

# 特徴選択における状態探索手法の比較検討

遠藤 瑛泰<sup>1,a)</sup> 永田 賢二<sup>2,3</sup> 庄野 逸<sup>1,b)</sup>

**概要:** 特徴選択とは、識別や回帰といった目的に対して特徴集合のすべての特徴を用いるのではなく、最適な特徴の組み合わせを用いることである。一般に、特徴集合に対して考えられる特徴の組み合わせをすべてを網羅的に探索する全状態探索を行うことで、最適な特徴の組み合わせを見つけることが可能である。しかし、考えられる組み合わせは特徴の総数に対して指数的に増加していくため、そのすべてを確認することは困難になる場合がある。このような問題に対して先行研究では、MCMC法を用いた近似的な全状態探索や組み合わせのスパース性を仮定したES-Kといった状態探索手法が提案されているが、これらが全状態探索の結果と比較してどのような結果が見込まれるのかといった議論は行われていない。そこで本研究では、複数のデータセットに対して全状態探索と各種状態探索手法を適用し、各状態探索手法で得られた結果の比較検討を行った。

**キーワード:** 特徴選択, 全状態探索, 状態密度, 組み合わせ最適化

## Comparison of state search methods for feature selection

AKIHIRO ENDO<sup>1,a)</sup> KENJI NAGATA<sup>2,3</sup> HAYARU SHOUNO<sup>1,b)</sup>

**Abstract:** Feature selection is using most effective feature combinations rather than using all features for purposes such as classification and regression. In general, it is possible to find most effective combination by Exhaustive search (ES) which performs all combinations. However, number of combinations increase exponentially with respect to the total number of features, it may be difficult to confirm all of them. To solve these problems, Approximate Exhaustive Search (AES) using the MCMC method and ES-K assuming sparsity of combinations have been proposed in the previous research. But, there is no discussion on what kind of results are expected. Therefore, in our research, we applied ES, AES and ES-K to multiple data sets, and compared each search methods.

**Keywords:** Feature Selection, Exhaustive Search, Density of State, Combinational Optimization Problem

### 1. はじめに

従来のモデルの構築には、様々な特徴を用いた構築が一般的である。しかし、必ずしも用いた特徴のすべてがモデルを構築する上で重要であるとは限らず、不必要なものが含まれている場合がある。不必要な特徴が含まれているモ

デルを用いた場合、その特徴によって生じるノイズによって予測精度が低下する [3][4]。特徴選択とはモデルの構築を行う際に、持ちうるすべての特徴を用いて構築するのではなく、目的に有効である特徴の組み合わせを用いてモデル構築を行う手法である。このような特徴の取舍選択を行うことにより、不必要な特徴を排除したことによる予測精度の向上や、選択した特徴から構築したモデルの解釈が容易になることが報告されている [3][4]。

特徴選択の適用を考えたとき、目的に対して最適な組み合わせを考えられる組み合わせ集合から見つけ出す必要がある。このとき、最も確実に最適な組み合わせ集合を見つける方法は、考えられる組み合わせをすべて確認するよ

<sup>1</sup> 電気通信大学大学院情報理工学研究所

〒182-8585 東京都調布市調布ヶ丘 1-5-1

<sup>2</sup> 国立開発法人産業技術総合研究所 人工知能研究センター

〒135-0064 東京都江東区青海 2-4-7

<sup>3</sup> 科学技術振興機構「さきがけ」

〒102-0076 東京都千代田区五番町 7

<sup>a)</sup> e1830017@edu.cc.uec.ac.jp

<sup>b)</sup> shouno@uec.ac.jp

うな全状態探索 (Exhaustive Search: ES) を行うことである [3]. しかし, 考慮しなければならない特徴の組み合わせは持ちうる特徴の総数に対して指数オーダーで存在するため, 容易に ES に出ない状況が起こりうる [1].

