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　 ７ 月　 ５ 日現在

機関番号：３２６０６
研究種目：挑戦的萌芽研究
研究期間：2016～2019
課題番号：１６Ｋ１２４９１

研究課題名（和文）ビッグデータにスケールする一貫性指標に基づいた特徴分析

研究課題名（英文）Bigdata scalable feature analysis based on consistency measures

研究代表者
申　吉浩（Shin, Kilho）

学習院大学・付置研究所・教授

研究者番号：６０５２３５８７

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では、教師あり学習と教師なし学習の両面から、ビッグデータにスケールする高速性を有する実用的な特徴選択アルゴリズムの開発を行った。
教師あり学習では、従来から特徴選択の評価に使われていた相関量と特徴数の指標に対し、特徴選択後の機械学習アルゴリズムに影響を与えるノイズを新たに指標に追加し、体系的な評価方法を提案した。さらに、この三指標をバランスさせる高速なアルゴリズムとして、BornFSを提案した。
教師なし学習における特徴選択は、教師あり学習の場合に比して格段に難問であり、今まで知られているアルゴリズムは速度性能にかけていた。本研究では、非常に高速なアルゴリズムUFVSを提案した。

研究成果の学術的意義や社会的意義
特徴選択は機械学習の中心問題の一つであり、実用的にも、重要な役割を果たす。例えば、DNA配列から特定の疾病の原因となる塩基を決定する問題は、バイオインフォマティクスの観点から見れば、特徴選択の適用に他ならない。他にも、ネットワークに侵入したパケットの検知において、パケットヘッダーのどのフィールド値が証拠になるかを決定することも、特徴選択の適用で可能となる。また、特徴選択を行った後で、機械学習を行うことで、正確性と速度性能が改善されることも広く知られている。現実の問題では、データにラベルを付与することが容易でないが、教師なし学習における実用的な特徴選択に先鞭をつけた意義も有する。

研究成果の概要（英文）：This research project has developed two practical feature selection algorithms, BornFS and UFVS, with high time efficiency that can scale to bigdata.
BornFS, a feature selection algorithm in the supervised learning context, evaluates relevance, feature count and noise, which is a new measure to evaluate performance of feature selection introduced in our research, and is capable to features with an optimal balance among values of these three measures.
UFVS, a feature selection algorithm in the unsupervised learning context on the other hand, outperforms any known algorithms in the literature in time efficiency. In principle, feature selection under the unsupervised learning setting is known to be significantly difficult, and as a result, the known algorithms were very slow. In contrast, UFVS has time efficiency that can scale to bigdata. In the experiments, UFVS could select small numbers of effective features for datasets with class labels but without using the class labels.

研究分野：機械学習

キーワード：特徴選択　教師あり学習　教師なし学習

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

特徴選択は機械学習の古典的かつ中心的な研究テーマの一つであり、長期にわたり、熱心に研究されてきた。この分野の専門家はおそらく皆同意するように、特徴選択問題の難しさの理由には、特徴数nに対して解の候補が2^nと指数的であり、計算量的に全探索は不可能であること、現実のデータセットは多様かつ複雑であり、多項式時間アルゴリズムを設計するための共通の原理が存在しそうにもないこと、近年のビッグデータの分析には特徴選択が重要な役割を果たすことが予想されるものの、膨大なデータ量を処理し得る効率的なアルゴリズムの設計が困難である点などにある。その点、筆者らは、2015 年にビッグデータにも適用できる画期的な超高速特徴選択アルゴリズム superCWC を IEEE Bigdata で発表し、実用的な観点から、上記の問題の解決に向けての第一歩を標した。

