敵対的生成ネットワークを用いた，ユーザーに分かりやすい融資拒否の説明の生成

Ramya Malur Srinivasan　　Ajay Chander

あらまし

金融サービス業界内では，業界内の競争や規制の関係上，AI（人工知能）の導入を意識する機運が高まっている。この目標に向けた重要な側面は，AIによる決定事項を各種関係者に説明することである。最新のXAI（Explainable AI：説明可能なAI）システムは，概ねAIのエンジニア向けのものであり，他の関係者にはほとんど無価値である。このギャップを埋めるために，米国富士通研究所は，融資申請者に分かりやすい説明の代表的なデータセットを初めて開発した。また，目的に応じてユーザーに分かりやすい説明を生成する，より小規模なデータセットにも対応する斬新なGAN（Generative Adversarial Networks：敵対的生成ネットワーク）も設計した。

本稿では，融資申請者の学習に役立つ説明や，今後融資の資格を得るために取るべき対策に役立つ説明など，米国富士通研究所が開発したシステムが複数の目的に対応する説明をどのように生成するかを実証する。

1. まえがき

顧客の行動予測からID検証まで，AI（人工知能）はフィンテックの用途で幅広く導入されている[1]。このようにAIが広くビジネスに採用されている中で，顧客，意思決定者，技術者などの間では，AIのブラックボックス的要素への関心が特に高まっている。例えば，AIベースの信用評価システムによって，与信リスク評価モデルが規制されている市場では，このようなモデルを説明可能なものにする要望が高まっている [2]。

最近では，政府 [3] やさまざまな民間産業 [4，5] から，AIによる説明責任を果たすための構想がいくつか打ち出されているため，その信頼度も高まっている。しかし，説明を提供する最新の方法の多くは，AIエンジニアを対象としている [6，7]。そのため，AIによる決定事項をもっと幅広いユーザー群に説明できるようなAIシステム構築のニーズが高まっている。

これを実現するためには，多目的に対応できる効果的な説明を生成する必要がある。そのような説明によって，ユーザーとの信頼関係も確立され，AIシステムがより堅牢になるための一助ともなる。また，意思決定者の学習にも役立ち，彼らの視野が広がり，適切な対策を選択できるようにもなる。

以上を背景として，米国富士通研究所では，AIベースの融資に関する意思決定，特に融資拒否の決定のユースケースを検討した。まず，ユーザーにとって分かりやすい説明の代表的なデータセットを収集し作成した。次に，前述の説明を生成できる機械学習を設計した。本手法では，融資申請者の学習に役立つ説明や，今後融資の資格を得るために取るべき対策に役立つ説明など，複数の目的に対応する説明を生成できる。これらの機能によって，AIシステムの信頼性が高まる。

2. 関連研究

本章では，金融業界を中心に関連する取り組みを紹介する。次に，AIエンジニアとエンドユーザーの視点から見た説明能力に関する関連研究を紹介する。

2.1 金融業界におけるXAIの取り組み

XAIに向けた様々な政府の構想の中でも，新たなEU一般データ保護規則（GDPR），ISO/IEC 27001，およびアメリカ国防高など研究計画局（DARPA）の説明可能なAI（XAI）プログラム [3] が特に注目に値する。また，いくつかの業界団体がAIの説明可能性に関する問題に取り組んでいる。与信分析企業のFICO（Fair Isaac Corporation）は，最近XAIツールキット [8] をリリースし，機械学習における説明可能性への対応の一部を概説している。このように，XAIは近い将来，金融業界においても，急成長が期待されている。

2.2 AIエンジニアに対する説明

AIエンジニアの視点から見た説明可能性に関しては，[9] および [10] に総括されている。しかし，AIによる説明の多くはAIエンジニアのためのもの [4] である。AI以外の専門家にとっては，AIの決定事項を理解する上でも，モデルのデバッグ [11] を行う上でも役に立つとは言えない。一方 [12] では，AIシステムによる意思決定に至るプロセスで用いられている主要素について，また要素の変更によって意思決定が変化する仕組みについて論じている。この種の説明は，AIエンジニアによるデバッグに役立つ。AIエンジニアの支援は重要だが，これらはエンジニア以外の幅広いユーザーが利用できるものではない。

