様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月　９日現在

機関番号：１７１０２
研究種目：基盤研究（B）
研究期間：2009～2012
課題番号：21300053
研究課題名（和文）　動的表現バイアスに基づくマルチタスクデータマイニング

研究課題名（英文）　Multi-task Data Mining Based on Dynamic Representation Bias

研究代表者
鈴木　英之進（SUZUKI EINOSHIN）
九州大学・システム情報科学研究院・教授
研究者番号：10251638

研究成果の概要（和文）：
互いに関連する複数のパターン発見タスクに効果的に対処するために，データとパターンの表現形式を自動的に変更する新しいデータマイニング手法を開発し，計算機システムとして実装して人工・実データでその有効性を示した．顕著な成果は，共通辞書を許容するように拡張した MDL 原理を用いるマルチタスク分類学習手法，コルモゴロフ複雑性に基づく情報量距離の拡張版を用いるマルチタスククラスタリング手法，マルチタスクデータマイニング用の次元縮退手法である．
(200 字程度)

研究成果の概要（英文）：
To effectively cope with multiple pattern discovery tasks related each other, we have developed novel data mining methods each of which automatically modifies representations of data and patterns, implemented them as computer systems, and demonstrated their effectiveness with synthetic and real data. Remarkable achievements are multi-task classification method which employs an extended MDL principle to allow a common dictionary, a multi-task clustering method which employs an extension of information distance based on Kolmogorov complexity, and a dimension reduction method for multi-task data mining.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	3,400,000	1,020,000	4,420,000
２０１０年度	4,300,000	1,290,000	5,590,000
２０１１年度	2,300,000	690,000	2,990,000
２０１２年度	2,300,000	690,000	2,990,000
年度			
総　計	12,300,000	3,690,000	15,990,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：動的表現バイアス、マルチタスクデータマイニング、分類学習，クラスタリング

１．研究開始当初の背景

研究代表者は，データマイニング研究分野において，興味深い例外性を表す構造ルール発見，情報理論に基づく事例発見，新規理論や新規手法に基づく各種応用，発見パターンや

発見手法の評価，大量データの圧縮による効率的発見などに関する研究を通して，研究コミュニティで一定の信頼を得るに至っていた．
その過程において，「データマイニング＝データからパターンを発見する孤立した単一タスク」という見方に強い疑問を抱くようになった．実際のデータマイニングでは，互いに関連する複数個のタスクを行い，各タスクにおいてパターンを発見するのが一般的である．そのようなマルチタスクデータマイニングでは，データとパターンの表現形式を改良することが可能であり，この機能は効果的な発見にとってきわめて重要である．
申請者は，そのような表現形式の改良法が，領域知識を理論的基盤に基づいて有効活用する情報圧縮性基準に基づくべきだと考えた．通常のデータマイニング手法は，最低 2-3 個のパラメータを指定する必要があり，データへの過適合などの欠点が起きる場合があることが当時から知られていた．情報圧縮可能性に基づくデータマイニングは，コルモゴロフ複雑性や最小記述長(MDL)原理などの理論的背景を有し，指定パラメータが０個か１個であり，過適合の危険性が大幅に低いことが知られている．もっとも，データやパターンの表現形式が適切であることが条件であり，領域知識の統一的枠組みに基づく利用法は不明であった．

２．研究の目的

本研究では，互いに関連しあう複数個のパターン発見を，データとパターンの表現形式を変更しながら効果的に行う動的表現バイアスに基づくマルチタスクデータマイニング手法を考案・開発し，計算機システムとして実装して複数の人工・実データでその有効性を示すことを目的とした．

３．研究の方法

動的表現バイアスに基づくマルチタスクデータマイニングを，情報圧縮可能性に関する基盤理論，表現形式改良の駆動方式，知識発見手法，データマイニング応用の４要素に分割し，並列して取り組んだ．
情報圧縮性に関する基盤理論として，ルール群発見用に開発済みの拡張 MDL 原理，およびコルモゴロフ複雑性に基づく情報距離を選択し，研究目標用に汎用化していった．
表現形式改良の手がかりとして，データ・仮説・知識駆動方式を選択し，基盤理論と知識発見手法との相性を考慮しつつ開発した．
知識発見手法としては，分類学習，クラスタリング，次元縮退，ルール発見を選択し，研究目標用に拡張あるいは新規開発した．
データマイニング応用としては，主にウェブマイニングとバイオインフォマティクスに取り組み，ベンチマークデータや人工データを用いた系統的実験も行った．

