The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.5.717
27(5), 717–729

Comparison of Variable Importance Measures in Tree-based Classification

Na-Young Kim[a] · Eun-Kyung Lee[a,1]

[a]Department of Statistics, Ewha Womans University

(Received July 11, 2014; Revised September 17, 2014; Accepted September 29, 2014)

Abstract

Projection pursuit classification tree uses a 1-dimensional projection with the view of the most separating classes in each node. These projection coefficients contain information distinguishing two groups of classes from each other and can be used to calculate the importance measure of classification in each variable. This paper reviews the variable importance measure with increasing interest in line with growing data size. We compared the performances of projection pursuit classification tree with those of classification and regression tree(CART) and random forest. Projection pursuit classification tree are found to produce better performance in most cases, particularly with highly correlated variables. The importance measure of projection pursuit classification tree performs slightly better than the importance measure of random forest.

Keywords: Classification, variable selection, tree-based classification, projection pursuit.

1. 서론

분류분석(classification)은 설명변수들을 이용하여 관측값을 미리 정해진 그룹 중의 하나로 분류하는 분석방법으로 모수적인 방법으로는 판별분석, 로지스틱 회귀분석 등이 있고 비모수적인 방법으로는 의사결정나무, 랜덤 포레스트 등의 나무구조를 이용한 방법들이 있다. 모수적인 방법은 자료공간에 대한 가정이 필요한 반면 나무구조의 분류방법은 의사결정규칙을 이용하여 자료를 그룹으로 분류하고 예측하는 분석방법으로 의사결정나무의 경우 분석과정을 나무구조로 표현할 수 있어 결과를 이해하고 해석하기 쉽다는 장점이 있다. 재표본 기법을 이용한 랜덤 포레스트 방법 (Breiman과 Cutler, 2012)은 의사결정나무모형에 비해 예측력이 뛰어나다는 장점이 있으나 의사결정규칙을 설명하기 힘들다는 단점이 있다. 이 두 방법 모두 하나의 마디에서 하나의 변수만을 이용하게 되므로 변수들 간의 상관관계가 큰 경우 이를 모형에 반영하지 못하게 된다. 이를 보완하기 위하여 개발된 사영추적분류나무 (Lee 등, 2013)는 하나의 마디에서 변수들의 선형결합을 이용하여 그룹을 분리한다. 이 과정에서 분류에 대한 변수의 중요도를 파악할 수 있게 된다.

본 연구에서는 분류분석에서 사영추적분류나무를 이용하여 자료의 크기가 방대해짐에 따라 중요한 문제로 대두되고 있는 변수 중요도에 대하여 고찰해 보고자 한다. 먼저 사영추적분류나무의 분류과정에서

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2009-0093827).
[1]Corresponding author: Department of Statistics, Ewha Womans University,52, Ewhayeodae-gil, Seodaemun-gu, Seoul, 120-750, Korea. E-mail: lee.eunk@ewha.ac.kr

계산되는 사영추적계수를 이용하여 분류를 위한 변수의 중요도를 계산하고 이들의 특성을 살펴본다. 이를 같은 형태의 나무모형방법중 일반적으로 널리 쓰이고 있는 CART (Breiman 등, 1984), 랜덤 포레스트의 결과와 비교 분석하여 사영추적분류나무의 특성을 살펴보고 분류를 위한 변수 중요도의 성능을 비교해 보고자 한다.

2장에서는 나무구조를 이용한 분류방법들의 특성과 장, 단점들을 살펴보고 3장에서는 Fish 자료에 사영추적분류나무를 적용하여 분석해 보고 변수 중요도를 계산하는 방법에 대하여 논의한다. 또한 모의시험자료와 Fish 자료를 이용하여 중요변수를 선택하고 이 결과를 CART와 랜덤 포레스트의 결과와 비교한다. 4장에서는 사영추적분류나무의 성능과 변수 중요도의 성능을 다른 방법들과 좀 더 심도있게 비교하기 위하여 8 개의 실제 자료에 적용시키고 결과를 비교, 분석한다.

2. 나무구조를 이용한 분류방법들

2.1. CART

CART(Classification And Regression Tree)는 Breiman 등 (1984)이 개발한 의사결정나무방법의 하나로 범주형 반응변수의 경우 분류(classification)의 형태로, 연속형 반응변수의 경우 회귀(regression) 의 형태로 추정을 하며 두 경우 모두 나무 형태의 그래프로 표현할 수 있다. CART는 나무의 성장, 가지치기, 타당성 평가, 그리고 해석 및 예측의 단계를 거쳐 만들어진다. 분류작업이 용이하고 연속형과 범주형 반응변수 모두에 편리하게 이용할 수 있는 방법이나 자료의 변화에 민감하게 반응하는 불안정성을 가지고 있으며 또한 회귀나무의 경우 연속형 반응변수에 대한 다양한 모형들에 비하여 상대적으로 예측력이 떨어지는 단점을 가지고 있다. 불안정성 등의 심각한 단점이 있음에도 불구하고 CART가 널리 쓰이는 가장 큰 이유는 모형의 해석이 쉽고 분류 및 예측 작업에 필요한 중요한 변수를 알아내는 데에 유용하기 때문이다.

2.2. 랜덤 포레스트(Random forest)

랜덤 포레스트는 다수의 나무를 생성시켜 이용하는 방법으로 재표본 방법을 이용하여 CART 의 불안정성을 보완하고 있으며 정확한 분류를 할 수 있는 학습 알고리즘 중 하나로 간주된다. 랜덤 포레스트의 가장 큰 장점은 변수에 노이즈가 있더라도 이에 영향을 크게 받지 않고 분류를 잘 수행할 수 있는 특징이 있다. 또한 재표본 기법을 이용함으로써 변수의 개수에 비해 자료의 수가 작은 경우의 문제(large p small n)를 자동적으로 다룰 수 있는 이점이 있다.

랜덤 포레스트는 알고리즘의 한 부분으로 각 설명변수의 중요도를 측정하는 값(importance measure)을 계산한다. 이는 각 마디에서 변수 값들이 분류의 정확성에 어느 정도 영향을 미치는가를 근거로 측정되는 것으로 변수의 중요도를 파악하는데에 유용하게 쓰인다.

