碩士學位論文

앙상블 기법을 이용한 커널 능형 로지스틱 회귀분류법의 효율성에 관한 연구

A study on efficiency of kernel ridge logistic regression classification using ensemble method

指導 李 碩 浩 敎授

韓國外國語大學校 大學院
統 計 學 科
黃 星 潤

碩士學位論文

앙상블 기법을 이용한 커널 능형 로지스틱 회귀분류법의 효율성에 관한 연구

A study on efficiency of kernel ridge logistic regression classification using ensemble method

指導 李 碩 浩 敎授

이 論文을 碩士學位請求論文으로 提出합니다.

2016年 10月

韓國外國語大學校 大學院
統 計 學 科
黃 星 潤

이 論文을 黃星潤의 碩士學位論文으로 認定함.

2016年 月 日

審 査 委 員 鄭碩午 (인)

審 査 委 員 이성은 (인)

審 査 委 員 양승은 (인)

韓國外國語大學校 大學院

감사의 글

진로에 대한 고민과 방황 끝에 시작했던 2년간의 대학원 석사과정 생활이 어느덧 본 논문의 완성과 함께 종점을 향해가고 있다는 것이 믿기지 않습니다. 본 논문이 완성되기까지 정말 많은 분들의 도움과 조력을 받았습니다. 그 분들의 도움이 있었기에 지금 이 자리까지 올 수 있었고 막연한 꿈이기만 했던 석사학위까지 취득하게 되었다고 생각합니다. 마지막 학기가 끝나가는 이 시점에 일일이 직접 찾아뵙고 인사드리는 것이 예의이겠지만 여건 상 그렇게 하기가 쉽지 않다는 것이 아쉬울 따름입니다. 대신 본 감사의 글을 통해서 그동안 도움을 주신 모든 분들에 대한 고마움과 감사함을 표하고자 합니다.

연구실 생활을 하면서 듣게 된 여러 가지 조언들 중 지도교수님, 그리고 부모님께서 지금 연구하고 보내는 시간이 결코 헛된 시간이 아니고 앞으로 제 자신에게 주어지는 많은 일들을 유연하게 대처하는 능력을 길렀으면 좋겠다고 말씀하셨던 것이 가장 기억에 남습니다. 무엇을 해야 할지 갈피를 잡지 못하고 있던 제 자신에게 가장 와 닿았던 조언이었고 결과적으로 석사과정 생활을 무사히 마칠 수 있도록 해 준 버팀목이 되었습니다. 그리고 연구실 생활을 통해서 인내력, 책임감, 동료들과 갈등 없이 의사소통하는 방법, 조직사회에 대한 적응력, 그리고 자신이 해야 할 공부를 스스로 찾아서 하는 능력 등 인생을 살아가는 데 있어 반드시 필요한 여러 자세들을 익힐 수 있었습니다. 이를 통해 나약하기만 했던 제 자신의 태도를 180°바꾸게 되었고 제 2의 삶을 시작하는 계기도 마련할 수 있었습니다.

돌이켜보니 본 논문을 작성하는 동안 제 인생의 큰 작품들 중 하나를

만든다는 생각을 가지고 모든 열정을 다 쏟았던 것 같습니다. 그렇게 작성했던 논문을 마무리함과 동시에 제가 몸담았던 연구실 자리를 정리하면서 이제야 마음의 짐을 내려놓는다는 생각을 하고 있으니 만감이 교차합니다. 논문을 처음 시작할 때는 그저 막막하기만 했었는데 어느덧 논문을 발표해야 하는 시점을 맞이하게 되니 아쉬움이 많이 남습니다.

이 자리에 서기까지 인내와 사랑으로 보살펴주신 부모님께 진심으로 감사드립니다. 저답지 않게 철없이 행동해서 걱정만 끼쳐드린 것 같아 마음이 편치 않습니다. 제가 잘못된 길로 가려고 할 때마다 항상 바른 길로 인도해주셨기 때문에 지금의 제가 있는 것이라고 생각합니다. 믿고 지켜봐주신 만큼 이제는 제가 든든한 기둥이 되어서 지켜드리겠습니다. 이제 사회초년생으로서 부모님께서 항상 자랑스러워할 수 있는 외동아들이 되겠습니다.

그리고 본 석사논문이 완성될 때까지 아낌없는 지도와 함께 연구자로서 가져야 할 자세에 대한 여러 조언을 해주셨던 이석호 교수님, 저에게는 미지의 영역이었던 통계학이라는 학문의 세계에 눈을 뜨게 해주신 통계학과 교수님들께도 깊이 감사드립니다. 결코 짧지만은 않았던 긴 시간 동안 통계학이라는 학문과 씨름하면서 여러 장애물들을 뛰어넘어야 했습니다. 그동안 많은 도움을 주셨기에 그 장애물들을 뛰어넘고 이 자리에 설 수 있었다고 생각합니다.

같은 연구실 내에서 한솥밥을 먹으며 생활하면서 많은 도움을 준 윤미씨, 주영이형, 효진이, 그리고 선우한테도 고맙다는 말 해주고 싶습니다. 처음 연구실 생활을 시작할 때부터 여러 조언도 해주고 힘들 때마다 친형제처럼 이끌어줘서 정말 고맙습니다. 특히 갑작스런 요청이었음에도 불구하고 본인의 석사논문을 기꺼이 참고문헌으로 허락해준 선우한테 더

감사할 따름입니다. 훗날 모두 이루고자 하는 일들 다 성취하고 만날 수 있었으면 좋겠습니다. 그리고 이제 막 대학원 석사과정 생활을 시작하게 된 저의 첫 연구실 후배인 선화에게 선배로서 잘 챙겨주지 못한 것 같아서 미안한 마음이 듭니다. 결코 쉽지 않은 과정이겠지만 본인 나름대로의 뜻을 가지고 그 과정에 발을 들여놓은 만큼 잘 적응해나갔으면 좋겠습니다.

대학원 영어시험을 준비할 때 많은 도움을 주신 이지희 선생님께도 감사의 인사를 전합니다. 사실 남들보다 영어실력이 뛰어난 편이 아니었기 때문에 이 자격시험에 의해 학위과정 중 발목이 잡히지는 않을 지 많이 걱정했었습니다. 직접 인사를 드리고 싶었는데 시간이 허락하지 않은 것이 조금 아쉽습니다. 앞으로도 하시는 일 잘 이루셨으면 좋겠습니다.

이 외에도 제가 대학원에서 석사과정을 마칠 때까지 물심양면으로 도움을 주신 모든 분들이 이 글 안에 모두 나열할 수 없을 정도로 너무나 많습니다. 그 분들께 감사의 인사를 이 글을 통해서나마 전하고 싶습니다. 그런 도움이 있었기에 무사히 학위과정을 마칠 수 있었다고 생각합니다. 현재 원하는 분야로의 취업 이후 MBA 과정 진학을 준비하고 있습니다. 항상 응원해주시고 믿어주시는 분들에게 부끄럽지 않은 사람이 되도록 하겠습니다.

본 글을 마치기 전에 인생을 살아가는 데 있어서 꼭 필요한 '10-10-10 법칙' 에 관하여 잠깐 언급하고자 합니다. 어떤 일을 선택해야하는 상황이 생길 경우 과연 10일 뒤, 10개월 뒤, 그리고 10년 뒤에 내 자신이 어떻게 될 지를 생각해보고 항상 심사숙고해서 결정하라는 의미를 담고 있는 법칙입니다. 평생을 배우며 살아가야 하는 미완성의 인격체인 사람들에게 던지는 뼈 있는 말이라고 생각합니다. 이제 모

든 일에 스스로 책임을 져야하는 만큼 앞으로 겪게 될 수많은 선택의 기로 앞에서 이 법칙이 던지는 메시지를 꼭 한번 되새길 생각입니다.

마지막으로 바쁘신 와중에도 본 석사논문 심사를 위해서 참석해주신 교수님들께 깊이 감사드립니다. 그리고 사랑으로 보살펴주신 부모님께도 다시 한 번 감사드립니다. 끝으로 제가 재미있게 본 드라마에서 나왔던 명대사 한 구절을 남기고자 합니다. 저의 평소 신조인 자신이 해야 할 일에 대한 책임감과 신념이 묻어나는 말이어서 개인적으로 마음에 들었던 대사입니다. 감사합니다.

2016년 10월 통계학과 대학원 조교 연구실에서

아랍군 : *Look, Captain.*

You'd better know exactly what you're doing.

유시진 대위 : *You do your job. The doctor will save a patient.*

And I will protect what I have to.

– KBS 드라마 ‘태양의 후예’ 中

요 약

본 연구는 분류문제에 자주 사용되는 로지스틱 회귀분류법(logistic regression classification)에 커널기법(kernel trick)과 능형 회귀분석(ridge regression) 방법을 적용한 커널 능형 로지스틱 회귀분류법(kernel ridge logistic regression classification)에 관한 것이다. 그리고 이 분류법에 배깅(bagging) 및 랜덤포레스트(random forests)와 같은 앙상블 기법(ensemble method)을 적용하여 추정량의 분산을 줄임으로써 분류의 정확성과 신뢰성을 높이는 방법을 제안하고자 한다. 컴퓨터 모의실험을 통하여 다양한 상황에서의 분류문제에서 일반적인 커널 능형 로지스틱 회귀분류법과 본 연구에서 제안하는 방법을 비교, 분석하였다. 이를 통해 본 연구에서 제안하고자 하는 방법이 일반적인 방법론보다 우수한 성능을 보임을 확인하였다.

