様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 ２０ 日現在

機関番号：１２６０１
研究種目：基盤研究(A)（一般）
研究期間：2011～2015
課題番号：２３２４００１９

研究課題名（和文）潜在的ダイナミックスの情報論的学習理論の研究

研究課題名（英文）Study on Information theoretic Learning Theory of Latent Dynamics

研究代表者
山西　健司（Yamanishi, Kenji）

東京大学・情報理工学（系）研究科・教授

研究者番号：９０５４９１８０

交付決定額（研究期間全体）：（直接経費）　　36,200,000円

研究成果の概要（和文）：大量のデータから価値ある情報を抽出するデータマイニング技術はますます重要になっている。従来技術では、データの表層的な関係性を抽出することが主であった。しかし、実際には、データの表面に現れない潜在的情報の動きを発見することが、より重要な知識発見をもたらす。そこで、本研究では、データに内在する潜在的情報とその変化（これを「潜在的ダイナミクス」と呼ぶ）を抽出するための技術の数理的基盤を構築した。特に、記述長最小原理に基づく情報論的学習理論の立場から統一的な理論体系を構築した。この理論を実際のデータ（セキュリティ、SNS、マーケティング、医療、教育等）に適用して、実世界での有効性を検証した。

研究成果の概要（英文）：Data mining technologies for extracting valuable information from big data are significantly important nowadays. The main purpose of conventional data mining technologies is to extract superficial statistical patterns from data. We are rather concerned with the issue of how to learn latent information behind data and how to detect its changes, which we call “latent dynamics.” We construct a theory for learning latent dynamics from a unifying view of information-theoretic learning theory, specifically on the basis of the minimum description length principle. We also demonstrate its effectiveness through the applications to real world data (e.g. security, SNS, marketing, healthcare, education, etc.).

研究分野：データマイニング

キーワード：情報論的学習理論　データマイニング　潜在的ダイナミクス　機械学習　ビッグデータ

１．研究開始当初の背景

（１）大量のデータから価値ある知識を発見することがますます重要になっている。そのような技術は「データマイニング」とよばれる。従来のデータマイニングでは、データの表層的な統計的パタンや相関関係といった顕在的情報を抽出することが主であった。しかし、実際には、データの表面に現れない潜在的情報のしかもその動的な変化を発見することが、より本質的に重要な知識発見をもたらすと考えられる。例えば、マーケットにおける購買傾向の変化は、ユーザ層の潜在的な構造の変化を知ることにより捉えることができる。しかしながら、そのような潜在的知識を抽出するためのデータマイニング技術は未発達であった。

（２）潜在的情報を表現するために、潜在変数モデルという知識表現手段が存在していた。例えば、混合モデル、トピックモデル、非負値行列因子分解といったモデルである。しかしながら、これらのモデルにおいてデータから最適な潜在的構造のモデル（最適な潜在変数の数など）を選択するための技術は未発達であった。また、従来、変化点検知技術は存在したが、データから最適な潜在的構造のモデル変化（潜在変数の数の変化や、クラスター構造の変化など）を抽出する技術もまた未発達であった。

２．研究の目的

（１）本研究では、上記の潜在的知識の構造的変化を「潜在的ダイナミクス」と呼ぶ。本研究の第一の目的は、潜在的ダイナミクスをデータから抽出するための機械学習理論を構築することである。

（２）第二の目的は、上記「潜在的ダイナミクス」を抽出する理論を現実のデータ解析に適用し、理論の有効性を実証することである。具体的には、マーケティング、セキュリティデータ、医療、SNS、教育などのデータを通じて、「潜在的ダイナミクス」抽出の事例を蓄積することである。

３．研究の方法

（１）本研究では、情報理論の立場から、潜在的ダイナミクスの機械学習にアプローチする。これを「潜在的ダイナミクスの情報論的学習理論」と呼ぶ。特に、情報理論の一大モデル選択原理である、「記述長最小原理(Minimum Description Length(MDL) Principle）に基づく方法論を基に展開する。MDL 原理とは、最もデータ圧縮できるモデルが最適なモデルであるという、モデル選択原理である。従来、MDL 原理は定常的なデータからの潜在変数を含まないモデルに対して適用されてきた。「潜在的ダイナミクス」の学習に MDL を適用するには､従来の MDL 原理を、非定常かつ、潜在変数モデルを対象にしたものに拡張しなければならない。これは決して自明ではない。非定常な状況に対しては、変化自体の情報を含めた記述長最小化が必要になる。また、潜在変数モデルは一般に確率モデルとしては非正則（パラメータと分布が一対一に対応しない）であるという問題がある。これらの問題を克服して、MDL 原理を拡張することにより、潜在的ダイナミクスの情報論的学習理論を構築する。