2.3 エンドユーザーに対する説明

近年，人間によるAIシステムの解釈について理解するための取り組みがいくつかなされている。[13] では，人間によるAIシステムの解釈のための分類法を提示している。また[14]や[15]では，ユーザー中心の説明の効果的な観点が示されている。これらの論文では，ユーザーが納得できる説明を強調している。[16] では，ユーザーの視点から見た双方向性についての見解を模索している。[17] では，機械学習による説明を人間が調査を通じて理解する仕組みについて論じている。解釈可能性の性能は，応答時間と応答の精度を基準としている。これらの取り組みは，人間による解釈可能性を定量化するうえで非常に重要だが，本稿で述べ

ているユーザーに分かりやすい説明を生成するまでには至っていない。

3. データセットの構築

従来，ユーザーに分かりやすい説明の代表的なデータセットは，どれも利用価値に乏しかった。そのため，米国富士通研究所では初めてこの点を考慮したデータセットを構築した。このデータセットを「X-Net」と呼ぶ。

X-Netを構築するために，まずAmazon Mechanical Turk（Mturk）[18]での調査を行った。この調査では，MTurkの作業者に融資申請のシナリオを提供し，自身が融資申請者であることを想定するよう求めた。

まず，これらの作業者に，融資拒否の理由を強調した説明文を提供した。次に，これらの説明文を，作業者が構文と意味の観点から修正・編集した。更に語学者が，拒否された融資ごとに広義の理由と具体的な理由にそれぞれ対応する注釈を各説明に付けた。例えば，広義の理由を「職業」とした場合，具体的な理由は「就業経験がない」「不安定な職業である」「職務経歴が少ない」「職務経歴がない」「職務経歴が安定していない」となる。この作業によって，最終的に2,432もの対応する広義の理由と具体的な理由の文からなるデータセットが構築された。

しかし，これらの理由を解析した結果，重複しない文は100にも満たないことが分かった。これらの結果から米国富士通研究所は，ユーザーに分かりやすい説明は比較的少数であるという所見を得た。更に，MTurkの作業者が提示した理由がデータセットの特徴として現れることが少ないという所見も得た。最も頻繁に表れる広義の理由は，与信，職業，収入，負債などであった。これら以外には，履歴の確認ミスや申請の不備など，言及されることの少ない語句がわずかにあっただけであった。

次に，融資申請者の学習に役立つ説明や，今後金融の分野で取るべき対策に役立つ説明，多目的に対応する説明を生成するために，2,432の文からなるデータセットから，学習と対策という2つの目的に対応するペアを収集した。この作業は語学者（修辞法の専門家）と協力して行った。このデータセットを「拡張X-Net」という。拡張X-Netからサンプリングした学習と対策の説明のペアを以下に示す。

「この申請に関する財務状況の記録には，多額のローンの支払いが残っていることが示されています」（学習）。

「新規ご融資を申し込まれる前に，ローンの残額を完済してください」（対策）。

データセットの詳細については，[19] を参照されたい。

4. ユーザーに分かりやすい説明の生成方法

本研究の第1の目標は，X-Netのようにユーザーに分かりやすい説明を自動的に生成することである。更に，生成方法をコントロールするために，生成された説明が融資拒否の具体的な理由と一致するなどの条件付きで説明を生成することである。第2の目標は，学習に役立つ説明や，取るべき対策に役立つ説明など，多目的に対応できる説明を生成することである。

最近のGANの成功を受けて，以上の目標に対応できる条件付きのGANを設計した。一般的にGANは，画像やテキストなどの各種データを生成できる。GANは，Generator（データ生成ネットワーク）とDiscriminator（真偽判別ネットワーク）という二つのニューラルネットワークからなる（図-1）。まず，Generatorがノイズを入力信号として取り込み，次にDiscriminatorがその真偽を予測してデータを生成する。

条件付きGANの場合，特定の条件をGeneratorに指定し，それを融資申請の要件に応じてデータ生成の指針とする。また，GANモデルにStyle Transfer（説明スタイルの変換）のメカニズムを

