説明性の高いニューラルモデルの予測確信度に関する分析

佐藤俊 †,1　大内啓樹 ‡,2　佐々木翔大 ‡,†,3　塙一晃 ‡,†,4　乾健太郎 †,‡,5
† 東北大学　‡ 理化学研究所
{[1]shun.sato, [5]inui}@ecei.tohoku.ac.jp
{[2]hiroki.ouchi, [3]shota.sasaki.yv, [4]kazuaki.hanawa}@riken.jp

1　はじめに

ニューラルネットワークを用いたモデルによって，自然言語処理の各タスクにおける予測性能は飛躍的に向上した．一方で，「**モデルがなぜそのような予測をしたのか**」を理解することは，人間にとって極めて困難であることが指摘されている [1]．そのような状況で，k 近傍法のような，学習事例との類似度にもとづいて予測を行うモデルが注目を集めている．この種のモデルでは予測への貢献度の高い学習事例を提示することが容易であり，機械学習の専門知識を持たないユーザにとってもモデルの挙動を直感的に理解可能な場合が少なくない．

これに加え，「**モデルがどの程度確信を持ってそのような予測をしたのか**」がわかるなら，人間の意思決定や人間と機械の協働がさらに促進されると考えられる．たとえばモデルが確信度の低い予測をした場合，人間による確認を入れるといった対策が打てる．このような背景から，多様な確信度計算手法が提案されており，特に近年では，学習事例との類似度にもとづいてモデルの予測確信度を算出する手法が注目を集めている [5]．

我々はニューラルネットワークモデルを使う際に**モデルの予測の根拠**と**予測確信度**が両方が同時にわかることが重要であると考えている．本研究ではこの 2 つの条件を満たすモデルの挙動に対する理解を深めるため，モデルが k 近傍法を用いて予測を行う場合に，訓練方法やデータの類似尺度の違いが確信度計算や近傍事例に与える影響について実験，分析を行った．本研究の貢献は以下の 2 点である．

- ニューラルネット上で k 近傍法を用いて予測を行う場合に，訓練手法 (Cross Entropy Loss による訓練，Triplet Loss による訓練) やデータの類似尺度 (L2 距離, コサイン距離, 内積) を変えて性能評価を行い，各条件における確信度の性能の違いを明らかにした．
- 多くの評価事例の近傍に重複して現れる訓練データ（ハブ）の出現度合いと確信度計算手法 Trust Score の関係を分析し，類似尺度が L2 距離やコサイン距離の時には，ハブが多く発生している予測に対して低い確信度が割り当てられることを明らかにした．

2　説明性の高いモデルの実現

この節では 1 節で述べたように，予測の根拠と予測確信度の両方がわかるような説明性の高いニューラルネットワークモデルの実現方法について述べる．

2.1　モデルの予測確信度

ニューラルネットワークの予測確信度に関する研究は数多くなされており，その研究の方向性のとしては大きく 2 つの方向性がある．1 つ目はモデルの予測確率を直接予測確信度として使えるように補正を行う研究である．具体的な手法には，temperature scaling[4] が挙げられる．temperature scaling は過剰に高い値がつく傾向にあるニューラルネットワークモデルの予測確率に対して温度パラメータという新しいパラメータを導入し，開発セットを用いて温度パラメータの値を変えながら予測確率が適切な確信度となるように補正を行う．

2 つ目は予測時に追加の計算を行うことで予測確率とは異なる予測確信度を計算する研究である．代表的な手法としては Trust Score[5] が挙げられる．Trust Score は評価データと訓練データの間の距離を用いて確信度を計算する手法であり，k 近傍法による予測とも自然に組み合わせることが可能である．