2.3. 사영추적분류나무(Projection pursuit classification tree: PPtree)

사영추적분류나무(Lee 등, 2013)는 사영추적(projection pursuit)방법을 이용한 나무모형의 분류방법이다. 사영추적방법은 고차원 자료의 관심있는 낮은 차수의 사영을 찾기 위해 사영추적지수(projection pursuit index)를 정의하고 이를 최적화시키는 선형 사형을 찾는 방법이다. 이는 탐색적 자료분석 기법 중 하나로 사영추적이란 용어를 만든 Kruskal (1969), 그리고 Freidman과 Tukey (1974)에 의해 유래되었다. 다양한 사영추적지수들이 개발되어 있으며 그 중 LDA 지수 (Lee 등, 2005)와 PDA 지수 (Lee와 Cook, 2010)는 자료의 그룹에 대한 정보를 이용하여 그룹 간 차이를 잘 나타내는 저차원의 사영을 찾

는 지수로 사영추적분류나무에 쓰인다. LDA 지수는 자료의 수가 변수의 수의 2배 이상이 되는 경우 적절하게 쓰일 수 있으나 변수의 개수가 자료의 개수에 비해 큰 경우에는 데이터 적재 현상(data piling problem) 을 보인다. PDA 지수는 이를 보완한 것으로 변수의 개수가 자료의 개수에 비해 큰 경우 유용하게 쓰인다.

사영추적분류나무는 각 마디에서 자료를 잘 분류할 수 있는 하나의 변수를 찾는 대신 그룹 간 차이를 잘 나타내주는 1차원의 사영을 찾아 변수들의 선형결합형태로 새로운 변수를 만들고 이를 각 마디에서의 분류에 이용하게 된다. 사영추적분류나무의 또 하나의 특징은 각 마디에서 그룹들을 두 군으로 분류할 수 있는 규칙이 제공되며 최종마디에는 각각의 그룹들이 중복 없이 할당이 된다. 즉, 두 그룹이 있는 자료의 경우 하나의 분류마디만 존재하고 2개의 최종마디가 각각의 그룹에 할당이 되어 나무의 깊이는 1 이 된다. 그러므로 사영추적분류나무의 최종마디의 개수는 분류하고자 하는 그룹의 개수와 같게 되고 사영추적분류나무의 깊이는 최대 그룹의 개수-1 이 된다. 그러므로 기존의 분류나무에서 필요한 가지치기 과정이 필요가 없어지며 각 마디에서 사용한 1차원의 사영계수를 통하여 나누어지는 그룹의 분류에 중요한 영향을 미치는 변수들이 어떤 변수인지를 파악할 수 있게 된다. 본 연구에서는 이들 각 마디에서의 사영계수 정보들을 종합하여 전체 분류에 대한 변수의 중요도를 측정할 수 있는 측도를 좀 더 자세히 고찰하고 이를 CART, 랜덤 포레스트의 결과와 비교, 분석하고자 한다.

3. 사영추적분류나무의 변수중요도

3.1. 각 마디별 정보

사영추적분류나무에서 각 마디에 이용된 사영계수들은 길이가 p인 벡터로 크기 1인 단위벡터로 표준화될 수 있다. 표준화된 사영벡터에서 각 변수에 해당하는 사영계수는 각 마디에서 분류규칙을 찾기 위한 변수들의 선형결합을 계산하는데 쓰이므로 절대값이 큰 사영계수를 갖는 변수는 해당 마디에서 큰 영향력을 갖는 반면 절대값이 작은 사영계수를 갖는 변수는 영향력도 작아지게 된다. 그러므로 각 마디에서의 사영계수의 절대값이 해당 분류에 대한 변수의 중요도를 나타내는 하나의 측도가 된다.

사영계수의 절대값이 변수의 중요도를 나타내기는 하지만 이는 해당 자료의 형태, 특히 변수의 개수(p)에 따라 달라지므로 이를 위한 명확한 기준을 제시하기는 어렵다. 본 연구에서는 모든 변수가 같은 중요도를 가질 때의 계수값인 $\sqrt{1/p}$을 각 마디에서의 변수의 중요도 기준으로 삼고자 한다. $\sqrt{1/p}$보다 큰 사영계수의 절대값을 갖는 변수는 평균 이상의 중요도를 갖는 것으로 해석할 수 있다.

사영추적분류나무의 각 마디에서의 변수 중요도의 역할을 좀 더 자세히 살펴보기 위해 Fish 자료를 이용한다. 이 자료는 Journal of Statistical Education Data Archive로부터 얻은 자료로 핀란드의 Laengelmavesi 호수에서 잡힌 159마리의 물고기에 대한 자료이다. 이들은 7종류의 물고기 - Bream(35마리), Parkki(11마리), Perch(56마리), Pike(17마리), Roach(20마리), Smelt(14마리), 그리고 Whitewish(6 마리) - 중 하나로 각각의 물고기로부터 6가지의 측정(weight, length1, length2, length3, height, 그리고 width)과 성별을 기록한 것이다. 본 연구에서는 성별을 제외한 6가지의 측정변수를 이용하여 7개의 그룹을 구분해 보고자 한다. Figure 3.1는 LDA 지수를 이용한 사영추적분류나무의 결과를 그림으로 나타낸 것이다. Table 3.1는 사영추적분류나무의 각 마디에서의 표준화 사영계수를 크기별로 나열한 것이다.

Figure 3.2은 첫번째 마디에서 찾은 1차원의 변수에 대한 히스토그램을 그린 것이다. 히스토그램 위의 숫자들은 각 자료의 그룹에 해당되는 정보를 나타낸다. 첫번째 마디(N1)에서는 Bream(1), Parkki(2)와 나머지 종류를 완벽히 구별해 낸다. 이 때 큰 영향을 미치는 변수는 height와 length3이다. 두번째 마디(N2) 에서는 Bream(1)과 Parkki(2)를 구분해 낸다 (Figure 3.3). 이 때 가장 큰

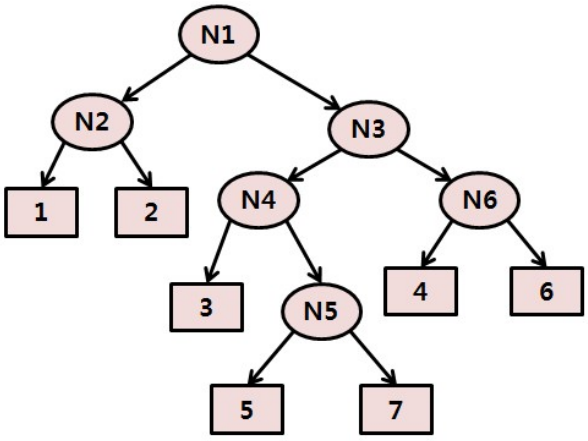

Figure 3.1. Fish data: The result of PPtree

Table 3.1. Fish data: The standardized projection coefficients of each node

	Node 1		Node 2		Node 3		Node 4		Node 5		Node 6	
	variable	표준화 사영계수	variable	표준화 사영계수	variable	표준화 사영계수	variable	표준화 사영계수	variable	표준화 사영계수	variable	표준화 사영계수
1	**Hgt**	−0.773	**L3**	−0.824	**L2**	−0.747	**L2**	−0.737	**L2**	0.801	**L3**	−0.850
2	**L3**	−0.412	**L2**	0.436	**L1**	0.604	**L3**	0.670	**L1**	−0.544	**L1**	0.495
3	Wdt	0.353	L1	0.361	Hgt	−0.245	L1	0.082	L3	−0.247	Wgt	0.153
4	L1	0.327	Wdt	0.025	L3	0.122	Hgt	0.024	Wgt	0.020	L2	0.088
5	L2	0.041	Hgt	0.013	Wgt	0.042	Wdt	−0.007	Hgt	0.012	Wdt	0.029
6	Wgt	−0.005	Wgt	−0.001	Wdt	−0.029	Wgt	−0.002	Wdt	−0.001	Hgt	−0.021

