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

平成　３０　年　　６　月　１３　日現在

機関番号：　３２６０６
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2017
課題番号：　２６２８００９０

研究課題名（和文）フィルタ型特徴選択法の統一理論と高性能アルゴリズム

研究課題名（英文）The theory of filter based feature selection and high-performance algorithms

研究代表者
久保山　哲二（Kuboyama, Tetsuji）

学習院大学・計算機センター・教授

研究者番号：８０３０２６６０

交付決定額（研究期間全体）：（直接経費）　　12,200,000円

研究成果の概要（和文）：本研究ではクラスラベルの付与された多次元の特徴空間を持つカテゴリカルデータからクラスラベルに関連する極小の特徴集合を抽出する特徴選択アルゴリズムに着目した。本研究の対象とするフィルター型の特徴選択アルゴリズムを構成する2つの要素、特徴集合とクラスラベル間の関連性を表す尺度（一貫性尺度）、および極小特徴集合の探索戦略について理論的および実験的な解析を行った。これらの解析に基づき、既存アルゴリズムを改良し、同種の既存手法の中では最速かつ規模耐性のある高精度なアルゴリズムを開発・実装した。また、開発した特徴選択アルゴリズムをTwitterからのトピック抽出と、グラフ構造からのパターン抽出に応用した。

研究成果の概要（英文）：We focus on feature selection algorithms that extract minimal subsets of features relevant to class labels from categorical data with high dimensional feature space. Filter-based feature selection consists of two important components; consistency measures between feature sets and class labels, and search strategies for minimal feature sets . Through theoretical and empirical analysis on these two components, we designed and implemented a very fast feature selection algorithm with high accuracy and scalability. We applied this algorithm to two applications; topic extraction from tweets, and pattern acquisition from graph-structured data.

研究分野：　情報科学

キーワード：　特徴選択　カテゴリカルデータ　一貫性指標

１．研究開始当初の背景

(1) 特徴選択は機械学習の中心的な研究領域のひとつである。例えば、人間の遺伝子は２万以上あるが、そのなかから特定の疾病と関連のある遺伝子を抽出する問題は特徴選択の問題として捉えることができる。しかし、このような問題を厳密に解こうとすると、組合せ爆発により計算困難となる。

(2) 本研究では、遺伝子のような特徴集合に対して、疾患の有無のようなラベルが与えられているような特徴選択問題（教師付き特徴選択）を扱う。特徴選択により抽出された特徴はその後の分類等の機械学習で利用されるが、この観点から①特定の機械学習アルゴリズムの最適化を目的とするラッパーアプローチ・埋め込みアプローチと、②特定の機械学習アルゴリズムとは独立にデータセットの統計特徴から、特徴を抽出するフィルターアプローチとがある。

(3) ラッパーアプローチに基づくカテゴリカル・データを対象とした特徴選択アルゴリズムCWCが分担者の申により提案されている。CWC は、その単純さにもかかわらず、予測性能、実行速度ともに、既存手法の中で群を抜く性能を示すことが大規模データを対象とした評価実験によりわかっている。しかし、なぜ CWC の性能が際立って良いのか、理論的には未解明である。

２．研究の目的

(1) 本研究では、CWC の数理的な解析と改良を通して、特徴選択法の中でも汎用性の高い**フィルタアプローチ**によるアルゴリズムの振る舞いを数理的に解明する。

(2) アルゴリズムの振る舞いを分析し、より高速なアルゴリズムの開発を目指す。

(3) 開発した高速な特徴選択アルゴリズムを実装し、特徴選択ツールとして公開する。

３．研究の方法

理論とアルゴリズム開発、および実装の側面から並行して研究を進める。

(1) 理論面では、CWC をはじめとする既存のフィルターアプローチに基づく型特徴選択アルゴリズムの性質を洗い出し、その背後にある共通の数理構造を分析する。

(2) アルゴリズム開発では、既存アルゴリズムの性能向上と実装の公開を目指して研究を進める。特に、CWC の機能拡張と性能改善を進めてソフトウェアを公開する。

４．研究成果

(1) データ集合 D において、特徴集合 $\{F_1, \ldots, F_n\}$ に**一貫性がある**とは、集合中の各特徴 F_i について各々同じ値を持つ任意のインスタンスが、同じクラスラベルを持つことをいい、特徴集合がクラスラベルに対して一貫していることを示す尺度を**一貫性指標**とよぶ。特徴間の相互作用を考慮しつつクラスラベルを説明できる特徴集合を一貫性指標を基準にして抽出する様々なアルゴリズムがこれまでに提案されている。本研究のベースとなる特徴選択アルゴリズム CWC と、既存のさまざまな特徴選択アルゴリズムで用いられている一貫性指標を比較整理した。また、もっともシンプルで厳しい一貫性指標に基づく CWC が、様々なベンチマークデータで平均的に高い性能が得られることを示した。