목 차

표 목 차

그 림 목 차

1 서 론

1.1 연구의 배경 및 목적

최근 들어 인터넷 등의 IT 기술의 발달로 인하여 스미싱, 보이스피싱 등의 신종사기 범죄가 증가하고 있다. 이에 따라 우리나라를 포함한 전 세계의 수많은 기업들이 범죄로 인한 막대한 피해를 사전에 예방하기 위하여 여러 가지 다양한 대책들을 세우고 있으며 이를 통해 고객들의 기업에 대한 신뢰도를 보다 향상시키기 위해 노력하고 있다. 특히 고객에 대한 신용도 예측은 많은 기업들이 사용하고 있는 방법들 중의 하나이며 이를 위해 지지도 벡터 기계 기법(Support Vector Machine : SVM), K-근접 이웃 기법(K-Nearest Neighbor : KNN), 가중치 K-근접 이웃 기법(Weighted K-Nearest Neighbor : KKNN), 선형 판별분석(Linear Discriminant Analysis : LDA), 이차형 판별분석(Quadratic Discriminant Analysis : QDA), 나무기반모형(Tree-based Model) 등의 다양한 통계적인 분류(classification)법들이 제안되고 있다. 이 분류법들 중 대표적으로 많이 쓰이고 있는 방법들 중의 하나가 바로 로지스틱 회귀분류법(Logistic Regression Classification)이며 실제로 많은 기업들이 빅데이터 분석을 통해서 고객에 대한 신용도를 평가하고자 하는

경우에 이 방법을 자주 사용하고 있다. 기업의 입장에서 봤을 때 고객은 보호해야 할 큰 자산인 셈이다. 이 때문에 고객들의 기업에 대한 신뢰도는 기업의 생존문제와 큰 관련이 있을 수밖에 없다. 고객들의 불신을 얻게 되면 그 기업은 성장할 수 없게 된다. 그만큼 신뢰도라는 요인이 기업에게는 중요한 것이다.

하지만 정보의 홍수 속에서 등장하는 대다수의 데이터들은 다루기 쉽도록 정형화(standardization)되어 있지 않은 것이 사실이며 관측치의 개수와 설명변수의 개수 또한 증가하고 있는 추세이기 때문에 단순한 방법론만으로는 의미 있는 분석결과를 얻기 어려운 것이 사실이다. 이러한 점들을 극복하기 위해서는 데이터의 비선형(non-linear)성에 따른 변수변환(variable transformation)이나 설명변수들 간의 상호작용(interaction effect) 등을 고려해야 한다. 하지만 이에 따라 설명변수들 간의 높은 연관성으로 인하여 나타나게 되는 다중공선성(multicollinearity problem)의 문제가 생길 가능성이 높아지게 된다. 다중공선성의 문제가 발생하게 되면 추정량의 분산이 과도하게 커지게 되어 안정성이 줄어들게 되며 분석에 대한 신뢰성이 떨어지는 문제를 낳게 된다. 따라서 본 논문에서는 이와 같은 문제점을 보완할 수 있는 방법을 제안하고자 하고 이에 관한 연구를 수행하였다.

1.2 연구방법 및 구성

본 연구에서는 분류문제에 자주 사용되는 로지스틱 회귀분류법(logistic regression classification)에 커널기법(kernel trick)과 능형 회귀분석(ridge regression) 방법을 적용한 커널 능형 로지스틱 회귀분류법(kernel ridge logistic regression classification)에 관하여 다룬다. 커널기법은 설명변수의 변환이 필요한 경우 변환함수를 미리 가정하지 않고 커널을 이용하여 이를 자동으로 해결할 수 있는 변환기법으로 통계적 기계학습(statistical machine learning)에서 자주 다루어지는 기법이다. 이 기법을 이용하게 되면 데이터 분석 시 변환을 위해 설정해야 하는 변환함수(transformation function)를 미리 가정하지 않더라도 주어진 데이터를 기반으로 하여 자동으로 변환이 이루어지게 할 수 있다. 이 때문에 커널기법을 적용한 커널 능형 로지스틱 회귀분류법은 일반적인 로지스틱 회귀분류법에 비해 비교적 정확하게 데이터의 분류를 수행해내는 특성을 가지고 있으며 최근 들어 분류와 관련된 문제에서 자주 언급되고 있는 추세이다. 또한 설명변수들 간의 연관성에 의해서 나타나는 다중공선성의 문제를 해결하고자 앙상블 기법(ensemble method)을 적용한다. 본 연구에서는 앙상블 기법에 해당하는 방법들 중 동일한 분포적인 성질을 가지는 다수의 서브모형(submodels)들에 의한 결과들을 평균하여 분산이 작아지는 효과를 얻을 수 있는 배깅(bagging), 그리고 모

든 설명변수들 중 일부만을 선택하여 다수의 서브모형들을 생성하고 그 결과들을 평균하여 연관성을 줄일 수 있는 랜덤포레스트(random forests)를 사용한다. 이러한 2가지 방법을 사용하여 커널 능형 로지스틱 회귀분류법의 분류정확도를 높이는 방안을 제시한다.

본 논문은 총 6장으로 구성되어 있다. 2장에서는 커널 능형 로지스틱 회귀분류법에 관하여 소개하며 3장에서는 앙상블 기법을 언급한다. 4장에서는 커널 능형 로지스틱 회귀분류법에 앙상블 기법을 적용하는 과정을 설명하며 5장에서는 제안하고자 하는 분석방법의 효율성을 입증하기 위해 실시한 모의실험과 실증분석, 그리고 이에 따른 결과를 설명한다. 마지막으로 6장에서는 결론으로 본 연구의 결과를 요약하고 마무리한다.

2 커널 능형 로지스틱 회귀분류법

2.1 능형 회귀분석 (ridge regression)

다중회귀분석(multiple regression analysis)은 하나의 반응변수(response variable)가 여러 종류의 설명변수(explanatory variable)들과 선형적(linear)인 관계를 나타내고 있다고 여겨질 때 그 관계를 모형화(modeling)하여 분석하는 통계적인 분석방법이며 실제적으로 경제학, 사회학 등의 여러 분야에서 다양하게 사용되고 있다(한선우, 2016). 예를 들어 관측치의 개수가 n이고 설명변수의 개수가 k인 데이터 $(y_1,x_1),(y_2,x_2),\cdots,(y_n,x_n)$ 이 있다고 가정해보자. 이러한 경우에 다중회귀분석에서는 다음과 같은 회귀모형을 가정하게 된다. 물론 이 회귀모형은 데이터 내에 있는 반응변수와 설명변수들 사이에 선형적인 관계가 존재한다는 가정 하에 만들어진 것이다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \ , \ i = 1,2,\cdots,n$$

여기에서 ϵ_i는 오차(error)로써 보통 평균이 0이고 분산이 σ^2인 동일한 정규분포(normal distribution)를 따르면서 서로 독립(identical and independently distributed)이라고 가정하게 된다. 위에 있는 회귀모형식을 다음과 같이 행렬(matrix)의 형태로 바꿔서 쓸 수 있다.

$$y = X\beta + \epsilon \ , \ \epsilon \sim N_n(0, \sigma^2 I_n)$$

여기에서 $y = (y_1, y_2, \cdots, y_n)^T$는 길이가 n인 반응변수로 이루어진 벡터이고 $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$는 길이가 n인 오차벡터를 나타낸다. 그리고 $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$는 길이가 $p(=k+1)$인 회귀계수 벡터이며

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

는 크기가 $n \times p$인 자료행렬이다. 또한 0은 길이가 n인 영벡터이며 I_n은 크기가 $n \times n$인 단위행렬이다. 회귀계수벡터의 추정량 $\hat{\beta}$은 보통 최소제곱추정법(least squares estimation : LSE)을 이용하여 찾게 된다. 이 방법을 적용하게 되면 다음과 같은 정규방정식(normal equation)을 얻을 수 있다.

$$\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta) \\ \frac{\partial Q}{\partial \beta} &= -2X^T (y - X\beta) = 0 \\ (X^T X)\hat{\beta} &= X^T y \end{aligned}$$

만약 행렬 X가 완전계수(full-rank)의 성질을 만족하는 경우, 다시 말해서 행렬 $X^T X$의 역행렬(inverse matrix)이 존재하는 경우 β의 추정량 $\hat{\beta}$은 다음과 같이 하나의 해로 결정된다. 그리고

이 추정량은 β에 대한 비편향추정량(unbiased estimator)이다.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
$$E(\hat{\beta}) = \beta \ (\because y \sim N_n(X\beta, \sigma^2 I_n))$$

능형 회귀분석(ridge regression) 또는 Tikhonov regularization은 러시아의 수학자 Andrey Nikolayevich Tikhonov(1906~1993)에 의하여 제안된 방법으로 라소 회귀분석(lasso regression)과 함께 데이터 분석 시 다중회귀분석을 적용하고자 할 때 다중공선성의 문제가 있다고 여겨지거나 관측치의 개수보다 설명변수의 개수가 더 많아서 추정량이 유일하게 결정되지 않는 경우에 자주 사용하는 방법이다. 이 방법의 핵심은 추정량을 구하는 데 있어 어느 정도의 편의(bias)를 허용하는 대신 분산(variance)을 큰 폭으로 줄이는 방식을 취한다는 점에 있다. 이를 통해 좀 더 신뢰성이 높은 추정량을 얻을 수 있게 된다. 추정량은 다음과 같은 방식으로 양의 실수 λ가 곱해진 이차형식 형태의 벌점함수(penalty function) $\lambda \| \beta \|^2 = \lambda \beta^T \beta$를 적용하여 구해지게 된다.

$$\hat{\beta} = \mathrm{argmin}_\beta \{ \| y - X\beta \|^2 + \lambda \| \beta \|^2 \}$$
$$= \mathrm{argmin}_\beta \{ (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \} \ , \ (\lambda > 0)$$

여기에서 λ는 능형모수(ridge parameter)로써 만약 이 값이 0에 가깝게 되면 추정량의 편의는 0에 가깝게 되지만 분산은 커지게 된다. 하지만 이 값이 증가하게 되면 편의가 커지는 대신 분산이 줄어들게 된다. 그러므로 능형 회귀분석에서는 적절한 모수 λ값을

선정하는 것이 중요한 문제라고 할 수 있는데 보통 CV(cross validation)를 이용하여 test MSE(mean squared error)의 추정값을 가장 작게 만들어주는 λ값을 최적의 조건으로 판단하고 선택하게 된다. 이 분석방법에 의한 β의 추정량 $\hat{\beta}$은 다음과 같이 주어지게 된다.

$$Q^{*} = (y - X\beta)^{T}(y - X\beta) + \lambda\beta^{T}\beta$$
$$\frac{\partial Q^{*}}{\partial \beta} = -2X^{T}(y - X\beta) + 2\lambda\beta = 0$$
$$(X^{T}X + \lambda I_{n})\hat{\beta} = X^{T}y$$
$$\hat{\beta} = (X^{T}X + \lambda I_{n})^{-1}X^{T}y$$

위의 식에서 행렬 $X^{T}X + \lambda I_{n}$ 은 능형모수 λ에 의해서 반드시 역행렬을 가지게 되므로 추정량 $\hat{\beta}$은 λ값에 따라 유일하게 하나의 값으로 주어지게 된다.