２．研究の目的

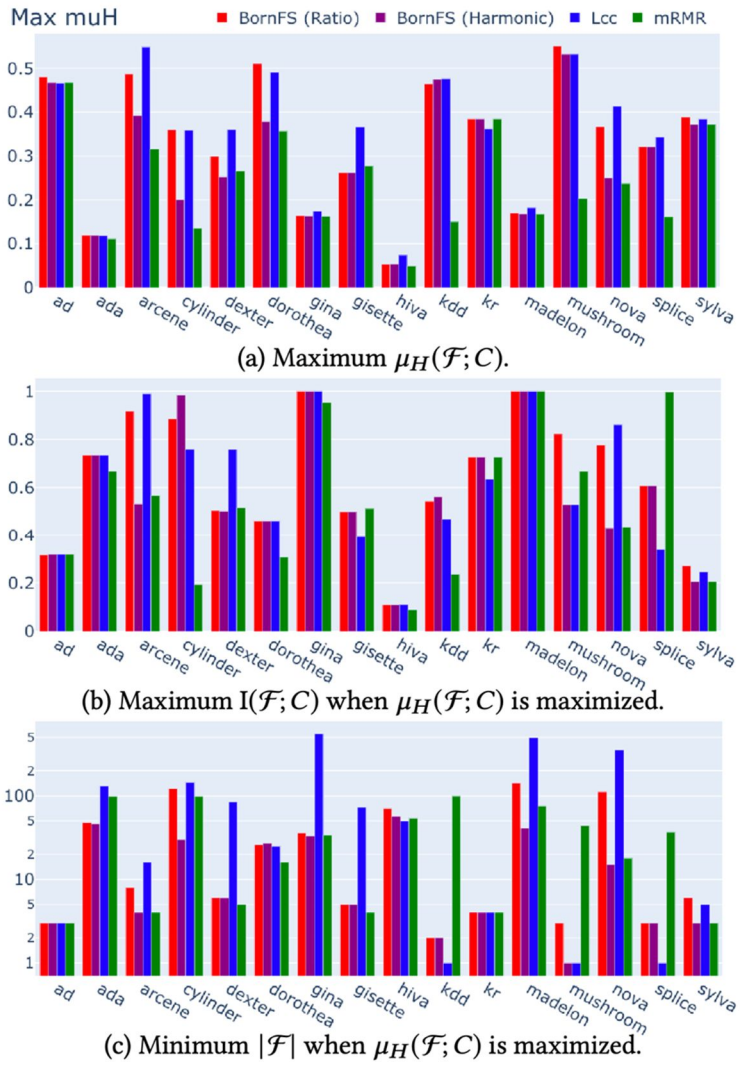
ビッグデータに適用可能な高速性能を有しながら、クラス相関・選択特徴数で定義される質的性能も良好であり、かつ、ハイパーパラメータの値によって異なる視点からの特徴集合を出力するような実用的アルゴリズムを提案することが目的である。

３．研究の方法

特徴選択は長い研究の歴史を持つが、「背景」で述べたように実用における要求が高いこと、取り扱う問題の多様性が高いことが理由となって、アルゴリズムの評価指標も確立されたものは存在しない。当研究では、従来研究で用いられていた分類器に依存した分類正確度ではなく、情報理論的な観点からの評価指標の発見を目標の一つとしている。評価指標の研究と並行して、superCWC や既存の特徴選択アルゴリズムの実用面での改良を研究することで、実用的なアルゴリズムを設計する上で必要となる知見の蓄積を図った。この車軸の両輪が揃った時点で、実用的に有効な特徴選択アルゴリズムの開発を行う計画とした。

４．研究成果

情報理論的な特徴選択アルゴリズムの評価指標として、従来から取り上げられていた、クラス相関、内部冗長性、相互干渉性、選択特徴数に関して、内部冗長性と相互干渉性の指標はアルゴリズム依存であり、最終的にはクラス相関に帰着されることを示した。一方、選択した特徴集合に含まれるノイズが重要な指標であり、クラス相関・ノイズ・選択特徴数により評価を行うことを提案した。クラス相関は相互情報量$I(F;C)$によって評価することができ、ノイズは条件付き情報エントロピー$H(F|C)$によって評価できる。これらの指標は互いに独立であるが、一方で、アルゴリズム設計においては、複数の指標を同時に扱うことは困難である。そこで、$\frac{I(F;C)}{I(E;C)}$と$\frac{I(F;C)}{H(F)}$の調和平均として、$\mu_H(F;C)$を定義し、$\mu_H(F;C)$をクラス相関とノイズの統合指標として使用することを提案した。ただし、Fは選択された特徴集合、Eはデータセットを記述する全特徴の集合、C はクラスラベルを記述する確率変数であるとする。



(a) Maximum $\mu_H(\mathcal{F};C)$.

(b) Maximum $I(\mathcal{F};C)$ when $\mu_H(\mathcal{F};C)$ is maximized.

