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　 ７ 日現在

機関番号：１７１０２
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８００８５

研究課題名（和文）自動データ選択とパターン抽出の統合による巨大データ集合体からの知識発見

研究課題名（英文）Knowledge Discovery from Huge Data Ensemble by an Integration of Automatic Data Selection and Pattern Extraction

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科（研究院・教授

研究者番号：１０２５１６３８

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：多数の巨大データ集合から知識を発見するために，自動データ選択とパターン発見を統合した新規性が高いデータマイニング手法を4種類考案・開発・実装した．クラスタ分布メタパターン群を発見する手法は，高ノイズに汚染されクラスタ境界が曖昧で重なり合う悪条件下でも高い再現率と適合率を示し，高速であった．方向性非零重みメタパターン群を発見する手法は，スパースモデリングに基づくマルチタスク分類学習手法に基づき，Kinectで計測した顔表情データなどで実用性を示した．線形分類子の階層クラスタリング手法とそれぞれ特定のデータ集合で成立する一般分類ルール群を評価・発見する手法も，種々の人工・実データで有効性を示した．

研究成果の概要（英文）：To discover knowledge from a large number of huge data sets, we invented, developed, and implemented 4 highly novel methods that integrate automatic data selection and pattern discovery. The method that discovers cluster distribution meta-patterns exhibited high recalls and precisions under difficult conditions of high noise contamination and ambiguous and mutually overlapping cluster boundaries and was proved to be time-efficient. The method that discovers directional non-zero weight meta-patterns, which is based on multi-task classification based on sparse modeling, showed its practicability on various kinds of data including facial expression data measured with Kinect. The method that hierarchically clusters linear classifiers and the method that evaluates and discovers general classification rules each holding true in its respective data set showed their effectiveness on various synthetic and real data.

研究分野：データマイニング

キーワード：巨大データ集合　自動データ選択　パターン抽出　データマイニング

１．研究開始当初の背景

巨大なデータから有用な知識の発見を目的とするデータマイニングは，基礎と応用の両面において種々の成功を収めてきた．データマイニングの手順は，図１に示すように，巨大データから選択・前処理・変換を経て主に表形式の変換データを得，そこからパターンを抽出した後に解釈・評価を経て知識を得る手順を，運用と問題形式化を含めて試行錯誤的に反復する KDD プロセスモデルとしてモデリングされる[Fayyad 1996]．種々のデータマイニング応用において労力の 8 割が，最もよく研究されているパターン抽出ではなく，選択・前処理・変換に要することが大きな問題となっていた．

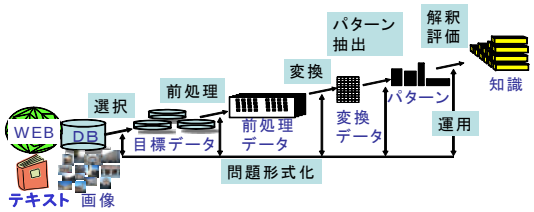

図１　KDD プロセスモデル

巨大データは，まったく構造を持たない場合は稀であり，きわめて多数のデータ集合から構成される巨大データ集合体と見なせる場合が多い．これまでのデータマイニングでは，このような巨大データ集合体に内在する構造はユーザが選択と前処理の過程でアドホックに利用し，その是非は抽出されたパターン群をユーザが解釈・評価することで判明していた．多大な労苦はこのような過度の人手依存が原因と見なせ，ユーザがより高度な知的作業に専念できるよう，自動データ選択とパターン抽出を統合することが望ましいと考えた．

２．研究の目的

本研究では，きわめて多数の巨大データ集合から興味深く有用な知識を発見するために，自動データ選択とパターン発見を統合した新規性が高い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人工・実データでその有効性を示す．この手法は，一部のデータ集合の各々からパターン群を発見してメタパターン群としてまとめ，残りのデータ集合をそれらと照合してそこから新たにパターンを発見するデータ集合を自動選択し，メタパターン群を洗練することにより，巨大データ集合体から興味深く有用な知識をきわめて高速に発見する．巨大データ集合体が得やすい人見守りとウェブマイニングに関する応用，および人工データ・ベンチマークデータでの検証に取り組み，提案手法を発見知識の興味深さ，有用性，時間・領域的効率，ユーザへの負荷などの観点から評価する．