４．研究成果

初期仮説を許容するように拡張した MDL 原理を用いるルール群発見手法，共通辞書を許容するように拡張した MDL 原理を用いるマルチタスク分類学習手法，コルモゴロフ複雑性に基づく情報量距離の拡張版を用いるマルチタスククラスタリング手法，マルチタスクデータマイニング用の次元縮退手法を考案・実装し，多くの種類の人工データと実データに適用して有効性を示した．その結果として，情報圧縮基準と動的表現バイアスに基づくマルチタスクデータマイニングを確立した．特に，後 3 件の研究は国際ジャーナル Knowledge and Information Systems (KAIS) に論文が採択され，質の高い国際的な成果発信が達成できたと考える．他に，動的マルチタスク学習や高ノイズデータからの転移学習など関連する課題・応用に関しても成果を挙げた．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２６件）
注意：DOI も URL もない論文は，
http://www.informatik.uni-trier.de/~ley/pers/hd/s/Suzuki:Einoshin.html からアクセス可能．
1. Thach Nguyen Huy, Hao Shao, Bin Tong, Einoshin Suzuki: "A Feature-Free and Parameter-Light Multi-Task Clustering Framework", Knowledge and Information Systems, An International Journal, Springer (accepted for publication). DOI 10.1007/s10115-012-0505-x
2. Bin Tong, Junbin Gao, Thach Nguyen Huy, Hao Shao, Einoshin Suzuki: "Transfer Dimensionality Reduction by Gaussian Process in Parallel", Knowledge and Information Systems, An International Journal, Springer (accepted for publication). DOI 10.1007/s10115-012-0601-y
3. Thach Nguyen Huy, Bin Tong, Hao Shao, Einoshin Suzuki: "Transfer Learning by Centroid Pivoted Mapping in Noisy Environment", Journal of Intelligent Information Systems, Springer (accepted for publication). DOI 10.1007/s10844-012-0226-3