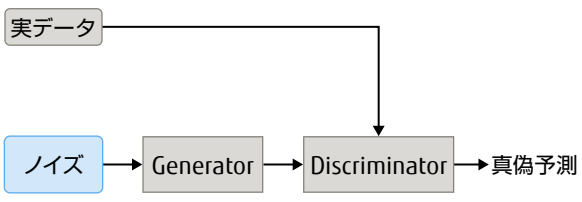

図-1　一般的なGANシステムアーキテクチャーのブロック図

組み込み，多目的に対応できる説明を生成する。

ただし，ここで直面する最大の課題は，学習データの数が限られていることである。本稿で示した例では，2,432ある理由の文の中で，一意のものは100しかない。そこで，米国富士通研究所の解決戦略の主な側面を以下に示す。詳細については，[20] を参照されたし。

限られた学習データを用いて条件付き説明生成の第1の目標を達成するために，ARAE（Adversarially Regularized Autoencoder：敵対的に正規化されたオートエンコーダー）[21] のアーキテクチャーを採用する。特に，以下に列挙するとおり，米国富士通研究所では，このアーキテクチャーに対する三つの改良案を提示する。

（1）GANモデル [22] の基本条件をモデル化するために，正規分布の混合を検討する。十分な数の正規分布の成分を与えれば，混合モデルによって複雑な任意の分布に近づけられる。そのため，従来型の分布ではなく正規分布の混合を用いることで，モデルの表現性を向上させられる。

（2）条件を階層化して組み込む。これは，子供がごく限られたデータから階層的に学習するという事実にヒントを得ている [23]。このことから，広義の理由（例：収入）と具体的な理由（例：収入が不安定）という2段階の条件付け方法に基づき，融資拒否の条件を組み込む。

（3）より関連性のある文を生成するために，LabelerとAntilabelerという二つの新たな損失関数 [24] を導入し，実際の文と生成された文をそれぞれの理由に基づき分類する。

ARAEGANをベースに構築した米国富士通研究所が提案するシステムアーキテクチャーを図-2に示す。詳細については，[20] を参照されたし。

ここでの目標は，多目的に対応できる説明，つまり融資申請者が拒否について学習できる説明や，今後取るべき対策に役立つ説明を生成することである。この研究でも，学習データが限られていることが大きな課題である。そこで，学習データにある「学習」と「対策」に対応する説明のペアを検討し，ガウスノイズの混合と合わせてARAEGANモデルを用いる。特に米国富士通研究所では，以下の二つの設定を検討する。一つは，学習に役立つ説明と今後取るべき対策に役立つ説明に対応する，一致した説明のペアが存在する状態で，これを「一致する事例」と呼ぶ。もう一つが「一致しない事例」で，学習データに対応する説明のペアが存在しない状態である。