(c) Minimum $|\mathcal{F}|$ when $\mu_H(\mathcal{F};C)$ is maximized.

図 1　クラス相関・ノイズ・選択特徴数の比較

実用的なアルゴリズムの設計に関しては、superCWC にハイパーパラメータを付与した superLCC を開発し、その有効性を実験により示した。superLCC は、ハイパーパラメータを最適化することにより、superCWC の課題であった選択特徴数が多くなりすぎるという問題を解

決し、また、分類器による分類正確度を向上することができる。これにより、従来技術のベンチマークである mRMR に対する実行速度の圧倒的優位性を維持しつつ、選択した特徴の分類正確度においても mRMR を凌ぐアルゴリズムとなった。superLCC は、特徴数、インスタンス数が数十万にものぼるビッグデータのデータセットに対しても適用可能な高速性を有する。この成果は論文誌で発表している。
加えて、広くから受け入れられているが、計算量が大きいことが課題である特徴選択アルゴリズムである、mRMR、及び、CWC の高速化に関する提案を行った。この高速化により 10 倍程度の速度の改善が得られ、その適用範囲は格段に拡大するが、ビッグデータに適用できるレベルには至らなかった。この研究により、これらのアルゴリズムの基本構造が高速計算に適していないことが明らかとなり、速度性能に関する superLCC の根本的優位性を実証することにもなった。この研究は、私が指導していたキューバからの文科省国費留学生である博士課程学生 Adrian Pino との共同研究として実施した。この研究も、論文誌で発表した。
このように、分類器に依存する分類正確度に代わり、情報理論に依拠する指標の考案に成功し、かつ、superLCC の開発によって、速度性能とハイパーパラメータによる選択の多様性を両立させるアルゴリズムフレームワークを獲得したため、当初の計画に従って、クラス相関・ノイズ・選択特徴数の評価指標の最適化問題を近似的に探索し、かつ、superLCC におけるバイナリ探索を利用した高速選択技法を取り入れた新規のアルゴリズムの開発に着手した。その結果、BornFS (Balance-Optimized Relevance and Noise Feature Selection)を開発し、superLCC 及び mRMR に値する優位性を、多数のデータセットを用いて実証した（図１）。BornFS は現在国際会議に投稿中であるが、ビッグデータに適用できる高速性を持ちながら、質的性能で mRMR に対して三つの指標全てで優位であり、かつ、superLCC に対して三つの指標のバランスにおいて優位である（図２）ことから、現時点において実用上最も優れたアルゴリズムであると自負している。

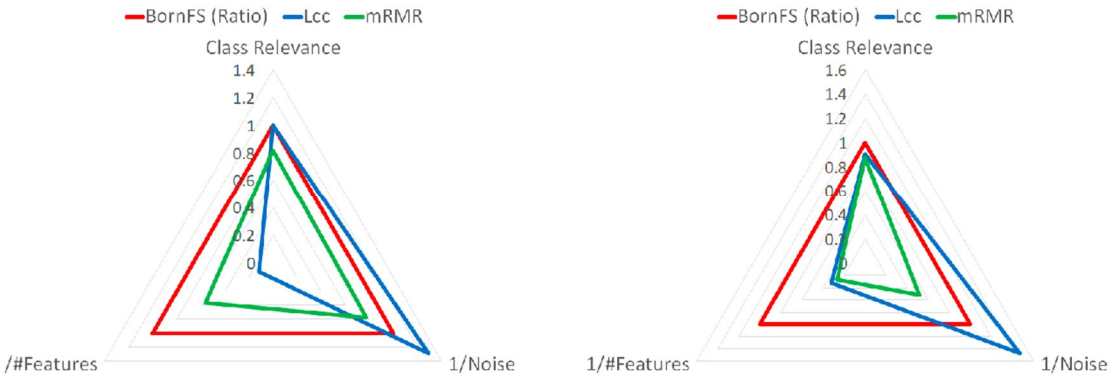


図 2 三指標のバランス

ここまで述べた研究成果は、教師あり学習におけるものであるが、教師なし学習に関しても、研究に着手した。教師なし学習の主要な目的はクラスタリングであり、教師なし学習において特徴選択を研究することにより、より少ない特徴に基づいてクラスタリングを行うことが可能となり、以下のようなメリットがあると考えられるからである。