4. Shin Ando, Einoshin Suzuki: "Time-sensitive Classification of Behavioral Data", Proc. Thirteenth SIAM International Conference on Data Mining (SDM 2013), pp. 458-466, 2013.
5. Hao Shao, Bin Tong, Einoshin Suzuki: "Extended MDL Principle for Feature-Based Inductive Transfer Learning", Knowledge and Information Systems, An International Journal, Vol. 35, No. 2, pp. 365-389, Springer, 2013. DOI 10.1007/s10115-012-0505-x
6. Bin-Hui Chou, Einoshin Suzuki: "RoClust: Role Discovery for Graph Clustering", Web Intelligence and Agent Systems, An International Journal, Vol. 11, No. 1, pp. 1-20, IOS Press, 2013. DOI 10.3233/WIA-130259
7. Bin-Hui Chou, Einoshin Suzuki: "Detecting Academic Plagiarism with Graphs", Extraction et Gestion des Connaissances (EGC'2013), pp. 293-304, 2013. http://www.egc.asso.fr/に問い合わせること.
8. Hao Shao, Bin Tong, Einoshin Suzuki: "Query by Committee in a Heterogeneous Environment", Advanced Data Mining and Applications (ADMA 2012), pp. 186-198, LNAI 7713, Springer, 2012.
9. Bin Tong, Hao Shao, Bin-Hui Chou, Einoshin Suzuki: "Linear Semi-Supervised Projection Clustering by Transferred Centroid Regularization", Journal of Intelligent Information Systems, Vol. 39, No. 2, pp. 461-490, Springer, 2012. DOI 10.1007/s10844-012-0198-3
10. Bin Tong, Weifeng Jia, Yanli Ji, Einoshin Suzuki: "Linear Semi-Supervised Dimensionality Reduction with Pairwise Constraint for Multiple Subclasses", IEICE Transactions on Information and Systems, Vol. E95-D, No. 3, pp. 812-820, 2012. DOI 10.1587/transinf.E95.D.812
11. Shin Ando, Einoshin Suzuki: Role-Behavior Analysis from Trajectory Data by Cross-Domain Learning, Proc. Eleventh IEEE International Conference on Data Mining (ICDM 2011), pp. 21-30, 2011.
12. Hiroshi Hirai, Bin-Hui Chou, Einoshin Suzuki: A Parameter-Free Method for Discovering Generalized Clusters in a Network, Discovery Science (DS 2011), LNAI 6926, Springer, pp. 135-149, 2011.
13. Hao Shao, Bin Tong, Einoshin Suzuki: "Compact Coding for Hyperplane Classifiers in Heterogeneous Environment", Machine Learning and Knowledge Discovery in Databases (ECML/PKDD 2011), Part III, LNCS 6913, Springer, pp. 207-222, 2011.
14. Thach Nguyen Huy, Shao Hao, Bin Tong, Einoshin Suzuki: "A Compression-Based Dissimilarity Measure for Multi-Task Clustering", Foundations of Intelligent Systems, LNAI 6804 (ISMIS 2011), pp. 123-132, Springer, 2011.
15. Shin Ando, Theerasak Thanomphongphan, Daisuke Hoshino, Yoichi Seki, Einoshin Suzuki: "ACE: Anomaly Clustering Ensemble for Multi-Perspective Anomaly Detection", Proc. Eleventh SIAM International Conference on Data Mining (SDM 2011), pp. 1-12, 2011.
16. Hao Shao, Einoshin Suzuki: "Feature-Based Inductive Transfer Learning through Minimum Encoding", ibid, pp. 259-270, 2011.
17. Bin Tong, Junbin Gao, Thach Nguyen Huy, Einoshin Suzuki: "Gaussian Process for Dimensionality Reduction in Transfer Learning", ibid, pp. 783-794, 2011.
18. Bin-Hui Chou, Einoshin Suzuki: "Role Discovery for Graph Clustering", Web Technologies and Applications (APWeb 2011), pp. 17-28, LNCS 6612, Springer, 2011.
19. Einoshin Suzuki: "Novel Statistical Rule Discovery for Understanding Behaviours of Swarm Robots", Proc. Fifth International Meeting on Statistical Implicative Analysis (A.S.I. 5). pp. 452-454, 2010. http://sites.univ-lyon2.fr/asi5/resum/ResASI5_EN001Poster.pdf
20. Bin Tong, ZhiGuang Qin, Einoshin Suzuki: "Topology Preserving SOM with Transductive Confidence Machine", Discovery Science (DS 2010), LNAI 6332, Springer, pp. 27-41, 2010.
21. Bin Tong, Hao Shao, Bin-Hui Chou, Einoshin Suzuki: "Semi-Supervised Projection Clustering with Transferred Centroid Regularization", Machine Learning and Knowledge Discovery in Databases (ECML/PKDD 2010), Part III, LNCS 6323, Springer, pp. 306-321, 2010.
22. Bin Tong, Einoshin Suzuki: "Subclass-Oriented Dimension Reduction with Constraint Transformation and Manifold Regularization", Advances in Knowledge Discovery and Data Mining (PAKDD 2010), Part II, LNAI 6119, Springer, pp. 1-13, 2010.
23. JianBin Wang, Bin-Hui Chou, Einoshin Suzuki: Finding the k-Most Abnormal

Subgraphs from a Single Graph, Discovery Science, Lecture Notes in Artificial Intelligence 5808 (DS), Springer, pp. 441-448, 2009.
24. Daisuke Ikeda, Einoshin Suzuki: Mining Peculiar Compositions of Frequent Substrings from Sparse Text Data using Background Texts, Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), Vol. 1, LNAI 5781, Springer, pp. 596-611, 2009.
25. Einoshin Suzuki: "Compression-Based Measures for Mining Interesting Rules", Next-Generatation Applied Intelligence (IEA/AIE), LNAI 5579, pp. 741-746, Springer, 2009 (invited talk at a special session).
26. Einoshin Suzuki: "Interestingness Measures - Limits, Desiderata, and Recent Results -", Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models Workshop (QIMIE'09), pp. 1-3, in conjunction with The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2009. http://conferences.telecom-bretagne.eu/data/qimie09/suzuki_QIMIE_2009.pdf