４．研究成果。

（１）混合モデルのモデル選択とクラスタリング構造変化検知：　ガウス混合分布とは、1 つのガウス分布をクラスターとし、複数のクラスターが線形結合しているモデルである。ガウス混合モデルは、データのクラスターへの割り当てを潜在変数とする潜在変数モデルである。データが与えられたとき、最適なクラスターの数を決定する問題は重要であり、この問題を解決する新しい手法を提案した。鍵となるアイデアは、完全変数化正規化最尤符号長と呼ばれる規準に基づいてこれを最小にするクラスター数を求めることにある。本手法においては、潜在変数モデルを完全変数化することにより、非正則性の問題を回避しつつ、MDL 原理に基づいてモデル選択することが可能になった。そのモデル選択手法は他のいかなるモデル選択手法を凌駕する性能を達成した、本成果は、ISIT2011 （学会⑲）で発表し、IEEE Transactions on Information Theory （201 に掲載された（雑誌⑦)。

また、多変量時系列データからガウス混合分布のクラスター数が変化する際に、この変化を検知するアルゴリズムを提案した。これは動的モデル選択(Dynamic Model Selection (DMS)とよばれる手法に基づいている。DMS とは、MDL 原理に基づいて変化するモデル系列を求める手法である。本提案手法では、DMS を潜在変数モデルに拡張することによって得られた。本提案アルゴリズムをビール購買層のクラスタリング変化検知に応用して有効性を実証した。本成果は KDD2012 にて発表した(学会⑰)。

（２）ネットワーク構造変化検知と広告効果測定への応用：多次元時系列データから、各変数の依存関係の変化を検知するアルゴリズムを開発した。本手法のポイントは、多変数の同時分布をベイジアンネットワークとよばれるグラフ形式で表現し、その構造変化を捉えることで、変数間の依存関係の変化を捉える点である。その際、データが与えられる毎に逐次的に、変化を検知する必要がある。この要請に対応すべく、逐次的 DMS アルゴリズムを開発し、逐次的にベイジアンネットワークの変化を検知することを可能にした。

本成果を、広告効果測定に応用して有効性を実証した。そこでは、特定の商品の売り上げ、広告出稿数、SNS での出現頻度などの多変量時系列データが与えられたもとで、広告出稿前後の変数の依存関係の変化を捉える

ことに成功した。本成果は ICDM2012 で発表し（学会⑯）、Data Mining and Knowledge Discovery 誌に掲載された（雑誌⑤）。

（３）Twitter からのトピック出現検知：Twitter 上の新しい話題の出現を早期に発見するアルゴリズムを開発した。通常、話題の検知には自然言語処理を用いて頻出語句を検知する手法を用いるが、ソーシャルネットでは自然言語以外のメディアを用いて話題を記述することも多く、自然言語以外の情報を用いて話題出現を検知する必要がある。そこで、本研究では、ユーザのリンク関係に着目し、リンク関係の変化を検知することにより話題の出現を検知する手法を開発した。曾その際、MDL 原理に基づいて忘却型逐次的正規化最尤符号長を変化スコアとするアルゴリズムを構築した。これを Twitter 解析に適用し、自然言語を利用するよりも 2 時間以上早く話題を検知できる場合があることを実証した。本成果は ICDM2011 で発表し（学会⑱）、IEEE Transactions on Knowledge and Data Engineering 誌に掲載された(雑誌⑧)。

（５）関係データのモデル選択：複数のアイテム間の関係を行列で表現したデータを関係データと呼ぶ。関係データの表現モデルとして確率ブロックモデルと呼ばれる潜在変数モデルを考え、データから最適なブロック数を選択するための新しい手法を開発した。これは、正則なモデルを選択するための MDL 原理を、非正則な潜在変数モデルのモデル選択に拡張して得られたものである。本手法を用いたモデル選択が AIC, BIC, ICL などの他のモデル選択規準に較べて高い精度を実現することを実証した。本成果は BigData2013 にて発表した（学会⑬）。図１参照。