３．研究の方法

(1)各例が数値属性で記述されるクラスなしデータアンサンブルから，クラスタ分布メタパターン群を発見する新しい問題とそれを解決する効率的なアルゴリズムを考案し，人工データと実データに適用してその有効性を調べた．クラスタ分布メタパターンは，特定データ集合の各データ集合が発見クラスタ群の類似線形重み群として表されることを示す．効率的アルゴリズムは，データスカッシングに基づく 1 スキャンクラスタリング手法である BIRCH [Zhang 97]に基づいてマイクロクラスター群を生成し，類似マイクロクラスター群の反復的にマージした後に，KL 情報量基準に基づいてクラスタ分布メタパターン群を発見する．人工データは，クラスタ境界がはっきりしない上に互いに重なり合う，きわめて高ノイズに汚染された場合を含み，関連パラメータを変化させて種々の状況をシステマチックに調べることができる．実データは総例数約 26 万超であり，Kinect で計測した 100 人の 25 種類の顔表情に関する．

(2)各例が数値属性で記述されるクラスつきデータアンサンブルから，方向性非零重みメタパターン群を発見する新しい問題とそれを解決する方式を考案し，実データに適用してその有効性を調べた．クラスつきデータアンサンブルは，マルチタスク分類学習問題と見なすことができ，データ集合間の相互関係をとらえる潜在空間を求め，興味深いメタパターンを考案することを考えた．その手段として，スパースモデリングに基づくマルチタスク分類学習手法 ELLA [Ruvolo 2013]を選択し，例外的に非零となる基底重みの符号に着目した．方向性非零重みメタパターンは，特定データ集合の各データ集合が特定の非零重みの同じ符号を共有することを示す．実データは，(1)で用いた実データの内，表情をクラスとしてラベリングした総例数 62,500 の部分データ集合に相当する．

(3) 各例が数値属性群で記述されるクラスつきデータアンサンブルから，各データ集合から学習される線形分類子を階層型クラスタリングする新しい問題とそれを解決する方式を考案し，人工データ，ベンチマークデータ，新規に計測した実データに適用してその有効性を調べた．線形分類子の類似性はコサイン類似度に基づく基準で判断し，3 クラス以上の場合などのために 1 対残り方式と 1 対 1 方式という 2 種類の拡張も提案した．人工データとしては，[Scott 1999]の方法にしたがい，18 万例から構成されるデータアンサンブルを生成した．ベンチマークデータとしては，UCI 機械学習レポジトリ中の母音データを[Widmer 1996]の文脈属性を用いてデータアンサンブルとした．実データの 1 種類目は(2)で用いた部分データ集合であり，意図的表情だけを含む．2 種類目は新規に測定し

た 2 名の実験参加者に関する意図的表情と自然表情の両方を含むデータである．

(4)データ集合と信念が指定された場合における一般分類ルールの興味深さ規準とその規準を用いた効率的な発見アルゴリズムを考案し，ベンチマークデータに適用してその有効性を調べた．まず，一般分類ルールが圧縮する情報量を表す J 値[Smyth 1992]を，入力がデータアンサンブルでありかつ同じ結論部を持つ信念も入力された場合に拡張した．次に，メタパターン発見のための深さ優先探索において，前提部が特殊化される際の上限値を求め，この上限値を用いる分岐限定法に基づく効率的な発見アルゴリズムを考案した．ベンチマークデータとしては，(3)で用いた UCI 機械学習レポジトリ中から，投票データとマッシュルームデータを選択し，各属性を文脈属性[Widmer 1996]として用い，前提部 1 の任意の強い規則性を信念と見なした場合を網羅的に調べた．

４．研究成果

(1) クラスタ分布メタパターンについては，図 1 に人工データを用いた実験結果の一部を示す．きわめて悪い条件下にも関わらず，提案手法は驚くほど高い再現率と適合率を示した．比較対象としてクラスタリングにおいて広く用いられる k-平均法をランダムリスタート数 10，正解のクラスタ数 k=27 で採用した．提案手法は k-平均法よりも約 40 倍高速である上，k -平均法が実際上無力となる総例数 7 千万個以上の場合でも実行時間 7 分間未満と実用的であった．図 2 に実データに関する実験結果の一部を示す．実行時間は最長 8.5 分であり，発見メタパターン数は 684 個から約 47 万個であった．

