，一つの尺度だけを用いて機械学習の評価を実施することはできない．また，閾値 k の変化によってそれぞれの尺度の値が変化することが問題である．そこで，トレードオフの関係にある複数の尺度を一度に評価したい．また，閾値 k に依存しない評価をしたいという要求が発生する．これらの要求に応えるため，以下に示す ROC 曲線，および，PR 曲線が用いられている．

2·3 ROC 曲線 (ROC curve)

2 クラス分類器の評価尺度として，ROC 曲線 (Receiver Operating Characteristic curve: ROC curve) や ROC-AUC が広く用いられている [Davis 06, Pietraszek 07, Tortorella 05]．ROC 曲線は，横軸に偽陽性率 $R_{FP}(k)$，縦軸に真陽性率 $R_{TP}(k)$ を閾値 k を 0 から 1 まで変化させてプロットしたグラフである (図 2(a))．

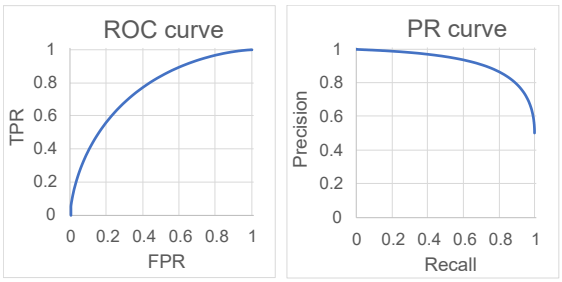
グラフを参照することにより，変化する閾値 k のすべての値を反映した結果を見ることができる．そのため，閾値 k に依存しない評価が可能となる．また，トレードオフの関係にある真陽性率と偽陽性率の両方をグラフ上で見ることができる．偽陽性率は低いほうが良いため左に行くほど良い．真陽性率は高いほうが良いため上に行くほど良い．複数の曲線がある場合，左上に存在する曲線が良いと判定することができる．

問題は，互いに交差する複数の曲線を比較する場合である．この場合のため，曲線の下側の面積を評価値として利用する方法が AUC(Area Under Curve) [Bandos 06, Landgrebe 06, Le Capitaine 14] である．面積 (AUC) が大きいほど良い分類器であると判定する．この AUC(ROC-AUC) を用いることで閾値 k に依存しない単一の値で比較することができる．

この ROC 曲線は真陽性率と偽陽性率のトレードオフ関係を加味した評価をする目的に用いるものであり，確認修正コストを反映していない．そのため，本論文の目的である確認修正コストに基づく機械学習評価には利用することができない．

2·4 PR 曲線 (PR curve)

ROC 曲線とともに，PR 曲線 (Precision-Recall curve: PR curve) が広く用いられている．PR 曲線は，横軸に再



(a) ROC 曲線 (b) PR 曲線

図 2: ROC 曲線と PR 曲線の例．(a)ROC 曲線: 横軸に偽陽性率，縦軸に真陽性率を閾値 k を変化させてプロットした曲線．(b)PR 曲線: 横軸に再現率，縦軸に適合率を閾値 k を変化させてプロットした曲線．

現率 $R_{TP}(k)$，縦軸に適合率 $R_P(k)$ を閾値 k を 0 から 1 まで変化させてプロットした曲線である (図 2(b))．

PR 曲線は，ROC 曲線と同様，閾値 k のすべての値を反映しているため，閾値 k に依存しない評価ができる．また，トレードオフの関係にある適合率と再現率の両方をグラフ上で見ることができる．良いモデルは適合率も再現率も高くなる傾向にあるため，右上に行くほど良い分類器といえる．また，PR 曲線においても，ROC 曲線と同様に AUC(PR-AUC) を定義することができる．

この PR 曲線は，再現率と適合率のトレードオフ関係を加味した評価をする目的に用いるものであり，確認修正コストを反映していない．そのため，本論文の目的である確認修正コストに基づく機械学習評価には利用することができない．

2·5 Risk-Coverage 曲線 (RC curve)

PR 曲線に類似する曲線である Risk-Coverage 曲線 (以下 RC 曲線，RC curve) が，[El-Yaniv 10] において利用されている．この曲線では，適合率に関連する値 $1 - R_P(k)$ をリスク (Risk) として縦軸にプロットする．また，[MHLW 18] のように，適合率 $R_P(k)$ を縦軸にプロットするケースもある．さらに，Accuracy-Rejection 曲線 (ARC) と呼ばれる曲線 [Fumera 02, Giacinto 00, Le Capitaine 14, Nadeem 09, Sousa 13] も広く利用されている．これらの曲線は RC 曲線と同じか，あるいは，軸の上下左右を反転した構造を持つものであり，実質的には RC 曲線と同じ情報を提示する．

PR 曲線との違いとして，RC 曲線では横軸にリジェクト器におけるアクセプト率 (Coverage) を採用している点を挙げることができる．本アクセプト率は確認コストを反映する値であり，確認修正コストに基づく機械学習評価を実施できる可能性がある．

しかしながら，RC 曲線は縦軸に適合率を採用しているため，修正後のエラー率が反映されない点と，後述するように曲線の形状が不安定になるという課題を有する．本論文では，この RC 曲線をさらに改良した評価曲線の提案を行う．提案手法との比較は次節以降で改めて実施する．

2·6 コスト曲線 (Cost Curve)

リジェクト時のコストを表現可能な曲線として，コスト曲線 [Abbas 19, Hanczar 19] が提案されている．文献 [Hanczar 19] では，相対リジェクトコスト λ_R を利用して，縦軸にコスト $L = (1 - \lambda_R)E + \lambda_R R$ をプロットするコスト曲線を提案している (ただし，リジェクト率を R，$E = 1 - R_P(k)$ とする)．文献 [Abbas 19] では，2 次元の ARC にコスト軸を加えた 3 次元の曲線を提案している．

相対リジェクトコスト λ_R を算出するためには，後述の 3·1 節で例示されるコスト定数を用いる．そのため，これらのコスト曲線を描画するためには，コスト定数を予め定めておく必要がある．ところが，コスト定数は，機械学習の利用者や利用者に依存する運用形態が決定した後に定まる定数である．

機械学習を OCR に適用する例を挙げる．この場合，リジェクト時には人間が文字認識結果を確認する．このときのコスト (リジェクトコスト) は，認識結果を確認する人間 (確認者) の人件費などに依存する．人件費を確定するためには，例えば，確認者を雇用する国などを確定する必要がある．そのため，確認者の雇用国などの運用形態の決定後に，初めてコスト定数が定まることになる．

機械学習を利用者個別にカスタマイズする場合は別であるが，一般には複数の利用者が利用可能な汎用システムとして機械学習が行われることが多い．こうした状況では，機械学習の評価時では利用者や利用者に依存する運用形態は確定しておらず，コスト定数を定めておくことはできない．そのため，本コスト曲線を用いた機械学習評価は現実的には使いにくいものとなる．コスト定数に影響されない評価曲線が望ましい．

3. 提 案 手 法

コスト定数に影響されない，かつ，確認修正コストを反映することのできる曲線の提案を行う．

3·1 確認修正コストに基づく評価尺度

曲線提案の準備として確認修正コストに基づく評価尺度を定義する．

§1 準 備

機械学習システムの目的の一つは人間業務の代替である．リジェクトモデルの目的も同様である．この文脈では，機械学習システムの評価尺度は人間作業の代替量を表現可能であることが望ましい．本節ではこの観点に基づく評価尺度を考察する．

リジェクトモデルにおいて，人間作業はリジェクトした場合に発生する．その場合に発生する業務は，下記の 2 種類に分けることができる．

- 確認: 出力ベクトル $\boldsymbol{v}$ が正解かどうかを確認する作業.
- 修正: 出力ベクトル $\boldsymbol{v}$ を修正する作業.