この問題に対して, スパース推定を利用したアプローチ [5] によって計算量を抑えて特徴選択を行う手法などが用いられているが, 最適な組み合わせが多数存在する場合などでは適切な特徴の組み合わせが選択されない場合がある [7]. 先行研究では, 全状態探索が行えないような問題に対して, 交換 MCMC 法といったサンプリング手法を用いた近似的全状態探索 (Approximate Exhaustive Search: AES) や, 適切な特徴が  $K$  個であることを仮定して  $K$  個の特徴の組み合わせについて網羅的に探索を行う  $K$  スパース全状態探索 (ES- $K$ ) とした手法が提案されている [4][6]. しかしながら, これらの提案された手法は全状態探索が不可能な場合に適用されてきた手法であるが, 全状態探索と比較してどのような結果を期待できるのかといった議論は行われていない.

本研究では, ES が可能なデータに対して, ES および AES, ES- $K$  を適用し, 特性を確認した.

## 2. 手法

本研究では,  $D$  個の特徴を持つ入力  $\mathbf{x} = \{x_0, x_1, \dots, x_{D-1}\}$  から, ラベルを  $y = \{-1, 1\}$  の 2 値分類を行う線形識別問題を考える.  $N$  個の入力データを  $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}\}$ , 対応するラベルデータを  $\mathbf{y} = \{y_0, y_1, \dots, y_{N-1}\}$  とする. 特徴選択における状態探索手法では, 特徴の選択, 識別器の構築, 識別器の評価の 3 つの手順を行うことで選択した特徴の組み合わせを評価を行い, 最良の評価結果が得られる組み合わせを選択することで特徴の選択を行う. 図 1 に状態探索による特徴選択の枠組みを示す.

はじめに, 特徴の選択では  $D$  個の特徴全てに対して, その特徴を使うか使わないかを定めることが目的となる.  $D$  個の特徴が使われているときを 1, 使われていないときを 0 で表現したものを状態  $\mathbf{s}$  とするとき, 以下のように表すことができる.

$$\mathbf{s} = (s_0, s_1, \dots, s_{D-1}) \in \{0, 1\}^D \quad (1)$$

次に与えられた  $\mathbf{s}$  をもとに, 識別器を構築する. このときの入力を  $\mathbf{x}_n(\mathbf{s})$  とする. 線形識別器を用いた時,  $\mathbf{x}_n(\mathbf{s})$  の予測ラベル  $\hat{y}_n(\mathbf{s})$  を特徴次元に対応した重み  $\mathbf{w}(\mathbf{s})$  とバイアス項  $b$  を用いて以下のように表現することができる.

$$\hat{y}_n(\mathbf{s}) = \text{sgn}(\mathbf{w}(\mathbf{s})^T \mathbf{x}_n(\mathbf{s}) + b) \quad (2)$$

構築した識別器に対して, その識別器の性能を正答率や交差検証法によるスコアなどを用いて評価する. 識別器の性能は  $\mathbf{s}$  によって決まるため, その評価値を  $\mathbf{s}$  の関数  $E(\mathbf{s})$  として考えることで  $\mathbf{s}$  の評価を行うことができる.

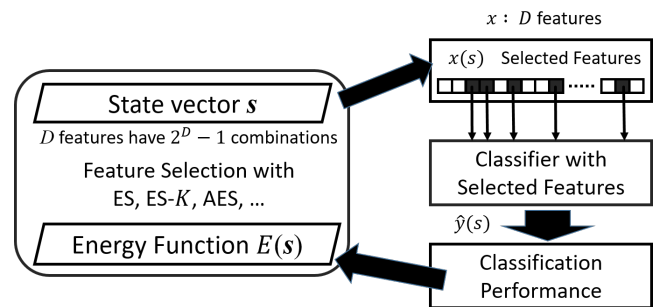


図 1 状態探索による特徴選択の枠組み.

$E(\mathbf{s})$  が小さい値を取るほど性能の良い識別器の構築が出来たと評価したとき, 最も  $E(\mathbf{s})$  が小さくなるような  $\mathbf{s}$  を見つけることが状態探索手法の目的となる.

### 2.1 全状態探索 (Exhaustive Search: ES)

$D$  個の特徴を持つデータの全状態探索 (Exhaustive Search: ES) を行うことを考えた時, 確認する必要がある組み合わせの数は  ${}_D C_1 + {}_D C_2 + \dots + {}_D C_D$  個, つまり  $2^D - 1$  状態すべてに対して評価を行うことで最適な組み合わせを見つけることができる [3][4][7].