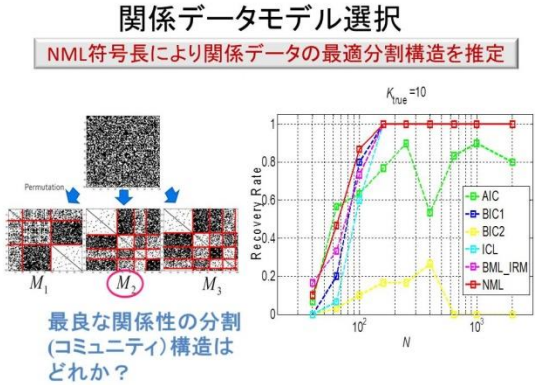

図１．関係データのモデル選択

（６）緑内障進行予測：緑内障は時間と共に視野が欠損する病気である。視野は 74 点で明るさが測定されているとする。緑内障進行予測とは、個々の患者の過去の視野データから将来の視野を予測する問題である。各患者のデータが十分にあれば、線形回帰を用いて将来の視野を高い精度で予測することができる。しかし、現実には、各患者のデータ数が予測に十分なほど多くなく、高い予測精度ができないという問題がある。そこで、時空間的な進行パターンをクラスタリングし、同じクラスターに属する患者のデータを全て用いて線形回帰で予測する手法を提案した。この手法により、単独の患者の 2 回分の測定データから、本来 8 回必要な予測精度を達成できることを実証した。本成果は ICDM2013 にて発表した（学会⑫）図 2 参照。

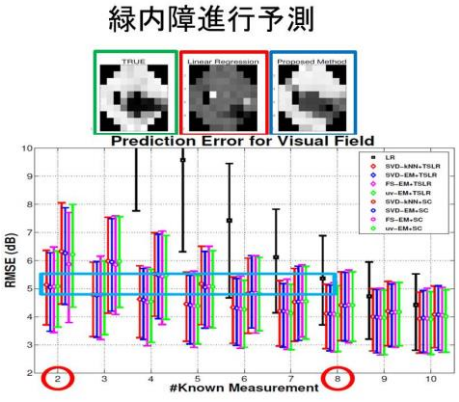

図２．緑内障進行予測

（７）教育データからのスキル構造変化検知：教育の現場では、大量の試験の成績データが時系列として取得され、生徒のスキル向上を正確に把握することが求められている。本研究では、オンライン非負値行列因子分解により潜在的なスキルの向上や停滞を把握する技術を開発した。本成果は EDM2013 及び EDM 2014 で発表した（学会⑮、⑧）。

５．主な発表論文等

〔雑誌論文〕（計　８　件）

①梶村俊介、馬場雪乃、梶野洸、鹿島久嗣：“列挙型クラウドソーシングタスクのための品質管理法”，人工知能学会論文誌 Vol. 31 (2016), No. 2 pp. K-F79_1-9.
DOI: http://doi.org/10.1527/tjsai.K-F79（査読有）．

②Shota Saito, Ryota Tomioka, and Kenji Yamanishi: “Early detection of persistent topics in social networks." Social Network Analysis and Mining, pp:5-19, Dec.2015. DOI: 10.1007/s13278-015-0257-1 (査読有).

③山西健司: 「複雑データからのディープナレッジの発見」 日本機械学会誌, Vol.118, No.1163, pp:8--11, 2015 年 10 月号（招待論文、査読無）.

④山西健司：「異常検知：外れ値検知と変化検知」日本信頼性学会誌,Vol.37, No.3,pp:134－－145, 2015 年 5 月号(招待論文､査読無).

⑤ Yu.Hayashi and Kenji.Yamanishi: “Sequential network change detection with its applications to ad impact relation

analysis," Data Mining and Knowledge Discovery: Volume 29, Issue 1 (2015), pp: 137-167, DOI: 10.1007/s10618-013-0338-6 (査読有).

⑥山西健司：「潜在的ダイナミクスの学習理論」電子情報通信学会誌、2014 年 5 月号、pp:422—426, (招待論文、査読無).