(2) 一貫性指標を数学的に厳密に定式化し、基礎の一貫性指標間の順序関係を、理論的に解析した。一貫性指標 μ を特徴集合 $\boldsymbol{F}$ により一意にクラスラベルが定まるとき、$\mu(\boldsymbol{F})=0$ となる性質(**決定性**)、および、$\boldsymbol{F}\subseteq\boldsymbol{G}$ のとき $\mu(\boldsymbol{F})\geqq\mu(\boldsymbol{G})$ となる性質(**単調性**)を持つ指標として定義する。このとき、ベイズリスク（またはIRC)、ラフセット一貫性指標、非一貫的ペア指標、条件付きエントロピー、CWC で用いられる２値一貫性指標などが一貫性の定義を厳密に満たすことがわかっている。これらの指標間の強弱の漸近的性質を解析し、「非一貫的ペア指標＞ベイズリスク>ラフセット一貫性指標, 条件付きエントロピー＞２値一貫性指標」のような半順序関係による階層があることを示した。この順序は、大きいほど一貫性のない場合にも寛容であり、小さいほど非寛容であることを表す。また、ベンチマークデータにより、階層と分類精度の間に相関があることを示した。

(3) 特徴選択の指標は、クラスラベルを一意に正しく決定できる〈最適な特徴集合〉からの距離として捉えることができる。この観点

から、データセットの確率分布を p-norm ($1 < p \leqq \infty$)による数列空間の要素とみなすことで、データセットと特徴集合の対を距離空間に埋め込み、最適特徴集合と任意の特徴集合との距離を特徴選択指標として定義した。さらに、この距離が p=1 のとき、よく知られているベイズリスクと一致することを示した。また、p >1 では、これまでに知られていない新しい指標となる。これらの指標によって得られた特徴集合を用いた分類精度は、ベイズリスクよりも強い相関を示すことがわかった。つまり、従来広く使われてきたベイズリスクよりも、よりよい指標となる可能性がある。

(4) 特徴選択アルゴリズムの開発には、クラスラベルに対して特徴集合の良さを測る一貫性指標のような尺度に加えて、この尺度を用いて効率的に特徴部分集合を探索する戦略が必要となる。CWC では、特徴集合の部分集合の包含関係を順序とする束上を、すべての特徴を含む集合から順に 1 つずつ特徴を削除して、深さ優先で探索し、極小な集合を発見する（図 1 参照（特徴集合 5 つの要素を含む場合））。これに対して、より高速な探索戦略を

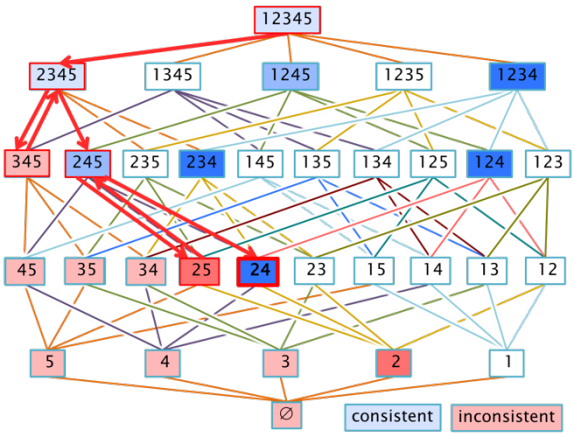

図 1. CWC による特徴空間探索

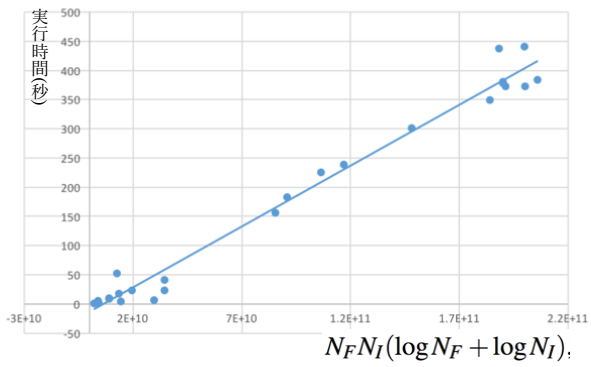

図 2. 実行時間(実測値と理論値)

導入したアルゴリズム sCWC を開発した。sCWC では、逐次的に 1 つ 1 つ特徴を削除せず、二分探索を用いて不要な特徴部分集合をまとめて削除してゆき、極小な特徴集合を探索する。sCWC の時間計算量は、N_F をデータの特徴数、N_I データのインスタンス数としたとき、$O(N_F N_I(\log N_F + \log N_I))$ であり、既存の同種の特徴選択アルゴリズム（フィルタ型、カテゴリカル・データ対象、一貫性指標に基づくアルゴリズム）の中では、理論計算量および実測値ともに最速であることを示した（図 2 参照）。