2.2 로지스틱 회귀분류법 (logistic regression classification)

로지스틱 회귀분류법(logistic regression classification)은 분류문제에서 자주 다루어지는 통계적인 분석방법으로 1942년부터 1944년 사이에 특정한 사건이 발생할 확률(probability)을 모형화(modeling)하는 것과 관련된 일반화선형모형(Generalized

Linear Model : GLM)들 중 프로빗 링크(probit link), 보완적 로그-로그 링크(complementary log-log link), 로짓 링크(logit link)를 사용하는 모형에 대한 연구로부터 제안되었으며 이 후 1958년에 D. R. Cox 에 의해서 재정립되었다. 이 분류법은 3가지 이상의 집단을 분류하는 경우에도 사용이 가능하며 비교적 해석이 쉽다는 장점을 가지고 있다. 본 연구에서는 2개의 집단을 분류하는 상황과 관련한 경우만을 다루게 되므로 이에 해당하는 모형에 대해서만 언급하도록 하겠다. 로지스틱 회귀분류법에 대한 자세한 내용은 McCullagh and Nelder (1989)를 참조하기 바란다. 참고로 위에서 언급한 3가지 링크의 형태는 다음과 같다.

probit link :

$$\Phi^{-1}(p_i)\,,\ \Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}dt=P(Z\le x)\,,\ Z\sim N(0,1^2)$$

complementary log-log link : $\log\{-\log(1-p_i)\}$

logit link : $\log\left(\frac{p_i}{1-p_i}\right)$

예를 들어 분석해야 할 데이터에 포함된 반응변수가 이항변수(binary variable)로써 0 또는 1의 2가지 값으로만 이루어져 있다고 가정해보자. 이러한 경우에 2.1절에서 언급했던 다중회귀분석방법을 직접 적용하게 된다면 반응변수의 예측값이 0 또는 1

이외의 값으로 나타나는 관측치가 나올 수 있기 때문에 문제가 발생할 수 있다. 이러한 문제점을 보완하기 위해 2개의 집단을 분류해야 하는 경우 로지스틱 회귀분류법에서는 다음과 같은 회귀모형을 제안하고 있다. 물론 이 회귀모형도 다중회귀분석에서의 경우와 마찬가지로 데이터 내에 있는 반응변수와 설명변수들 사이에 선형적인 관계가 존재한다는 가정 하에 만들어진 것이다.

$$\log\left(\frac{p(x_i;\beta)}{1-p(x_i;\beta)}\right)= x_i^T\beta \ , \ i=1, \cdots, n$$
$$(\ 0 < p(x_i;\beta) < 1 \ )$$

위의 식에서 좌변에 있는 수식을 보통 로짓(logit)이라고 부른다. 그리고 $p(x_i;\beta)$는 반응변수의 실제 값이 1일 확률로써 보통 이 값이 0.5보다 크거나 같으면 반응변수 y_i의 예측값은 1로 주어지게 되고 그렇지 않은 경우에는 반응변수 y_i의 예측값이 0으로 주어지게 된다. 또한 $p(x_i;\beta)$는 0과 1 사이의 값을 취하기 때문에 로짓의 값의 범위는 실수 전체임을 알 수 있다. 즉, 설명변수들을 이용하여 실수 전체의 범위를 가지는 로짓에 대한 추정량을 먼저 구하고 이를 통해서 확률의 추정량을 구하는 방법을 적용하기 때문에 다중회귀분석방법의 문제점을 보완할 수 있는 것이다. 위에 있는 회귀모형식을 확률에 관하여 다시 정리해서 표현하면 다음과 같이 나타낼 수 있다.

$$p(x_i;\beta) = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)} \ , \ 0 < p(x_i;\beta) < 1$$

다시 말해서 2개의 집단을 분류하는 것과 관련한 로지스틱 회귀분류법에서는 반응변수 y_i의 예측값 $\hat{y_i}$을 직접 구하는 대신 반응변수의 실제 값이 1일 확률에 대한 추정량 $\hat{p}(x_i;\hat{\beta})$을 구해서 이 값이 0.5보다 크거나 같은 경우에는 $\hat{y_i}=1$, 그렇지 않으면 $\hat{y_i}=0$ 으로 정하게 된다.

확률에 대한 추정량을 구하기 위해서는 회귀계수벡터의 추정량인 $\hat{\beta}$을 구해야 한다. 로지스틱 회귀분류법에서는 설명변수 x_i가 주어졌을 때 반응변수 y_i의 조건부분포(conditional distribution)가 다항분포(multinomial distribution)의 형태를 만족한다는 가정하에 최대가능도추정법(maximum likelihood estimation : MLE)을 통해서 추정량 $\hat{\beta}$을 구하게 된다. 여기에서 반응변수 y_i는 독립적으로 추출된다고 가정한다. 만약 본 연구에서처럼 반응변수가 2가지의 값만을 나타내는 경우에는 베르누이분포(Bernoulli distribution)를 적용하면 될 것이다. 이를 식으로 표현해보면 다음과 같다.

$$y_i|x_i \sim Bernoulli[p(x_i;\beta)] \quad , \quad i=1,2,\cdots,n$$

위와 같은 가정에 의하여 가능도함수(likelihood function) $L(\beta)$는 다음과 같이 표현된다.

$$L(\beta)=\prod_{i=1}^{n}\{p(x_i;\beta)\}^{y_i}\{1-p(x_i;\beta)\}^{1-y_i}$$

그리고 이 식의 양변에 로그를 취함으로써 다음과 같은 로그가능

도함수(log-likelihood function) $l(\beta)$를 얻을 수 있다.

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n}\{y_i \log(p(x_i;\beta)) + (1-y_i)\log(1-p(x_i;\beta))\}$$
$$= \sum_{i=1}^{n}\{y_i(x_i^T\beta) - \log(1+\exp(x_i^T\beta))\}$$

추정량 $\hat{\beta}$은 로그가능도함수 $l(\beta)$를 최대화하는 방법을 통해서 구하게 된다. 이를 위해 함수 $l(\beta)$를 β에 관하여 미분하게 되면 다음과 같은 스코어방정식(score equation)을 얻을 수 있다.

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i\{y_i - p(x_i;\beta)\} = 0$$

이 방정식의 해는 뉴튼-랩슨 방법(Newton-Raphson method)을 적용한 반복적 재가중치 최소제곱 추정법(iteratively reweighted least squares estimation)을 통하여 이차형식 형태(quadratic form)의 함수를 최소화하는 문제로 근사시켜서 수치적(numerical)으로 구하게 되는데 다음과 같은 형태의 기울기 벡터(Gradient vector)와 헤시안 행렬(Hessian matrix)을 이용하는 알고리즘(algorithm)을 반복하여 $\hat{\beta}$의 값을 계산하게 된다. 이 과정을 통해서 찾은 $\hat{\beta}$의 값을 대입하여 얻을 수 있는 $l(\hat{\beta})$의 값을 이용해서 이 값이 정해진 수렴조건(convergence criterion)을 만족하게 되는 시점에서 얻어지게 되는 $\hat{\beta}$의 값을 벡터 β에 대한 최종적인 추정량으로 사용하게 된다.

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i \{y_i - p(x_i;\beta)\} \quad : \text{Gradient vector}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} x_i x_i^T \{p(x_i;\beta)\}\{1-p(x_i;\beta)\} \quad : \text{Hessian matrix}$$

$$\begin{aligned}\hat{\beta} &= \beta_{old} - \left[\frac{\partial^2 l(\beta_{old})}{\partial \beta_{old} \partial \beta_{old}^T}\right]^{-1}\left[\frac{\partial l(\beta_{old})}{\partial \beta_{old}}\right] \\ &= \beta_{old} + (X^T W X)^{-1} X^T (y-p)\end{aligned}$$

$$\begin{aligned}\hat{\beta} &= \text{argmax}_\beta \{l(\beta)\} \\ &= \text{argmin}_\beta (z - X\beta)^T W (z - X\beta) \\ l^*(\beta) &= (z - X\beta)^T W (z - X\beta) \\ \frac{\partial l^*(\beta)}{\partial \beta} &= -2X^T W (z - X\beta) = 0 \\ \hat{\beta} &= \beta_{old} + (X^T W X)^{-1} X^T (y-p) \\ &= (X^T W X)^{-1} X^T W z \\ z &= X\beta_{old} + W^{-1}(y-p)\end{aligned}$$

만약 행렬 $X^T WX$의 역행렬이 존재하는 경우 위와 같은 과정을 통해서 추정량을 구할 수 있다. 결과적으로 추정량 $\hat{\beta}$은 이차형식의 함수 $l^*(\beta)$의 값을 최소로 만들어주는 β의 값이라고 말할 수 있는 것이다. 여기에서 β_{old}는 $\hat{\beta}$이 구해지기 바로 이전 시점에서 구한 추정량을 의미한다. 그리고 p는 i번째 원소가 $p(x_i;\beta_{old})$인 길이가 n인 확률벡터이며 W는 i번째 대각원소가 $\{p(x_i;\beta_{old})\}\{1-p(x_i;\beta_{old})\}$이고 나머지 원소들은 모두 0인 크기 $n \times n$의 대각행렬(diagonal matrix)을 나타낸다. 본 연구에서는