⑦So Hirai and Kenji Yamanishi: "Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering," IEEE Transactions on Information Theory Vol.59, Issue: 11, Aug.2013, DOI: 10.1109/TIT.2013.2276036 (査読有).

⑧Toshimitsu Takahashi, Ryota Tomioka, Kenji Yamanishi: "Discovering emerging topics in social streams via link anomaly detection," IEEE Transactions on Knowledge and Data Engineering, Vol.26, Issue: 1, pp:120 – 130, 2012, DOI: 10.1109/TKDE.2012.239 (査読有).

〔学会発表〕(計 ２０ 件)

① Yu Ito, Shin-ichi Oeda, and Kenji Yamanishi: "Rank selection for non-negative matrix factorization using normalized maximum likelihood coding," Proceedings of 2016 SIAM International Conference on Data Mining, Miami, US, May 4—8, 2016 (査読有).

②寺園泰、山西健司：「過完備辞書の再構成条件のブロックスパースモデルへの拡張」第 24 回情報論的学習理論と機械学習研究会（IBISML), 立川市、東京、統計数理研究所, 2016 年 3 月 17－18 日 (査読無).

③山西健司: 「進化する MDL－MDL の基礎から最近の発展まで - 」電子情報通信学会総合大会, 福岡、九州、九州大学、2016 年 3 月 17 日 (招待講演).

④Kohei Miyaguchi and Kenji Yamanishi: "On-line detection of continuous changes in stochastic processes," Proceedings of 2015 IEEE/ACM International Conference on Data Science and Advanced Analytics (DSAA'2015), pp:1--9, Paris, France, Oct.2015.（査読有）.

⑤ Shunsuke Kajimura, Yukino Baba, Hiroshi Kajino, and Hisashi Kashima: "Quality control for crowdsourced POI collection," Proceedings, Part II of the 19th Pacific-Asia Conference(PAKDD 2015), pp: 255-267 Ho Chi Minh City, Vietnam, May 19-22, 2015, (査読有).

⑥ Yoshiki Sakai and Kenji Yamanishi: "Data fusion using restricted Boltzmann machines ", Proceedings of IEEE International Conference on Data Mining (ICDM2013), pp: 953 - 958, Shenzen, China, Dec. 16th, 2014. (査読有).

⑦Shota Saito, Ryota Tomioka, and Kenji Yamanishi, "Early detection of persistent topics in social networks", Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM2014), pp: 417 – 424, Beijin, China, Aug.17th-20th, 2014. (査読有).

⑧ Shinichi Oeda, Yu Ito, and Kenji Yamanishi, "Extracting latent skills from time series of asynchronous and incomplete examinations", Proceedings of the 7th International Conference on Educational Data Mining (EDM2014), pp:367—368, London, UK, July 4th-7th, 2014. (査読有).

⑨山西健司：「潜在空間からのディープナレッジの発見」 応用統計学会 2014 年度大会, 立川、東京、統計数理研究所, 2014 年 5 月 22 日 （招待講演）.

⑩Jingling Wang, Satoshi Oyama,Masahito Kurihara, Hisashi Kashima: "Learning an accurate entity resolution model from crowdsourced labels," Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, Siem Reap, Cambodia, Jan.9th-10th, 2014, (査読有).

⑪ Shoichi Sato and Kenji Yamanishi: "Graph partitioning change detection using tree-based clustering," Proceedings of 2013 IEEE International Conference on Data Mining (ICDM2013)1, pp:1169-1174,Dallas, TX, US, Dec.8th-10th, 2013, (査読有).

⑫Zenghan Liang, Ryota Tomioka, Hiroshi Murata, Ryo Asaoka, and Kenji Yamanishi: "Quantitative prediction of glaucomatous visual field loss from few measurements," Proceedings of 2013 IEEE International Conference on Data Mining (ICDM2013), pp:1121-1126, Dallas, TX, US, Dec.8th-10th, 2013, (査読有).

⑬Yoshiki Sakai and Kenji Yamanishi: "An NML-based model selection criterion for general relational data modeling," 2013 IEEE International Conference on BigData (BigData2013), Santa Clara, CA, US,

Oct.6th-9th, 2013, (査読有).