本研究では k 近傍法による予測と Trust Score による確信度計算を組み合わせることで，モデルの予測に対して「なぜその予測が行われたのか」「その予

測がどれくらい信用できるのか」の 2 点が明示的にわかるニューラルネットワークモデルを実現する．

2.1.1 Trust Score による予測確信度の算出

予測確信度として用いる Trust Score の算出方法について述べる．クラスラベルの集合を $C = \{l_1, l_2, ..., l_M\}$ とし，入力 x_i とそのラベル $y_i \in C$ からなる n 個の訓練データ集合を $\mathbf{X}_{\text{train}} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ とする．また，入力 x の文ベクトル $\mathbf{h}_x \in \mathbb{R}^{d\times 1}$ を計算するエンコーダを $f(\cdot)$ とする．訓練済みのエンコーダに $\mathbf{X}_{\text{train}}$ を入力して得られた文ベクトルの集合を $\mathcal{H} := \{\mathbf{h}_{x_1}, \mathbf{h}_{x_2}, ..., \mathbf{h}_{x_n}\}$ とし，評価データ x_{test} に対する文ベクトルを $\mathbf{h}_{x_{\text{test}}}$ とする．ここで，$\mathbf{h}_{x_i} = f(x_i)$，$\mathbf{h}_{x_{test}} = f(x_{test})$ である．また，評価データ x_{test} に対するモデルの予測が $\hat{y}_{test} = l$ である時，集合 $\mathcal{H}$ の中で正解ラベルが l である集合を $\mathcal{H}_l = \{\mathbf{h}_j \in \mathcal{H} | 1 \le j \le n \wedge y_j = l\}$ とする．この時に評価データ x_{test} に対する Trust Score $S(x_{\text{test}}, \mathcal{H})$ は以下の式で算出される．

$$S(x_{\text{test}}, \mathcal{H}) = \frac{d_{\text{np}}(x_{\text{test}}, \mathcal{H})}{d_{\text{p}}(x_{\text{test}}, \mathcal{H}) + d_{\text{np}}(x_{\text{test}}, \mathcal{H})} \tag{1}$$

ただし

$$d_{\text{p}}(x_{\text{test}}, \mathcal{H}) = \min_{\mathbf{h}\in\mathcal{H}_l} d(\mathbf{h}_{x_{\text{test}}}, \mathbf{h}) \tag{2}$$

$$d_{\text{np}}(x_{\text{test}}, \mathcal{H}) = \min_{\mathbf{h}\in(\mathcal{H}\setminus\mathcal{H}_l)} d(\mathbf{h}_{x_{\text{test}}}, \mathbf{h}) \tag{3}$$

ここで $d(\mathbf{h}_{x_{\text{test}}}, \mathbf{h})$ は $\mathbf{h}_{x_{\text{test}}}, \mathbf{h}$ の間の距離を表す．Trust Score は予測したクラスの代表点までの距離が近いほど，また予測していないクラスの代表点までの距離が遠いほど大きな確信度が付けられる．またこの計算式 (1) は，確信度が 0 から 1 までの範囲に収まるように Trust Score を提案している論文 [5] 内における式に修正を加えたものである．

3 分析を行う項目

本研究では 2 節で述べた説明性の高いニューラルネットワークに関して分析を行う．具体的には k 近傍法による予測と Trust Score の計算に大きな影響を及ぼす (1) データ間の類似尺度と (2) データ同士の位置関係の 2 つの項目について変化させた時の Trust Score の挙動について分析を行う．

(1) のデータ間の類似尺度については，Trust Score では L2 距離が用いられることが多いが，その他の類似尺度を用いた際の Trust Score 挙動についてはまだ明らかになっていない．具体的には L2 距離以外にもコサイン距離や内積を類似尺度として用いることで，Trust Score の挙動が変化すると考えられる．

(2) のデータ同士の位置関係はモデルの訓練方法に大きく依存する．通常の M クラスの分類問題においては損失関数として次の **Cross Entropy Loss** が用いられることが多い．

$$L_{\text{CrossEntropy}}(x) = -\mathbf{t}\log(\text{Softmax}(\mathbf{W}\mathbf{h}_x + \mathbf{b})) \tag{4}$$

ただし，$\mathbf{h}_x = f(x)$ とし，$\mathbf{W} \in \mathbb{R}^{M\times d}$ は重み行列，$\mathbf{b} \in \mathbb{R}^{M\times 1}$ はバイアス項，$\mathbf{t} \in \mathbb{R}^{1\times M}$ は x の正解ラベルの箇所に 1 がたつ 1-hot ベクトルである．またデータ間のラベルの違いを明示的に距離関係として学習させる **Triplet Loss**[10] を用いることで，Cross Entropy Loss を用いた際のデータ間の位置関係とは大きく異なった位置関係になると考えられる．Triplet Loss はある訓練データ x に対して，同じクラスの訓練データ x_{p} と異なるクラスの訓練データ x_{n}，マージン m を用いて算出される．

$$L_{\text{Triplet}}(x) = \text{ReLU}(d(\mathbf{h}_x, \mathbf{h}_{x_{\text{p}}}) - d(\mathbf{h}_x, \mathbf{h}_{x_{\text{n}}}) + m) \tag{5}$$

