様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １９ 日現在

機関番号：２０１０３
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００３１０

研究課題名（和文）不均一データストリームに対応した深層学習

研究課題名（英文）Deep Learning for Imbalanced Data Stream

研究代表者
新美　礼彦（Niimi, Ayahiko）

公立はこだて未来大学・システム情報科学部・准教授

研究者番号：８０３４７１７９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：深層学習とIoT の発展により、時間軸を持った大規模データ（ストリームデータ）のデータマイニングによる知識発見システムを短期間で構築したいというニーズが高まってきている。しかし、扱うデータによっては不均一分布を考慮するやパラメータチューニングの必要がある。本研究は、並列分散環境で深層学習によるストリームデータマイニングを行う際に不均一分布を考慮したデータマイニングのフレームワークを構築することを目的とした。本研究で，深層学習における不均一データの影響を検討し，不均一データのための深層学習を提案した.

研究成果の学術的意義や社会的意義
本研究の学術的意義は、大規模並列分散環境での超巨大データ分析システムの開発の効率化につながるものとなる。現在、並列分散環境としてグリッド・コンピューティングが一般化してきた。また、動的に変化するストリームデータに対する分析要求も上がってきている。本研究の結果はこの動きを加速させ、分散マイニングシステムの扱いに関して実社会に還元できる可能性がある。特に、Twitter やセンサデータのリアルタイムデータ分析などに有用である。

研究成果の概要（英文）：Due to the development of deep learning and IoT, there is an increasing need to build a knowledge discovery system by data mining of large-scale data with a time axis (stream data) in a short period of time. However, depending on the data to be handled, it is necessary to tune the parameters and to consider imbalanced dataset. The purpose of this research is to construct a data mining framework that considers imbalanced data when performing stream data mining by deep learning in a parallel and distributed environment. In this research, we examined the effect of imbalanced data in deep learning and proposed deep learning for imbalanced data.

研究分野：データマイニング

キーワード：データマイニング　深層学習　不均一データ　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年のデータマイニングは、時間的に変化するデータを川を流れるデータとしてとらえて扱うストリームマイニングや大規模グラフを対象とするグラフマイニングなど、大規模で時間的に変化する構造を持ったデータを対象とするようになり、VLDB や ICDM などのデータベースやデータマイニングに関するトップレベルの国際会議でも Streaming Data や Graph and Pattern Matching などのセッションが設けられ、様々な発表が行われている。2015 IEEE International Conference では Apache Spark の開発者が Spark の特徴として、ストリームデータを高速に扱えることを取り上げている[a]。

提案者は今までの研究で、以下のデータ形式に対し、大規模データに対応した知識発見手法を実装し、大規模ストリームデータを効率よくマイニングできるという結果を得た。

表 1: これまでの研究データと提案手法

データ形式	規模	提案手法
ストリームデータ (識別ルール)	110 万件	ストリームカーネル法とパラレルブースティング
ストリームデータ (決定木)	2500 万件	偏りのある分布をデータに対する拡張 VFDT によるオンライン学習
グラフ系列 (変化ルール)	12 万件	階層型クラスタリングと時間重み付きたたみ込みグラフカーネル
グラフ系列 (PageRank)	32 万件	グラフクラスタリングと並列分散処理

深層学習(Deep Learning)は各種フレームワークによって導入しやすくなったが、活用するには二つの問題がある。一つは、深層学習で不均一データストリームを学習すると、少数データに対応したクラスの学習が行えないという問題である。このことを、解決するのが本研究の第 1 の目的である。もう一つは、多数のパラメータを適切にチューニングしないと性能を発揮できないことである[b]。この問題に対し、自動パラメータチューニングを用いて解決するのが本研究の第 2 の目的である。

現在までの研究で、数値データや記号データを扱うデータマイニングにおいて、複数手法の組み合わせ手法を提案してきた。また、この手法を拡張し、テキストマイニングを行う際にデータ解析者が経験的・試行錯誤的にシステム構築するという作業を遺伝アルゴリズムを用いることにより、自動的に行うシステムを提案した。（これに関して、平成 17～18 年度の科学研究費補助金の交付を受けた。）この結果を受けて、提案システムを Web 上からの情報統合というアプリケーションとして実装したシステムを構築し、システム構築の簡便化に成功したという結果を得た。