다음과 같은 수렴조건(convergence criterion)을 사용하였다.

$$\left| l(\hat{\beta}) - l(\beta_{old}) \right| < 10^{-4} \quad \text{or} \quad r = 100$$

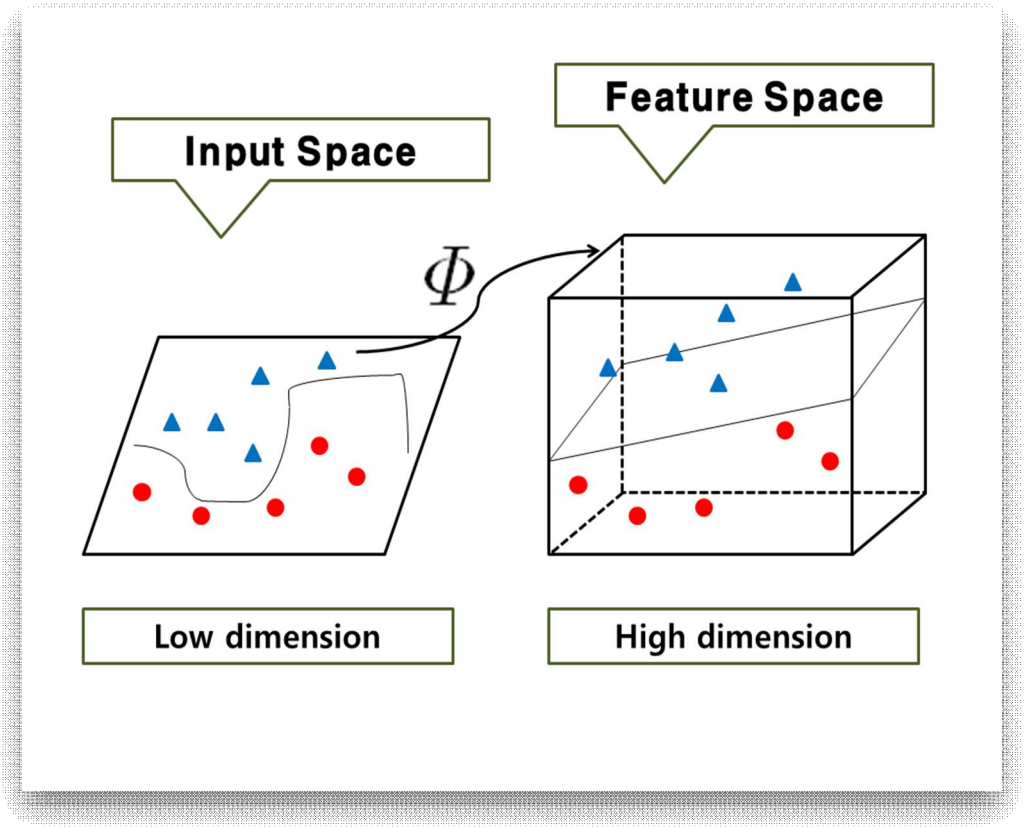
여기에서 r은 알고리즘을 반복한 횟수를 의미한다.

2.3 커널 능형 로지스틱 회귀분류법

앞서 소개한 능형 회귀분석(ridge regression)이나 로지스틱 회귀분류법(logistic regression classification)은 반응변수와 설명변수들 사이의 관계가 선형(linear relationship)인 경우를 가정한 분석방법이다. 하지만 일반적으로 얻어지는 자료들은 그렇지 못한 경우가 대다수이다. 즉, 비선형(nonlinear relationship)구조를 보이거나 상호작용(interaction effect) 등을 고려해야 하는 경우에는 적절한 변환함수를 찾아야 할 것이지만 이를 미리 파악하는 것은 매우 어려운 것이 사실이다. 이러한 경우에 적용할 수 있는 방법이 커널트릭 기법(kernel-trick method)이다. 이 기법은 복잡한 형태를 나타내는 비선형 데이터에 대해 적절한 사상함수(mapping function) Φ를 적용해 p차원의 설명변수 공간에 있는 데이터를 변형하여 고차원의 힐버트 공간(Hilbert space)으로 이동시키는 기법이다. 이를 이용해 데이터를 특성에 맞게 변형시키게 되면 자동적으로 그 데이터의 형태에 알맞은 변환함수를 적용한 것과 같은 효과를 얻을 수 있으며 변형된 데이터에 대해 선형

모형을 적합하여 분석을 실시하게 된다. 아래에 있는 그림 1은 커널트릭 기법의 핵심을 간략하게 표현한 것이다.

<그림 1 : 커널기법을 통한 데이터 변환의 예>



커널트릭 기법을 다음과 같이 2개의 집단을 분류하는 것과 관련한 로지스틱 회귀분류법에 적용할 수 있다. 관측치의 개수가 n이고 설명변수 공간이 p차원인 훈련자료(training data) X가 있다

고 하자. 이 자료에 다음과 같은 관측치들이 포함되어 있다고 가정해보자.

$$(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)$$
$$y_i \in \{0,1\}\ ,\ i = 1, \cdots, n$$

이에 대해 사상함수 Φ를 이용하여 다음과 같이 훈련자료 X에 대한 n개의 설명변수 데이터들을 변환시킨다.

$$x_1, x_2, \cdots, x_n \ \rightarrow\ \Phi(x_1), \Phi(x_2), \cdots, \Phi(x_n)$$

사상함수 Φ를 이용하여 변환된 설명변수 데이터들, 다시 말해서 $\Phi(x_1), \Phi(x_2), \cdots, \Phi(x_n)$은 힐버트 공간(Hilbert space)이라고 불리는 고차원의 특성공간(feature space with high dimension)에 놓이게 된다. 이를 이용하여 다음과 같은 모형을 적합시킬 수 있다.

$$\log\left(\frac{p(\Phi(x);d)}{1-p(\Phi(x);d)}\right) = \Phi(x_1)d_1 + \Phi(x_2)d_2 + \cdots + \Phi(x_n)d_n$$

위와 같은 모형을 통해서 길이가 n인 회귀계수벡터 $d = (d_1, d_2, \cdots, d_n)^T$를 설정하도록 한다. 이는 커널트릭을 통한 공간변환에 관하여 개념적으로 설명한 것이다. 이러한 변환과정은 실제적으로 커널함수 $k(\bullet, \bullet)$에 의한 계산을 통해서 이루어지게 된다. 실제적으로 $\Phi(x_1), \Phi(x_2), \cdots, \Phi(x_n)$ 의 선형결합(linear combination) $v = d_1\Phi(x_1) + \cdots + d_n\Phi(x_n)$ 에 대한 $\Phi(X)$의 사영(projection)은 다음과 같이 계산된다.

$$\sum_{j=1}^{n} \langle \Phi(x_i), \Phi(x_j) \rangle d_j = \sum_{j=1}^{n} k_{i,j} d_j$$

$$k_{i,j} = \langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j)$$

여기에서 $k_{i,j}$는 행렬 $K = \left(k_{i,j}\right)_{\substack{i=1,\cdots,n \\ j=1,\cdots,n}}$의 (i,j)번째 원소이다. 이러한 과정을 통해서 얻을 수 있는 행렬 K와 회귀계수벡터 d를 이용해서 반응변수 y를 설명할 수 있게 되는 것이다. 이에 대한 자세한 설명은 Schölkopf and Smola (2002) 및 Huh (2015)를 참조하기 바란다. 본 연구에서는 다음과 같은 형태의 가우시안 커널(Gaussian kernel)을 적용하였다.

$$k_{i,j} = \exp\left(-\sigma \parallel x_i - x_j \parallel^2\right) = \exp\left\{-\sigma (x_i - x_j)^T (x_i - x_j)\right\},$$
$$\sigma > 0\,,\ \sigma = \frac{1}{p}\ \ ,\ \ i = 1,\cdots,n\ \ ,\ \ j = 1,\cdots,n$$

여기에서 p는 변환하기 전 데이터에 대한 설명변수 공간의 차원이다. 가우시안 커널은 관측치의 설명변수로 구성된 벡터 사이의 유클리디안 거리(Euclidean distance)에 의존한다. 그렇기 때문에 이 커널을 적용하여 생성한 행렬 K는 대칭행렬(symmetric matrix)의 특성을 나타내게 된다. 이와 같은 방법에 의한 커널 변환을 통해서 다음과 같은 형태의 모형을 얻을 수 있다.

$$\theta_i = \log\left(\frac{p(k_i;d)}{1-p(k_i;d)}\right) = k_i^T d\,,\ i=1,\cdots,n$$

$$\Theta = \begin{pmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_n\end{pmatrix} = \begin{pmatrix}k_1^T\\ k_2^T\\ \vdots\\ k_n^T\end{pmatrix} d = Kd$$

$$k_i = (k_{i,1}, k_{i,2}, \cdots, k_{i,n})^T$$

위와 같은 모형에 일반적인 로지스틱 회귀분류법에서 사용되는 최대가능도추정법을 적용하게 되면 다음과 같은 형태의 가능도함수 $L(d)$와 이 함수에 로그를 취한 형태의 함수인 로그가능도함수 $l(d)$를 얻을 수 있다.