(5) 本研究で開発した高速な特徴選択アルゴリズムを以下の 2 つの問題に適用した

① 大規模なツイートからのトピック語抽出: ツイートから得られる単語文書行列をもとに、文書クラスタリングを行い、クラスタラベルを各文書のクラスラベルとする。特徴選択により、各文書クラスを判別する最小限の単語集合を抽出し、これをトピック語として用いる。東日本大震災前後のツイートを対象に、30 分おきにツイートをスライスし、各々、数百万ツイート、数十万ユーザからなる文書を対象とした。その結果、従来手法と比べて、より明確に解釈しやすい単語集合として、時系列毎のトピックを抽出することができた。

② 化学化合物からのパターン抽出: 正負の分類ラベルがついた化学化合物を外平面的グラフとして表現し、このグラフデータから正事例のグラフ構造に共通したパターンを遺伝的プログラミングを用いて進化的に獲得する問題に sCWC による特徴選択を適用した。進化計算において、正事例と負事例を判別するために必要なパターンが、正事例パターン発見のためには本質的であることから、特徴選択を用いることにより、効率的に候補となるパターンのみを進化計算の次世代候補として抽出した。特徴選択アルゴリズムの組み込みによって従来手法よりも少ない世代で高い適合度に到達することを示した。

(6) sCWC の疎行列対応版を Scala により実装し、github 上に公開した。

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 16 件）

① Naoya Higuchi, Yasunobu Imamura, Tetsuji Kuboyama, Kouichi Hirata, Takeshi, Nearest Neighbor Search using Sketches as Quantized Images of Dimension Reduction, Proc. of 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM), LNCS 10857, 2018, pp.356—363, 査読有
DOI: 10.5220/0006585003560363

② Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto, Dave Shepard, sCwc/sLcc: Highly Scalable Feature Selection Algorithms, Information, 8(4)-159, 2017, pp.1—26, 査読有
DOI: 10.3390/info8040159

③ Fumiya Tokuhara, Tetsuhiro Miyahara, Tetsuji Kuboyama, Yusuke Suzuki, Tomoyuki Uchida, A Context-Aware Fitness Function Based on Feature Selection for Evolutionary Learning of Characteristic Graph Patterns, Proc of 9th Asian Conference on Intelligent Information and Database Systems, LNCS 10191, 2017, pp.748—757, 査読有
DOI: 10.1007/978-3-319-54472-4_70

④ Takako Hashimoto, Hiroshi Okamoto, Tetsuji Kuboyama, Kilho Shin, Topic life cycle extraction from big Twitter data based on community detection in bipartite networks, Proc. of IEEE International Conference on Big Data, 2017, pp.2740—2745, 査読有
DOI: 10.1109/BigData.2017.8258238

⑤ Takao Hashimoto, Tetsuji Kuboyama, Hiroshi Okamoto, Kilho Shin, Topic Extraction on Twitter Considering Author's Role Based on Bipartite Networks, Proc. of 20th International Conference on Discovery Science (DS), LNCS 10558, 2017, pp.239—247, 査読有
DOI: 10.1007/978-3-319-67786-6_17

⑥ Takao Hashimoto, Tetsuji Kuboyama, Hiroshi Okamoto, Kilho Shin, Topic Extraction from Millions of Tweets Based on Community Detection in Bipartite Networks, Proc. in Information Modelling and Knowledge Bases XXIX, 27th International Conference on Information Modelling and Knowledge Bases (EJC), 2017, pp.395—408, 査読有
DOI: 10.3233/978-1-61499-834-1-395

⑦ Adrian Pino Angulo, Kilho Shin, Improving Classification Accuracy by Means of the Sliding Window Method in Consistency-Based Feature Selection, Proc. of 20th International Conference on Discovery Science (DS), LNCS 10558, 2017, pp. 155—170, 査読有
DOI: 10.1007/978-3-319-67786-6_12

⑧ Takako Hashimoto, Dave Shepard, Tetsuji Kuboyama, Kilho Shin, Topic Extraction Method from Millions of Tweets Based on Fast Feature Selection Technique CWC, Proc. IEEE International Conference on Data Mining Workshops, IEEE Comp. Soc. Ord. No. E6018, 2016, pp.724-731, 査読有
DOI: 10.1109/ICDMW.2016.0107

⑨ Kilho Shin, Seiya Miyaza, A Fast and Accurate Feature Selection Algorithm Based on Binary Consistency Measure, Computational Intelligence, 32(4), 2016, pp.646-667, 査読有
DOI: 10.1111/coin.12072