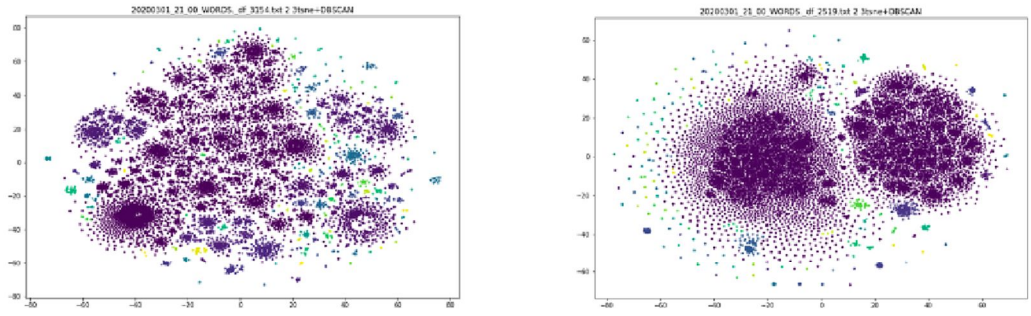
- 通常のクラスタリングでは、クラスターの意味をしるためには、後付けで解釈をしなければならない。特徴値選択に基づくクラスタリングの場合、クラスターの意味を知ることが容易かもしれない（次元が下がっているから）。
- 上の記述を逆から見ると、通常のクラスタリングでは、予め意味を特定できないが、複数の特徴値集合を選択して、それぞれでクラスタリングすることで、意味をある程度特定したクラスタリングを行うことができる。

理論的に言っても、特徴選択をしないクラスタリングの問題は明らかである。
例えば、インスタンスを共有する二つのラベル付きデータセットがあるとする。インスタンスは共通でも、特徴集合は交わりがないと仮定し、更に、ラベルは確率変数として互いに独立だとする。この二つのデータセットを接合して、ラベルを省いて、クラスタリングすると、例えば、One hot encoding で数値化して、k-means などでクラスタリングすると考えるとわかりやすいが、独立のデータセットからクラスを省いてクラスタリングした場合の有効なクラスタ数がそれぞれ nとすると、接合したデータセットの有効なクラスタ数はおおよそn^2となるが、さらに、k個のデータセットを接合するとすると、有効クラスタ数はn^k個となり、容易にインスタンス数を超えてしまう。このkというのは、接合したデータセットの独立した「正当な」解釈であるので、つま

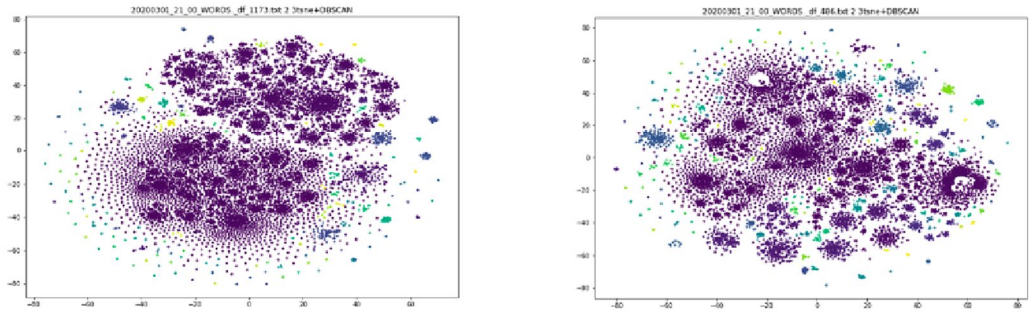
り、

- 少ないクラスタ数でクラスタリングすると、k 個の解釈が混じりあい、クラスタの意味が不明になり、
- 有効クラスタ数を考えると、インスタンス数を容易に超えてしまうことが予想され、クラスタリングそのものが不可能となる。