영향을 미치는 변수는 length3이며 length2와 length도 어느정도 영향을 미치고 있다. 세번째 마디(N3)에서는 length2와 length1이 중요 역할을 하여 Perch(3), Roach(5), Whitewish(7)와 Pike(4), Smelt(6)를 구별해 내며 (Figure 3.4) 네번째 마디(N4)에서는 Perch(3)와 Roach(5), Whitewish(7)를 length2와 length3의 중요 역할로 분류해 낸다 (Figure 3.5). 다섯번째 마디(N5)에서는 Roach(5)와 Whitewish(7)를 구별해 내며 (Figure 3.6) 이때는 세번째 마디에서와 마찬가지로 length2와 length1이 큰 역할을 한다. 여섯번째 마디(N6)에서는 length3와 length1의 역할로 Pike(4)와 Smelt(6)를 구별해 낸다 (Figure 3.7). 첫번째로 Bream, Parkki와 나머지를 분류할 때를 제외하고는 모두 length1, length2, length3이 큰 역할을 하고 있다. 이 사영추적분류나무로는 오분류없이 Fish 자료의 모든 관측값들을 분류할 수 있다.

Figure 3.8는 CART의 결과를 그림으로 나타낸 것이다. Fish 자료의 경우 CART의 결과가 사영추적분류나무의 결과에 비해 조금 더 단순한 형태로 나타난다. Bream, Parkki와 나머지를 분류하는데에는 height 변수가 이용되었으며 Bream과 Parkki를 분류하는데에는 length3이 이용되었다. Perch, Roach와 Pike, Smelt를 분리하는 데에는 height 변수가 쓰이고 있으며 Perch와 Roach에는 width 변수가, Pike와 Smelt에는 weight 변수가 쓰이고 있다. 그러나 CART의 경우 자료의 수가 작은 Whitewish 그룹을 분리해 내지 못하고 있고 이에 따라 오분류가 21개로 13%의 오분류율을 나타내고 있다.

사영추적분류나무와 CART의 결과에서 각 그룹을 분류하는데에 중요하게 쓰이는 변수가 다르게 나타나고 있음을 알 수 있다. CART의 경우 하나의 마디에서 하나의 변수만을 택하게 되나 사영추적분류나

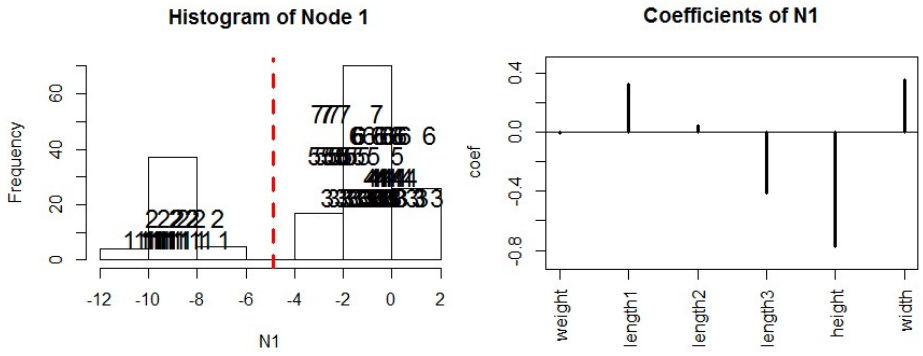

Figure 3.2. Fish data: Node 1 in PPtree, Bream(1), Parkki(2), Perch(3), Pike(4), Roach(5), Smelt(6), White wish(7)

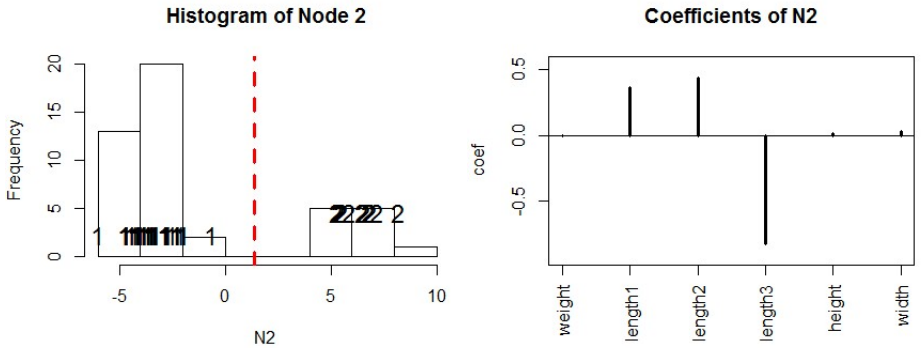

Figure 3.3. Fish data: Node 2 of PPtree

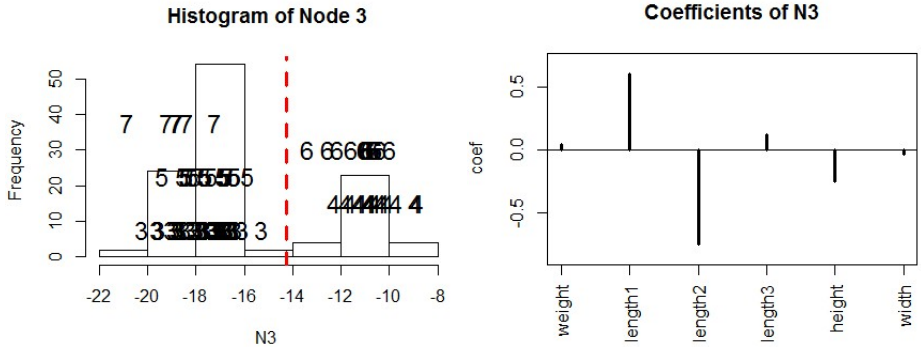

Figure 3.4. Fish data: Node 3 of PPtree

무의 경우 모든 변수를 이용하게 되므로 좀 더 정확한 분류를 할 수 있게 된다. 또한 각 마디마다의 사영계수를 이용하여 세부 그룹들이 분리되는데에 중요한 역할을 하는 변수들을 파악할 수 있게 된다.