しかし, ES において考慮すべき状態数は  $D$  の値に対して指数的に増加するため, 対象となるデータの特徴数が増えるとき, そのすべての状態に対して評価を行うことは困難を極める [1].

### 2.2 $K$ スパース全状態探索 (ES- $K$ )

全状態探索は最も確実に最適な組み合わせを見つける状態探索手法であるが, 対象となるデータが全状態探索が行えないような高次元データであるようなことは容易に起こりうる. そのような高次元データに対して, 五十嵐らは最適な特徴の組み合わせが  $K$  個の特徴で構成された組み合わせであると仮定し, 考えられる組み合わせを網羅的に探索する  $K$  スパース全状態探索 (ES- $K$ ) を提案している [6]. ES- $K$  では,  $D$  個の特徴で考えられる  $2^D - 1$  状態のうち,  $K$  個で構成された組み合わせ  ${}_D C_K$  状態に対して評価を行う.

### 2.3 近似的全状態探索 (Approximate Exhaustive Search: AES)

全状態探索および ES- $K$  が困難な場合に, サンプリング手法を用いて効率的かつ網羅的な状態探索を行う近似的全状態探索 (Approximate Exhaustive Search: AES) が Nagata らによって提案されている [4]. 本研究では MCMC 法の拡張である交換 MCMC 法による AES を考える. 交換 MCMC 法による AES は状態  $\mathbf{s}$  を評価値  $E(\mathbf{s})$  の分布からサンプリングすることによって全状態探索の近似とする. 以下のボルツマン分布  $p_{\beta}(\mathbf{s})$  を複数準備し, 評価値  $E(\mathbf{s})$  から状態  $\mathbf{s}$  のサンプリングをする.

$$p_{\beta}(s) = \frac{1}{Z_{\beta_m}} \exp(-\beta E(s)) \quad (3)$$

交換 MCMC 法では、 $\beta_m$  は逆温度であり、 $0 < \beta_1 < \dots < \beta_m < \dots < \beta_{M-1}$  で設定した  $M$  個のレプリカを並列のサンプリングとレプリカ間の状態  $s$  の交換を交互に行う。サンプリングと状態交換を交互に行うことは、局所解からの脱出と状態空間の広範囲からサンプリングを目的とし、効率的かつ効果的なサンプリングを行うことが出来る。

## 2.4 交差検証 (Cross Validation: CV) 法

状態探索における評価指標として、交差検証 (Cross Validation: CV) 法を用いた評価を行った。CV 法はデータセットを訓練データと検証データに分割し、訓練データを用いて識別器の構築を行い、検証データを用いて識別器の性能評価を行う手法である。

本研究では、データセットを  $V$  個に分割し、 $V-1$  個のデータセットを訓練データ、残り 1 つをテストデータとした評価を  $V$  パターン行う  $V$ -fold CV 法を行った。このときの評価値を  $CVscore(s)$  と表す。 $CVscore(s)$  は 0~1 の範囲で値を取り、 $CVscore(s)$  が高いほど、識別に生じるノイズが小さく汎化性能の高い識別器を構築できたと考えることが出来る。

状態探索の枠組みに  $CVscore(s)$  を当てはめるとき、状態  $s$  に関する評価関数  $H(s)$  として正負を反転させた  $CVscore(s)$  を定義することで、最小の  $H(s)$  となる  $s$ 、すなわち最高の  $CVscore(s)$  となるような  $s$  の探索を行った。

## 3. 実験

本研究では、UC Irvine Machine Learning Repository で公開されているデータセットのうち、ワインの分類問題を行う Wine Data Set (Wine) と糖尿病性網膜症の診断を行う Diabetic Retinopathy Debrecen Data Set (DRD) に対して状態探索による特徴選択を適用した [2].