$$L(d) = \prod_{i=1}^{n} \{p(k_i;d)\}^{y_i} \{1-p(k_i;d)\}^{1-y_i}$$

$$l(d) = \log L(d) = \sum_{i=1}^{n} \{y_i \log(p(k_i;d)) + (1-y_i)\log(1-p(k_i;d))\}$$

$$= \sum_{i=1}^{n} \{y_i (k_i^T d) - \log(1+\exp(k_i^T d))\}$$

그리고 변환을 통해서 얻어진 행렬 K가 항상 역행렬을 가진다는 보장이 없다는 점을 보완하기 위해 능형 회귀형태의 벌점함수 $\lambda d^T K d$와 음의 로그가능도함수 $-l(d)$의 합의 형태로 이루어진 함수인 $l^*(d)$를 최소화하는 방법을 통해 벡터 d를 추정한다. 이를 위해 함수 $l^*(d)$를 벡터 d에 관하여 미분하게 되면 다음과 같은 스코어방정식(score equation)을 얻을 수 있다.

$$l^*(d) = -l(d) + \lambda d^T K d$$
$$\frac{\partial l^*(d)}{\partial d} = -\sum_{i=1}^{n} k_i \{y_i - p(k_i;d)\} + 2\lambda K d = 0$$

이 방정식의 해는 뉴튼-랩슨 방법(Newton-Raphson method)을 활용한 반복적 재가중치 최소제곱 추정법(iteratively reweighted least squares estimation)을 사용하여 수치적으로 구하게 된다. 2.2절에서 언급했던 일반적인 로지스틱 회귀분류법에서 사용하는 수치적 접근방법을 비슷하게 적용하게 되면 함수 $l^*(d)$의 최소화 문제를 다음과 같은 이차형식 형태(quadratic form)의 함수를 최소화하는 반복 알고리즘 문제로 근사시킬 수 있다. 이 과정을 통해서 찾은 $\hat{d}$의 값을 대입하여 얻을 수 있는 $l^*(\hat{d})$의 값을 이용해서 이 값이 정해진 수렴조건을 만족하게 되는 시점에서 얻어지게 되는 $\hat{d}$의 값을 벡터 d에 대한 최종적인 추정량으로 사용하게 된다.

$$\begin{aligned}
&\hat{d} = \operatorname{argmin}_d \{-l(d) + \lambda d^T K d\} \\
&= \operatorname{argmin}_d \{(z - Kd)^T W (z - Kd) + \lambda d^T K d\}, \\
&l_p(d) = (z - Kd)^T W (z - Kd) + \lambda d^T K d, \\
&z = K d_{old} + W^{-1}(y - p), \ (\lambda > 0)
\end{aligned}$$

위의 과정을 통해서 얻을 수 있는 이차형식의 함수 $l_p(d)$를 벡터 d에 관하여 미분하게 되면 다음과 같은 추정량 $\hat{d}$에 대한 최종적인 표현식을 얻을 수 있다.

$$\frac{\partial l_p(d)}{\partial d} = -2KW(z-Kd)+2\lambda Kd = 0$$

$$\hat{d} = \left(K+\lambda W^{-1}\right)^{-1} z$$

여기에서 d_{old}는 $\hat{d}$이 구해지기 바로 이전 시점에서 구한 추정량을 의미한다. 그리고 p는 i번째 원소가 $p(k_i;d_{old})$인 길이가 n인 확률벡터이며 W는 i번째 대각원소가 $\{p(k_i;d_{old})\}\{1-p(k_i;d_{old})\}$이고 나머지 원소들은 모두 0인 크기 $n\times n$의 대각행렬이다. 본 연구에서는 추정량 $\hat{d}$에 대한 수렴조건(convergence criterion)을 다음과 같이 정했다.

$$\left|l^*(\hat{d})-l^*(d_{old})\right| < 10^{-4} \quad \text{or} \quad r = 100$$

여기에서 r은 알고리즘을 반복한 횟수를 의미한다. 그리고 최적의 능형모수 λ의 값은 $k-$fold cross validation을 통하여 선정한다. 본 연구에서는 $k=5$로 설정하였고 최적의 λ값을 선정 시 음의 로그가능도함수(minus log-likelihood function) $-l(\hat{d})$의 평균값이 최소인 경우를 가장 바람직한 경우로 판단하고 선택하였다. 이러한 과정을 통해서 다음과 같은 형태의 로짓(logit)추정량 $\hat{\Theta}$을 구할 수 있다.

$$\hat{\theta}_i = \log\left(\frac{\hat{p}(k_i;\hat{d})}{1-\hat{p}(k_i;\hat{d})}\right) = k_i^T\hat{d}\ ,\ i=1,\ \cdots,n$$

$$\hat{\Theta} = \begin{pmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \vdots\\ \hat{\theta}_n\end{pmatrix} = \begin{pmatrix}k_1^T\\ k_2^T\\ \vdots\\ k_n^T\end{pmatrix}\hat{d} = K\hat{d}$$

이에 따른 확률(probability)추정량 $\hat{p}$은 로짓과 확률 사이의 관계식에 의해서 다음과 같이 계산된다.

$$\hat{p}(k_i;\hat{d}) = \frac{\exp(k_i^T\hat{d})}{1+\exp(k_i^T\hat{d})}\ \ ,\ \ i=1,2,\ \cdots,n$$

$$\hat{p} = \left(\hat{p}(k_1;\hat{d}),\hat{p}(k_2;\hat{d}),\ \cdots,\hat{p}(k_n;\hat{d})\right)^T$$

만약 훈련자료(training data) X를 이용하여 검증자료(test data) X^*에 대한 평가를 실시하고자 한다면 커널함수 $k(\bullet,\bullet)$를 사용해서 $\Phi(x_1),\Phi(x_2),\ \cdots,\Phi(x_n)$ 의 선형결합인 $v = d_1\Phi(x_1) + \ \cdots\ + d_n\Phi(x_n)$ 에 대한 $\Phi(X^*)$의 사영(projection)을 계산한 다음 이를 바탕으로 검증자료 평가를 위해서 사용할 행렬 K^*를 얻는 과정을 우선적으로 거쳐야 한다. 이에 대한 계산과정을 표현해보면 다음과 같다. 단, n은 훈련자료에 대한 관측치의 개수, s는 검증자료에 대한 관측치의 개수를 나타낸다. 이렇게 구별해서 표기한 이유는 훈련자료의 크기와 검증자료의 크기를 비교했을 때 서로 다른 경우가 대부분이기 때문이다.

$$\sum_{j=1}^{n} \langle \Phi(x_i^*), \Phi(x_j) \rangle d_j = \sum_{j=1}^{n} k_{i,j}^* d_j$$
$$k_{i,j}^* = \langle \Phi(x_i^*), \Phi(x_j) \rangle = k(x_i^*, x_j)$$

그리고 본 연구에서는 다음과 같은 형태의 가우시안 커널(Gaussian kernel)을 적용하였다.

$$k_{i,j}^* = \exp\left(-\sigma \parallel x_i^* - x_j \parallel^2\right) = \exp\left\{-\sigma (x_i^* - x_j)^T (x_i^* - x_j)\right\},$$
$$\sigma > 0\,,\ \sigma = \frac{1}{p}\ ,\ \ i = 1, \cdots, s\ ,\ \ j = 1, \cdots, n$$

여기에서 $k_{i,j}^*$는 행렬 $K^* = \left(k_{i,j}^*\right)_{\substack{i=1,\cdots,s \\ j=1,\cdots,n}}$의 (j,i)번째 원소이며 x_i^*는 검증자료 X^*에 포함되어 있는 i번째 관측치에 대한 설명변수 데이터이다. 이러한 과정을 통해서 얻을 수 있는 행렬 K^*, 그리고 훈련자료 X를 통해서 구한 추정량 $\hat{d}$을 이용해서 검증자료 X^*에 대한 평가를 실시하게 된다. 결과적으로 행렬 K^*와 추정량 $\hat{d}$을 이용하는 계산을 통해서 다음과 같은 형태의 훈련자료를 통한 검증자료에 대한 로짓(logit)추정량 $\widehat{\Theta}^*$을 구할 수 있다.

$$K^{*T}=\begin{pmatrix} k_1^{*T} \\ k_2^{*T} \\ \vdots \\ k_s^{*T} \end{pmatrix}$$

$$\widehat{\Theta^*}=K^{*T}\hat{d}=\left(\hat{\theta}_1^*,\hat{\theta}_2^*,\cdots,\hat{\theta}_s^*\right)^T$$

$$\hat{\theta}_i^*=\log\left(\frac{\hat{p}\left(k_i^*;\hat{d}\right)}{1-\hat{p}\left(k_i^*;\hat{d}\right)}\right)=k_i^{*T}\hat{d}$$
$$k_i^*=\left(k_{i,1}^*,k_{i,2}^*,\cdots,k_{i,n}^*\right)^T$$
$$i=1,2,\cdots,s$$

이러한 로짓추정량 $\widehat{\Theta^*}$을 이용하여 로짓(logit)과 확률 사이의 관계식을 적용하는 계산을 실시하게 되면 다음과 같은 형태의 훈련자료를 통한 검증자료에 대한 확률(probability)추정량 $\hat{p^*}$을 계산할 수 있다.

$$\hat{p}(k_i^*;\hat{d})=\frac{\exp(\hat{\theta}_i^*)}{1+\exp(\hat{\theta}_i^*)}=\frac{\exp(k_i^{*T}\hat{d})}{1+\exp(k_i^{*T}\hat{d})}\quad,\quad i=1,2,\cdots,s$$
$$\hat{p^*}=\left(\hat{p}(k_1^*;\hat{d}),\hat{p}(k_2^*;\hat{d}),\cdots,\hat{p}(k_s^*;\hat{d})\right)^T$$

위의 결과를 통해 훈련자료에 의해서 얻어지는 검증자료의 반응변수에 대한 예측값 $\hat{y^*}=\left(\hat{y_1^*},\hat{y_2^*},\cdots,\hat{y_s^*}\right)^T$는 다음과 같이 판정하게 된다.

$$\hat{y_i^*}=\begin{cases}1\text{ , if }\hat{p}(k_i^*;\hat{d})\geq 0.5\\ 0\text{ , if }\hat{p}(k_i^*;\hat{d})<0.5\end{cases}\quad,\quad i=1,2,\cdots,s$$

이러한 과정을 통해서 얻어진 예측값 $\hat{y}^{*} = \left(\hat{y}_1^{*}, \hat{y}_2^{*}, \cdots, \hat{y}_s^{*}\right)^T$와 검증 자료의 반응변수 $y^{test} = \left(y_1^{test}, y_2^{test}, \cdots, y_s^{test}\right)^T$를 비교하여 다음과 같은 형태의 최종적인 검증 오분류율(test misclassification rate) $\hat{\pi}^{*}$을 계산한다.

$$\hat{\pi}^{*} = \frac{1}{s}\sum_{i=1}^{s} I\left(y_i^{test} \neq \hat{y}_i^{*}\right)$$

여기에서 $I(\bullet)$는 지시함수(indicator function)로써 괄호 안의 조건을 만족하면 1, 그렇지 않으면 0의 값을 가진다. 즉, 이 함수는 검증자료의 반응변수와 이에 대한 예측값이 서로 다른 경우에 한해서만 1의 값을 가지고 나머지는 모두 0의 값을 가지게 된다. 그러므로 $\hat{\pi}^{*}$은 검증자료의 관측치들 중 잘못 분류된 경우에 대한 비율이라고 할 수 있다.