3.2. 변수의 중요도

사영추적분류나무는 각 마디에서 그룹들을 두 군으로 나누게 되므로 각 마디에서의 사영계수들은 해당

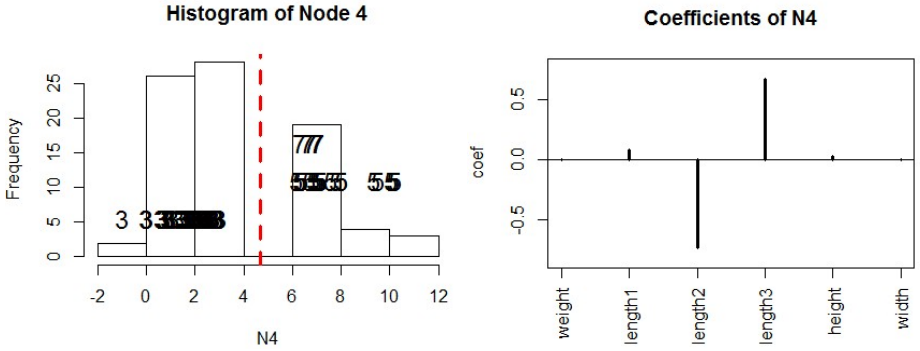

Figure 3.5. Fish data: Node 4 of PPtree

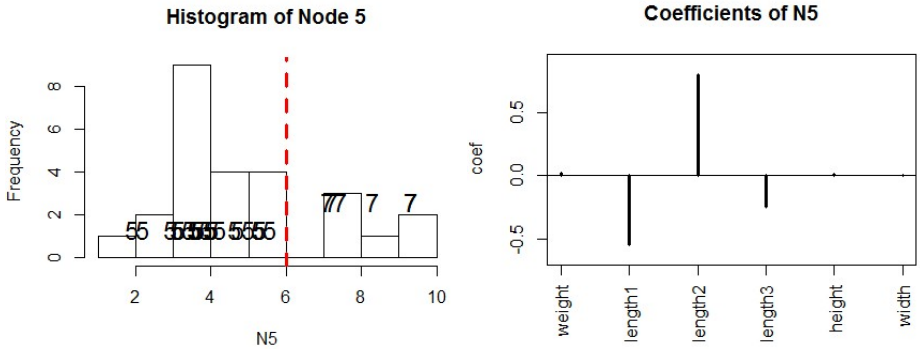

Figure 3.6. Fish data: Node 5 of PPtree

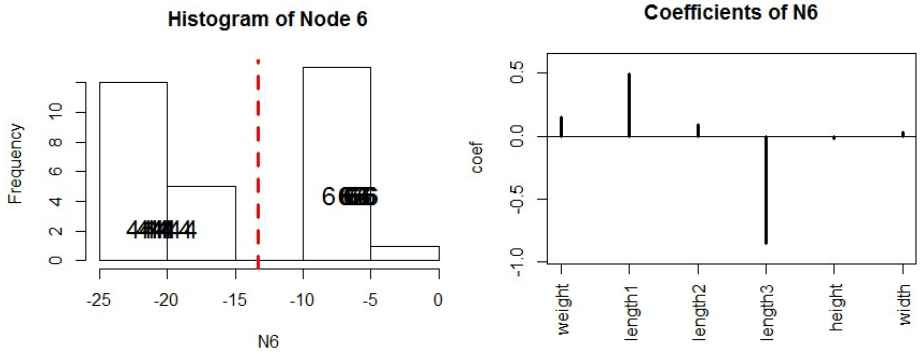

Figure 3.7. Fish data: Node 6 of PPtree

그룹들을 두 군으로 나눈데에 대한 변수들의 중요도를 나타낸다. 이를 종합하여 전체 분류에서의 중요도를 나타내는 측도를 만들 수 있다. Lee 등 (2013)은 각 마디에서 나누어지는 그룹의 수를 가중치로 하여 사영계수들의 가중합을 각 변수의 중요도로 이용하고 있다. 이는 그룹 내의 자료 개수는 다르지만 모든 그룹들이 동일한 정보를 가지고 있다는 가정을 고려하여 만들어진 것으로 Fish 자료의 경우 관측값이 6개 뿐인 Whitewish 그룹을 56마리의 관측값이 있는 Perch 그룹과 동일하게 취급하여 분류에 대한 중요도를 계산하는 것이다.

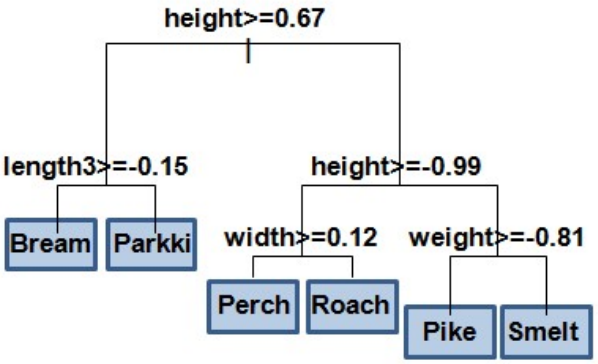

Figure 3.8. Fish data: The result of CART

Table 3.2. Fish data: The importance measures of PPtree and random forest

	PPtree		PPtree		Random Forest	
	variable	중요도1	variable	중요도2	variable	중요도
1	**length2**	0.705	**length2**	0.679	height	42.85960
2	**length3**	0.561	**length3**	0.446	width	25.24001
3	length1	0.243	length1	0.230	length3	17.20940
4	weight	0.078	height	0.104	weight	15.17031
5	height	0.062	weight	0.074	length2	11.94645
6	width	0.030	width	0.042	length1	11.84341

반면 각 그룹보다는 관측값 각각이 동일한 정보를 가지고 있다는 가정 하에서 각 마디에 해당되는 자료의 수를 가중치로 이용할 수 있다. 즉, Perch 그룹의 한 관측값과 Whitewish 그룹의 한 관측값에 동일한 비중을 두는 것으로 이 경우는 전체의 오분류을 좀 더 고려하여 중요도를 계산하게 되는 것이다. 그룹마다 관측값의 개수가 같다면 이 두 측도는 같게 나타난다. 사영추적분류 중요도에 대한 기준은 각 마디별 기준으로 이용했던 $\sqrt{1/p}$을 사용한다.