## 4. 結果

### 4.1 Wine Data Set

Wine Data Set はワインの化学分析の結果から 3 つの品種への分類が目的であり、最大で 13 の成分を特徴として用いることが出来る。識別器には線形サポートベクターマシンを用いた。一般に線形 SVM は 2 値分類を行う識別器であるため、3 クラス分類である Wine に対しては各クラスごとに一対他識別器の構築を行った。

#### 4.1.1 ES

Wine に対して ES を適用した。class0 において  $CVscore$  が高い組み合わせを左から順に列挙した結果を図 2 に示す。縦軸には各特徴ごとに使われているならば黒、使われていないならば白を表示している。結果より各クラスにおいて、 $CVscore$  が高い組み合わせでは 3 番目の特徴や 12

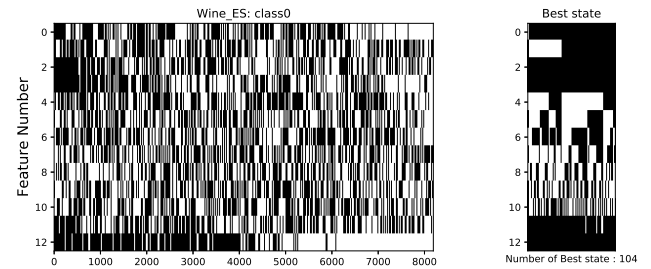


図 2 Wine の class0 における全組み合わせ列挙. 左から順に  $CVscore$  が高い組み合わせの列挙した. 図中右には最も  $CVscore$  が高い組み合わせを列挙している.

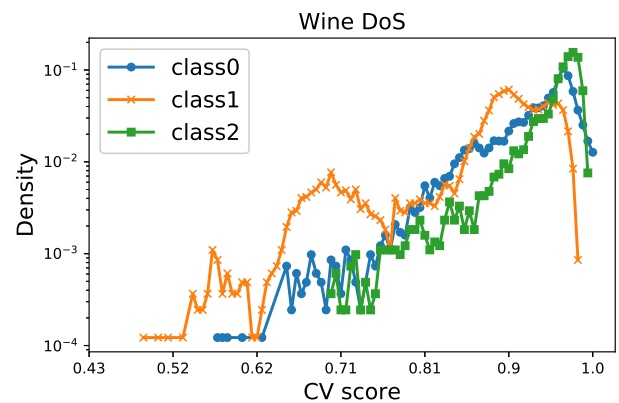


図 3 ES の結果より求めた Wine 各クラスの状態密度.

番目のように優先的に使用されている特徴があることが確認できる。また、最も  $CVscore$  が高い組み合わせが 104 状態 (約 1.4%) 存在していることがわかった。これら複数の最適解でも共通して使われている特徴が存在しており、これらが識別において重要な特徴であることが確認できる。

さらに各クラスにおいて、ES の結果より状態密度を求めた。状態密度とは、全状態に対する評価値の度数分布である。横軸には今回の評価値である  $CVscore$ 、縦軸には横軸で対応する  $CVscore$  が全状態のどのくらいの割合を占めているのかを数値化した密度を示している。図 3 に 3 つのクラスごとの状態密度を示す。どのクラスの状態密度からも最高の  $CVscore$  に近い準最適解が多数存在していることが確認できる。前述の組み合わせの列挙の結果より、 $CVscore$  の高い組み合わせには特有の特徴の組があることが確認できた。この結果を踏まえると、このような識別に起因する特徴の組を親として、親から派生して出来た組み合わせが最適解や準最適解といった十分に高い  $CVscore$  を取りうる状態として集中したと考えられる。

#### 4.1.2 ES-K

Wine に対して ES-K を適用した。K の推移と K における  $CVscore$  の最高値を確認した。図 4 に class0 における推移を示す。横軸に K、縦軸に  $CVscore$  を取っている。class0 では K が 5 から 11 の間で、 $CVscore$  が最大を取るような組み合わせが形成されている。さらに K を大きくしていく