ただし $\mathbf{h}_x = f(x), \mathbf{h}_{x_{\text{p}}} = f(x_{\text{p}}), \mathbf{h}_{x_{\text{n}}} = f(x_{\text{n}})$ とし，$d(\cdot)$ は L2 距離，コサイン距離，内積のいずれかを表す．

以上を踏まえて本研究では，データ間の類似尺度 (L2 距離/コサイン距離/内積) とモデルの訓練方法 (Cross Entropy Loss/Triplet Loss) をそれぞれ変化させた場合の Trust Score の挙動の分析を行う．

4 実験

4.1 タスク設定

今回 Trust Score の分析に用いるタスク設定としては文書分類タスクを用いる．予測確信度に関する先行研究 [2, 3, 5] では確信度の性能評価を画像データやテキストデータに対する分類タスクを用いて論じており，本研究でもそれに従う．

データセットには **20 Newsgroups** データセット [7] を用いる．このデータセットは 20 種類の異なるニュース記事を収集したデータセットであり，20 クラスの文書分類問題として用いる．データ数は約 20,000 であり，今回そのうち 10182 個を訓練データ，1132 個を開発データ，7532 個を評価データとして分割した．

4.2 評価指標

予測確信度の評価指標としては E-AURC (Excess-Area Under the Risk-coverage Curve)[3] を用いる．ま

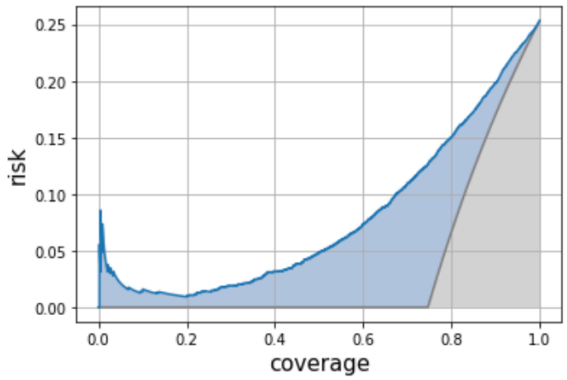

図 1 E-AURC の説明図．AURC が risk-coverage 曲線に囲まれる面積 (青色+灰色) であり，そこから現在のモデルの予測性能で達成しうる最も小さい risk-coverage 曲線の面積 (灰色) を除いた部分の面積 (青色) が E-AURC の値となる．なお図は訓練方法を Triplet Loss, データの類似尺度をコサイン距離とした場合のものである．

ず AURC (Area Under the Risk-coverage Curve) について述べる．AURC は図 1 のように risk-coverage 曲線に囲まれる面積として求められる．risk-coverage 曲線は全ての評価データをそのデータに対するモデルの予測確信度について降順に 1 つずつ見ていき，見ているデータまでの予測の累計の誤り率をプロットして描かれる曲線である．この AURC が小さいほど，ある確信度を閾値とした時に正しい予測と誤った予測が明確に分離されており，優れた確信度計算手法であることを意味する．

しかし，この AURC は，モデルの予測性能に依存しており，異なるモデル間での性能比較が困難である．E-AURC は AURC (図 1 の灰色部分+青色の部分) からを各モデルが達成しうる最小の risk-coverage 曲線 (図 1 の灰色部分) を取り除いて正規化を行うことで，異なるモデル同士の確信度としての性能比較を行えるようにした指標である．E-AURC の詳細な計算方法については付録 A.1 を参照されたい．今回の実験では，訓練方法や予測方法が異なるモデル間での，確信度の性能を比較するため，E-AURC を性能評価の指標として採用した．また結果の視認性の向上のため E-AURC の値は全て 1000 倍して表示した．

4.3 モデル設定

今回は文書分類モデルを入力として訓練済みの単語ベクトル，系列のモデリングに畳み込みニューラルネットワーク (CNN)[11] を用いて実装した．訓練済みの単語ベクトルには 200 次元の Glove[8] を用い，訓練中に単語ベクトルの値の更新も行った．CNN にはカーネルサイズを 3, フィルタのサイズを 3,4,5 とした Max Pooling を行った．損失の最適化には Adam [6] を用い，学習率の初期値は $\rho = 0.001$ とした．

