ISSN 2186-5647

OS-3

組み合わせ爆発した決定木集合上のルールマイニングアルゴリズムの提案と多様な領域への応用事例

○大前 佑斗[1]，森 雅也[1]，柿本陽平[1]
[1]日本大学 生産工学部

1. はじめに

本稿では、arXiv における Machine Learning (cs.LG) セクションに投稿したプレプリント arXiv:2310.02633 の紹介を行う1)。これは、ビッグデータからルールマイニングを行うための機械学習アルゴリズムについて記載されたものである。

2. 決定木について

半世紀近く前に、Quinlan によって提案された ID3 アルゴリズムによる決定木は、先人たちの創意工夫により多少の改良は行われつつも、エントロピーないしは Gini 係数最小化による分岐探索という本質的な処理は変わらずに使用され続けている。

事前に集めたデータに対し上述のアルゴリズムを適用すると「降水確率が90%以上で、予想気温が30度以上ならば、惣菜は売れない」のように、人にとってわかりやすいルールが抽出される。このルールをまとめたものが決定木であり、これが今なお使用され続ける大きな理由は、解釈の容易性の高さからであろう。

決定木を構築するアルゴリズムは複数あるものの、そのすべてにデータを明瞭に分割できる特徴量を探索する処理が導入されている。例えば、降水確率、予想気温、予想湿度、予想風速、…その他色々…、のような多数の特徴量から「売れるか、売れないか」を推定できる能力があるものが探索される。これによって、例えば降水確率が1番データを明瞭にわけることができると判断された場合、それによるデータ分割が行われる。その後、分割後のデータに対しても同様の処理を行うことで、木が成長していく。

これはとてもシンプルで、結果的に人にとってわかりやすいルールを取り出すことができる。このように必要最低限のルールのみを取り出す思想のことを、the principle of Occam's Razor（オッカムの剃刀）と呼ぶ。

3. 決定木が抱える問題

非常に良さそうに見える決定木は、いくつかの問題を抱えている。1つは、探索により確定した特徴量でデータを分岐する際に、データサイズが（期待値的に）半分になることである。例えば、データ数が1,000くらいある状況を考える。このとき、深度1の決定木のリーフノードのサンプル数は500、深度2なら250、深度3なら125と、どんどん減っていく。つまり、1つの特徴量を木構造に反映させるという単純な作業に対し、データ数を半分も消費してしまう。これは非常にもったいないように思える。

もう1つの問題は、1番有効な特徴量のみで分岐を生成してしまう、というものである。前の例で説明すると、降水確率が1番重要だから降水確率で分岐を作るわけだが、2番目や3番目に重要と判断された特徴量は、果たしてどこにいくかというと、それらは木の中には反映されない。下位のノードとして採用されることもなくはないが、あくまでもその後の最適化問題の解としてそうなるだけでしかない。このような考えが「必要最低限のルールを取り出す」という特性に直結することになり、多くの場合これはメリットとして説明されるが、しかし、他にも多数のルールが眠っているのにも関わらず、1番ではないという理由だけで捨ててしまうこの考えは、見方を変えるとデメリットのように感じられる。

4. ランダムな方法の問題

すべての特徴量を使って決定木を構築すると1つの木しか取り出されないため、結果として少数のルールしか抽出されない。逆にいえば、ランダムに取り出された特徴量で決定木を作る作業を反復すれば、色々なルールを取り出すことができるだろう（ランダムフォレストがこれに近いかもしれない）。

この考えは悪くはないように思えるが、大きな欠点を2つ抱えている。1つ目は、決定木を構築するアルゴリズムが、あまりノイズに対し

Rule mining algorithm for combinatorially exploded decision trees and introducing some application examples