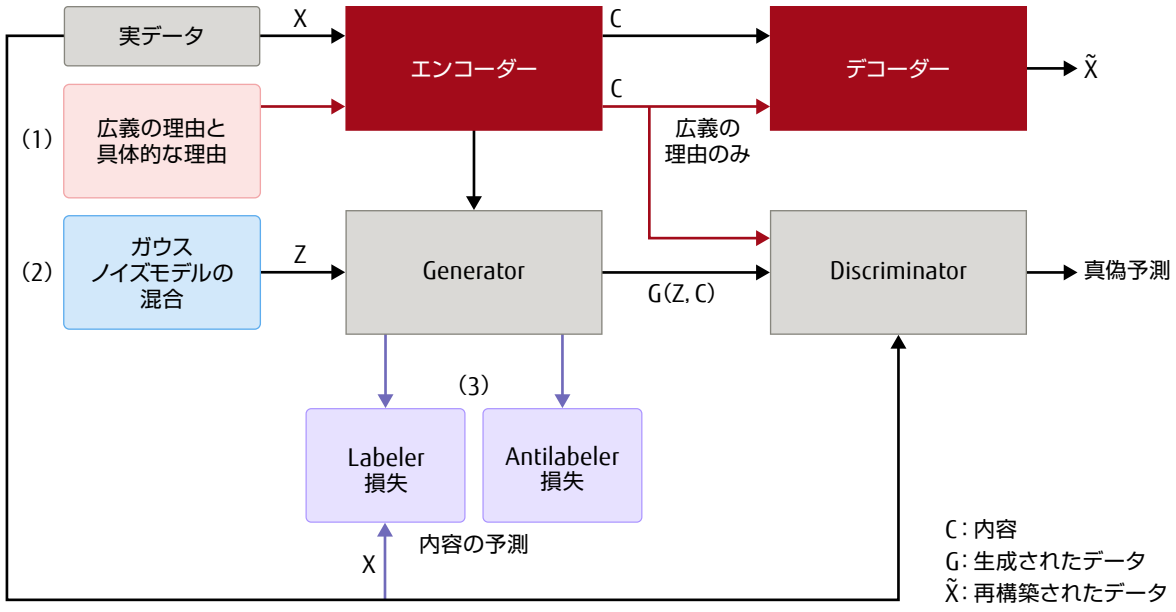

図-2　ARAEGANベースのシステムアーキテクチャー案のブロック図

表-1　多目的に対応できる説明の生成

説明文	モデル	教育を重視する説明スタイルから 対策を重視する説明スタイルへの変換	対策を重視する説明スタイルから 教育を重視する説明スタイルへの変換
1	学習データに対応する参照文	不安定なローン **返済** の記録があります。	より **収入** に見合った **額** の融資への申し込みを再検討してみてください。
2	生成された文： ARAEGANモデルと一致しなかった事例	より ***収入*** に見合った額の融資への申し込みを再検討してみてください。	申請者の現在の ***在職*** 期間が短すぎます。
3	生成された文： ARAEGANモデルとガウスノイズモデルの混合モデルと一致しなかった事例	***与信*** の改善方法について金融機関と相談してください。	残念ながら，今回の申請に関連付けられた ***与信*** レベルが，今回の融資の対象として検討できるレベルに達していません。
4	生成された文： GANモデルと一致した事例	今後はローンの **返済** を定期的に行ってください。	今回の申請で記載された **収入** のレベルが，ご希望の融資 **額** のレベルと一致しません。

5. 結果

本章では，生成された様々な説明を例示する。学習済みの分類コードを評価基準として使用し，意味と関連性のある文を生成する際に，米国富士通研究所モデルが理由の情報を用いる能力を評価する。モデルのハイパーパラメーターと実験の設定に関する詳細については，[20] を参照されたし。

生成された説明を表-1に示す。説明文1の太字のテキストは，融資拒否の理由を表す。説明文2と3の太字かつ斜体のテキストは，生成された文の間違った理由を表す。説明文4の太字かつ下線付きのテキストは，生成された文の正しい理由を表す。この表から分かるとおり，一致しないARAEGANモデルでは，多目的に対応できる説明を生成することができなかった。一方，一致するGANモデルでは，意味のある文を生成するだけでなく，参照文（太字のテキスト）からの理由を保存することもできた。

6. むすび

本稿では，ユーザーの視点から見た説明可能性の問題を探った。特に，融資拒否の説明に関するユースケースを検討し，ユーザーに分かりやすい説明の代表的なデータセットを初めて構築した。機械学習のデータセットの特徴としては，ユーザーが納得できる理由はほとんど表示されないため，モデル中心の説明はユーザーにとってあまり有効でないことが分かった。この問題に対応するために，データセットで指定した理由に基づき，ユーザーに分かりやすい説明を生成できる，新たな条件付きGANを設計した。また，このGANアーキテクチャーにStyle Transferを組み込んで拡張し，学習に役立つ説明や今後取るべき対策に役立つ説明など，多目的に対応できる説明を生成するようにした。

説明を求めるより広範なユーザー群のニーズに対応し，かつ多目的に対応できる複数の説明を生成することによって，この作業が金融業界における研究と実践のギャップを埋める一助になることを期待している。

参考文献・注記

[1] Global fintech investment robust on back of strong VC funding: KPMG. Digital News Asia, 2017.
https://www.digitalnewsasia.com/digital-economy/global-fintech-investment-robust-back-strong-vc-funding-kpmg

[2] FICO：Machine Learning and FICO Scores: An Evolution in ML innovations that helps both lenders and consumers. White Paper: Business Technology Overview 2018.
https://www.fico.com/en/latest-thinking/white-paper/machine-learning-and-fico-scores

[3] D. Gunning：Explainable Artificial Intelligence (XAI). DARPA/I2O，2017.