前章で示した ROC 曲線と PR 曲線は，これら確認修正工数を反映できない点が課題である．そこで，方式提案の準備としてリジェクトモデルにおける確認修正コストを定義する．以下では，確認修正コストを混同行列の諸指標を用いて表す.

まず，確認コストを算出するため，下記の (1) 式でアクセプト率 (Acceptance Rate)$X(k)$ を定義する:

$$X(k) = [T_A(k) + F_A(k)]/\Omega. \tag{1}$$

ただし，Ω は全事象数であり，下記で定義する:

$$\Omega = T_A(k) + F_A(k) + T_R(k) + F_R(k). \tag{2}$$

アクセプト率は全体の中でリジェクト器がアクセプトした割合である．RC 曲線における Coverage[El-Yaniv 10] と同じものである．リジェクト器がアクセプトした場合は，確認を実施しない．そのため，アクセプト率=「自動化率」と考えることができる．そこで，以下 $X(k)$ を自動化率とも称する．また，$X(k)$ を用いて，確認割合 $R_R(k)$ は下記のように算出できる:

$$R_R(k) = 1 - X(k). \tag{3}$$

次に，修正が必要となるのは，リジェクト器でリジェクトされ，かつ，誤っている場合である．修正割合 $R_C(k)$ は下記である:

$$R_C(k) = T_R(k)/\Omega. \tag{4}$$

さらに，全体における誤りの見逃し割合 (全体に対する，アクセプトかつ予測器の結果が誤りである割合) をエラー率 $R_E(k)$ とする．エラー率は下記となる:

$$R_E(k) = F_A(k)/\Omega. \tag{5}$$

加えて，出力ベクトル $\boldsymbol{v}$ の認識率 β(下記) を考えておく:

$$\beta = [T_A(k) + F_R(k)]/\Omega. \tag{6}$$

これは予測器の正解率と言い換えてもよい．上式の分母は総数であるため，閾値 k に依存しない．また分子は予測器が正しく予測できた数であるため，これも閾値 k に依存しない．そのため，この認識率 β は k に依存しない.

式 (2)(4)(5)(6) より，認識率 β，エラー率 $R_E(k)$，修正割合 $R_C(k)$ の関係は下記で表すことができる:

$$\beta + R_E(k) + R_C(k) = 1. \tag{7}$$

§2 評価尺度の定義

以上の準備を基に評価尺度の定義を行う．コスト定数として，リジェクト時確認コストを C_R，リジェクト時修正コストを C_C，誤りを見逃した場合のコスト (エラーコスト) を C_E とすると，確認修正コスト $L(k)$ は

$$L(k) = C_R R_R(k) + C_C R_C(k) + C_E R_E(k) \tag{8}$$

となる [Abbas 19, De Stefano 00, Kimura 17]．ただし，各コスト C_R，C_C，C_E は正の定数であり，この確認修正コストが小さい機械学習システムが「良い」システムである．ここで (3) 式を用いると，(8) 式は下記となる:

$$L(k) = C_R[1 - X(k)] + C_C R_C(k) + C_E R_E(k). \tag{9}$$

以下，(9) 式をできるだけ単純化する．まず，(7) 式を用いて $R_C(k)$ を消去する:

$$L(k) = C_R + C_C - C_R X(k) - C_C\beta + (C_E - C_C)R_E(k). \tag{10}$$

次に，(10) 式において，$C_R + C_C$ を左辺に移項し，さらに，C_C で除算すると下記となる:

$$\frac{1}{C_C}[L(k) - (C_R + C_C)] = -\frac{C_R}{C_C}X(k) - \beta + \left(\frac{C_E}{C_C} - 1\right)R_E(k). \tag{11}$$

ここで，(11) 式の左辺を $L'(k)$ とし，$L'(k)$ を確認修正コストとして新たに定義する:

$$L'(k) = \frac{1}{C_C}[L(k) - (C_R + C_C)]. \tag{12}$$

また,

$$\gamma = C_R/C_C, \delta = C_E/C_C - 1 \tag{13}$$

とすると，(12) 式，(13) 式より，(11) 式は下記となる:

$$L'(k) = -\gamma X(k) - \beta + \delta R_E(k). \tag{14}$$

(14) 式に示されるように，確認修正コストを，自動化率 $X(k)$，認識率 β，エラー率 $R_E(k)$，および，コストの比で定義される k に依存しない定数 γ，δ の式として単純化できる.

ここで，確認修正コストに基づく機械学習評価尺度 $W(k)$ を考える．確認修正コストは小さいほうが良い値となるが，評価尺度 $W(k)$ としては大きいほうが良い値とするほうが分かりやすい．そこで，(14) 式の正負を反転した値を $W(k)$ とする:

$$W(k) = \gamma X(k) + \beta - \delta R_E(k). \tag{15}$$

(15) 式に示されるように，評価尺度 $W(k)$ は，自動化率 $X(k)$，認識率 β，エラー率 $R_E(k)$ の評価尺度の和か

ら構成されている．自動化率は高いほうが良い．認識率は高いほうが良い．エラー率は低いほうが良いという一般的な結果を示していることを確認することができる．エラー率に関して補足しておく．見逃した場合のコスト C_E のほうが，機械学習システム内での修正コスト C_C よりも通常は大きい．そのため，一般的には，$C_E \geq C_C$，よって，$\delta \geq 0$ となる．このとき，(15) 式ではエラー率は低いほうが良いことになる．

3·2 ARAC 曲線 (ARAC curve)

§1 ARAC 曲線の提案

前節で構築した評価尺度 $W(k)$ は k の関数であるため，k によって値が変化し，機械学習の評価を簡易に実施することができない．そこで，閾値 k に依存しない評価を実施するため，横軸に自動化率 $X(k)$ をプロットする曲線を考える．