⑩ Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto, Dave Shepard, Super-CWC and super-LCC: Super fast feature selection algorithms, Proc. of IEEE International Conference on Big Data, IEEE Cat.No. CFP15BGD-USB, 2015, pp.61—67, 査読有
DOI: 10.1109/BigData.2015.7363742

⑪ Tomoya Yamazaki, Akihiro Yamamoto, Tetsuji Kuboyama, Tree PCA for Extracting Dominant Substructures from Labeled Rooted Trees, Discovery Science, LNCS 9356, 2015, pp.316—323, 査読有
DOI: 10.1007/978-3-319-24282-8_27

⑫ Takako Hashimoto, Dave Shepard, Tetsuji Kuboyama, Kilho Shin, Event Detection from Millions of Tweets Related to the Great East Japan Earthquake Using Feature Selection Technique, Proc. of IEEE International Conference on Data Mining Workshop, IEEE Comp. Soc. Ord. No. E5653, 2015, pp.7—12, 査読有
DOI: 10.1109/ICDMW.2015.248

⑬ Kilho Shin, Adrian Pino Angulo, A Geometric Theory of Feature Selection and Distance-Based Measures, Proc. of IJCAI, 2015, pp.3812—3819, 査読有

⑭ Adrian Pino Angulo, Kilho Shin, Fast and Accurate Steepest-Descent Consistency-Constrained Algorithms for Feature Selection, Machine Learning, Optimization, and Big Data, LNCS 9432, 2015, pp.293—305, 査読有
DOI: 10.1007/978-3-319-27926-8_26

⑮ Noriaki Kawamae, Real Time Recommendations from Connoisseurs, Proc. of the ACM SIGKDD, 2015, pp.537—546, 査読有

⑯ Takako Hashimoto, Tetsuji Kuboyama, Basabi Chakraborty, Topic extraction from millions of tweets using singular value decomposition and feature selection, Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE Catalog No. 36228, 2015, pp. 1145—1150, 査読有
DOI: 10.1109/APSIPA.2015.7415451

〔学会発表〕(計８件)

① Tetsuji Kuboyama, Polishing Big Data for Interpretable Results and Simple Algorithm Design (Panel Discussion on New Research Challenges), 10th Asian Conference on Intelligent Information and Database Systems (ACIIDS), (**招待講演**), 2017年4月

② 紫藤佑介, 山本章博, 小林靖明, 久保山哲二, モジュラリティを基準とした関係データに対する特徴選択, 第103回人工知能基本問題研究会(SIG-FPAI), 2017年3月

③ 山崎朋哉, 山本章博, 久保山哲二, Tree PCAによる任意形状の木構造を抽出するアルゴリズム, 人工知能学会 第99回人工知能基本問題研究会(SIG-FPAI), 2016年1月

④ Adrian Pino Angulo, 申吉浩, A Novel Hybrid Feature Selection Algorithm for Intrusion Detection, 人工知能学会 第100回人工知能基本問題研究会(SIG-FPAI), 2016年3月

⑤ 申吉浩, Angulo Adrian Pino, 久保山哲二、距離ベースの特徴選択指標、第97回 人工知能基本問題研究会(SIG-FPAI), 2015年3月22日

⑥ Basabi Chakraborty, Evolutionary Algorithms and various Evaluation Measures for Feature Subset Selection, Proc. of International Conference on Electronic Design, Computer Networks & Automated Verification EDCAV 2015(**招待講演**), 2015年1月

⑦ Tetsuji Kuboyama, Kilho Shin, De Morgan Property of Bayes Risk as A Feature Selection Measure, Workshop on Graph-Based Algorithms for Big Data and Its Application (GABA2014), 査読有, 2014年11月23日

⑧ 久保山哲二、申吉浩、特徴選択指標について、第94回 人工知能基本問題研究会(SIG-FPAI)、2014年07月24日

〔その他〕

Tetsuji Kuboyama, sCWC: very fast feature selection for nominal data, sCWC のソースコード: https://github.com/tkub/scwc/

６．研究組織

(1)研究代表者

久保山 哲二（Kuboyama, Tetsuji）
学習院大学・計算機センター・教授
研究者番号：80302660

(2)研究分担者

申 吉浩（Shin, Yoshihiro）
兵庫県立大学・応用情報科学研究科・教授
研究者番号：60523587

橋本 隆子（Hashimoto, Takako）
千葉商科大学・商経学部・教授
研究者番号：80551697

チャクラボルティ バサビ（Chacraborty, Basabi）
岩手県立大学・ソフトウェア情報学部・教授
研究者番号：90305293

川前 徳章（Kawamae, Noriaki）
東京電機大学・未来科学部・研究員
研究者番号：30447031