(2)方向性非零重みメタパターンについては，図 3 に実データに関する実験結果の一部を示す．発見されたメタパターンを前提部の長さに関して種類分けし，低被覆（1-32 データ集合），中被覆（33-56 データ集合），高被覆（57-100 データ集合）についてのパターン数も調べた．図より，パターン数は 1,000 個未満と許容範囲であり，特に重要と考える高被覆の場合にはその数が少なく良いことが分かった．

(3) 各々のデータ集合から学習される線形分類子群の階層型クラスタリングについて，人工データとベンチマークデータを用いた実験は良好な結果を示した．図 4 に，意図的表情だけを含む 1 種類目の実データを用いた実験結果を示す．提案手法のグループ内距離は比較的小さく，グループ間距離は比較的大きい．

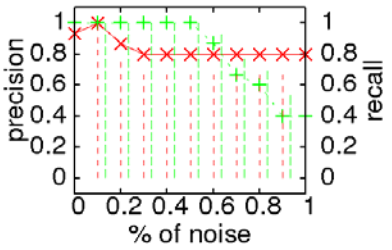

図 1　クラスタ分布メタパターンについて，人工データを用いた実験結果の一部．赤は提案手法，緑は比較手法．折れ線グラフが適合率，棒グラフが再現率

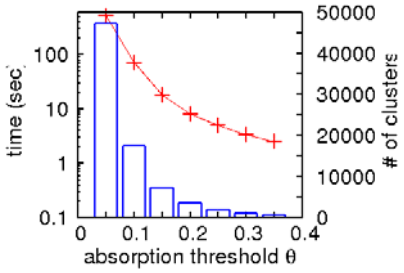

図 2　クラスタ分布メタパターンについて，実データを用いた実験結果の一部．折れ線グラフが実行時間，棒グラフがクラスタ数

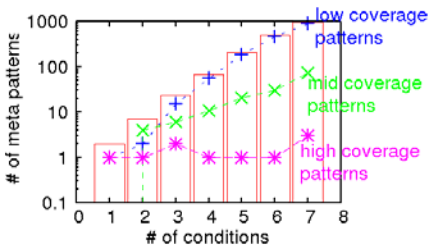

図 3　方向性非零重みメタパターンの個数について，実データを用いた実験結果の一部

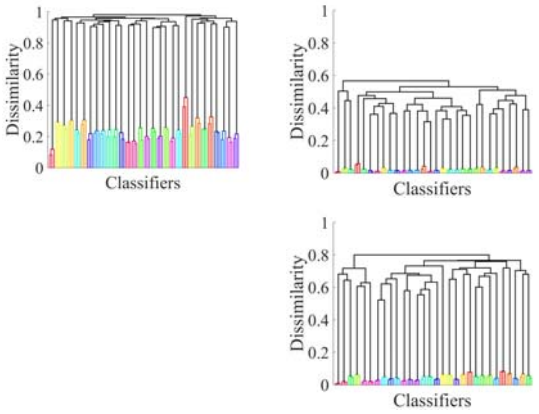

図 4　各々から学習される線形分類子の類似度に基づくデータ集合の階層型クラスタリングについて，意図的表情を含む 1 種類目の実データを用いた実験結果の一部．左上から時計回りに，[Tsoumakas 2004]の比較手法，1 対残り方式を用いた提案手法，1 対 1 方式を用いた提案手法

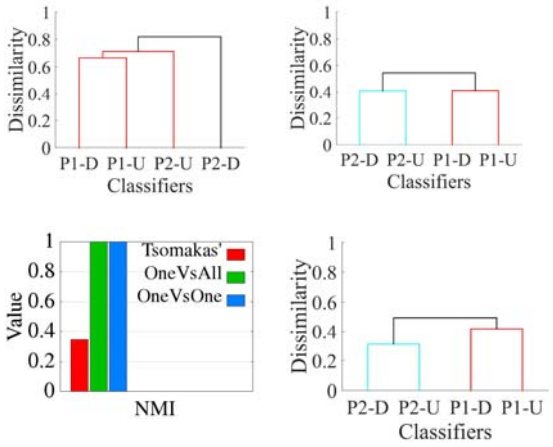

図５　各々から学習される線形分類子の類似度に基づくデータ集合の階層型クラスタリングについて，意図的表情と自然表情を含む２種類目の実データを用いた実験結果の一部. 左上から時計回りに, [Tsoumakas 2004]の比較手法, 1 対残り方式を用いた提案手法, 1 対 1 方式を用いた提案手法，３手法の NMI