さらに，認識率 β とエラー率 $R_E(k)$ を表現するため，縦軸には下記の $Y(k)$ を導入する:

$$Y(k) = [T_A(k) + T_R(k) + F_R(k)]/\Omega. \tag{16}$$

$Y(k)$ は確認修正した後の修正後正解率 (Accuracy after Correction) を表す数値である．上記の定義より，縦軸 $Y(k)$ はエラー率 $R_E(k)$ を $1-Y(k)$ として表現可能である．また，$k=0$ のときにはすべての結果をアクセプトすることになるため，$Y(0)=\beta$ が成り立つ．そのため，縦軸 $Y(k)$ は β も表現可能である．

本曲線 (横軸をアクセプト率 $X(k)$，縦軸を修正後正解率 $Y(k)$ とする曲線) を，アクセプト率-修正後正解率曲線 (Acceptance Rate-Accuracy after Correction curve) と称する．以下では，ARAC 曲線 (ARAC curve) と略して記載する．本 ARAC 曲線を用いることで，ROC 曲線，PR 曲線，および RC 曲線と同様，閾値 k に依存しない評価ができる．

ARAC 曲線において，横軸のアクセプト率 (自動化率)，縦軸の修正後正解率共に，大きい値であるほど良好な評価値となる尺度である．そのため，ARAC 曲線は右上に行くほど良い評価となる曲線である．この曲線を描くことにより，確認修正コスト (自動化率，認識率，エラー率) に基づく評価尺度を視覚的に判断できる．ARAC 曲線の例を図 3 に示す．

この ARAC 曲線を描画するには，表 1 に示した混同行列の情報だけが必要である．そのため，コスト定数を利用せずに描画可能という特徴を持つ．また，前述のように ARAC 曲線は右上に行くほど良い評価となる曲線である．

§2 許容エラー率が与えられた場合の評価手法

以下，ARAC 曲線を用いた評価尺度 W の取得方法を見ていく．ARAC 曲線が描画されているグラフの横軸を X 軸，縦軸を Y 軸とする．ARAC 曲線は，XY 平面において，$(0,0)$〜$(1,1)$ の矩形領域で描画される．

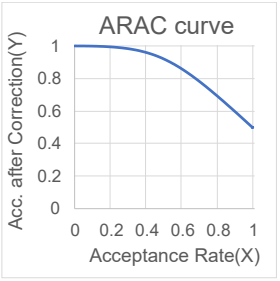


図 3: ARAC 曲線．横軸にアクセプト率 (自動化率)，縦軸に修正後正解率を閾値 k を変化させてプロットした曲線．

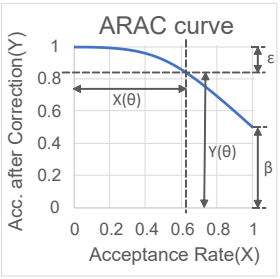


図 4: ARAC 曲線を用いた評価尺度算出．ARAC 曲線を用いて許容エラー率 ε から自動化率 $X(\theta)$ と，認識率 β を求めることができる．

機械学習システムの運用として，最初にシステムの許容エラー率 ε を定め，それに基づいて閾値などのパラメタ調整を実施する場合がある．本項では，許容エラー率が与えられた場合の，ARAC 曲線を用いた評価尺度取得方法を示す．

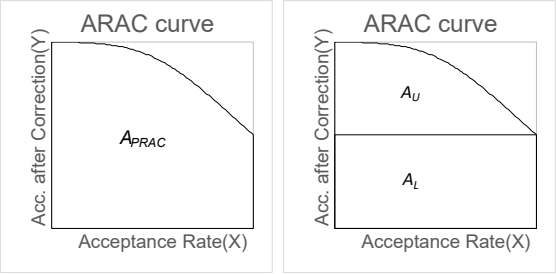
まず，エラー率 ε を満たす閾値を θ，すなわち，$R_E(\theta) = \varepsilon$ とする．$\varepsilon = R_E(\theta) = 1 - Y(\theta)$ であるため，図 4 に示すように，$Y = 1 - \varepsilon$ となる点を ARAC 曲線上で探索し，その点の X 座標値を求めることで，閾値 θ に対応するアクセプト率 $X(\theta)$ を求める．ARAC 曲線を用いて，閾値 θ を介さずに ε から $X(\theta)$ を直接求めることができる．

次に，認識率 β は $k=0$ としたときの修正後正解率 $Y(0)$ に対応している．これは，ARAC 曲線が $X=1$ と交わる点をグラフから取得することで得ることができる．曲線上の β を図 4 に示す．

以上で，$X(\theta)$，β が取得できたため，許容エラー率 ε を設定した場合の評価値 $W(\theta)$ は，(15) 式を用いることで $W(\theta) = \gamma X(\theta) + \beta - \delta\,\varepsilon$ のように算出することができる．

3·3 ARAC-AUC

ARAC 曲線 (ARAC curve) の下部面積を用いた機械学習評価尺度 ARAC-AUC，および，ARAC-AUC を改良した正規化 ARAC-AUC を提案する．



(a) ARAC-AUC(A_{ARAC})　(b) ARAC-AUC の上部領域 (A_U) と下部領域 (A_L)

図 5: ARAC-AUC とその分割．(a)ARAC-AUC: ARAC 曲線 (ARAC curve) の下側の領域の面積が ARAC-AUC である．(b)ARAC-AUC の分割: 上部領域面積 A_U は，平均自動化率に対応する．下部領域面積 A_L は，認識率 β に対応する．

§1　ARAC-AUC の定義

ARAC 曲線においても，ROC 曲線や PR 曲線と同様に曲線下部の面積として，AUC(ARAC-AUC) を定義する．ARAC-AUC を用いて閾値 k に依存しない単一値を用いた評価を実施することができる．ARAC-AUC の様子を図 5(a) に示す．以下，ARAC-AUC の数値を A_{ARAC} として示す．

§2　正規化 ARAC-AUC 算出の準備 1(AUC の領域分割)

次に，ARAC-AUC を改良した正規化 ARAC-AUC の考察を行う．まず，正規化 ARAC-AUC 算出の準備として，ARAC-AUC を図 5(b) のように 2 つの領域に分離する．

最初に，ARAC-AUC の $Y < \beta$ の領域 (図 5(b) の下部領域) を考える．この領域は縦 β，横 1 の長方形であるため，面積は $A_L = \beta$ である．

次に，ARAC-AUC の $Y \geq \beta$ の領域 (図 5(b) の上部領域)A_U を考えると，ARAC-AUC(A_{ARAC}) は

$$A_{ARAC} = A_U + A_L \tag{17}$$

と表すことができる．ただし，

$$A_U = \int_{\beta}^{1} X(k)dY, A_L = \beta \tag{18}$$

である．面積 A_U は ARAC 曲線を用いて数値積分で求める．また，β は ARAC 曲線と $X = 1$ の交点から求めることができる．

§3　正規化 ARAC-AUC 算出の準備 2(AUC に関する考察)