3 앙상블 기법 (ensemble method)

앙상블 기법(ensemble method)은 통계적 기계학습(statistical machine learning)의 나무모형 기법(tree-model method)에서 나온 분석방법이다. 본 연구에서는 배깅(bagging)과 랜덤포레스트(random forests)를 사용한다. 이 2가지 방법의 공통적인 핵심은 서로 독립적인 부트스트랩(bootstrap) 표본들을 추출한 다음 이 표본들을 이용하여 얻은 결과들에 대한 평균(mean)을 추정량으로 사용한다는 것이다. 이러한 방법을 통해서 얻어진 추정량은 일반적인 방법을 통해서 얻어진 추정량에 비해 분산의 크기가 매우 작다는 특징이 있다. 이를 통해 보다 더 정확한 예측이나 분류를 할 수 있다.

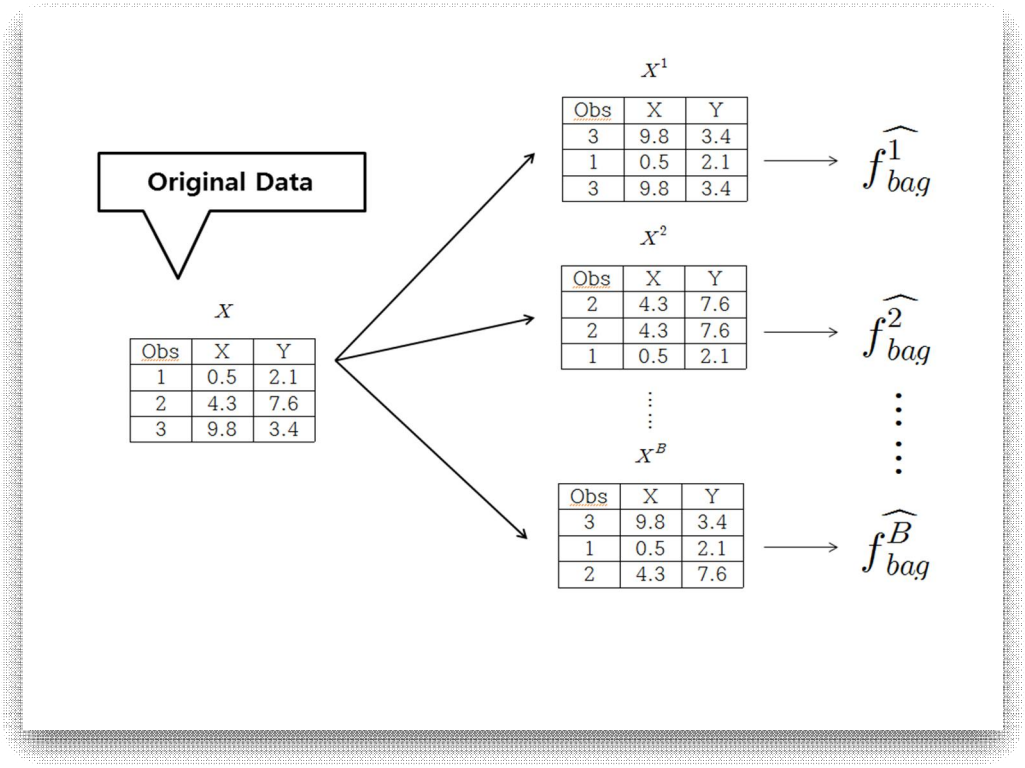
예를 들어 다음과 같이 서로 독립이고 각각의 분산이 σ^2인 n개의 표본 $W_1, W_2, \cdots, W_n$ 이 있다고 한다면 이들의 평균 $\overline{W}= \frac{1}{n}\sum_{i=1}^{n} W_i$ 의 분산은 $\frac{\sigma^2}{n}$으로 크게 줄어들게 된다. 즉, 평균화의 방법을 통해서 얻은 추정량은 일반적인 방법을 통해서 얻은 추정량에 비해 분산의 크기가 매우 작기 때문에 신뢰성 측면에서 더 바람직한 성능을 보이게 된다. 그리고 표본의 개수 n과 분산 $\frac{\sigma^2}{n}$은 서로 반비례한다. 그러므로 신뢰성 측면에서 좀 더 바람직한 추정량을 구하고자 한다면 더 많은 개수의 표본들을 추출해서

분산의 값을 큰 폭으로 줄이면 될 것이다.

3.1 배깅 (bagging)

배깅(bagging) 또는 bootstrap aggregation은 1994년 미국의 통계학자 Leo Breiman(1928~2005)에 의하여 제안된 방법으로 하나의 훈련자료(training data)에서 반복적인 복원추출(sampling with replacement)을 통하여 훈련자료와 비교했을 때 관측치의 개수가 동일한 여러 개의 표본들을 추출하는 부트스트랩(bootstrap)을 응용한 기법이다. 배깅 기법과 부트스트랩에 대한 자세한 내용은 Hastie et al. (2011) 및 James et al. (2014)를 참조하기 바란다. 그림 2는 부트스트랩 기법에 관하여 표현한 것이다. 간단한 예제이기는 하지만 부트스트랩의 핵심을 제대로 표현하고 있다고 필자는 생각한다.

<그림 2 : 부트스트랩 기법에 대한 간단한 예제>



예를 들어 다음과 같이 부트스트랩을 통해서 얻은 표본들을 이용해서 구한 B개의 추정량들이 있다고 가정하자.

$$\widehat{f^1_{bag}}(x), \widehat{f^2_{bag}}(x), \cdots, \widehat{f^B_{bag}}(x)$$

배깅 기법에 의한 추정량은 이 B개의 추정량들을 평균내서 구하게 된다. 이를 통해 상대적으로 분산이 작은 다음과 같은 추정량을 만들 수 있게 된다.

$$\widehat{f_{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f^b_{bag}}(x)$$

3.2 랜덤포레스트 (random forests)

랜덤포레스트(random forests)는 배깅 기법의 문제점을 보완하기 위해 제안된 기법으로 Leo Breiman과 Adele Cutler의 공동연구에 의해서 정립되었다. 이 기법에서도 배깅 기법과 동일하게 부트스트랩 기법을 이용하여 여러 개의 추정량들을 만들고 이들을 평균내서 최종적인 추정량을 얻는 과정을 거치게 된다. 하지만 각각의 표본들에 대해 모든 설명변수들을 다 사용하지 않고 이들 중 일부만을 선택해서 사용한다는 점이 예측력을 높이는 큰 차이를 준다고 할 수 있다. 모든 설명변수들을 다 사용하는 배깅 기법을 적용하게 되면 얻어진 서브모형들끼리 서로 연관성을 지니고 있을 수 있기 때문에 문제가 될 수 있다. 이 점을 보완하기 위해 모든 설명변수들 중 일부만을 선택해서 사용하게 되면 그 연관성을 낮출 수 있게 된다. 물론 이러한 과정에 의해 편의(bias)가 발생하게 되지만 연관성을 줄임으로써 분산을 좀 더 큰 폭으로 줄이게 되므로 그 효과를 상쇄시킬 수 있다. 이를 통해 좀 더 정확한 예측, 분류를 할 수 있게 된다.

예를 들어 설명변수의 개수가 p인 훈련자료가 있다고 가정하자. 이 자료를 이용하여 부트스트랩을 통해서 B개의 표본들을 추출한다. 여기에서 각각의 표본을 추출할 때마다 p개의 모든 설명변수들 중 m개만을 선택해서 포함시켜야 한다는 점에 주의할 필요가 있다. 일반적으로 선택되는 설명변수의 개수는 $m \approx \sqrt{p}$ 또는

$m \approx \frac{p}{3}$로 정하게 된다. 본 연구에서는 $m \approx \sqrt{p}$를 적용하였으며 m의 값이 자연수의 형태로 나타나지 않을 경우에는 소수점 이하 반올림을 사용하였다. 그리고 각각의 부트스트랩 표본에 포함시킬 m개의 설명변수들은 부트스트랩 표본을 추출할 때마다 다르게 선택해야 한다는 점도 유념해야 할 것이다. 이러한 과정을 통해서 다음과 같은 B개의 추정량들을 얻을 수 있다.

$$\widehat{f_{rf}^{1}}(x), \widehat{f_{rf}^{2}}(x), \cdots, \widehat{f_{rf}^{B}}(x)$$

랜덤포레스트 기법에 의한 추정량은 이 B개의 추정량들을 평균내서 구하게 된다. 이를 통해 좀 더 정확도가 높은 우수한 추정량을 다음과 같은 형태로 얻을 수 있다.

$$\widehat{f_{rf}}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f_{rf}^{b}}(x)$$

이러한 방법을 통해서 얻은 추정량은 배깅 기법에 의해서 만들어진 추정량보다 정확성과 신뢰성 측면에서 더 좋은 성능을 가지게 된다. 배깅이나 랜덤포레스트 같은 앙상블 기법을 커널 능형 로지스틱 회귀분류법에 적용하는 방법은 4장을 통해서 소개하도록 하겠다.