２．研究の目的

本研究では、時間軸を持ち、分布が偏った大規模データ（ストリームデータ）に対し、分布の偏りを考慮した分散処理可能な深層学習(Deep Learning)アルゴリズムを確立することを目的とする。

この目的のために、以下の 3 つの問題解決を図ることを研究目的とする。

(1) 不均一分布を持ったデータに対する、分散処理可能な深層学習(Deep Learning)のためのアルゴリズムの構築
(2) ストリーム性を考慮したマイニングアルゴリズムの開発
(3) 上記アルゴリズムの大規模データへの適用

３．研究の方法

本研究は、平成 29 年度から 3 年間の計画であった。

初年度は、主として、(1) 不均一分布を持ったデータに対する、分散処理可能な深層学習のためのアルゴリズムの構築を行った。これまで提案者が開発してきたアルゴリズムでの知見を元に、不均一分布を持ったデータに対する分散処理可能な深層学習(Deep Learning)のためのアルゴリズムを構築した。深層学習(Deep Learning)は訓練データに過学習してしまう問題がある。これは特に不均一分布を持つデータを学習する際は問題である。本研究では、不均一分布を持つデータに対しても過学習を起こさないアルゴリズムを検討する。実験データとしては、不均一データ

分布の特徴を持つクレジットカードトランザクションデータを利用する。まず、クレジットカードトランザクションデータの特徴を元に人工データを作成し、アルゴリズムの検討と構築を行った。次に、実データである実際のクレジットカードトランザクションデータを用いて、性能評価を行った。

2 年目は，主として(2) ストリーム性を考慮したマイニングアルゴリズムの開発を行った。これまで提案者が開発してきたアルゴリズムでの知見を元に、前年度に構築した，不均一分布を持ったデータに対する分散処理可能な深層学習(Deep Learning)のためのアルゴリズムをストリームデータに対応させるアルゴリズムの検討を行った。不均一分布を持ったデータをサンプリングとデータ生成により、精度を向上させる手法を提案し，ベンチマーク用データセットに適用し、提案手法の有効性を検証した。ストリームデータとしてネットワークパケットを元にした不正侵入検知を取り上げ，検知手法を検討した。また、自然言語処理を対象にしたトピックモデルによる話題推定手法の多義語と新語への対応、ソースコード特有の近傍単語の影響を考慮した word2vec を用いた類似コード片推薦手法を提案した。

3 年目は，主として(3)これまでに開発したアルゴリズムの大規模データへの適用，(4)提案する不均一データストリームに対応した深層学習について研究成果を普及させるための活動をおこなった．(3)に関して，データマイニングの手法として，深層学習(Deep Learning)を取り上げ，不均一分布を持ったデータに対応したフレームワークを提案した．深層学習では，画像データに対する深層学習を取り上げた.また，不均一分布を持つデータとしてネットワーク攻撃検知問題を取り上げた． 1) 生成モデルによる minority クラスのデータ生成，2) 深層学習を用いた画像処理，3) オンライン学習と構造変化検知を利用したパケット分析によるネットワーク攻撃検知手法の提案，の 3 つのテーマを行った．

４．研究成果

初年度の研究成果をまとめて，3 つの国際会議(6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017), The 4th International Conference on Fuzzy Systems and Data Mining (FSDM 2018), World Congress on Internet Security (WorldCIS-2017))にて発表を行った。

2 年目の研究成果をまとめて,1 つの雑誌論文(JITST)と 5 つの国際会議(IEA-AIE 2018, FSDM2018, SCIS&ISIS with ISWS2018, WorldCIS-2018, IMCIC 2019)にて発表を行った。国内研究会でも発表を行った.本研究を受けて，日本知能情報ファジィ学会の論文誌の特集を企画した．