ここで，ROC-AUC や PR-AUC の持つ意味を考察する．ROC-AUC は平均真陽性率あるいは平均偽陽性率を算出したものと考えることができる．また，PR-AUC は平均適合率を算出したものといえる．

真陽性率，偽陽性率，適合率等の各指標は閾値 k の関数であるため，単純に考えると閾値 k を積分変数として積分することで平均値を算出することができる．閾値 k を積分変数としない理由は，確信度 t(および確信度 t の閾値 k) が分類器によって異なる非線形の挙動を示すためである．分類器によって異なる非線形な座標軸上での積分結果は共通な指標として採用することができない．一方，真陽性率，偽陽性率，再現率等は，分類器に共通な指標として使える．つまり，分類器共通に利用できる，真陽性率，偽陽性率，再現率などの指標を軸として積分し，平均値 (面積) を取得したものが AUC であるということができる．

§4　改良 ARAC-AUC

上記の考察を基に，ARAC-AUC の改良を実施する．

まず，(18) 式において，A_U の積分範囲が 1〜β であるため，A_U の値が認識率 β に依存してしまう点が問題である．そこで，A_U を $(1-\beta)$ で割って正規化する．

次に，A_U は自動化率 $X(k)$ の平均値 (に関連する値) を示しているが，自動化率 $X(k)$ と認識率 β それぞれの評価尺度 $W(k)$ における重みを反映していない点が問題である．そこで，(15) 式を用いて，評価尺度 $W(k)$ における自動化率 $X(k)$ と認識率 β の重み係数 γ を導入する．具体的には，平均自動化率 A_U に重み係数 γ を乗じることで，重みを反映させることができる．

上記の考えを用いて，改良 ARAC-AUC(A'_{ARAC}) を下記のように定義する:

$$A'_{ARAC} = \frac{\gamma}{1-\beta}A_U + A_L. \tag{19}$$

ただし，$\beta = 1$ のときは，$A'_{ARAC} = \gamma + 1$ とする．

ここで，ARAC-AUC を示す (17) 式と，改良 ARAC-AUC を示す (19) 式を比較する．両者の違いは A_U の係数である．ARAC-AUC((17) 式) の場合は，β が 1 に近い値の場合，A_U の面積が小さい値となってしまい，自動化率が AUC に正しく反映されない．一方，改良 ARAC-AUC((19) 式) の場合は，$(1-\beta)$ で除算して正規化することにより，β の値に依存せず，一定の割合で自動化率が AUC に反映される．つまり，改良 ARAC-AUC は，ARAC-AUC と比較して，確認修正コストをより忠実に反映する評価指標であるといえる．

§5　正規化 ARAC-AUC(改良 ARAC-AUC の正規化)

一般に，AUC の利点として最大値が 1 となる点が挙げられる．相対的な比較を実施することなく，値を見ただけでモデルの良し悪しが判断できるためである．ARAC-AUC もこの性質 (最大値が 1) を持つ．

そこで，本項では改良 ARAC-AUC を正規化した正規化 ARAC-AUC を定義する．正規化 ARAC-AUC では，最大値が 1 となり，一般的な AUC と同様の利点を得ることができる．

まず，A_U は，$Y \geq \beta$ の領域の面積であることより，A_U の最大値は $1-\beta$ である．$A_U = 1-\beta$ と $A_L = \beta$ を

(19) 式に代入することで，A'_{ARAC} の最大値を $\gamma+\beta$ と求めることができる．

さらに，認識率 β の値域が，$0 \leq \beta \leq 1$ であることを考慮すると，A'_{ARAC} の最大値は $\gamma+1$ となる．改良 ARAC-AUC を本最大値で除算し正規化することで，下記の正規化した改良 ARAC-AUC(A''_{ARAC}) を得る:

$$A''_{ARAC} = \frac{\frac{\gamma}{1-\beta}A_U + A_L}{\gamma+1}. \tag{20}$$

ただし，$\beta=1$ のときは，$A''_{ARAC}=1$ とする．

上記の (20) 式で定義した A''_{ARAC} を,「正規化 ARAC-AUC」と称する．

3·4 RC 曲線との比較

ARAC 曲線と RC 曲線 [El-Yaniv 10] の比較を行う．RC 曲線は縦軸が小さいほど高い評価となる曲線であるため ARAC 曲線との比較がし難い．そこで，ここではまず RC 曲線の縦軸の上下を反転した曲線を，アクセプト率-適合率曲線 (Acceptance Rate-Precision curve: ARP curve，ARP 曲線) と定義する．ARP 曲線は横軸がアクセプト率，縦軸が適合率の曲線であり，右上に行くほど高い評価となるため，ARAC 曲線との比較が容易となる．本 ARP 曲線は，[MHLW 18] で提示されている曲線と同じものである．以下，比較を実施する．

ARP 曲線は下記の利点を ARAC 曲線と同様に持つ．

- 曲線から自動化率を取得することができる．
- AUC(ARP-AUC) が定義可能である．
- 横軸は ARAC 曲線と同じである．また，$k=0$ のときの縦軸の適合率 $R_P(0)=\beta$ より，$X=1$ の直線と ARP 曲線との交点が (1, β) となる点も同じである．そのため，正規化 ARAC-AUC と同様に正規化 ARP-AUC を定義することができる．

ARAC 曲線との比較上，ARP 曲線は下記の欠点を持つ．

- ARP 曲線は，自動化率が低い領域では適合率算出時の分母の値が小さくなるため，適合率値が不安定となる．そのため，ARP 曲線では曲線が不安定となる．
- ARP 曲線は，縦軸の適合率が許容エラー率とは直接関係していない．そのため，許容エラー率を用いて自動化率を直接取得することはできない．また，AUC は確認修正コストと直接には関連していない．

3·5 各曲線の定義まとめ

ここで，表 2 に各曲線の定義をまとめておく．表では各曲線の縦軸および横軸を混同行列の指標を用いて統一的に表現した．

4. 実　　験

提案手法の効果を実験で確かめる．以下，本章においては，$\gamma=1, \delta=0$ とした．

表 2: 各曲線の定義まとめ．ARAC 曲線，ROC 曲線，PR 曲線，RC 曲線，ARP 曲線の定義を混同行列の指標を用いて示す．

Name	Horizontal Axis	Vertical Axis
ARAC	$\frac{T_A(k)+F_A(k)}{\Omega}$	$\frac{T_A(k)+T_R(k)+F_R(k)}{\Omega}$
ROC	$\frac{F_A(k)}{T_R(k)+F_A(k)}$	$\frac{T_A(k)}{T_A(k)+F_R(k)}$
PR	$\frac{T_A(k)}{T_A(k)+F_R(k)}$	$\frac{T_A(k)}{T_A(k)+F_A(k)}$
RC	$\frac{T_A(k)+F_A(k)}{\Omega}$	$\frac{F_A(k)}{T_A(k)+F_A(k)}$
ARP	$\frac{T_A(k)+F_A(k)}{\Omega}$	$\frac{T_A(k)}{T_A(k)+F_A(k)}$