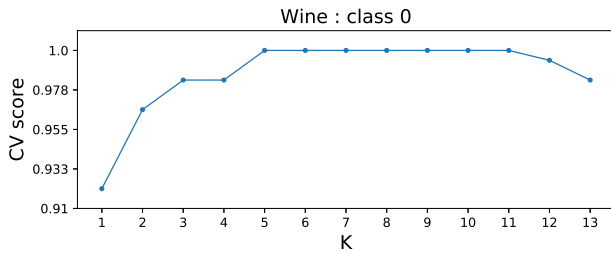


図4 Wine の class0 における ES-K の  $K$  の変化と CVscore の最高値の推移。

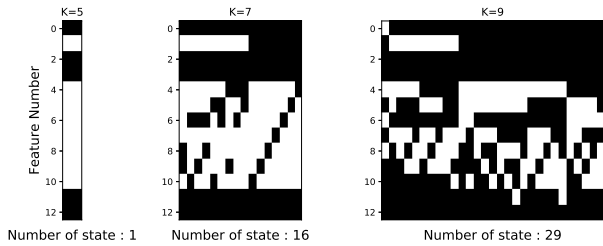


図5 Wine の class0 における  $K = 5, 7, 9$  のときに CVscore が最大となる組み合わせの列挙。

と、 $K = 12$  になったとき CVscore が低下していくことが確認できる。これは unnecessary 特徴を含んでいるような組み合わせを用いて識別器を構築してしまうと、汎化性能が低下することを示している。

図5に class0 において  $K = 5, 7, 9$  のときに CVscore が最も高かった組み合わせを示す。 $K = 5$  のときに初めて CVscore が 1 となる組み合わせが形成され、以降の  $K$  ではそれを親とするような組み合わせが CVscore が 1 となる組み合わせになっていることがわかる。

#### 4.1.3 AES

Wine に対して交換 MCMC 法による AES を適用した。レプリカを 20 個準備し、合計で 1000 状態をサンプリングした。

class0 におけるサンプリングの結果を図6に示す。AES によるサンプリングによって、様々な組み合わせがサンプリングされていることがわかる。また、すべての組み合わせを確認しなくても、最適解や準最適解で見られるような特徴の偏りを確認することができる。

今回は 1000 状態のサンプリングを行ったところ、35 状態の最適解をサンプリングすることが出来た。これはすべての最適解のうち、約 34% の組み合わせを見つけることが出来ており、約 12% のサンプリングにおいて十分効果的に最適解を見つけられたといえる。

#### 4.1.4 従来手法との比較

従来の特徴選択の手法として、Recursive Feature Elimination (RFE) による特徴選択手法を適用し、比較した。本研究では識別器を構築したときの特徴毎の重みをその特徴の重要度とし、ランク付けをもとに、1つずつ特徴を減らし

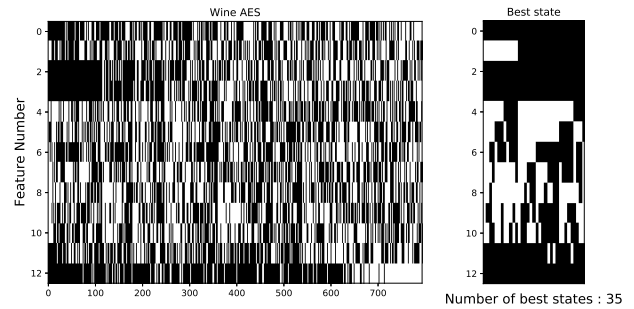


図6 Wine の class0 における AES による組み合わせ列挙と最適解。

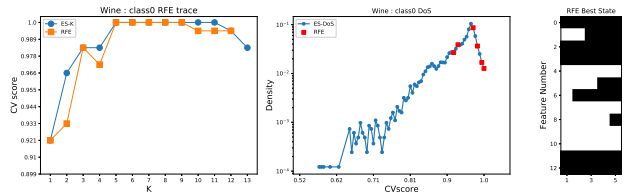


図7 Wine の class0 における、RFE による特徴選択と状態探索による特徴選択の結果。

ていく。 $D$  個の特徴に対して RFE を適用したとき、 $K$  を  $D - 1$  から 1 まで変化させた場合の  $D - 1$  状態の CVscore を評価した。図7に RFE で特徴を 1つずつ減らしていったときの ES-K の最高値との比較 (図中左)、状態密度上での選択された組み合わせの位置 (図中中央)、および RFE で CVscore が最大の組み合わせ (図中右) を示す。重みによって評価して特徴を減らしていくと  $K = 9$  のとき、はじめて CVscore が 1 となるような組み合わせが選択される。以降、CVscore が低下しない状態まで特徴を減らしたとき、最小である  $K = 5$  の組み合わせにたどり着いた。最終的に RFE では減らす特徴の数によって、5つの最適解となる組み合わせが候補として挙げられることがわかった。Wine の class0 に対しては RFE を適用しても最適解を求めることが出来た。しかし、これは全 104 状態の最適解のうちの 5つが見つかっただけであり、 $K = 10, 11$  の最適解を見つけることはできていない。これは個々の特徴の重要度以上に、特徴の組み合わせが最適解を形成していることを示している。