3 年目の研究成果をまとめて，2 つの査読付き学術論文(IJICR 2 本)と 3 つの国際会議(FSDM2019,WorldCIS-2019,Computing Conference 2019)にて発表を行った．国内研究会でも 2 件の発表を行った．本研究を受けて，"Data mining with generating data to improve data imbalance problem"のタイトルで国際会議(Big Data Analytics and Data Science)の Plenary Speech を行った．また、1 本の解説記事(古くて新しいデータマイニング --- 不均衡データ問題とプライバシー保護データマイニングと画像処理によるデータ処理，知能と情報（日本知能情報ファジィ学会誌）)を行った．

【参考文献】

[a] I. Stocia, "Conquering Big Data with Spark," Big Data (Big Data), 2015 IEEE International Conference on, Santa Clara, CA, 2015, pp. 3-3.
[b] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures", CoRR, abs/1206.5533, 2012.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Niimi Ayahiko、Takahata Koki	４．巻 11
２．論文標題 Attack Detection Approach by Packet Analysis Using Online Learning with Kernel Method and Correlation Change Method	５．発行年 2020年
３．雑誌名 International Journal of Intelligent Computing Research	６．最初と最後の頁 1033～1040
掲載論文のDOI（デジタルオブジェクト識別子） 10.20533/ijicr.2042.4655.2020.0125	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Niimi Ayahiko	４．巻 10
２．論文標題 Data Anonymization Using Imbalanced Data for Deep Learning with Uppersampling and Undersampling	５．発行年 2019年
３．雑誌名 International Journal of Intelligent Computing Research	６．最初と最後の頁 971～976
掲載論文のDOI（デジタルオブジェクト識別子） 10.20533/ijicr.2042.4655.2019.0118	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 新美礼彦	４．巻 32
２．論文標題 古くて新しいデータマイニング --- 不均衡データ問題とプライバシー保護データマイニングと画像処理によるデータ処理	５．発行年 2020年
３．雑誌名 知能と情報	６．最初と最後の頁 9-12
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ayahiko Niimi	４．巻 6
２．論文標題 Majority Rule Approach to Deep Learning for Large Benchmark Data and Real Credit Card Transaction Data	５．発行年 2018年
３．雑誌名 Journal of Internet Technology and Secured Transaction (JITST)	６．最初と最後の頁 541-547
掲載論文のDOI（デジタルオブジェクト識別子） 10.20533/jitst.2046.3723.2018.0067	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 齋藤 尊、新美 礼彦、伊藤 恵	4．巻 35
2．論文標題 過去の情報を用いたPBL向け工数見積り手法の提案	5．発行年 2018年
3．雑誌名 コンピュータ ソフトウェア	6．最初と最後の頁 1_117～1_123
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.11309/jssst.35.1_117	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計20件（うち招待講演　0件／うち国際学会　12件）

1．発表者名 Ayahiko Niimi, Kosuke Sakamoto
2．発表標題 Generating Data to Alleviate Data Imbalance Problems in Machine Learning
3．学会等名 The 5th International Conference on Fuzzy Systems and Data Mining (FSDM2019), Kitakyushu City, Japan, (国際学会)
4．発表年 2019年

1．発表者名 新美礼彦, 兼目 真生
2．発表標題 画像分割と深層学習を用いた不動産外観画像からの築年数推定
3．学会等名 第35回 ファジィ システム シンポジウム
4．発表年 2019年

1．発表者名 Ayahiko Niimi, Koki Takahata
2．発表標題 Attack Detection Method by Packet Analysis Using Online Learning Method and Correlation Change Method
3．学会等名 World Congress on Internet Security (WorldCIS-2019), London, UK (国際学会)
4．発表年 2019年

１．発表者名
Ayahiko Niimi

２．発表標題
Data mining with generating data to improve data imbalance problem

３．学会等名
International Conference on Big Data Analytics and Data Science, Las Vegas, USA（国際学会）

４．発表年
2019年

１．発表者名
Keisuke Tanaka, Ayahiko Niimi

２．発表標題
Word Topic Prediction Model for Polysemous Words and Unknown Words Using a Topic Model