表 3: 入力ケース (実験 1-1)．予測器出力データの正予測分布と誤予測分布のセットを 3 ケース用意した．

Case No.	ω_P	μ_P	σ_P	ω_N	μ_N	σ_N
1	0.75	0.6	0.1	0.25	0.4	0.1
2	0.25	0.6	0.1	0.75	0.4	0.1
3	0.25	0.7	0.1	0.75	0.3	0.1

4·1 人工データを用いた比較

予測器が出力するデータにおいて，特定のケースでは ARAC 曲線の効果が高くなる．そのようなケースを疑似的に生成して効果を示す．ここでは，確信度 t における予測器の出力データの確率密度関数 $f(t)$ を下記のように作成した:

$$f(t) = \omega_P f_P(t) + \omega_N f_N(t).$$

ここで，ω_P と ω_N はそれぞれ予測器における正予測データと誤予測データの比率を制御する変数である ($\omega_P+\omega_N=1$ とする)．また，t の定義域は $t\in[0,1]$ であり，$f_P(t)$，$f_N(t)$ はそれぞれ正予測データと誤予測データの確率密度関数 (下記) である:

- $f_P(t) = \mathcal{N}(t;\mu_P,\sigma_P)/\int_0^1 \mathcal{N}(x;\mu_P,\sigma_P)dx$,
- $f_N(t) = \mathcal{N}(t;\mu_N,\sigma_N)/\int_0^1 \mathcal{N}(x;\mu_N,\sigma_N)dx$.

ただし，$\mathcal{N}(x;\mu,\sigma)$ は，変数 x，平均 μ，標準偏差 σ の正規分布の確率密度関数 (下記) とした:

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

§1 実験 1-1

表 3 に示す 3 つのケースにおける実験を実施した．表 4 に実験結果を示す．表 4 には確認修正コストに基づく評価尺度として，$\varepsilon=0.01$ および 0.05 の場合の $W(\theta)$ を付記する．$W(\theta)$ 値は 3·2 節に示した方法で取得したものである．

本実験の結果，下記が分かる．以下，不等号は各実験ケースの番号を示す．

- 確認修正コストに基づく評価尺度の順は，$1>3>2$ である ($\varepsilon=0.01, 0.05$ 共に).

表 4: AUC 算出結果 (実験 1-1)．ケースごとに ROC-AUC, PR-AUC, ARAC-AUC, 正規化 ARAC-AUC を算出した結果を示す．ε =0.01, 0.05 の場合の評価尺度を付記する．

Case No.	ROC-AUC	PR-AUC	ARAC-AUC	正規化 ARAC-AUC	$W(\theta)$ ε = 0.01	$W(\theta)$ ε = 0.05
1	0.921	0.970	0.954	0.783	1.21	1.46
2	0.921	0.822	0.704	0.423	0.36	0.47
3	0.998	0.994	0.718	0.437	0.50	0.55

表 5: 正規化 ARAC-AUC 効果確認用入力ケース (実験 1-2)．予測器出力データの正予測分布と誤予測分布のセットを 2 ケース用意した．

Case No.	ω_P	μ_P	σ_P	ω_N	μ_N	σ_N
4	0.95	0.55	0.1	0.05	0.45	0.1
5	0.95	0.7	0.1	0.05	0.3	0.1

- ARAC-AUC および正規化 ARAC-AUC の順は，$1 > 3 > 2$ であり，評価尺度 $W(\theta)$ の順 $1 > 3 > 2$ と等しい．
- ROC-AUC の順は，$3 > 1 = 2$ であり，評価尺度 $W(\theta)$ の順 $1 > 3 > 2$ とは異なる．
- PR-AUC の順は，$3 > 1 > 2$ であり，評価尺度 $W(\theta)$ の順 $1 > 3 > 2$ とは異なる．

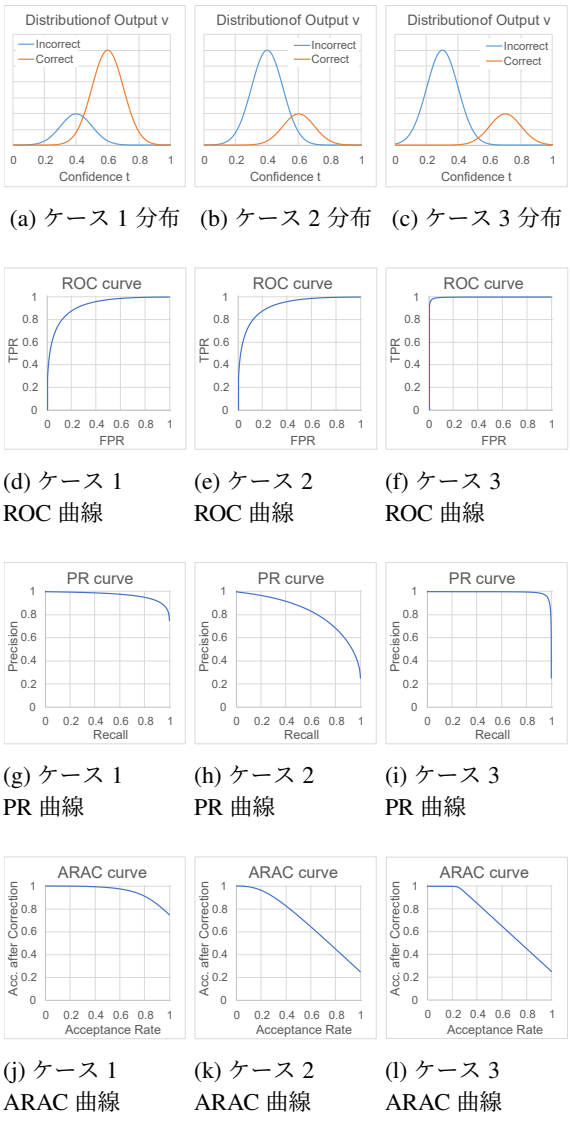
以上より，ARAC-AUC および正規化 ARAC-AUC は，ROC-AUC，PR-AUC と比較して確認修正コストに基づく評価尺度をより忠実に反映できる場合があることが分かる．さらに，図 6 に入力分布の形状と各曲線の形状を示す．図 6 から下記が分かる．

- 図 6(d) と (e) に示すように，ROC 曲線は予測器の認識率が変化した場合 (正予測と誤予測の比率が変化した場合) でも，形状が変わらない．本論文で対象としているような評価には使えない．
- PR 曲線は ARAC 曲線に近い挙動を示す．しかしながら，図 6(i) に示すように，予測器で正予測したデータの比率が小さく認識率が悪い場合でも，曲線の形状が極端に右上に張り付くような場合 (PR 曲線としては高い評価となってしまう場合) があり得る．このようなケースがあるため，本論文で対象としているような評価には使えない．
- ARAC 曲線は右上に行くほど高い評価となる．AUC を用いることなく，ARAC 曲線の形状を比較することで $1 > 3 > 2$ の順序を取得することも可能である．

