


学 位 論 文 の 要 旨

専攻名	環境工学専攻	ふりがな 氏 名	なかむら まさとし 中 村 将 俊	
学位論文題目	Inference and Evaluation Based on Garrote Trees as Regression Analysis (回帰解析における Garrote Trees による推測とその評価)			
<p>In regression analysis, stochastic models are often constructed to model relationships between outcomes and explanatory variables. We derive statistical interpretation about the underlying structure of data based on these models. When we use a linear regression model and the model provides good fitting to the data, it is straightforward to interpret the relation. However, there are cases where it may be difficult to formulate a linear model reflecting actual characteristics in detail. We must depart from the traditional paradigm of linearity in creating models, without any prescribed functional form of covariates for the underlying base model and relaxing assumptions imposed in those analyses. Especially, we need a flexible approach which constructs a model that is as easy to interpret as a regression model and identifies interactions between factors automatically based on the data.</p> <p>As alternative methods fulfilling these requirements, we recommend tree-structured approaches that develop one/some tree structure/s adaptively without any prescribed base structure and derive interpretations of the data from the tree/s. A tree-structured model is constructed using the subspaces of outcomes divided by some cut-off values of explanatory variables together with corresponding estimates in each subspace. This model can be expressed as a structure with nodes determined by values of selected explanatory variables and terminal nodes that contain the estimates in subspaces of outcomes. We can regard this structure as a tree, looking the terminal nodes as leaves and the edges between adjacent nodes as stems of a plant. The tree structure gives a graphical interpretation of the potential relationship between outcomes and explanatory variables for the given data, and visualizes the process of classifying the data into terminal nodes, according to the rules represented in the structure. In addition, we can derive insights related to their interactions affecting the outcome and non-linear relationship between outcome and explanatory variables.</p> <p>CART method is a foundation of studies for tree structured approaches. Noteworthy, ensemble learning methods, which have been studied primarily in connection with machine learning, were incorporated into the tree-structured approach. Random Forest (RF) involves an ensemble learning method based on the trees and can predict outcomes more precisely. However, RF cannot provide a tree-structured model for interpreting the data. In this paper, we aim the improvement of regression analysis based on RF. In terms of predictive accuracy of RF, there exist some tree-based approaches which have higher</p>				

predictive performance than RF, and we expect some alternative improvement of predictive accuracy of RF. Furthermore, in terms of descriptive feature of RF, it is difficult to interpret the data based on trees in the ensemble, and we would like to make some interpretable and meaningful trees.

To attain predictive accuracy and interpretability based on the above two viewpoints, we propose Garrote Trees (GT) as a new tree-structured model, and formulate GT as an adjustment of RF based on NNG. In GT, entire trees are removed or weighted by the NNG-type penalty, and a few useful trees selected from RF can be visualized for interpreting data. The GT provides an alternative adjustment of RF in terms of predictive accuracy as well as capability of interpreting data.

Our simulation studies show that the proposed method is highly accurate predictively and provides a potential ability to interpret the data from new meaningful standpoints. Two case studies of diabetes and prostate cancer data illustrate predictive accuracy and descriptive features of GT. Case studies and simulations elucidated the merits of GT. GT has better performance with regard to prediction error and can visualize a few trees. Especially, simulation studies describe how GT can recapture tree-structure via adaptation of representative trees. These trees provide new potentially meaningful insights regarding the data.

【604 語】

(注) 和文 2,000 字又は英文 800 語以内

続紙 有 無

学位論文審査結果の要旨

専攻	環境工学 専攻	氏名	中村 将俊
論文題目	Inference and Evaluation Based on Garrote Trees as Regression Analysis		
主査	中島 誠		
審査委員	古家 賢一		
審査委員	西野 浩明		
審査委員	越智 義道		
審査委員			
審査結果の要旨 (1000 字以内)			
<p>統計的なデータの分析において、説明変数によって反応変数の挙動を説明し、その予測等のために用いられる代表的な方法として線形回帰分析がある。事前情報に乏しく、探索的に反応の構造を調査する際に、この方法が用いられることがある。しかし、数理的構造に強い仮定をおくことから、この方法では現実の現象を十分に記述できない場合がある。その代替法の一つが CART(Breiman, 1984)に代表される樹木構造接近法である。CART では、樹木構造による柔軟な表現が可能で、説明変数のプロファイルによって現象の説明が可能であるが、十分な予測精度を得ることが困難な場合がある。その予測性能を強化する方法として、多数の樹木を構成し、それらを統合するアンサンブル学習を基礎とした Boosting(Freund, 1995)や Random Forest(RF)(Breiman, 2001)などが提案されている。しかし、これらの手法は多くの樹木を統合しているために、現象の解釈上の困難さを抱えている。この解釈上の問題点を克服するために、本論文では縮小推定法の一つである non-negative garrote(Breiman, 1995)を RF に適用した Garrote Trees(GT)を新たに提案している。GT では、RF に比し、予測精度を低減することなく、樹木を削減でき、さらに構造理解に有用な代表的樹木を選別可能なことを、シミュレーションにより確認した。さらに GT では CART では捉えることが出来ない構造的な特徴を代表樹木として示すことが可能であることをシミュレーションならびに 2つの事例検討により示している。同様な研究として縮小推定法 Lasso(Tibshirani, 1996)を用いて RF の樹木を削減する研究(LassoRF: Nakamura, 2013)もあるが、GT と LassoRF とは予測精度の観点でほぼ同等の性能を示し、共に RF よりも優れていることを確認している。ただし、LassoRF では代表的な樹木を選び出すには至っていない。以上のことから、本研究は現象の解明につながる探索的なデータ解析の手法の提案として、有用であり、新規性も高いものと考えられる。</p> <p>論文審査会や公聴会における質問に対しても明確かつ的確に回答が行なわれた。また、審査の過程で指摘された分析結果の安定度、予測精度の評価に関わる課題についても精査が行われ、論文中に適切に修正が反映されており、論文の完成度は十分である。</p> <p>以上の審査結果から、本論文は博士(工学)の学位に値するものと審査委員会全員一致して判定した。</p>			