4 앙상블 기법을 이용한 커널 능형 로지스틱 회귀분류법

3장에서 소개한 배깅 기법과 랜덤포레스트 기법은 분산을 큰 폭으로 줄여서 추정의 정확도를 높일 수 있는 방법들이다. 이를 커널 능형 로지스틱 회귀분류법에 적용한다면 보다 더 정확한 분류결과를 얻을 수 있을 것이라는 예상을 할 수 있을 것이다. 본 장에서는 앙상블 기법을 어떠한 방식으로 2개의 집단을 분류하는 것과 관련한 커널 능형 로지스틱 회귀분류법에 적용할 수 있는지 소개한다.

4.1 배깅 기법을 이용한 커널 능형 로지스틱 회귀분류법

관측치의 개수가 n이고 설명변수 공간이 p차원인 훈련자료(training data) X가 있다고 하자. 이 자료에서 부트스트랩을 통해 다음과 같은 관측치들을 포함하고 있는 b번째 표본 X^b를 추출하였다고 가정해보자.

$(b = 1, 2, \cdots, B)$

$$(y_1^b, x_1^b), (y_2^b, x_2^b), \cdots, (y_n^b, x_n^b)$$
$$y_i^b \in \{0,1\} \ , \ i = 1, \cdots, n$$

이에 대해 사상함수 Φ를 이용하여 다음과 같이 부트스트랩 훈련

자료 X^b에 대한 n개의 설명변수 데이터들을 변환시킨다.

$$x_1^b, x_2^b, \cdots, x_n^b \rightarrow \Phi(x_1^b), \Phi(x_2^b), \cdots, \Phi(x_n^b)$$

사상함수 Φ를 이용하여 변환된 설명변수 데이터들, 다시 말해서 $\Phi(x_1^b), \Phi(x_2^b), \cdots, \Phi(x_n^b)$는 힐버트 공간(Hilbert space)이라고 불리는 고차원의 특성공간(feature space with high dimension)에 놓이게 된다. 이를 이용하여 다음과 같은 모형을 적합시킬 수 있다.

$$\log\left(\frac{p(\Phi(x^b);d^b)}{1-p(\Phi(x^b);d^b)}\right)=\Phi(x_1^b)d_1^b+\Phi(x_2^b)d_2^b+\cdots+\Phi(x_n^b)d_n^b$$

위와 같은 모형을 통해서 길이가 n인 회귀계수벡터 $d^b=(d_1^b, d_2^b, \cdots, d_n^b)^T$를 설정하도록 한다. 이러한 변환과정은 실제적으로 커널함수 $k(\bullet,\bullet)$에 의한 계산을 통해서 이루어지게 된다. 실제적으로 $\Phi(x_1^b), \Phi(x_2^b), \cdots, \Phi(x_n^b)$ 의 선형결합(linear combination) $v^b=d_1^b\Phi\left(x_1^b\right)+\cdots+d_n^b\Phi\left(x_n^b\right)$ 에 대한 $\Phi(X^b)$의 사영(projection)은 다음과 같이 계산된다.

$$\sum_{j=1}^{n}\left\langle\Phi\left(x_i^b\right),\Phi\left(x_j^b\right)\right\rangle d_j^b=\sum_{j=1}^{n}k_{i,j}^b d_j^b$$

$$k_{i,j}^b=\left\langle\Phi\left(x_i^b\right),\Phi\left(x_j^b\right)\right\rangle=k(x_i^b,x_j^b)$$

여기에서 $k_{i,j}^b$는 행렬 $K^b=\left(k_{i,j}^b\right)_{\substack{i=1,\cdots,n\\ j=1,\cdots,n}}$의 (i,j)번째 원소이다. 이러한 과정을 통해서 얻을 수 있는 행렬 K^b와 회귀계수벡터 d^b를

이용해서 반응변수 y^b를 설명하게 된다. 본 연구에서는 다음과 같은 형태의 가우시안 커널(Gaussian kernel)을 적용하였다.

$$k_{i,j}^b = \exp\left(-\sigma \parallel x_i^b - x_j^b \parallel^2\right) = \exp\left\{-\sigma (x_i^b - x_j^b)^T (x_i^b - x_j^b)\right\},$$
$$\sigma > 0\,,\ \sigma = \frac{1}{p}\ \ ,\ \ i = 1,\cdots,n\ \ ,\ \ j = 1,\cdots,n$$

여기에서 p는 변환하기 전 데이터에 대한 설명변수 공간의 차원을 나타낸다. 이와 같은 방법에 의한 커널 변환을 통해서 다음과 같은 형태의 모형을 얻을 수 있다.

$$\theta_i^b = \log\left(\frac{p\left(k_i^b; d^b\right)}{1 - p\left(k_i^b; d^b\right)}\right) = k_i^{bT} d^b\,,\ i = 1,\cdots,n$$
$$\Theta^b = \begin{pmatrix} \theta_1^b \\ \theta_2^b \\ \vdots \\ \theta_n^b \end{pmatrix} = \begin{pmatrix} k_1^{bT} \\ k_2^{bT} \\ \vdots \\ k_n^{bT} \end{pmatrix} d^b = K^b d^b$$
$$k_i^b = (k_{i,1}^b, k_{i,2}^b, \cdots, k_{i,n}^b)^T$$

위와 같은 모형에 일반적인 로지스틱 회귀분류법에서 사용되는 최대가능도추정법을 적용하게 되면 다음과 같은 형태의 가능도함수 $L(d^b)$와 이 함수에 로그를 취한 형태의 함수인 로그가능도함수 $l(d^b)$를 얻을 수 있다.

$$L(d^b) = \prod_{i=1}^{n} \left\{ p(k_i^b; d^b) \right\}^{y_i^b} \left\{ 1 - p(k_i^b; d^b) \right\}^{1-y_i^b}$$

$$l(d^b) = \log L(d^b) = \sum_{i=1}^{n} \left\{ y_i^b \log\left(p\left(k_i^b; d^b\right)\right) + \left(1 - y_i^b\right) \log\left(1 - p\left(k_i^b; d^b\right)\right)\right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i^b (k_i^{bT} d^b) - \log\left(1 + \exp(k_i^{bT} d^b)\right)\right\}$$

그리고 변환을 통해서 얻어진 행렬 K^b가 항상 역행렬을 가진다는 보장이 없다는 점을 보완하기 위해 음의 로그가능도함수 $-l(d^b)$와 능형 회귀형태의 벌점함수 $\lambda^b d^{bT} K^b d^b$의 합의 형태로 이루어진 함수인 $l^*(d^b)$를 최소화하는 방법을 통해 벡터 d^b를 추정한다. 이를 위해 함수 $l^*(d^b)$를 벡터 d^b에 관하여 미분하게 되면 다음과 같은 스코어방정식(score equation)을 얻을 수 있다.

$$l^*(d^b) = -l(d^b) + \lambda^b d^{bT} K^b d^b$$

$$\frac{\partial l^*(d^b)}{\partial d^b} = -\sum_{i=1}^{n} k_i^b \left\{ y_i^b - p(k_i^b; d^b) \right\} + 2\lambda^b K^b d^b = 0$$

이 방정식의 해는 뉴튼-랩슨 방법(Newton-Raphson method)을 활용한 반복적 재가중치 최소제곱 추정법(iteratively reweighted least squares estimation)을 사용하여 수치적으로 구하게 된다. 이 방법을 적용하게 되면 함수 $l^*(d^b)$의 최소화 문제를 다음과 같은 이차형식 형태(quadratic form)의 함수를 최소화하는 반복 알고리즘 문제로 근사시킬 수 있다. 이 과정을 통해서 찾은 $\hat{d^b}$의 값을 대입하여 얻을 수 있는 $l^*(\hat{d^b})$의 값을 이용해서 이 값이 정해

진 수렴조건을 만족하게 되는 시점에서 얻어지게 되는 $\hat{d^b}$의 값을 벡터 d^b에 대한 최종적인 추정량으로 사용하게 된다.

$$\begin{aligned}
\hat{d^b} &= \mathrm{argmin}_{d^b}\{-l(d^b)+\lambda^b d^{bT}K^b d^b\} \\
&= \mathrm{argmin}_{d^b}\left\{(z^b-K^b d^b)^T W^b(z^b-K^b d^b)+\lambda^b d^{bT}K^b d^b\right\}, \\
l_p(d^b) &= (z^b-K^b d^b)^T W^b(z^b-K^b d^b)+\lambda^b d^{bT}K^b d^b, \\
z^b &= K^b d^b_{old}+(W^b)^{-1}(y^b-p^b) \quad , \quad (\lambda^b>0)
\end{aligned}$$

위의 과정을 통해서 얻을 수 있는 이차형식의 함수 $l_p(d^b)$를 벡터 d^b에 관하여 미분하게 되면 다음과 같은 추정량 $\hat{d^b}$에 대한 최종적인 표현식을 얻을 수 있다.

$$\frac{\partial l_p(d^b)}{\partial d^b} = -2K^b W^b(z^b-K^b d^b)+2\lambda^b K^b d^b = 0$$

$$\hat{d^b} = (K^b+\lambda^b(W^b)^{-1})^{-1} z^b$$

여기에서 d^b_{old}는 $\hat{d^b}$이 구해지기 바로 이전 시점에서 구한 추정량을 의미한다. 그리고 p^b는 i번째 원소가 $p(k_i^b;d_{old}^b)$인 길이가 n인 확률벡터이며 W^b는 i번째 대각원소가 $\{p(k_i^b;d_{old}^b)\}\{1-p(k_i^b;d_{old}^b)\}$이고 나머지 원소들은 모두 0인 크기가 $n\times n$인 대각행렬이다. 본 연구에서는 추정량 $\hat{d^b}$에 대한 수렴조건(convergence criterion)을 다음과 같이 정했다. 단, r은 알고리즘을 반복한 횟수를 나타낸다.