§2 実験 1-2

次に，β が大きい場合 ($\beta = 0.95$) を例に採り，正規化 ARAC-AUC の効用を示す．下記 2 ケースでの実験を実施した．表 5 には入力データ内容を示す．表 6 に実験結果を示す．

ケース 4 とケース 5 に対する評価尺度 $W(\theta)$ の差は大きいにもかかわらず，ARAC-AUC の差はほとんどない．一方，正規化 ARAC-AUC では，実際の評価尺度のケース 4 とケース 5 の差をより良く表しているといえる．図 7 に，入力分布と各曲線の形状を示しておく．図 7(g) と (h) を比較すると，入力分布が異なっていても，ARAC 曲線の違いはグラフ上ではほとんどない．このような場合は正規化 ARAC-AUC 算出の効用があると考えられる．



(a) ケース 1 分布 (b) ケース 2 分布 (c) ケース 3 分布

(d) ケース 1 ROC 曲線 (e) ケース 2 ROC 曲線 (f) ケース 3 ROC 曲線

(g) ケース 1 PR 曲線 (h) ケース 2 PR 曲線 (i) ケース 3 PR 曲線

(j) ケース 1 ARAC 曲線 (k) ケース 2 ARAC 曲線 (l) ケース 3 ARAC 曲線

図 6: 実験 1-1 結果．正予測 (Correct) および誤予測 (Incorrect) データの分布が異なる 3 ケースの実験を実施．各ケースに対する ROC 曲線，PR 曲線，ARAC 曲線を示す．

4.2 機械学習モデル選択実験

具体的な ARAC 曲線を用いて機械学習モデル選択を実施する実験 (実験 2) を実施した．ここでは，model1, model2, model3, model4 の 4 種類の機械学習モデルの選択実験を行う．入力が画像であり出力が文字コードと確信度である OCR を予測器として用いた．各機械学習モデルは 4 種類の学習結果に対応する．前節の人工データでは ARAC 曲線と ARP 曲線の形状の違いがほとんど出な

表 6: AUC 算出結果 (実験 1-2)．ケースごとに ROC-AUC, PR-AUC, ARAC-AUC, 正規化 ARAC-AUC を算出した結果を示す．ε =0.01, 0.02 の場合の評価尺度を付記する．

Case No.	ROC-AUC	PR-AUC	ARAC-AUC	正規化 ARAC-AUC	$W(\theta)$ ε = 0.01	$W(\theta)$ ε = 0.02
4	0.760	0.982	0.987	0.849	1.49	1.70
5	0.998	0.9999	0.999	0.961	1.91	1.92

表 7: AUC 算出結果 (実験 2)．model ごとに ARAC-AUC，正規化 ARAC-AUC，ARP-AUC，正規化 ARP-AUC，ROC-AUC，PR-AUC を算出した結果を示す．

No.	ARAC-AUC	正規化 ARAC-AUC	ARP-AUC	正規化 ARP-AUC	ROC-AUC	PR-AUC
1	0.9876	0.8575	0.9820	0.8052	0.7840	0.9828
2	0.9956	0.9317	0.9953	0.9291	0.9407	0.9966
3	0.9579	0.7941	0.9421	0.7521	0.8409	0.9541
4	0.9717	0.8307	0.9675	0.8196	0.9311	0.9843

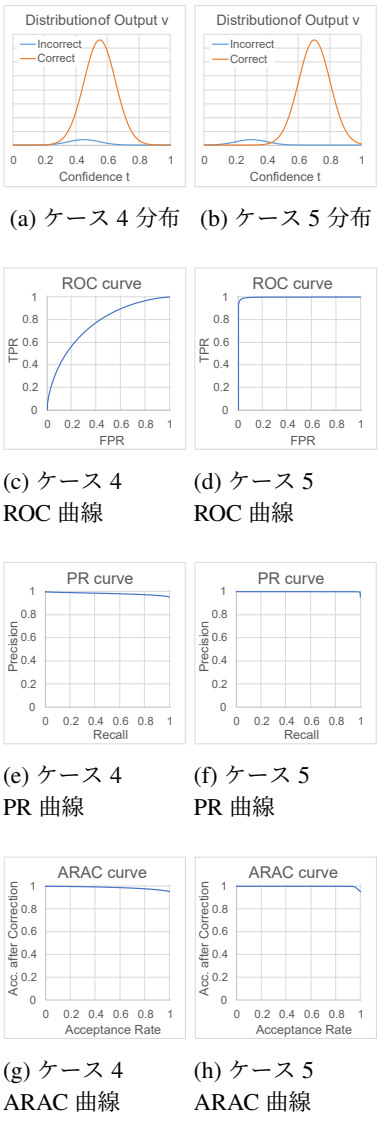
いため，ARP 曲線の提示を実施しなかったが，本節 (実データに対する実験) においては ARP 曲線との比較を行う．加えて，ROC 曲線と PR 曲線との比較も行う．

システムの許容エラー率が与えられている場合，あるいは，閾値 k が与えられている場合は，(15) 式を用いて評価尺度を算出し，最も高い値となったモデルを選択すればよい．しかしながら，機械学習システムとしては，許容エラー率や閾値などのパラメタは都度設定するケースがある．このような許容エラー率や閾値が与えられていない場合に対しても，予めより良いモデル選択を実施したい．そのために各種曲線，あるいは，各種 AUC を用いた評価の意味がある．

まず，図 8 を用いて，ARAC 曲線を目視で確認する．以下，不等式は model 番号を示す．明らかに，$2 > 1 > 3$, および，$2 > 4 > 3$ であることが分かる．問題は，model1 と model4 の比較である．そこで AUC を用いた比較を実施する．表 7 に AUC の比較結果を示す．ARAC-AUC の比較では model1 のほうが model4 より評価値が良い．本結果より，予め許容エラー率が規定されていない場合，機械学習モデルとしては，$2 > 1 > 4 > 3$ の順で選択すれば良いと判断する．

ここで，図 8(b), (d) を参照して ARP 曲線と PR 曲線の形状を見る．これらの曲線では縦軸に適合率を採用している．適合率の定義では，分母は全アクセプト数 $[T_A(k) + F_A(k)]$ である．ARP 曲線と PR 曲線のグラフの横軸が小さい値のとき (アクセプト率が低いとき)，分母が小さい値となるため適合率値は不安定となる．実際，図 8(b), (d) に示されるように，ARP 曲線と PR 曲線の形状がいびつである．このような不安定な曲線を基に算出する AUC の値は不安定となる．