## 4.2 Diabetic Retinopathy Debrecen Data Set

Diabetic Retinopathy Debrecen Data Set は最大で 19 個の特徴から糖尿病性網膜症の兆候の有無を診断する問題である。Wine 同様、線形 SVM を用いた識別器の構築を行い、Wine で得られた状態探索の知見が異なるデータに対しても考えられることであるかを検討した。

### 4.2.1 ES

DRD に対して ES を適用した。図8と図9に ES によって得られた状態密度と組み合わせ列挙の結果を示す。図8より、DRD ではどのような特徴選択を行っても、



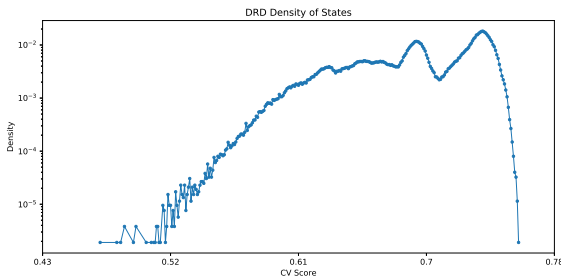


図 8 DRD の ES による状態密度.

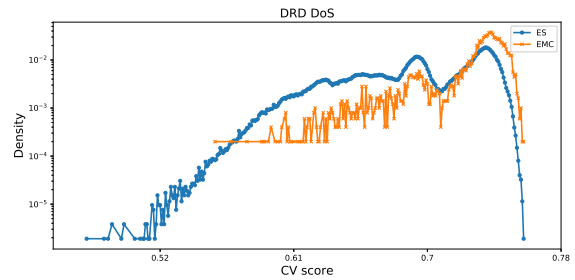


図 11 DRD の AES による状態密度.

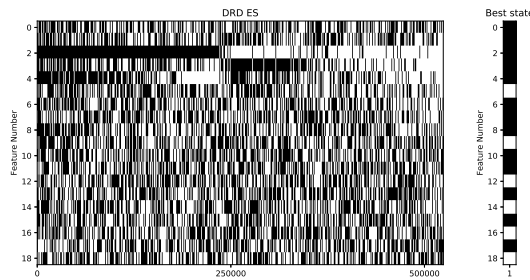


図 9 DRD の ES による組み合わせ列挙.

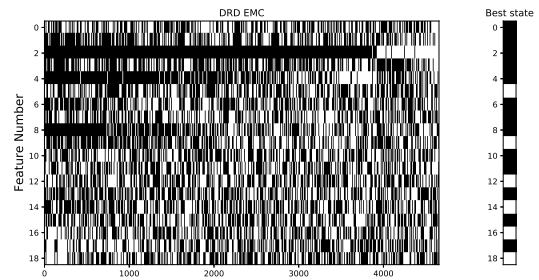


図 12 DRD の AES による組み合わせ列挙.

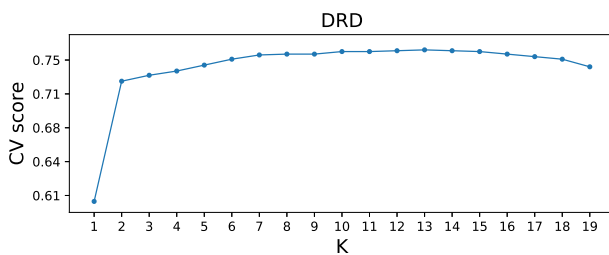


図 10 DRD における特徴数  $K$  に対する  $CVscore$  の最高値.