今回実験では損失関数には Cross Entropy Loss と Triplet Loss のいずれかを用いた．Triplet Loss のマージン m としては距離尺度が L2 距離，コサイン距離，内積の場合にそれぞれ $m = 1.0$, $m = 0.005$, $m = 0.01$ とした．このマージンの値は開発セット用いて最も確信度の性能が高くなった場合の値を用いた．

次にモデルの予測方法と確信度計算について述べる．各モデルの予測は各評価データについて近傍の訓練データに基づいて k 近傍法を用いて行う．今回予測に用いる近傍の訓練データの数は $k = 10$ とした．Trust Score の計算には先行研究 [5] に従い，モデルの最終層のデータの文ベクトルを用いて計算を行った．また全ての実験の計測はシード値の異なる 3 つのモデルを用いて行い，その平均値を記載した．

表 1 訓練方法，データ間の類似尺度を変えた時の予測性能と Trust Score の性能

訓練方法	Cross Entropy			Triplet		
類似尺度	L2	cos	dot	L2	cos	dot
Acc	0.747	**0.754**	0.615	0.720	0.741	0.637
E-AURC	48.29	47.12	109.76	48.67	**45.56**	70.11

4.4 実験結果

表 1 に訓練方法，データ間の類似尺度を変えた時のモデルの予測性能と Trust Score の確信度の性能を示す．表 1 から，類似尺度が内積の場合には L2 距離やコサイン距離の時に比べて，予測性能，確信度の性能ともに最も性能が低くなった．また訓練方法ごとに比較すると，類似尺度がコサイン距離の場合が最も確信度の性能が高くなった．全条件で比較すると Triplet loss で訓練を行い，類似尺度をコサイン距離とした場合に最も E-AURC の値が小さく，確信度として優れた性能を発揮することがわかった．Triplet loss で訓練を行い，類似尺度をコサイン距離とした場合の E-AURC は図 1 の青い部分の面積であり，その他のいくつかの条件の E-AURC の様子についても付録 A.3 に記載した．

5 分析

この節ではなぜ Triplet Loss を用いて訓練を行い，データの類似尺度をコサイン距離とした場合に他の条件に比べて Trust Score が確信度として優れた性能を発揮したのかを考察する．我々は原因を探るた

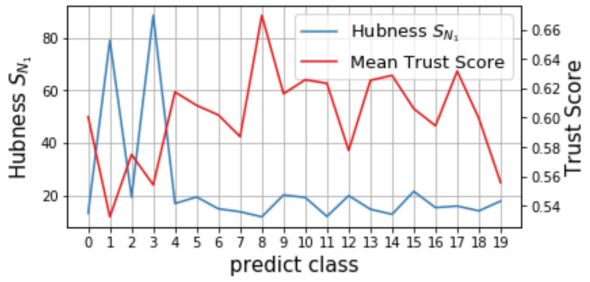

図 2 Triplet/cos モデルにおける予測クラスごとの平均 Trust Score とハブの出現度合い S_{N_1} の関係

め，Trust Score と k 近傍法において発生する**ハブ**という現象の間の関係について分析を行った．

5.1 近傍検索におけるハブの出現

k 近傍法を用いた予測においては，異なる評価データの近傍事例として同じ訓練データが重複して検索されてしまう現象が観測されており，そうした訓練データはハブと呼ばれている [9]．一般に予測の根拠として，同じ事例が何度も出てきてしまうことは，有意義な予測の根拠とは言えない．今回の我々が行った実験においても多くの条件でもこのハブが観測された．

5.1.1 ハブの出現度合い

先行研究 [9] に従い，ハブの出現度合いの計算には各評価事例の予測に用いられる k 個の訓練データの中に各訓練データが何回含まれるかという分布 N_k の歪度 S_{N_k} を用いて行う．この S_{N_k} の値が大きいほど，特定の訓練データが重複して近傍事例として選択されており，ハブが発生していることを表す．計算方法の詳細については付録 A.2 を参照されたい．

5.2 Trust Score とハブの関係

図 2 に Triplet Loss を用いて訓練を行い，類似尺度をコサイン距離とした時の予測クラスごとの平均 Trust Score と各予測の最近傍点におけるハブの発生度合い S_{N_1} の関係を示す．図 2 からハブが多く発生しているクラスでは，Trust Score が平均的に小さく，逆にハブのあまり発生してないクラスでは Trust Score が平均的に大きくなっていることがわかった．