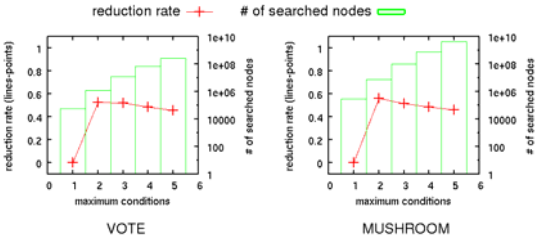

図６　特定のデータ集合において成立する一般分類ルールに関するメタパターン発見手法について，提案した分岐限定法の効果を示す実験結果の一部

図５には意図的表情と自然表情の両方を含む２種類目の実データを用いた実験結果を示す．このデータは２種類の表情を含むため背景にある真のデータを適切に表していないことが判明し，[Tsoumakas 2004]の比較手法は事例の予測差異に依存するため性能が悪い．一方，分類境界面を考慮する提案手法はより頑健であり，高い NMI（正規化相互情報量）を達成することができた．

(4)特定のデータ集合において成立する一般分類ルールに関するメタパターン発見手法について，提案した分岐限定法の効果を示す実験結果の一部を図６に示す．図より，約半数の探索ノードを発見結果を変えることなく枝刈りできたことが分かる．探索ノードが 1 億個を超える場合があることを考えると，この高速性は貴重である．

＜引用文献＞

[Fayyad 1996] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy: From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining, pp. 1-34, 1996.

[Ruvolo 2013] P. Ruvolo, E. Eaton: ELLA: An Efficient Lifelong Learning Algorithm, Proc. ICML, Vol. 1, pp. 507-515, 2013.

[Scott 1999] P. D. Scott, E. Wilkins: Evaluating Data Mining Procedures: Techniques for Generating Artificial Data Sets, Information and Software Technology, Vo. 41, No. 9, 579-587, 1999.

[Smyth 1992] P. Smyth, R. M. Goodman: An Information Theoretic Approach to Rule Induction from Databases, IEEE Trans. Knowl. Data Eng., Vol. 4, No. 4, pp. 301-316, 1992.

[Tsoumakas 2004] G. Tsoumakas, L. Angelis, I. P. Vlahavas: Clustering Classifiers for Knowledge Discovery from Physically Distributed Databases, Data Knowl. Eng., Vol. 49, No. 3, pp. 223-242, 2004.

[Widmer 1996] G. Widmer, M. Kubat: Learning in the Presence of Concept Drift and Hidden Contexts, Machine Learning, Vol. 23, No. 1, pp. 69-101, 1996.

[Zhang 97] T. Zhang, R. Ramakrishnan, M. Livny: BIRCH: An Efficient Data Clustering Method for Very Large Databases, Proc. SIGMOD Conference, pp. 103-114, 1996.

５．主な発表論文等

〔雑誌論文〕（計１７件）

1. Y. Deguchi, D. Takayama, S. Takano, V.-M. Scuturici, J.-M. Petit, E. Suzuki: Skeleton Clustering by Multi-Robot Monitoring for Fall Risk Discovery, Journal of Intelligent Information Systems, Springer （採録決定） DOI 10.1007/s10844-015-0392-1

2. S. Ando, E. Suzuki: Minimizing Response Time in Time Series Classification, Knowledge and Information Systems, Vol. 46, No. 2, pp. 449-476, 2016. DOI 10.1007/s10115-015-0826-7.