Fish 자료의 경우 N1에서는 전체자료인 159개의 7개 그룹을 나누며 N2에서는 2개 그룹의 46개 자료를, N3에서는 5개 그룹의 113개, N4에서는 3개 그룹의 82개, N5에서는 2개 그룹의 26개, N6에서는 2개 그룹의 31 개를 분리해 내므로 N1에서의 자료의 개수를 이용한 가중치는 $159/(159+46+113+82+26+31)$, 그룹수를 이용한 가중치는 $7/(7+2+5+3+2+2)$가 된다. 본 논문에서는 자료개수의 가중치를 이용한 중요도를 중요도1, 그룹 수의 가중치를 이용한 중요도를 중요도2로 나타낸다. Table 3.2는 Fish 자료에 대한 두 가지 사영추적분류 중요도와 랜덤 포레스트에서 계산된 변수중요도를 나타낸 표이다. 각 방법에 대하여 중요도 별로 순서화하여 나타내었다. Fish 자료의 경우 각 그룹마다 자료의 개수가 크게 다르지만 변수 중요도에 있어서 자료의 수를 이용한 경우와 그룹의 수를 이용한 경우는 크게 다르지 않다. 두 중요도 모두 length2와 length3이 가장 중요한 변수임을 나타내고 있다. 반면 랜덤 포레스트에서 계산된 변수 중요도는 height와 width가 가장 중요한 두 변수임을 나타내고 있다.

각 방법에서의 중요도 기준에 따라 선택된 변수들이 실제 분류에 어느 정도의 영향을 미치는지를 알아보기 위하여 각 방법의 중요도 순서에 따라 2개부터 6개까지 선택 변수의 수를 증가시켜 가면서 CART, 랜덤 포레스트, 판별분석, 그리고 사영추적분류나무를 이용하여 각각의 오분류율을 계산한 결과가 Table 3.3과 Table 3.4에 나타나 있다. Table 3.3은 사영추적분류나무의 중요도1의 순서에 따라 처음에는

Table 3.3. The performance of the importance measure of PPtree - the order of measures in PPtree

variables	CART	random forest	LDA	PPtree
length2, length3	0.308	0.358	0.145	0.157
length2, length3, length1	0.314	0.365	0.120	0.120
length2, length3, length1, weight	0.277	0.308	0.076	0.082
length2, length3, length1, weight, height	0.120	0.201	0.006	0.000
length2, length3, length1, weight, height, width	0.132	0.189	0.006	0.000

Table 3.4. The performance of the importance measure of PPtree - the order of measures in random forest

variables	CART	random forest	LDA	PPtree
height, weight	0.245	0.327	0.239	0.308
height, weight, length3	0.132	0.176	0.138	0.126
height, weight, length3, width	0.132	0.182	0.138	0.107
height, weight, length3, width, length2	0.132	0.195	0.019	0.006
height, weight, length3, width, length2, length1	0.132	0.195	0.006	0.000

length2와 length3만을 이용하여 분류분석을 시행하고 그 이후 length1, weight, height, width의 순으로 변수를 추가하면서 분류분석을 시행한 결과이다. 4가지 방법 모두 변수의 수가 증가함에 따라 오분류율이 낮아짐을 확인할 수 있다. 랜덤 포레스트의 경우 전반적으로 가장 낮은 성능을 보이고 있으며 판별분석은 사영추적분류나무와 비슷한 성능을 보이고 있다. 특히 2 ∼ 3개의 변수만을 이용한 경우 사영추적분류나무는 12%의 낮은 오분류율을 보이고 있다.

Table 3.4는 랜덤 포레스트의 중요도에 따라 변수를 추가하면서 분류분석을 시행한 결과이다. 처음 3개의 변수만을 사용한 경우 랜덤 포레스트의 오분류율은 17.6%로 사영추적분류나무의 중요도에 따라 선택한 3개의 변수만을 사용한 경우인 36.5%보다 훨씬 낮은 오분류율을 보이고 있으나 2개의 변수를 이용한 경우는 33%정도로 비슷한 정도의 오분류율을 보이고 있다. 처음 2개의 변수만을 사용한 경우 사영추적분류나무의 오분류율은 30.8%로 상당히 높게 나타나고 있으나 하나의 변수를 더 추가하여 3개의 변수를 이용한 경우에는 12.6%로 낮은 오분류율을 나타내고 있다.

랜덤 포레스트의 중요도에 따라 선택된 변수를 이용한 경우 사영추적분류나무의 중요도에 따라 선택된 변수를 이용한 경우보다 조금 더 높은 오분류율을 나타내고 있으나 랜덤 포레스트보다는 좋은 성능을 보이고 있다. 이를 통하여 사영추적분류나무의 중요도를 이용하여 선택된 변수들이 분류에 중요한 역할을 하고 있음을 확인함으로써 중요도 측도에 대한 성능을 파악할 수 있었다. 중요도1과 중요도2는 Fish 자료의 경우 큰 차이를 보이고 있지는 않지만 오분류율에 중심을 둔 분석의 경우에는 중요도1이 좀 더 유용하게 쓰인다. 본 논문에서는 이후의 논의에서 사영추적분류나무의 변수 중요도에 대한 측도로 중요도1을 사용한다.

3.3. 모의실험

사영추적분류나무의 변수 중요도와 랜덤 포레스트의 변수 중요도를 좀 더 심도있게 비교해보기 위하여 모의실험을 실시하였다. 표준정규분포로부터 생성된 서로 독립인 10개의 변수 $X1 \sim X10$을 이용하여 각각 50개씩의 관측값이 있는 그룹 A와 그룹 B를 생성하였다. 그룹 A는 처음 4개의 변수 $X1 \sim X4$를 각각 2, 1.5, 1, 0.5만큼 이동시키고 그룹 B는 변수 $X1 \sim X4$를 각각 -2, -1.5, -1, -0.5만큼 이동시켜 A와 B의 분류에 있어서 $X1$이 가장 중요한 변수가 되고, $X2$, $X3$, 그리고 $X4$가 순서대로 중요한 변수가 되도록 하였다. 이와같은 자료에서 사영추적분류나무와 랜덤 포레스트 방법이 $X1 \sim X4$ 변수의