他の 5 つの条件とも比較するため，予測クラスごとの最近傍点におけるハブの発生度合い S_{N_1} と Trust Score との相関，及び最近傍点におけるハブの発生度合いと予測の適合率の相関をピアソンの相関係数を用いて計測した結果が表 2 である．表 2 からは類似尺度が L2 距離とコサイン距離の場合に，各予測クラスごとの S_{N_1} と Trust Score，S_{N_1} と適合率の間に負の相関関係が読み取ることができる．この負の相関関係は全条件の中で Triplet Loss で訓練し，距離尺度をコサイン距離とした場合に最も強くなっていた．すなわち，この条件下では，ハブの発生している予測クラスでの予測は誤りやすくまた，その予測に低い確信度がつきやすい傾向が最も強くでたため，確信度として最も優れた性能を発揮したと考えられる．予測クラスごとの詳細な実験値は付録 A.3 を参照されたい．

表 2 各条件における予測クラスごとの S_{N_1} と Trust Score(TS), S_{N_1} と適合率の相関係数

訓練方法	Cross Entropy			Triplet		
距離尺度	L2	cos	dot	L2	cos	dot
S_{N_1}&TS	-0.47	-0.44	0.38	-0.29	-0.50	0.21
S_{N_1}&適合率	-0.52	-0.37	0.21	-0.35	-0.53	0.44

6 終わりに

本論文ではニューラルネットワークモデルの予測についてモデルの予測の根拠と予測確信度が両方同時にわかることが重要だと考え，予測を k 近傍法で行う場合の Trust Score の挙動について訓練方法やデータの類似尺度を変えて分析を行った．その結果，損失関数を Triplet Loss, データの類似尺度をコサイン距離とした場合に確信度として最も優れた性能を発揮することを明らかにした．

また分析の過程で，k 近傍法で検索される近傍事例において，同じ訓練データが重複して検索されるハブという現象に注目し，ハブが発生している場合に Trust Score による予測確信度が低くつく傾向があることを明らかにした．なぜこのような現象が起こるのかの解明は今後の研究課題としたい．

謝辞

本研究は JSPS 科研費 JP19H04425 の助成を受けたものです．

参考文献

[1]Marina Danilevsky et al. “A Survey of the State of Explainable AI for Natural Language Processing”. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 2020.

[2]Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep

Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. 2016.

[3]Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. "Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers". In: *Proceedings of the 7th International Conference on Learning Representations*. 2019.

[4]Chuan Guo et al. "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017.

[5]Heinrich Jiang et al. "To Trust Or Not To Trust A Classifier". In: *Advances in Neural Information Processing Systems*. 2018.

[6]Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations*. 2015.

[7]Ken Lang. "NewsWeeder: Learning to Filter Netnews". In: *Proceedings of the 12th International Machine Learning Conference*. 1995.

[8]Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014.

[9]Miloš Radovanovi; Alexandros Nanopoulos, and Mirjana Ivanovi; "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data". In: *Journal of Machine Learning Research* (2010).

[10]Jiang Wang et al. "Learning Fine-grained Image Similarity with Deep Ranking". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[11]Kim Yoon. "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014.

A 付録

A.1 E-AURC の計算方法

risk-coverage 曲線の面積である AURC は，各々のデータにつけられた確信度について降順に並べ替えた m 個の評価データ集合 $\{x_1, x_2, ..., x_m\}$ について 以下の式 (6) によって計算される．

$$\mathrm{AURC} = \sum_{i=1}^{m} \frac{\sum_{j=1}^{i} g(x_j)}{i \times m} \tag{6}$$

ただし，関数 g は評価データ x に対する分類器の予測が当たっている場合には 0，外れている場合には 1 を返す 2 値関数である

E-AURC を提案している論文 [3] では，AURC の正規化を行うための最小の risk-coverage 曲線の面積 $\mathrm{AURC}_{\mathrm{min}}$(図 1 の灰色部分) を求めるために以下の式 (7) のような近似計算を行なっている．

$$\mathrm{AURC}_{\mathrm{min}} \approx \int_0^{\hat{r}} \frac{x}{1-\hat{r}+x} dx = \hat{r} + (1-\hat{r}) log(1-\hat{r}) \tag{7}$$