〔学会発表〕(計２０件)
1. Shin Ando: "Time-sensitive Classification of Behavioral Data", Thireenth SIAM International Conference on Data Mining (SDM 2013), Austin, Texas, USA, 2013/5/2.
2. Bin-Hui Chou: "Detecting Academic Plagiarism with Graphs", 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2013), Toulouse, France, 2013/1/31.
3. Hao Shao: "Query by Committee in a Heterogeneous Environment", Eighth International Conference on Advanced Data Mining and Applications (ADMA 2012), Nanjing, China, 2012/12/16.
4. Shin Ando: Role-Behavior Analysis from Trajectory Data by Cross-Domain Learning, Eleventh IEEE International Conference on Data Mining (ICDM 2011), Vancouver, Canada, 2011/12/12.
5. Einoshin Suzuki: A Parameter-Free Method for Discovering Generalized Clusters in a Network, Fourteenth International Conference on Discovery Science (DS 2011), Espoo - Helsinki, Finland, 2011/10/5.
6. Hao Shao: "Compact Coding for Hyperplane Classifiers in Heterogeneous Environment", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2011), Athens, Greece, 2011/9/6.
7. Thach Nguyen Huy: "A Compression-Based Dissimilarity Measure for Multi-Task Clustering", 19th International Symposium on Methodologies for Intelligent Systems (ISMIS 2011), Warsaw, Poland, 2011/6/28.
8. Theerasak Thanomphongphan: "ACE: Anomaly Clustering Ensemble for Multi-Perspective Anomaly Detection", Eleventh SIAM International Conference on Data Mining (SDM 2011), Phoenix/Mesa, Arizona, USA, 2011/4/28.
9. Hao Shao: "Feature-Based Inductive Transfer Learning through Minimum Encoding", ibid, 2011/4/28.
10. Hao Shao: "Gaussian Process for Dimensionality Reduction in Transfer Learning", ibid, 2011/4/28.
11. Bin-Hui Chou: "Role Discovery for Graph Clustering", Thirteenth Asia-Pacific Web Conference (APWEB 2011), Beijing, China, 2011/4/18.
12. Einoshin Suzuki: "Novel Statistical Rule Discovery for Understanding Behaviours of Swarm Robots", Fifth International Meeting on Statistical Implicative Analysis (A.S.I. 5), Palermo, Italy, 2010/11/6.
13. Bin Tong: "Topology Preserving SOM with Transductive Confidence Machine", Thirteenth International Conference on Discovery Science (DS 2010), Canberra, Australia, 2010/10/8.
14. Bin Tong: "Semi-Supervised Projection Clustering with Transferred Centroid Regularization", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2010), Barcelona, Spain, 2010/9/23.
15. Bin Tong: "Subclass-Oriented Dimension Reduction with Constraint Transformation and Manifold Regularization", 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010), Hyderabad, India, 2010/6/23.
16. Thach Nguyen Huy : "A Symbolic Representation for Trajectory Data", 第24回人工知能学会全国大会, 長崎, 2010/6/9.
17. Einoshin Suzuki: Finding the k-Most Abnormal Subgraphs from a Single Graph, Twelfth International Conference on

Discovery Science (DS 2009), Porto, Portugal, 2009/10/4.
18. Daisuke Ikeda: Mining Peculiar Compositions of Frequent Substrings from Sparse Text Data using Background Texts, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2009), Bled, Slovenia, 2009/9/9.
19. Einoshin Suzuki: ″Compression-Based Measures for Mining Interesting Rules″, Twenty Second International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE 2009), Tainan, Taiwan, 2009/6/25 (invited talk at a special session).
20. Einoshin Suzuki: ″Interestingness Measures - Limits, Desiderata, and Recent Results -″, Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models Workshop (QIMIE'09), in conjunction with The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand, 2009/4/27 (keynote talk).

〔その他〕
ホームページ等
http://www.i.kyushu-u.ac.jp/~suzuki/kaken0912-j.html
http://www.i.kyushu-u.ac.jp/~suzuki/kaken0912.html

６．研究組織
(1)研究代表者
鈴木　英之進 (Suzuki Einoshin)

研究者番号：10251638

(2)連携研究者
童　彬　　(Tong Bin)
九州大学・システム生命科学府, 博士課程学生

邵　浩　　(Shao Hao)
九州大学・システム生命科学府, 博士課程学生

グエン　ヒ　テアク　　(Nguyen Huy Thach)
九州大学・システム生命科学府, 博士課程学生

菅谷　信介　　(Sugaya Shinsuke)
九州大学・システム情報科学府, 博士課程学生