Table 3.5. The result of simulation data with various correlation structures

	No corr		corr($X1, X2, X3$)		corr($X1, X2, X5$)		corr($X3, X4, X5$)	
	PPtree	RF	PPtree	RF	PPtree	RF	PPtree	RF
$X1$	1.14	1.04	1.01	1.02	1.01	1.01	2.60	1.03
$X2$	2.02	1.96	4.30	1.98	6.41	2.02	3.79	1.98
$X3$	3.11	3.00	1.99	3.08	2.82	2.97	1.54	2.99
$X4$	5.01	4.06	4.82	4.01	4.83	4.07	6.24	4.02
$X5$	7.35	7.44	7.21	7.57	3.03	7.37	2.14	5.22
$X6$	7.17	7.53	7.00	7.44	7.50	7.33	7.48	8.14
$X7$	7.44	7.23	7.30	7.44	7.35	7.55	7.67	8.07
$X8$	7.18	7.53	7.10	7.52	7.39	7.55	7.89	7.83
$X9$	7.24	7.27	6.91	7.32	7.16	7.81	7.66	7.60
$X10$	7.34	7.94	7.36	7.62	7.50	7.32	7.99	8.12
error rate(%)	0.19	0.90	0.64	2.99	0.58	1.91	0.06	0.88

중요도를 파악할 수 있는지를 알아보기 위하여 각 자료에서 중요도의 순위를 기록하였다. 이를 100번 반복하여 기록한 중요도 순위의 평균이 Table 3.5에 정리되어 있다. $X1$의 경우 평균이 1에 가까울수록 중요도를 잘 예측한 것이 되고, $X4$의 경우 평균이 4에 가까울수록 중요도를 잘 예측한 것이 된다. $X1 \sim X3$의 경우 사영추적분류나무와 랜덤 포레스트 모두 $1 \sim 3$에 가까운 평균을 보이고 있어 비슷한 성능을 보인다고 할 수 있다. 그러나 $X4$의 경우 랜덤 포레스트는 4에 가까운 평균을 보이는 반면 사영추적분류나무의 경우 5에 가까운 평균을 보여 랜덤 포레스트의 성능이 더 좋다고 할 수 있다.

변수들 간의 상관관계가 있는 경우의 성능을 비교하기 위하여 $X1$, $X2$, $X3$간의 상관계수가 각각 0.8이 되도록 하여 같은 실험을 반복하였다. 또한 분류에 영향을 주지 않는 변수와의 상관관계가 있는 경우의 성능을 비교하기 위하여 $X1$, $X2$, $X5$의 상관계수, 그리고 $X3$, $X4$, $X5$의 상관계수가 각각 0.8이 되도록 한 실험을 하였다. $X1$, $X2$, $X3$간의 상관관계가 있는 것은 중요 변수들간의 상관관계가 높은 경우를 의미하며 $X1$, $X2$, $X5$간의 상관관계가 있는 것은 가장 중요한 변수들과 중요하지 않은 변수와의 상관관계가 있는 경우를, $X3$, $X4$, $X5$간의 상관관계가 있는 것은 가장 중요하지는 않지만 어느정도 중요한 변수들과 중요하지 않은 변수와의 상관관계가 있는 경우를 보기 위한 것이다.

$X1$, $X2$, $X3$간의 상관관계가 있는 경우 사영추적분류나무의 경우 $X1$, $X3$이 높은 순위를 차지하여 중요한 변수로 여겨지고 있으며 $X2$와 $X4$는 상대적으로 낮은 순위를 보이고 있다. 랜덤 포레스트의 경우는 상관관계가 없는 경우와 큰 차이를 보이지 않고 있다. $X1$, $X2$, $X5$간의 상관관계가 있는 경우에도 랜덤 포레스트는 큰 차이를 보이고 있지 않으나 사영추적분류나무의 경우 그룹 A와 그룹 B의 차이를 나타내지 않는 $X5$ 변수가 $X2$나 $X4$ 변수보다 중요한 변수로 여겨지고 있다. 또한 $X3$, $X4$, $X5$간의 상관관계가 있는 경우에는 $X5$가 두번째로 중요한 변수로 나타나고 있다. 랜덤 포레스트는 여전히 상관관계가 없는 경우와 큰 차이를 보이고 있지 않다. 이는 랜덤 포레스트의 경우, 하나의 마디에서 하나의 변수만을 사용하게 되므로 변수간의 상관관계는 고려하지 않게 되므로 상관관계에 영향을 받지 않게 되는 것이다.

오분류율을 살펴보면 변수들간의 상관관계가 있는 경우의 상황을 좀 더 명확히 알 수 있다. 상관관계가 없는 경우나 $X1$, $X2$, $X3$간의 상관관계가 있는 경우, 그리고 $X1$, $X2$, $X5$의 상관관계가 있는 경우에는 랜덤 포레스트가 사영추적분류나무오분류율의 $3 \sim 5$배정도를 나타내고 있으나 $X3$, $X4$, $X5$의 상관관계가 있는 경우에는 14배 더 많은 오분류율을 보이고 있다. 이는 상관관계의 고려여부가 분류에 중요한 영향을 미치고 있으며 사영추적분류나무의 변수 중요도는 이를 반영한 결과라는 것을 알 수 있다 .

Table 4.1. Comparison of misclassification rates in random forest, CART and PPtree

dataset	자료수	변수 개수	그룹 개수	random forest	CART	PPtree
Glass	214	9	6	0.215	0.205	0.351
Wine	178	13	3	0.023	0.062	0.000
Cars	93	17	6	0.183	0.161	0.086
Image	2310	19	7	0.018	0.075	0.068
Parkinson	195	22	2	0.092	0.082	0.092
Australian-crab	200	5	4	0.095	0.075	0.040
Lymphoma	80	50	3	0.063	0.050	0.013
NCI60	61	30	8	0.426	0.377	0.049

4. 여러자료들의 분석비교

나무구조의 분류분석방법들의 성능비교에 많이 사용되고 있는 8개의 자료에 대하여 랜덤 포레스트, CART, 그리고 사영추적분류나무 분석을 시행하고 각 분석에서 분류를 위해 중요하게 사용된 변수들을 살펴보고 이를 비교, 분석해 보고자 한다. 8개의 자료에 대한 자료 수, 변수의 개수, 그리고 그룹 개수의 자료에 대한 기본 정보와 각 분류방법을 시행했을 때의 오분류율은 Table 4.1에 나타나 있다. Glass, Image, Parkinson의 세 자료에 대해서는 랜덤 포레스트 방법이 더 좋은 성능을 보이고 있으나 다른 5가지의 자료에 대해서는 사영추적분류나무가 더 좋은 성능을 보이고 있다. 특히 NCI60 자료의 경우 랜덤 포레스트와 CART 모두 40%에 가까운 오분류율을 보이고 있으나 사영추적분류나무는 5%로 아주 낮은 오분류율을 나타내고 있다.