特徴選択は、特に、複数の選択候補を出力できるように設計された場合、接合されたデータセットを分解する機能を提供できる可能性がある。
本研究では、多様な特徴集合を出力できるよう、高速性を有し、かつ、ハイパーパラメータにより出力である特徴集合を変動させることが可能な新規のアルゴリズムとして、EUFVS (Explainability-based Unsupervised Feature Value Selection)を開発し、国際学会で発表した。当論文は、Selected Paper に選ばれ、内容を拡大した論文が Springer の LNAI に掲載される予定である。
EUFVS が出力する相当数の特徴値集合を Jaccard 係数による距離（Jaccard 係数が正定値カーネルである事実を利用して再生核カーネルヒルベルト空間内で定義される距離）を使ってクラスタリングしてみたところ、閾値の取り方とクラスタリングが正確に対応する例を多数得ることができ、EUFVS でパラメータを変えていくことにより、意味のある特徴値のクラスタを連動して得ることができる可能性が示された。Jaccard 係数による距離を使ったのは方便で、Hamming 距離を使えば、距離はデータセットの情報に基づくことから、より正確な結果が得られることが期待できる。
千葉商科大学の橋本隆子教授が、EUFVS を用いて、新型コロナ感染拡大期における 25000 のインスタンス（ツイート）と 50000 の特徴（特徴単語）からなるビッグデータを解析したところ、図３に示されるようなクラスタリングを得た。クラスタリングアルゴリズムとしては、DBSCAN を使用している。
意味は現在解析中であるが、一見して、明確なクラスターが現れており、有効にクラスタリングが実行されたことが分かる。500 次元から 3000 次元でのクラスタリングの結果を二次元に次元削減した結果を図示したものであることから、元の次元では非常に明確にクラスターが分離していることが窺われる。
教師なし学習における特徴選択は、教師あり学習における特徴選択に比較して格段に難問であるとされているが、ビッグデータではクラスラベルを付与することが容易ではなく、教師なし学習における特徴選択の重要性は今後一層高まるものと考えられる。



(a) $\xi = 1.0, t = 0, |S| = 3154$　　(b) $\xi = 0.95, t = 20, |S| = 2519$



(a) $\xi = 0.9, t = 40, |S| = 1173$　　(b) $\xi = 0.95, t = 80, |S| = 486$

図 3　Twitter データに適用して得られるクラスタリング

５．主な発表論文等

〔雑誌論文〕　計4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 Adrian Pino Angulo, Kilho Shin	４．巻 49
２．論文標題 Mrmr+ and Cfs+ feature selection algorithms for high-dimensional data	５．発行年 2019年
３．雑誌名 Applied Intelligence	６．最初と最後の頁 1954-1967
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/s10489-018-1381-1	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 Adrian Pino Angulo, Kilho Shin	４．巻 7
２．論文標題 Improving the genetic bee colony optimization algorithm for efficient gene selection in microarray data	５．発行年 2018年
３．雑誌名 Progress in Artificial Intelligence	６．最初と最後の頁 399-410
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/s13748-018-0161-9	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 Kilho Shin, Tetsuji Kuboyama, Takako Hashimoto and Dave Shepard	４．巻 8 (4)
２．論文標題 sCwc/sLcc: Highly Scalable Feature Selection Algorithms.	５．発行年 2017年
３．雑誌名 Information	６．最初と最後の頁 159
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.3390/info8040159	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 Kilho Shin, Seiya Miyazaki	４．巻 32 (4)
２．論文標題 A Fast and Accurate Feature Selection Algorithm Based on Binary Consistency Measure.	５．発行年 2016年
３．雑誌名 Computational Intelligence	６．最初と最後の頁 646-667
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1111/coin.12072	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3件（うち招待講演　0件／うち国際学会　3件）

1．発表者名 Kilho Shin, Taichi Ishikawa
2．発表標題 Linear-time algorithms for the subpath kernel
3．学会等名 29th CPM 2018（国際学会）
4．発表年 2018年

1．発表者名 Adrian Pino Angulo, Kilho Shin
2．発表標題 Improving Classification Accuracy by Means of the Sliding Window Method in Consistency-Based Feature Selection.
3．学会等名 DS2017（国際学会）
4．発表年 2017年

1．発表者名 Kilho Shin, Kenta Okumoto David Shepard, Tetsuji Kuboyama, Takako Hashimoto, Hiroaki Ohshima
2．発表標題 A Fast Algorithm for Unsupervised Feature Value Selection
3．学会等名 ICAART2020（国際学会）
4．発表年 2020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