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf
[4] KYNDI：How ‘Explainability’ is Driving the Future of Artificial Intelligence. A Kyndi White Paper，2018.
https://kyndi.com/wp-content/uploads/2018/01/Kyndi-final-Explainable-AI-White-Paper.pdf
[5] PWC：Explainable AI: Driving business value through greater understanding. White paper, Intelligent Digital 2018.
https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf
[6] R. Selvaraju et al.：Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. The IEEE International Conference on Computer Vision，p.618-626，2017.
https://ieeexplore.ieee.org/document/8237336
[7] D. Park et al.：Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. Arxiv, 2018.
https://arxiv.org/abs/1802.08129
[8] A. Flint et al.：xAI Toolkit: Practical, Explainable Machine Learning. White Paper，2018.
https://www.fico.com/sites/default/files/2018-06/FICO_xAI_Toolkit-Practical_Explainable_Machine_Learning_4547WP_EN.pdf
[9] Z. Lipton：The Mythos of Model Interpretability. ICML Workshop，2016.
https://arxiv.org/abs/1606.03490
[10] D. Doran：What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. ArXiv，2017.
https://arxiv.org/abs/1710.00794
[11] A. Chandrasekaran et al.：Do Explanations make VQA Models more Predictable to a Human? EMNLP，2018.
https://arxiv.org/pdf/1810.12366.pdf
[12] F. Doshi-Velez et al.：Accountability of AI Under the Law: The Role of Explanation. ArXiv，2017.
https://arxiv.org/pdf/1711.01134.pdf
[13] F. Doshi-Velez et al.：Towards A Rigorous Science of Interpretable Machine Learning. ArXiv，2017.
https://arxiv.org/pdf/1702.08608.pdf
[14] T. Millers et al.：Explainable AI:Beware of Inmates Running the Asylum. ArXiv，2017.
https://arxiv.org/pdf/1712.00547.pdf
[15] B. Herman：The Promise and Peril of Human Evaluation for Model Interpretability. NIPS Workshop 2017.
https://arxiv.org/abs/1711.07414
[16] S. Amershi et al.：Power to the People: The Role of Humans in Interactive Machine Learning. AI magazine，p.105-120，2014.
https://www.aaai.org/ojs/index.php/aimagazine/article/view/2513
[17] M. Narayanan et al.：How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. 2018.
https://arxiv.org/pdf/1802.00682.pdf
[18] Amazonウェブサービスの1つ。コンピュータープログラムと人間の知能を組み合わせることで，コンピューターのみでは不可能な作業を処理することができる。
[19] A. Chander et al.：Creation of User Friendly Datasets: Insights from a Case Study concerning Explanation of Loan Denials. ICML HILL Workshop 2019.
https://arxiv.org/abs/1906.04643
[20] R. Srinivasan et al.：Generating User-friendly Explanations for Loan Denials using GANs. Neurips workshop on Challenges and Opportunities for AI in Financial Services: The Impact of Fairness, Explainability, Accuracy, and Privacy.
https://arxiv.org/pdf/1906.10244.pdf
[21] J. Zhao et al.：Adversarially Regularized Autoencoders. ArXiv，2018.
https://arxiv.org/pdf/1706.04223.pdf
[22] S. Gurumurthy et al.：DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data. 2017.
https://arxiv.org/pdf/1706.02071.pdf
[23] C. Kemp et al.：Learning overhypotheses with hierarchical Bayesian models. Developmental

Science，10:3，p.307-321，2007.
https://web.mit.edu/cocosci/Papers/devsci07_kempetal.pdf
[24]M. Kocaoglu et al.：CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. arXiv preprint arXiv:1709.02023,2017.
https://arxiv.org/pdf/1709.02023.pdf

著者紹介

Ramya Malur Srinivasan
Fujitsu Laboratories of America Inc.
Solutions for Augmented Intelligence Lab
XAIテクノロジーの研究開発に従事。

Ajay Chander
Fujitsu Laboratories of America Inc.
Solutions for Augmented Intelligence Lab
人間中心の最新テクノロジーおよび製品の研究開発に従事。