Figure 4.1은 Table 3.3과 Table 3.4에서 보여주었던 것과 같은 방식으로 사영추적분류나무의 중요도에 따라 변수의 수를 증가시켜가면서 선택한 변수들을 여러 분류분석 방법들에 적용하여 계산한 오분류율을 각각의 자료에 대하여 그림으로 나타낸 것이다. 그림의 X축은 변수의 중요도 순서에 따라 선택되어 분류에 사용된 변수의 개수를 나타내며 Y축은 오분류율을 나타낸다. 실선은 사영추적분류나무의 오분류율을 나타내며 점선은 랜덤 포레스트, 그리고 굵은 점선은 CART의 오분류율을 나타낸다. 각 자료의 그림에서 왼쪽은 사영추적분류 중요도를 이용한 그림이고 오른쪽은 랜덤 포레스트 중요도를 이용한 그림이다.

전체적으로 변수의 수가 늘어남에 따라 오분류율아 낮아지는 경향을 보이고 있으며 Table 4.1의 결과와 마찬가지로 Glass, Image, Parkinson 자료의 경우 랜덤 포레스트 방법이 더 좋은 성능을 보이고 있다. Lymphoma와 NCI60 자료 (Dudoit 등, 2002) 의 경우 랜덤 포레스트 방법은 일정 오분류율 이하로는 내려가지 않는 경향을 보이고 있으나 사영추적분류나무의 경우 15개 이상의 변수를 이용한 경우부터는 거의 0에 가까운 오분류율을 보이고 있다. 사영추적분류나무의 변수 중요도 순서에 따라 변수를 증가시킨 경우와 랜덤 포레스트의 변수 중요도 순서에 따라 증가시킨 경우는 큰 차이를 보이지 않고 있다.

5. 결과 및 논의

본 논문에서는 사영추정분류나무와 나무구조의 분류분석 방법인 CART, 랜덤 포레스트 방법들을 이용하여 자료의 분류에 중요한 역할을 하는 변수를 파악하고 분석 방법별로 어떤 차이가 있는지를 확인하였다. 분류방법의 비교에 많이 이용되는 8개의 자료를 이용하여 사영추적분류나무, CART, 랜덤 포레스트 방법으로 분석한 후 오분류율을 비교하고 사영추적분류나무에서 계산된 변수에 대한 중요도(importance measure)에 대해서도 살펴보았다.

세 방법 모두가 비슷한 결과를 낸 자료도 있는 반면, 사영추적분류나무의 결과와 랜덤 포레스트의 결과

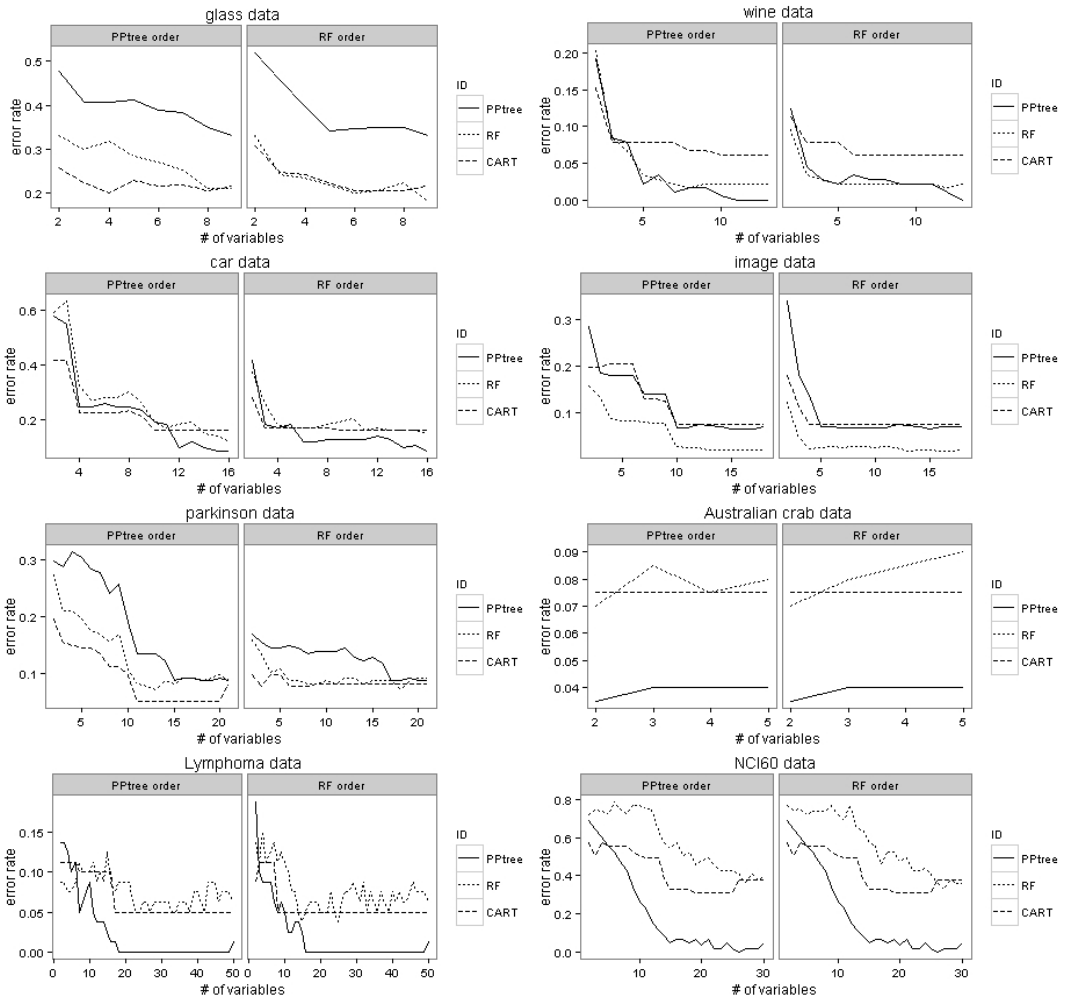

Figure 4.1. The performance of importance measures of PPtree with various data

가 다른 경우도 관찰되었다. Fish 자료에서 사영추적분류나무의 중요도 기준으로 세 변수를 선택한 경우 length1, length2, 그리고 length3이 선택이 되었고 이들만을 이용하여 분류를 한 경우 사영추적분류나무에서는 12%, CART에서는 31.4%, 그리고 랜덤 포레스트 방법에서는 36.5%의 오분류율을 나타내었다. 세 변수를 랜덤 포레스트의 중요도 기준으로 선택한 경우에는 height, weight, 그리고 length3으로 사영추적분류나무의 경우와는 다른 변수들이 선택되었고 사영추적분류나무에서는 12.6%, CART에서는 13.2%, 그리고 랜덤 포레스트 방법에서는 17.6%의 오분류율을 나타내었다. 랜덤 포레스트의 중요도 기준으로 선택한 경우 CART 와 랜덤 포레스트 방법 모두 이전보다는 훨씬 낮은 오분류율을 보이고 있으나 사영추적분류나무의 성능에는 미치지 못하고 있다.