次に図 8(c) と表 7 を参照して ROC 曲線の考察を行う．ROC 曲線，および，ROC-AUC では，model2 と model4 の評価値が高い．ROC 曲線が認識率 β を全く考慮しないため，認識率が極めて低いモデルであっても選択してしまう危険性がある．本論文で対象としているようなリジェクトモデルの評価には ROC 曲線は用いることができないことがわかる．



(a) ケース 4 分布　(b) ケース 5 分布

(c) ケース 4 ROC 曲線　(d) ケース 5 ROC 曲線

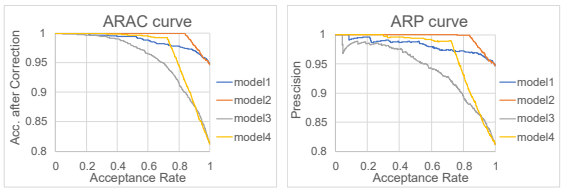
(e) ケース 4 PR 曲線　(f) ケース 5 PR 曲線

(g) ケース 4 ARAC 曲線　(h) ケース 5 ARAC 曲線

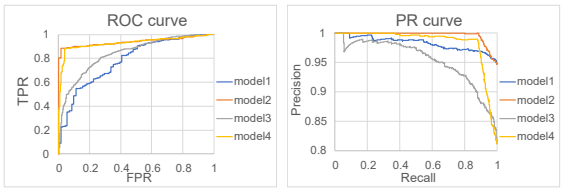
図 7: 実験 1-2 結果．正予測 (Correct) および誤予測 (Incorrect) データの分布が異なる 2 ケースの実験を実施．各ケースに対する ROC 曲線，PR 曲線，ARAC 曲線を示す．

5. ま と め

本論文では，リジェクトモデルに適用した場合の，確認修正コストに基づく機械学習評価手法の提案を行った．ROC 曲線，PR 曲線，RC 曲線などの従来手法では，許容エラー率及び確認修正コストを基準とした評価を実施できない点が課題となっていた．確認修正コストを反映するため，まず，確認修正コストに基づく機械学習評価尺度を定義した．次に，アクセプト率 (自動化率) を横軸と



(a) ARAC 曲線 (b) ARP 曲線



(c) ROC 曲線 (d) PR 曲線

図 8: 実験 2 結果．4 種の機械学習モデル model1〜model4 に対応する ARAC 曲線，ARP 曲線，ROC 曲線，PR 曲線を示す．

し，修正後正解率を縦軸としてプロットするアクセプト率-修正後正解率曲線 (ARAC 曲線) の提案を行った．本 ARAC 曲線を利用して，許容エラー率が与えられた場合の評価尺度取得方法を示した．さらに，ARAC 曲線の面積を用いた機械学習評価尺度 ARAC-AUC，および，正規化 ARAC-AUC を提案した．

この ARAC 曲線を用いて確認修正コストに基づいた機械学習評価を実施することができるが，ARAC 曲線の描画にはコスト定数は不要である．そのため，ARAC 曲線は，機械学習の運用形態の決定前に描画可能という特徴を持つ．

ARAC 曲線に関わる本論文の貢献内容は次に示す三点である．第一に，今回混同行列に基づき，ARAC 曲線を定義および提案した．混同行列の指標を用いることにより ROC 曲線，PR 曲線，RC 曲線との差異を明確とした．第二に確認修正コストに基づく機械学習評価尺度を定義し，ARAC 曲線と確認修正コストに基づく機械学習評価尺度との関連を示した．また，ARAC 曲線を用いた機械学習評価尺度取得方法を示した．第三に ARAC 曲線に関して AUC(ARAC-AUC) を定義した．また，確認修正コストをより忠実に反映する正規化 ARAC-AUC を提案した．

最後に提案手法の効果を実験で確かめた．ARAC-AUC および正規化 ARAC-AUC は，ROC-AUC，PR-AUC と比較して確認修正コストに基づく評価尺度をより忠実に反映するケースが存在することを確認した．また，提案手法を用いて実際の機械学習モデルの評価を行い，機械学習モデル選択が安定的に可能であることを示した．

本論文で提案した ARAC 曲線は，確信度に依存して処理方法を選択するシステムであれば適用可能である．閾値などの確信度の利用方法を予め明確に定めることができない時点で，システム全体，学習済みモデル，パラメタなどの評価を実施するときに効果を発揮する．確信度を出力する機械学習システムの評価方法における，ARAC 曲線の具体的な利用方法の検討が今後の課題である．

◇ 参 考 文 献 ◇

[Abbas 19] Abbas, M. R., Nadeem, M. S. A., Shaheen, A., Alshdadi, A. A., Alharbey, R., Shim, S.-O., and Aziz, W.: Accuracy rejection normalized-cost curves (ARNCCs): A novel 3-dimensional framework for robust classification, *IEEE Access*, Vol. 7, pp. 160125–160143 (2019)

[Bandos 06] Bandos, A. I., Rockette, H. E., and Gur, D.: Resampling methods for the area under the ROC curve, *ROC Analysis in Machine Learning*, pp. 1–8 (2006)

[Bradley 97] Bradley, A. P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, Vol. 30, No. 7, pp. 1145–1159 (1997)

[Chow 70] Chow, C.: On optimum recognition error and reject tradeoff, *IEEE Transactions on Information Theory*, Vol. 16, No. 1, pp. 41–46 (1970)

[Condessa 16] Condessa, F., Bioucas-Dias, J., and Kovačević, J.: Supervised hyperspectral image classification with rejection, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 9, No. 6, pp. 2321–2332 (2016)

[Davis 06] Davis, J. and Goadrich, M.: The relationship between precision-recall and ROC curves, in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240 (2006)

[De Stefano 00] De Stefano, C., Sansone, C., and Vento, M.: To reject or not to reject: That is the question-an answer in case of neural classifiers, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 30, No. 1, pp. 84–94 (2000)

[De Stefano 14] De Stefano, C., Fontanella, F., Marcelli, A., Parziale, A., and Di Freca, A. S.: Rejecting both segmentation and classification errors in handwritten form processing, in *2014 14th International Conference on Frontiers in Handwriting Recognition*, pp. 569–574 (2014)

[Eban 17] Eban, E., Schain, M., Mackey, A., Gordon, A., Rifkin, R., and Elidan, G.: Scalable learning of non-decomposable objectives, in *Artificial Intelligence and Statistics*, pp. 832–840 (2017)

[El-Yaniv 10] El-Yaniv, R., et al.: On the foundations of noise-free selective classification, *Journal of Machine Learning Research*, Vol. 11, No. 5 (2010)

[Fumera 00] Fumera, G., Roli, F., and Giacinto, G.: Reject option with multiple thresholds, *Pattern Recognition*, Vol. 33, No. 12, pp. 2099–2101 (2000)