⑭Kenji Yamanishi and Hiroki Kanazawa: "Stochastic complexity for piecewise stationary memoryless sources," 2013 Workshop on Information-theoretic Methods for Science and Engineering, Tokyo, Japan, 東大 , Aug.26th-29th, 2013, (招待講演).

⑮ Shin-ichi Oeda and Kenji Yamanishi: "Extracting time-evolving latent skills from examination time series," The 6th International Conference on Educational Data Mining(EDM2013), Memphis, TN, US, July 6th-9th, 2013 (査読有).

⑯ Yu Hayashi and Kenji Yamanishi: "Sequential network change detection with its applications to ad impact relation analysis," 2012 IEEE International Conference on Data Mining(ICDM2012), pp:280-289, Brissel, Belgium, Dec.10th-13th, 2012, (査読有).

⑰ So Hirai and Kenji Yamanishi: "Detecting changes of clustering structures using normalized maximum likelihood coding," Proceedings of 2012 ACM International Conference on Knowledge Discovery and Data Mining (KDD2012), pp:343-351, Beijing China, Aug.12th-16th , 2012, (査読有).

⑱Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi: "Discovering emerging topics in social streams via link anomaly detection," Proceedings of 2011 IEEE International Conference on Data Mining (ICDM2011), pp:1230-1235, Vancouver, Canada, Dec.11th-14th,2011, (査読有).

⑲ So Hirai and Kenji Yamanishi: "Efficient computation of normalized maximum likelihood coding for Gaussian mixtures with its applications to optimal clustering," Proceedings of 2011 IEEE International Symposium on Information Theory, pp:1031-1035, St. Petersburg, Russia, July 31st-Aug.5th, 2011, (査読有).

⑳Yasuhiro Urabe, Kenji Yamanishi, Ryota Tomioka, and Hiroki Iwai, "Real-time change-point detection using sequentially discounting normalized maximum likelihood coding", Proceedings Part II of 15th Pacific-Asia Conference, (PAKDD 2011), pp: 185-197, Shenzhen, China, May 24-27, 2011, (査読有).

〔図書〕(計 2 件)
①山西健司:「情報論的学習とデータマイニング」朝倉書店(数理工学ライブラリー3), 2015, 167.

②鹿島久嗣(共著):「ビッグデータマネジメント - データサイエンティストのためのデータ利活用技術と事例」NTS 出版、356

〔産業財産権〕
○出願状況(計 0 件)
○取得状況(計 0 件)

〔その他〕ホームページ等
山西研究室
http://ibis.t.u-toyo.ac.jp/yamanishiken
鹿島研究室
http://www.ml.ist.kyoto-u.ac.jp

6. 研究組織
(1)研究代表者
山西 健司 (YAMANIHSI, Kenji)
東京大学 情報理工学系研究科 教授
研究者番号:90549180

(2)研究分担者
鹿島 久嗣 (KASHIMA, Hisashi)
京都大学 情報学研究科 教授
研究者番号: 80545583
冨岡 亮太 (TOMIOKA、Ryota)
Microsoft Research, Researcher
研究者番号: 70518282

(3)連携研究者
(4)研究協力者
寺園 泰 (TERAZONO, Yasushi)
櫻井 瑛一 (SAKURAI, Ei-ichi)
浦部 泰宏 (URABE, Yasuhiro)
岩井 宏樹 (IWAI, Hiroki)
高橋 俊充 (TAKAHASHI, Toshimitsu)
平井 聡 (HIRAI, So)
早矢仕 裕 (HAYASHI, Yu)
金澤 宏紀 (KANAZAWA, Hiroki)
佐藤 翔一 (SATO, Sho-ichi)
坂井 良樹 (SAKAI, Yoshiki)
梁 曾漢 (LIANG, Zenghan)
大枝 真一 (OEDA, Shin-ichi)
朝岡 亮 (ASAOKA, Ryo)
村田 博史 (MURATA, Hiroshi)
斎藤 翔太 (SAITO, Shota)
伊藤 優 (Ito, Yu)
宮口 航平 (MIYAGUCHI, Kohei)
梶村 俊介 (KAJIMURA, Shun-suke)
馬場 雪乃 (BABA、Yukino)
小山 聡 (OYAMA, Satohi)
栗原 正仁 (KURIHARA, Masahito)
Jingling Wang(WANG, Jingling)