이와같이 전혀 다른 변수들이 선택되었으나 사영추적분류나무의 성능에는 크게 영향을 미치지 못하고 있다. 이는 사영추적분류나무의 경우 각 마디에서 변수들 간의 선형결합을 이용함으로써 모든 변수들의 관계를 고려하게 되기 때문이다. Fish 자료의 경우 length1, length2, 그리고 length3들 간의 상관계수

는 모두 0.9이상으로 거의 직선에 가까운 관계를 가지고 있다. 즉, 이 세 변수들은 조금씩 다르기는 하지만 대부분 비슷한 정보를 가지고 있는 것이다. CART나 랜덤 포레스트 방법에서는 한번에 하나의 변수만을 선택하게 되므로 이 세 변수들 중 하나만을 선택하여 사용하게 되는 것이다. 그러나 사영추적분류나무의 경우 이들의 선형결합을 이용하여 모두를 이용하게 되므로 이들의 중요도가 높아지게 되고 또한 오분류율도 낮아지게 되는 것이다. 이와같이 변수들 간의 상관계수가 높아짐에 따라 사영추적분류나무의 중요도와 랜덤 포레스트 방법의 중요도의 차이가 커진다. 이와같이 중요변수들 간의 상관관계가 높은 경우에는 사영추적분류나무가 예측에서 조금 더 좋은 성능을 보이는 경향이 있으며 이들 상관관계는 사영추적분류나무의 변수의 중요도를 측정하는데에도 반영되고 있다. 이는 모의실험을 통하여도 확인하였다.

본 논문에서 살펴본 바와 같이 사영추적분류나무의 변수 중요도는 분류를 위한 변수선택에 유용하게 쓰일 수 있다. 또한 각 마디에서 제공되는 사영계수들은 각 마디에서 분류되는 그룹들에 대한 정보를 가지고 있으므로 종합적으로 측정되는 중요도 이외에 좀 더 부분적이고 정확한 정보를 제공함으로써 좀 더 정확한 분류를 위한 정보들을 수집할 수 있게 된다. 본 연구에서는 자료의 개수가 변수의 수에 비해 상대적으로 큰 자료만을 다루었으므로 사영추적분류나무에서 모두 LDA 지수를 이용하였다. 유전자 자료와 같은 변수의 개수가 자료의 개수에 비해 큰 자료의 경우 PDA 지수를 이용하여 사영추적분류나무를 구성하면 좀 더 정확한 결과를 얻을 수 있다.

현재 개발되어 있는 사영추적분류나무의 경우 LDA를 기본으로 하고 있으므로 설명변수가 모두 연속형임을 가정하고 있다. 범주형 설명변수가 포함된 경우에는 가변수 형태로 변형하여 사영추적분류나무를 이용할 수 있으나 권장되지는 않는다. 범주형 설명변수만을 위한 사영추적 방법 (Caussinusand와 Ruiz-Gazen, 2006)은 개발되어 있으나 범주형과 연속형의 변수가 함께 있는 자료에 대한 연구는 많이 진행되어 있지 않다. 범주형 변수를 위한 사영추적과 연속형 변수를 위한 사영추적의 결과를 결합하여 이용하는 사영추적지수를 개발하여 사영추적분류나무에 적용하는 것은 앞으로의 좋은 연구주제가 될것으로 생각된다.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Belmont: Wadsworth.

Breiman, L. and Cutler, A. (2012). RandomForest: Breiman and Cutler's random forests for classification and regression, Available from http://cran.r-project.org/web/packages/randomForest/index.html.

Caussinusand, H. and Ruiz-Gazen, A. (2006). Projection-pursuit approach for categorical data, *Multiple Correspondence Analysis and Related Methods* (eds. Greenacre, M. and Blasius, J.), Chapman and Hall/CRC, 405–418.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77–87

Friedman, J. and Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers*, **23**, 881–890.

Kruskal, J. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new index of condensation, *Statistical Computing*, New York; Academic Press, 427–440.

Lee, E., Cook, D., Klinke, E. and Lumley, T. (2005). Projection pursuit for exploratory supervised classification, *Journal of Computational and Graphical Statistics*, **14**, 831–846.

Lee, E. and Cook, D. (2010). A projection pursuit index for large p small n data, *Statistics and Computing*, **20**, 381–392.

Lee, Y., Cook, D., Park, J. and Lee, E. (2013). PPtree: Projection pursuit classification tree, *Electronic Journal of Statistics*, **7**, 1369–1386.

나무구조의 분류분석에서 변수 중요도에 대한 고찰

김나영[a] · 이은경[a,1]

[a]이화여자대학교 통계학과

(2014년 7월 11일 접수, 2014년 9월 17일 수정, 2014년 9월 29일 채택)

요 약

본 연구에서는 나무구조의 분류분석에서 자료의 크기가 방대해짐에 따라 중요한 문제로 대두되고 있는 변수의 중요도에 대하여 사영추적분류나무를 중심으로 고찰하였다. 사영추적분류나무(projection pursuit classification tree)는 각 마디에서 사영추적을 이용하여 그룹을 잘 분리하는 변수들의 선형결합을 이용하는 방법으로 이때 사용되는 사영계수들은 각 마디에서의 분류에 대한 정보를 가지고 있다. 이를 종합하여 각 변수의 분류에 대한 중요도를 계산할 수 있다. 먼저 사영추적분류나무의 분류과정에서 계산되는 사영추적계수를 이용하여 분류를 위한 변수선택의 중요도를 계산하고 이들의 특성을 살펴보고 이를 같은 형태의 나무모형방법인 CART와 랜덤 포레스트의 결과와 비교 분석하여 사영추적분류나무의 특성을 살펴보고 비교, 분석하였다. 대부분의 자료에서 사영추적분류나무가 훨씬 좋은 성능을 보이고 있었으며 특히 상관계수가 높은 변수들이 포함되어 있는 경우에는 상대적으로 적은 수의 변수로도 잘 분류를 할 수 있음을 확인하였다. 랜덤 포레스트에서 제공하는 변수 중요도는 변수들 간의 상관관계가 높은 경우에는 사영추적분류나무의 변수중요도와 매우 다르게 나타나며 사영추적분류나무의 변수 중요도가 조금 더 나은 성능을 보이고 있음을 알 수 있다.

주요용어: 분류분석, 나무모형, 변수선택, 사영추적.

이 논문은 2009년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2009-0093827).

[1](120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: lee.eunk@ewha.ac.kr