$CVscore = 0.78$  以上の性能を示すような識別器を構築できないことがわかる。図 9 の右図は最も  $CVscore$  が高い組み合わせを表しているが、DRD においては 1 つの組み合わせが見つかった。DRD における最適解は 1 つのみだったが Wine 同様、図 8 より準最適解のような比較的高い  $CVscore$  を持つような状態がいくつも存在しており、図 9 からは最適解や準最適解が依存している特徴の組み合わせがあることも確認できる。

#### 4.2.2 $K$ スパース全状態探索

DRD に対して ES- $K$  を適用した。図 10 に  $K$  とそのときの ES- $K$  における最高の  $CVscore$  を示す。Wine 同様、適切な組み合わせが形成されていくとともに  $CVscore$  は上昇し、不要な特徴が含まれやすい大きな  $K$  では  $CVscore$  の低下が見られる。

#### 4.2.3 近似的全状態探索

DRD に対して交換 MCMC 法による AES を適用した。レプリカ数は 20 個準備し、全状態数の約 1% のサンプル数である 5000 状態のサンプリングを行った。

図 11 と図 12 に DRD に AES を適用したときの状態密度と得られた組み合わせの列挙を示す。図 11 より、少ないサ

ンプル数から ES の状態密度を捉えるようなサンプリングができていることがわかる。また、図 12 より、ES の組み合わせ列挙でも見られた 2,4,8 番目の特徴が高い  $CVscore$  で優先的に使われているようすが、AES によるサンプリングの結果からも確認することができる。

#### 4.2.4 従来手法との比較

DRD においても RFE による特徴選択を適用した。図 13 右側は RFE で特徴を減らしていったときの  $CVscore$  とそのときの  $K$  における  $CVscore$  の最大値を表している。結果より、RFE で特徴を減らしていったとき、そのときの  $K$  の  $CVscore$  の最大値と比べて低い組み合わせを選択している。これは RFE によって有効な組み合わせに必要な特徴を削ってしまったためと考えられる。また図 13 中央の ES から求めた状態密度上に RFE で取りうる組み合わせの  $CVscore$  を確認してみたところ、ある  $CVscore$  が集まってできた分布の山に多くの組み合わせが存在していることがわかる。

### 5. 考察

全状態探索を行うことで、最適解の数は必ずしも 1 つであるとは限らず、問題によっては複数個存在することを確認した。さらに、状態密度を求めることで最適解に次ぐ準最適解のような組み合わせもまた多く存在していることがわかった。これは有効な特徴を推定する上で、単純に最適解によって選ばれた特徴を有効な特徴として理解するのではなく、その中でもどの特徴に強く依存しているのかを確認する重要な手がかりであることを示唆している。