ここで $\hat{r}$ は評価データ全体に対する予測の精度 α について $\hat{r} = 1 - \alpha$ として定まる値である．

最終的に式 (6) で求まる AURC から，式 (7) で求めた最小の risk-coverage 曲線の面積を引くことで E-AURC の値を求めることができる．

A.2 ハブの出現度合いの計算

先行研究 [9] に従い，ハブの出現度合いの計算には式 (8) のように，各評価事例の予測に用いられる k 個の事例の中に各訓練データが何回含まれるかという分布 N_k の歪度 S_{N_k} を用いて行う．

$$S_{N_k} = \frac{\sum_{i=1}^{l} (N_k(i) - E[N_k])^3 / l}{Var[N_k]^{3/2}} \tag{8}$$

ここで l は訓練データの数であり，E, Var はそれぞれ期待値と分散を表す．式 (8) の値が大きいほど評価事例の k 近傍に頻繁に含まれる訓練データが存在し，ハブが出現していることを表す．

A.3 実験値の詳細

表 3 に Triplet loss を用いて訓練し，距離尺度をコサイン距離とした時の予測クラスごとの最近傍におけるハブの発生度合い S_{N_1} と，予測の適合率，Trust Score の平均値を記載した．

また Cross Entropy Loss で訓練を行い類似尺度をコサイン距離とした場合の E-AURC の結果を図 3 に，Triplet Loss で訓練を行いコサイン距離以外の類似尺度を用いた場合の E-AURC の結果を図 4,5 記載した．

表 3 Triplet/cos モデルにおける予測クラスごとの S_{N_1}，予測の適合率，平均 Trust Score

予測クラス	0	1	2	3	4	5	6	7	8	9
S_{N_1}	13.15	78.94	19.28	88.50	16.85	19.32	14.83	13.70	11.78	20.11
適合率	0.80	0.46	0.64	0.60	0.78	0.82	0.59	0.73	0.91	0.85
TS	0.62	0.54	0.58	0.56	0.63	0.63	0.58	0.60	0.67	0.63
予測クラス	10	11	12	13	14	15	16	17	18	19
S_{N_1}	19.14	11.92	19.86	14.66	12.73	21.48	15.27	15.87	14.05	17.76
適合率	0.93	0.87	0.65	0.88	0.86	0.80	0.67	0.95	0.74	0.41
TS	0.65	0.64	0.59	0.63	0.65	0.64	0.59	0.64	0.59	0.55

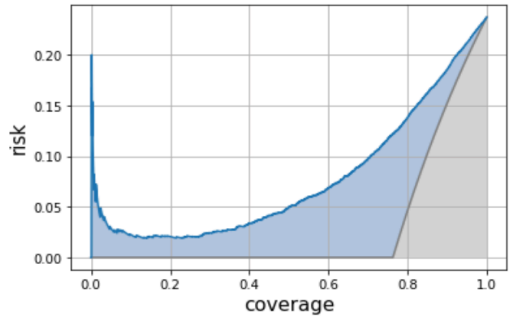

図 3 Cross Entropy/cos モデルにおける E-AURC の様子

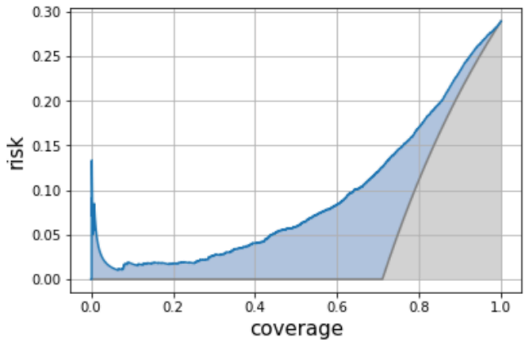

図 4 Triplet/l2 モデルにおける E-AURC の様子

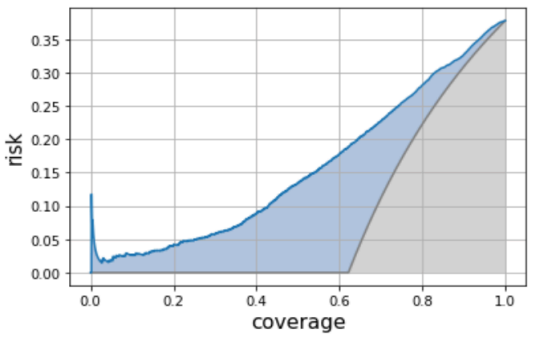

図 5 Triplet/dot モデルにおける E-AURC の様子