3. Y. Deguchi, E. Suzuki: Hidden Fatigue Detection for a Desk Worker Using Clustering of Successive Tasks, Ambient Intelligence (AmI 2015), LNCS 9425, Springer-Verlag, pp. 263-283, 2015. DOI 10.1007/978-3-319-26005-1_18

4. K. Zhao, E. Suzuki: Clustering Classifiers Learnt from Local Datasets Based on Cosine Similarity, Foundations of Intelligent Systems, LNCS 9384 (ISMIS 2015), Springer-Verlag, pp. 150-159, 2015. DOI 10.1007/978-3-319-25252-0_16

5. E. Suzuki: On the Feasibility of Discovering Meta-Patterns from a Data Ensemble, Discovery Science (DS 2015), LNAI 9356, Springer-Verlag, pp. 266-274, 2015. DOI 10.1007/978-3-319-24282-8_22

6. V.-Marian Scuturici, Y. Gripay, J.-M. Petit, Y. Deguchi, E. Suzuki: Continuous Query Processing over Data, Streams and Services: Application to Robotics, New Trends in Databases and Information Systems (ADBIS 2015), pp. 36-43, CCIS 539, Springer-Verlag, 2015. DOI 10.1007/978-3-319-23201-0_5

7. S. Boubou, A. H. Abdul Hafez, E. Suzuki: Visual Impression Localization of Autonomous Robots, Proc. 2015 IEEE International Conference on Automation Science and Engineering (CASE 2015), pp. 328-334, 2015. DOI 10.1007/s10844-014-0329-0

8. E. Suzuki, Y. Deguchi, T. Matsukawa, S. Ando, H. Ogata, M. Sugimoto: Toward a Platform for Collecting, Mining, and Utilizing Behavior Data for Detecting Students with Depression Risks, Proc. Eighth International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2015), 2015. DOI 10.1145/2769493.2769538

9. S. Ando, T. Thanomphongphan, D. Hoshino, Y. Seki, E. Suzuki: Ensemble Anomaly Detection from Multi-resolution Trajectory Features, Data Mining and Knowledge Discovery, Vol. 29, No. 1, pp. 39-83, 2015. DOI 10.1007/s10618-013-0334-x

10. S. Ando, E. Suzuki: Discriminative Learning on Exemplary Patterns of Sequential Numerical Data, Proc. 2014 IEEE International Conference on Data Mining (ICDM 2014), pp. 1-10, 2014. DOI 10.1109/ICDM.2014.122

11. R. Kondo, Y. Deguchi, E. Suzuki: Developing a Face Monitoring Robot for a Desk Worker, Ambient Intelligence (AmI 2014), LNCS 8850, Springer-Verlag, pp. 226-241, 2014. DOI 10.1007/978-3-319-14112-1_19

12. D. Takayama, Y. Deguchi, S. Takano, V.-M. Scuturici, J.-M. Petit, E. Suzuki: Multi-view Onboard Clustering of Skeleton Data for Fall Risk Discovery, Ambient Intelligence (AmI 2014), LNCS 8850, Springer-Verlag, pp. 258-273, 2014. DOI 10.1007/978-3-319-14112-1_21

13. A. Erna, L. Yu, K. Zhao, W. Chen, E. Suzuki: Facial Expression Data Constructed with Kinect and their Clustering Stability, Active Media Technology, LNCS 8610 (AMT 2014), Springer-Verlag, pp. 421-431, 2014. DOI 10.1007/978-3-319-09912-5_35

14. Y. Deguchi, E. Suzuki: Skeleton Clustering by Autonomous Mobile Robots for Subtle Fall Risk Discovery, Foundations of Intelligent Systems, LNCS 8502 (ISMIS 2014), Springer-Verlag, pp. 500-505, 2014. DOI 10.1007/978-3-319-08326-1_51

15. Y. Deguchi, D. Takayama, S. Takano, V.-M. Scuturici, J.-M. Petit, E. Suzuki: Multiple-Robot Monitoring System Based on a Service-Oriented DBMS, Proc. Seventh ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2014), 2014. DOI 10.1145/2674396.2674418

16. E. Suzuki, Y. Deguchi, D. Takayama, S. Takano, V.-M. Scuturici, J.-M. Petit: Towards Facilitating the Development of a Monitoring System with Low-Cost Autonomous Mobile Robots, Information Search, Integration and Personalization, pp. 57-70, CCIS 421, Springer-Verlag, 2014. DOI 10.1007/978-3-319-08732-0_5

17. D. Ikeda, E. Suzuki: Finding Peculiar Compositions of Two Frequent Strings with Background Texts, Knowledge and Information Systems, Vol. 41, No. 2, pp. 499-530, 2014. DOI 10.1007/s10115-013-0688-9

〔学会発表〕(計１６件)

1. E. Suzuki: Compression-Based Evaluation of a Meta-Pattern in Terms of a Belief and a Data Ensemble, Twelfth International Conference on Operations Research (ICOR 2016), Havana (Cuba), 2016年3月9日.