ES- $K$  では、 $K$  を 1 から順に確認することで、スパースな組み合わせでの最適解を確認した。しかし ES 同様、最

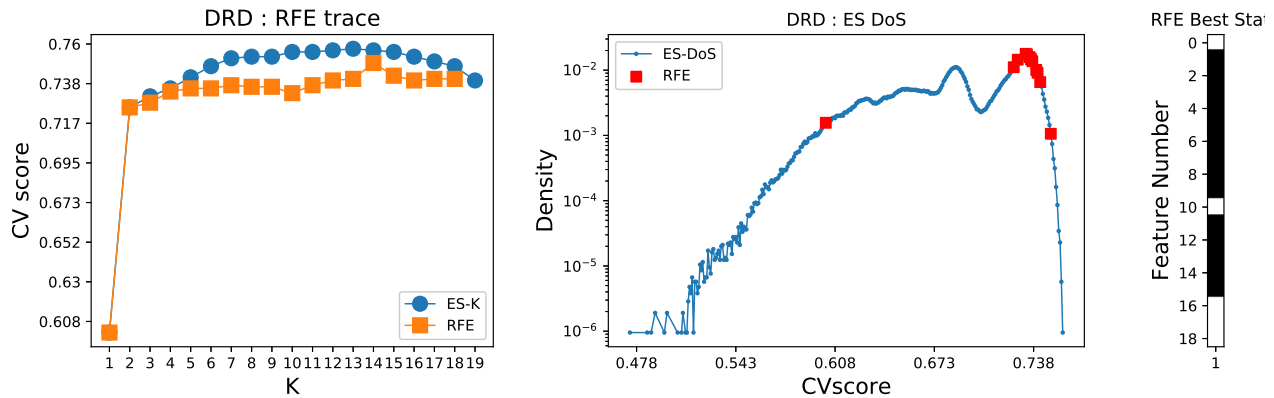


図 13 DRD における特徴選択と状態探索による特徴選択の結果の比較。

適解は 1 つとは限らなかった。ある  $K$  において複数解が得られたときに、どの解が全状態探索の最適解の親となる組み合わせであるのかを議論することは難しい。また、どの範囲で  $K$  の大きさを仮定するかという問題も生じる。今回は ES でおおよそどの特徴がいくつ識別に必要であるのかを大まかに見積もり、その結果と ES- $K$  の結果を比較することにより、全状態探索による推論の妥当性を示した。しかし、ES- $K$  そのものは ES が行えない場合を想定して提案された手法であり、 $K$  に対する見積もりとその妥当性を議論しなければならない。 $K$  を小さく見積もりすぎると、より奥に潜む構造を見逃す可能性があり、逆に多く見積もりすぎると全状態探索同様に計算が困難となってしまう。

AES では、ES のように最適解に加え準最適解といった組み合わせを数多く見つけるようなサンプリングであることを確認できた。さらに、交換 MCMC 法を用いることで ES で得られるような状態密度を推定することも出来た。しかし、ES や ES- $K$  のように想定している組み合わせに対してすべてを確認しているわけではないため、確実に最適解を見つける保証はなく、見たい構造を確認するためには適切なサンプリング回数が必要である。

また今回は従来の特徴選択の手法の 1 つとして、RFE による特徴選択を例に挙げて状態探索手法との比較を行った。Wine のような簡単な問題に対しては従来手法でも最適解を選び出すことができたが、多数存在している最適解のわずか一部しか見つけることができない。また、DRD に従来手法を適用したところ、各  $K$  における最適解を見つけることも難しいことがわかった。このような結果から、組み合わせが解釈の観点において、最適化によって選ばれた組み合わせから議論するよりも、状態列挙によって多様な組み合わせを確認することが妥当であると考えられる。

## 6. まとめ

本研究では、異なる 2 つのデータに対して ES を適用した結果の考察と、ES- $K$  や AES で得られる結果との比較を

行った。ES を用いることにより、最適解のみでなく準最適解を多数見つけることが出来るだけでなく、軸となるようなスパースな組み合わせが存在していることを確認し、これらの結果が ES- $K$  や AES を適用することでも十分に得られることを確認した。しかしながら、ES が困難な高次元のデータにおいて、ES- $K$  における  $K$  の見積もりや AES による適切なサンプリング数といった点に関しては考察時に考慮していかなければならない。今後は、より次元の高いデータに対して ES- $K$  や AES の適用し、従来の特徴選択アルゴリズムとの比較を行う。

## 参考文献

- [1] Cover, T. M. and Campenhout, J. M. V.: On the Possible Orderings in the Measurement Selection Problem, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 7, No. 9, pp. 657–661 (1977).
- [2] Dheeru, D. and Karra Taniskidou, E.: UCI Machine Learning Repository (2017).
- [3] Ichikawa, H., Kitazono, J., Nagata, K., Manda, A., Shimamura, K., Sakuta, R., Okada, M., Yamaguchi, M. K., Kanazawa, S. and Kakigi, R.: Novel method to classify hemodynamic response obtained using multi-channel fNIRS measurements into two groups: exploring the combinations of channels, *Frontiers in human neuroscience*, Vol. 8, p. 480 (2014).
- [4] Nagata, K., Kitazono, J., Nakajima, S., Eifuku, S., Tamura, R. and Okada, M.: An Exhaustive Search and Stability of Sparse Estimation for Feature Selection Problem, *IPSJ Online Transactions*, Vol. 8, pp. 25–32 (2015).
- [5] Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288 (1996).
- [6] 五十嵐康彦, 竹中 光, 中西 [大野] 義典, 植村 誠, 池田思朗, 岡田真人: 全状態探索による線形回帰のスパース変数選択, 人工知能学会全国大会論文集, Vol. JSAI2017, pp. 2I22–2I22 (2017).
- [7] 川端大貴, 市川寛子, 永田賢二, 永福智志, 田村了以, 岡田真人: ES-SVM の解空間の解析, 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 2L51in2–2L51in2 (2016).