[Fumera 02] Fumera, G. and Roli, F.: Support vector machines with embedded reject option, in *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002 Niagara Falls, Canada*, pp. 68–82 (2002)

[Fumera 03] Fumera, G., Pillai, I., and Roli, F.: Classification with reject option in text categorisation systems, in *12th International Conference on Image Analysis and Processing*, pp. 582–587 (2003)

[Fumera 04] Fumera, G., Pillai, I., and Roli, F.: A two-stage classifier with reject option for text categorisation, in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, SSPR 2004 and SPR 2004*, pp. 771–779 (2004)

[Geifman 17] Geifman, Y. and El-Yaniv, R.: Selective classification for deep neural networks, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4878–4887 (2017)

[Giacinto 00] Giacinto, G., Roli, F., and Bruzzone, L.: Combination of neural and statistical algorithms for supervised classification of remote-sensing images, *Pattern Recognition Letters*, Vol. 21, No. 5, pp. 385–397 (2000)

[Hanczar 08] Hanczar, B. and Dougherty, E. R.: Classification with reject option in gene expression data, *Bioinformatics*, Vol. 24, No. 17, pp. 1889–1895 (2008)

[Hanczar 19] Hanczar, B.: Performance visualization spaces for classification with rejection option, *Pattern Recognition*, Vol. 96, p.

106984 (2019)
[Hendrickx 24] Hendrickx, K., Perini, L., Plas, Van der D., Meert, W., and Davis, J.: Machine learning with a reject option: A survey, *Machine Learning*, Vol. 113, No. 5, pp. 3073–3110 (2024)
[Hogan 23] Hogan, J. and Adams, N. M.: On averaging ROC curves, *Transactions on Machine Learning Research* (2023)
[Huang 20] Huang, L., Zhang, C., and Zhang, H.: Self-adaptive training: beyond empirical risk minimization, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 19365–19376 (2020)
[Huber 05] Huber, R., Ramoser, H., Mayer, K., Penz, H., and Rubik, M.: Classification of coins using an eigenspace approach, *Pattern Recognition Letters*, Vol. 26, No. 1, pp. 61–75 (2005)
[Kimura 17] Kimura, S., Tanaka, E., Sekino, M., Sakurai, T., Kubota, S., So, I., and Koshi, Y.: A man-machine cooperating system based on the generalized reject model, in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1, pp. 1324–1329 (2017)
[Landgrebe 06] Landgrebe, T. C., Tax, D. M., Paclík, P., and Duin, R. P.: The interaction between classification and reject performance for distance-based reject-option classifiers, *Pattern Recognition Letters*, Vol. 27, No. 8, pp. 908–917 (2006)
[Le Capitaine 14] Le Capitaine, H.: A unified view of class-selection with probabilistic classifiers, *Pattern Recognition*, Vol. 47, No. 2, pp. 843–853 (2014)
[MHLW 18] MHLW(厚生労働省)：人工知能を活用した副作用症例報告の評価支援の基盤整備と試行的評価 (2018),（別添 4）H30 年度 ICT 研究分担報告書, 厚生労働科学研究成果データベース, 文献番号 201803015A, https://mhlw-grants.niph.go.jp/project/26986
[Nadeem 09] Nadeem, M. S. A., Zucker, J.-D., and Hanczar, B.: Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option, in *Machine Learning in Systems Biology*, pp. 65–81 (2009)
[Navarro-Cerdan 15] Navarro-Cerdan, J. R., Arlandis, J., Llobet, R., and Perez-Cortes, J.-C.: Batch-adaptive rejection threshold estimation with application to OCR post-processing, *Expert Systems with Applications*, Vol. 42, No. 21, pp. 8111–8122 (2015)
[Pietraszek 07] Pietraszek, T.: On the use of ROC analysis for the optimization of abstaining classifiers, *Machine Learning*, Vol. 68, No. 2, pp. 137–169 (2007)
[Pugnana 23] Pugnana, A. and Ruggieri, S.: AUC-based selective classification, in *International Conference on Artificial Intelligence and Statistics*, pp. 2494–2514 (2023)
[Quevedo 11] Quevedo, J. R., Bahamonde, A., Pérez-Enciso, M., and Luaces, O.: Disease liability prediction from large scale genotyping data using classifiers with a reject option, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, No. 1, pp. 88–97 (2011)
[Sotgiu 20] Sotgiu, A., Demontis, A., Melis, M., Biggio, B., Fumera, G., Feng, X., and Roli, F.: Deep neural rejection against adversarial examples, *EURASIP Journal on Information Security*, Vol. 2020, pp. 1–10 (2020)
[Sousa 13] Sousa, R. and Cardoso, J. S.: The data replication method for the classification with reject option, *AI Communications*, Vol. 26, No. 3, pp. 281–302 (2013)
[Tang 14] Tang, W. and Sazonov, E. S.: Highly accurate recognition of human postures and activities through classification with rejection, *IEEE Journal of Biomedical and Health Informatics*, Vol. 18, No. 1, pp. 309–315 (2014)
[Tortorella 05] Tortorella, F.: A ROC-based reject rule for dichotomizers, *Pattern Recognition Letters*, Vol. 26, No. 2, pp. 167–180 (2005)
[Zhang 12] Zhang, Y., Zhang, B., Coenenz, F., and Lu, W.: Highly reliable breast cancer diagnosis with cascaded ensemble classifiers, in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8IEEE (2012)
[Zhou 22] Zhou, L., Martínez-Plumed, F., Hernández-Orallo, J., Ferri, C., and Schellaert, W.: Reject before you run: Small assessors anticipate big language models, in *EBeM@ IJCAI* (2022)

〔担当委員：松井 考太〕

2024 年 10 月 15 日 受理

著 者 紹 介

木村 俊一

1989 年東京大学工学部卒業．1991 年同大学院工学系研究科修士課程修了．2003 年カリフォルニア大学サンタバーバラ校大学院修士課程修了．1991 年富士ゼロックス株式会社 (現 富士フイルムビジネスイノベーション株式会社) 入社．画像処理および文書処理の研究開発に従事．映像情報メディア学会優秀研究発表賞受賞．

久保田 聡

1992 年静岡大学工学部卒業．1995 年同大学院工学研究科修士課程修了．同年，富士ゼロックス株式会社 (現 富士フイルムビジネスイノベーション株式会社) 入社．画像処理および文書処理の研究開発に従事．

和田 直己

2020 年早稲田大学基幹理工学部卒業. 2022 年同大学院基幹理工学研究科修士課程修了．同年，富士フイルムビジネスイノベーション株式会社入社．画像処理および文書処理の研究開発に従事．

橋本 一成（正会員）

2000 年北海道大学工学部卒業．2002 年同大学院工学研究科修士課程修了．同年，富士ゼロックス株式会社 (現 富士フイルムビジネスイノベーション株式会社) 入社．自然言語処理，Semantic Web，知識グラフに関する研究に従事．