2. Y. Deguchi, E. Suzuki: Hidden Fatigue Detection for a Desk Worker Using

Clustering of Successive Tasks, 12th European Conference on Ambient Intelligence (AmI 2015), Athens (Greece), 2015年11月12日

3. E. Suzuki: Clustering Classifiers Learnt from Local Datasets Based on Cosine Similarity, 22nd International Symposium on Methodolodies for Intelligent Systems (ISMIS 2015), Lyon (France), 2015年10月22日.

4. E. Suzuki: On the Feasibility of Discovering Meta-Patterns from a Data Ensemble, 18th International Conference on Discovery Science (DS 2015), Banff (Canada), 2015年10月6日.

5. V.-M. Scuturici: Continuous Query Processing over Data, Streams and Services: Application to Robotics, 19th East-European Conference on Advances in Databases and Information Systems (ADBIS 2015), Poitiers (France) 2015年9月9日.

6. S. Boubou: Visual Impression Localization of Autonomous Robots, 2015 IEEE International Conference on Automation Science and Engineering (CASE 2015), Gothenburg (Sweden), 2015年8月25日.

7. E. Suzuki: Toward a Platform for Collecting, Mining, and Utilizing Behavior Data for Detecting Students with Depression Risks, Eighth International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2015), Corfu (Greece), 2015年7月2日.

8. S. Ando: Discriminative Learning on Exemplary Patterns of Sequential Numerical Data, 2014 IEEE International Conference on Data Mining (ICDM 2014), Shenzhen (China), 2014年12月17日.

9. R. Kondo: Developing a Face Monitoring Robot for a Desk Worker, Fifth International Joint Conference on Ambient Intelligence (AmI 2014), Eindhoven (Netherlands), 2014年11月13日.

10. Y. Deguchi: Multi-view Onboard Clustering of Skeleton Data for Fall Risk Discovery, Fifth International Joint Conference on Ambient Intelligence (AmI 2014), Eindhoven (Netherlands), 2014年11月13日.

11. E. Suzuki: Multi-Task Data Mining toward Automating the KDD Process, Sixth International Conference on Information Technology and Electrical Engineering (ICITEE 2014) + Regional Conference on Computer and Information Engineering 2014 (RCCIE 2014), Jogjakarta (Indonesia), 2014年10月8日（基調講演）.

12. 田之上伸吾：生涯学習の人物表情分類問題における実験的評価, 平成26年度(第67回)電気・情報関係学会九州支部連合大会, 鹿児島大学（鹿児島県・鹿児島市), 2014年9月19日.

13. A. Erna: Facial Expression Data Constructed with Kinect and their Clustering Stability, 2014 International Conference on Active Media Technology (AMT 2014), Warsaw (Poland), 2014年8月11日.

14. Y. Deguchi: Skeleton Clustering by Autonomous Mobile Robots for Subtle Fall Risk Discovery, 21st International Symposium on Methodologies for Intelligent Systems (ISMIS 2014), Roskilde (Denmark), 2014年6月25日.

15. Y. Deguchi: Multiple-Robot Monitoring System Based on a Service-Oriented DBMS, Seventh ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2014), Rhodes Island (Greece), 2014年5月29日.

16. E. Suzuki: Towards Facilitating the Development of a Monitoring System with Autonomous Mobile Robots, 2013 International Workshop on Information Search, Integration, and Personalization (ISIP 2013), Bangkok (Thailand), 2013年9月17日.

〔その他〕
ホームページ
1. 自動データ選択とパターン抽出の統合による巨大データ集合体からの知識発見
http://www.i.kyushu-u.ac.jp/~suzuki/kaken1315-j.html

2. Knowledge Discovery from Huge Data Ensemble by an Integration of Automatic Data Selection and Pattern Extraction
http://www.i.kyushu-u.ac.jp/~suzuki/kaken1315.html

６．研究組織

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