Yuto OMAE, Masaya MORI, and Yohei KAKIMOTO

て頑健ではないことである。先日、9個の特徴量からなるデータセットに対し、乱数生成させた無意味な特徴量を20個分生成し、データセットに加え、これによって得られた29個の特徴量を用いて決定木を構築してみた。当然、過学習を抑制するためのハイパラ探索を実施している。さて、この結果、無意味な特徴量が木の分岐に採用された割合はどれくらいだったかというと、およそ80%であった。つまり、無意味な特徴量が格納されたデータで決定木を作ると、ハイパラ探索を行ったとしても、おかしな分岐を獲得してしまうのである。これは、ランダム選択された特徴量で決定木を作ってしまうと、無意味なルールが生成されてしまうリスクがあることを意味している。

2つ目の問題は、取り出すことのできる特徴量のパターン数が、組合せ爆発を起こしていることである。例えば、100個の特徴量の中からランダムに10個を取り出すパターンの総数は、${}_{100}C_{10}$ ～ 20兆くらいになる。つまり、理論的に構成可能な決定木の数も、20兆くらいとなる。ランダム選択された特徴量で決定木を1,000個や2,000個作ったところで、データに隠れ潜むルールはほぼ取り出せないことになる。なお、全探索の場合は1,000年間くらいの計算時間が要求されるし、先ほど述べたように、その決定木の集まりの中には無意味な特徴量が多数埋め込まれることになる。

5. MAABO-MT アルゴリズム

この問題を解決するには、無意味な特徴量の採択を回避しながら、良いルールが眠っている決定木を戦略的に構築する手法が必要となる。

このため、最近人気のある近似解の探索手法である Parzen 推定に基づくベイズ最適化に注目する。Parzen 推定とは、non-para な確率分布により確率を近似する手法であり、ベイズ最適化の内部処理として適用される場合がある。特に探索対象が実数値ベクトルで表現される場合には、Gaussian カーネルが、カテゴリカルベクトルの場合には Aitchison-Aitken カーネルが利用される。しかし、本研究の探索対象は特徴量の組み合わせであるため、上述の関数では確率分布を構築することができない（厳密にいうとできなくもないのだが、数学的にきれいではない）。

そのため、集合を対象として確率分布を構成するための新たな関数として、Modified Aitchison-Aitken 関数（MAA関数）を提案した。これは、Kolmogorov の確率の公理を満たすように設計されており、つまり、これにより得られる関数は確率分布の定義を満たす。そのため、MAA関数を利用することで、特徴量の組み合わせ探索問題に対し、ベイズ最適化を適用することが可能となる。これは、少ない計算量で、戦略的に良い決定木を複数構築できることを意味している。このアルゴリズムは、MAA関数による Bayesian Optimization で Making Tree をするということで、MAABO-MT と名付けた。

6. GS-MRM アルゴリズム

上述の方法で、良い決定木を複数構築することができるが、その内部には大量のルールが存在することになる。例えば、木の深度が3のとき、1つの決定木は最大で $2^3 = 8$ 個のリーフノードを保有していることになる。この程度の規模の木が100個あるならば、リーフノードの総数は最大で800個となる。人がこのような大量のルールから何か考察することはとても難しく、すなわちルール削減・抽出が必要となる。

良い決定木に含まれるリーフノードは一見信頼して良いようにも思われるが、実のところそういうわけでもない。たとえば、クラスが明瞭に分類できていないリーフノードや、サンプルサイズが少ないリーフノードは、統計的に意味があるとはいえないため、信頼できないだろう。また、複数の決定木には類似性の高いリーフノードが多数含まれていることが想定される。そのため、これらのリーフノードをすべて削除し、クラスが明瞭に分類されており、サンプルサイズが十分で、かつ、すでに抽出されたルールと似ていないもののみを取り出すアルゴリズムを提案した。これは、Gini 係数と Simpson 係数を利用した Multi Rules Mining ということで、GS-MRM と名付けた。

7. おわりに

本シンポジウムでは、MAABO-MT と GS-MRM をいくつかの領域で適用した結果について説明する。また、数行のコードを書くだけで（高速に）ルールマイニングを実行できるデモを行う。

参考文献：

1) Y. Omae, et al., Multi-rules mining algorithm for combinatorially exploded decision trees with modified Aitchison-Aitken function-based Bayesian optimization, arXiv:2310.02633, URL: https://arxiv.org/abs/2310.02633