３．学会等名
Computing Conference 2019, London, UK（国際学会）

４．発表年
2019年

１．発表者名
丸尾海月，新美礼彦

２．発表標題
時系列データを用いたデータストリームマイニングアルゴリズムの性能評価手法の検討

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（第18回日本データベース学会年次大会）,DEIM Forum 2020

４．発表年
2020年

１．発表者名
鳴海雄登，新美礼彦

２．発表標題
カテゴリカル属性の自動判別方法の提案

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（第18回日本データベース学会年次大会）,DEIM Forum 2020

４．発表年
2020年

１．発表者名
Ayahiko Niimi

２．発表標題
Study on Data Anonymization for Deep Learning

３．学会等名
The 31st International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE 2018)（国際学会）

４．発表年
2018年

１．発表者名
Ayahiko Niimi

２．発表標題
Word Topic Prediction Model Using a Topic Model

３．学会等名
The 4th International Conference on Fuzzy Systems and Data Mining (FSDM2018)（国際学会）

４．発表年
2018年

１．発表者名
Takeru Uchiyama and Ayahiko Niimi

２．発表標題
Similar Code Fragment Recommendation Using Word2Vec

３．学会等名
SCIS&ISIS with ISWS2018（国際学会）

４．発表年
2018年

１．発表者名
Ayahiko Niimi

２．発表標題
Data Anonymization Using Imbalanced Data for Deep Learning

３．学会等名
World Congress on Internet Security (WorldCIS-2018)（国際学会）

４．発表年
2018年

１．発表者名
Ayahiko Niimi

２．発表標題
Data Anonymization and Sampling Algorithms Using Imbalanced Datasets for Deep Learning

３．学会等名
Proceedings of the 10th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2019（国際学会）

４．発表年
2019年

１．発表者名
阪本 宏輔, 新美 礼彦

２．発表標題
オートエンコーダを使用したサンプリング手法による不均衡データの再現度向上

３．学会等名
研究報告情報基礎とアクセス技術（IFAT）

４．発表年
2018年

１．発表者名
高畑 孝輝 , 新美 礼彦

２．発表標題
相関関係の変化を利用した内部ネットワークの異常検知手法

３．学会等名
研究報告情報基礎とアクセス技術（IFAT）

４．発表年
2018年

１．発表者名
Ayahiko Niimi, Kousuke Sakamoto

２．発表標題
Accelerated Bayesian Optimization For Deep Learning

３．学会等名
6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017), Sydney, Australia（国際学会）

４．発表年
2017年

１．発表者名
Ayahiko Niimi

２．発表標題
Deep Learning for Real Credit Card Data

３．学会等名
Fuzzy Systems and Data Mining III, Proceedings of FSDM 2017, Hualien, Taiwan（国際学会）

４．発表年
2017年

１．発表者名
Ayahiko Niimi

２．発表標題
Majority Rule Approach of Deep Learning for Real Credit Card Transaction Data

３．学会等名
World Congress on Internet Security (WorldCIS-2017), Church College, University of Cambridge, Cambridge, UK（国際学会）

４．発表年
2017年

１．発表者名
田中 桂介，新美 礼彦

２．発表標題
トピックモデルによる話題推定手法に応じた多義語の意味切り分け

３．学会等名
第9回データ工学と情報マネジメントに関するフォーラム（第15回日本データベース学会年次大会）DEIM2017

４．発表年
2017年

１．発表者名
内山 武尊，新美 礼彦

２．発表標題
ソースコード特有の近傍単語の影響を考慮したWord2Vecを用いた類似コード片推薦手法

３．学会等名
ソフトウェアエンジニアリングシンポジウム2017論文集，Vol.2017

４．発表年
2017年

１．発表者名 高畑 孝輝，新美 礼彦
２．発表標題 侵入検知システムへのカーネル法を用いたオンライン学習手法の適用
３．学会等名 Computer Security Symposium 2017 (CSS2017)
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考