$$\left| l^*(\hat{d^b}) - l^*(d^b_{old}) \right| < 10^{-4} \quad \text{or} \quad r = 100$$

그리고 최적의 능형모수 λ^b의 값은 훈련자료(training data)에서 임의의 B^*개의 부트스트랩 표본들을 추출하여 실시하는 out-of-bag(OOB)의 방법을 통하여 선정한다. 즉, 추출한 표본들을 새로운 훈련자료(training data)로 두고 이에 대해 표본 추출 시 뽑히지 않은 나머지 관측치들을 해당 표본들에 대한 평가자료(validation data)로 두어서 최적의 조건을 찾는 방식을 취한다. 여기에서 최적의 조건이란 각 λ^b의 값에 대하여 얻어지는 B^*개의 검증 오분류율(test misclassification rate)의 수치들에 대한 평균값이 가장 작게 나오는 경우를 의미한다. 이러한 과정을 통해서 다음과 같은 형태의 로짓(logit)추정량 $\widehat{\Theta^b}$을 구할 수 있다.

$$\hat{\theta_i^b} = \log\left(\frac{\hat{p}\left(k_i^b; \hat{d^b}\right)}{1 - \hat{p}\left(k_i^b; \hat{d^b}\right)} \right) = k_i^{bT} \hat{d^b}\ ,\ i = 1,\ \cdots, n$$

$$\widehat{\Theta^b} = \begin{pmatrix} \hat{\theta_1^b} \\ \hat{\theta_2^b} \\ \vdots \\ \hat{\theta_n^b} \end{pmatrix} = \begin{pmatrix} k_1^{bT} \\ k_2^{bT} \\ \vdots \\ k_n^{bT} \end{pmatrix} \hat{d^b} = K^b \hat{d^b}$$

이에 따른 확률(probability)추정량 $\hat{p^b}$은 로짓과 확률 사이의 관계식에 의해서 다음과 같이 계산된다.

$$\hat{p}(k_i^b;\hat{d^b}) = \frac{\exp(k_i^{bT}\hat{d^b})}{1+\exp(k_i^{bT}\hat{d^b})} \quad , \quad i=1,2,\cdots,n$$
$$\hat{p^b} = \left(\hat{p}(k_1^b;\hat{d^b}),\hat{p}(k_2^b;\hat{d^b}),\cdots,\hat{p}(k_n^b;\hat{d^b})\right)^T$$

만약 부트스트랩에 의해서 추출된 훈련자료(training data) X에 대한 b번째 표본 X^b를 이용하여 검증자료(test data) X^*에 대한 평가를 실시하고자 한다면 커널함수 $k(\bullet,\bullet)$를 사용해서 $\Phi(x_1^b),\Phi(x_2^b),\cdots,\Phi(x_n^b)$ 의 선형결합인

$v^b = d_1^b\Phi\left(x_1^b\right)+\cdots+d_n^b\Phi\left(x_n^b\right)$ 에 대한 $\Phi(X^*)$의 사영(projection)을 계산한 다음 이를 바탕으로 검증자료 평가를 위해서 사용할 행렬 K^{b*}를 얻는 과정을 우선적으로 거쳐야 한다 $(b=1,2,\cdots,B)$. 이에 대한 계산과정을 표현해보면 다음과 같다. 단, n은 훈련자료에 대한 관측치의 개수, s는 검증자료에 대한 관측치의 개수를 나타낸다.

$$\sum_{j=1}^{n}\left\langle\Phi(x_i^*),\Phi(x_j^b)\right\rangle d_j^b = \sum_{j=1}^{n}k_{i,j}^{b*}d_j^b$$
$$k_{i,j}^{b*} = \left\langle\Phi(x_i^*),\Phi(x_j^b)\right\rangle = k(x_i^*,x_j^b)$$

그리고 본 연구에서는 다음과 같은 형태의 가우시안 커널(Gaussian kernel)을 적용하였다.

$$k_{i,j}^{b*} = \exp\left(-\sigma\parallel x_i^*-x_j^b\parallel^2\right) = \exp\left\{-\sigma(x_i^*-x_j^b)^T(x_i^*-x_j^b)\right\},$$
$$\sigma>0\ ,\ \sigma=\frac{1}{p}\quad,\quad i=1,\cdots,s\quad,\quad j=1,\cdots,n$$

여기에서 $k_{i,j}^{b*}$는 행렬 $K^{b*}=\left(k_{i,j}^{b*}\right)_{\substack{i=1,\cdots,s\\ j=1,\cdots,n}}$의 (j,i)번째 원소이며 x_i^*는 검증자료 X^*에 포함되어 있는 i번째 관측치에 대한 설명변수 데이터이다. 이러한 과정을 통해서 얻을 수 있는 행렬 K^{b*}, 그리고 부트스트랩 훈련자료 X^b를 통해서 구한 추정량 $\hat{d}^b$을 이용해서 검증자료 X^*에 대한 평가를 실시하게 된다. 결과적으로 행렬 K^{b*}와 추정량 $\hat{d}^b$을 이용하는 계산을 통해서 각각의 부트스트랩 훈련자료를 통한 검증자료에 대한 로짓(logit)추정량 $\widehat{\Theta_{bag}^{b*}}$을 구할 수 있으며 이에 대한 계산과정은 다음과 같다.

$$K^{b*T}=\begin{pmatrix} k_1^{b*T} \\ k_2^{b*T} \\ \vdots \\ k_s^{b*T} \end{pmatrix}$$

$$\widehat{\Theta_{bag}^{b*}}=K^{b*T}\hat{d}^b=\left(\widehat{\theta_1^{b*}},\widehat{\theta_2^{b*}},\cdots,\widehat{\theta_s^{b*}}\right)^T$$

$$\widehat{\theta_i^{b*}}=\log\left(\frac{\hat{p}\left(k_i^{b*};\hat{d}^b\right)}{1-\hat{p}\left(k_i^{b*};\hat{d}^b\right)}\right)=k_i^{b*T}\hat{d}^b$$

$$k_i^{b*}=\left(k_{i,1}^{b*},k_{i,2}^{b*},\cdots,k_{i,n}^{b*}\right)^T$$

$$i=1,2,\cdots,s$$

이러한 과정을 B번 반복하여 $(b=1,2,\cdots,B)$ B개의 $\widehat{\Theta_{bag}^{b*}}$을 얻는다. 이를 평균하여 다음과 같은 최종적인 배깅 기법을 이용한 로짓추정량 $\widehat{\Theta_{bag}^{*}}$을 구할 수 있다.

$$\widehat{\Theta_{bag}^{*}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\Theta_{bag}^{b*}} = \left(\widehat{\theta_1^{bag}}, \widehat{\theta_2^{bag}}, \cdots, \widehat{\theta_s^{bag}}\right)^T$$

이러한 로짓추정량 $\widehat{\Theta_{bag}^{*}}$을 이용하여 로짓(logit)과 확률 사이의 관계식을 적용하는 계산을 실시하게 되면 다음과 같은 최종적인 배깅 기법을 통한 확률(probability)추정량 $\widehat{p_{bag}^{*}}$을 얻을 수 있게 된다.

$$\widehat{p_{bag}^{*}} = \left(\widehat{p_1^{bag}}, \widehat{p_2^{bag}}, \cdots, \widehat{p_s^{bag}}\right)^T$$

$$\widehat{p_i^{bag}} = \frac{\exp\left(\widehat{\theta_i^{bag}}\right)}{1+\exp\left(\widehat{\theta_i^{bag}}\right)} \ , \ i = 1, 2, \cdots, s$$

이러한 방법 이외에도 배깅 기법을 통한 추정량을 구하는 데 있어 각각의 부트스트랩 훈련자료를 통한 검증자료에 대한 확률추정량 $\widehat{p_{bag}^{b*}}$을 바로 계산해내는 방법을 생각해 볼 수도 있으며 이에 대한 계산과정은 다음과 같다.

$$\hat{p}(k_i^{b*};\widehat{d^b}) = \frac{\exp(\widehat{\theta_i^{b*}})}{1+\exp(\widehat{\theta_i^{b*}})} = \frac{\exp(k_i^{b*T}\widehat{d^b})}{1+\exp(k_i^{b*T}\widehat{d^b})} \ , \ i = 1, 2, \cdots, s$$

$$\widehat{p_{bag}^{b*}} = \left(\hat{p}(k_1^{b*};\widehat{d^b}), \hat{p}(k_2^{b*};\widehat{d^b}), \cdots, \hat{p}(k_s^{b*};\widehat{d^b})\right)^T$$

이러한 과정을 B번 반복하여 $(b = 1, 2, \cdots, B)$ B개의 $\widehat{p_{bag}^{b*}}$을 얻는다. 이를 평균하여 다음과 같은 최종적인 배깅 기법을 이용한 확률(probability)추정량 $\widehat{p_{bag}^{*}}$을 구할 수 있다.

$$\widehat{p^{*}_{bag}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{p^{b*}_{bag}} = \left(\widehat{p_1^{bag}}, \widehat{p_2^{bag}}, \cdots, \widehat{p_s^{bag}}\right)^T$$

이러한 방법들을 통해서 얻은 확률추정량 $\widehat{p^{*}_{bag}}$을 이용하여 반응변수의 예측값 $\widehat{y^{*}_{bag}} = \left(\widehat{y_1^{bag}}, \widehat{y_2^{bag}}, \cdots, \widehat{y_s^{bag}}\right)^T$는 다음과 같이 판정한다.

$$\widehat{y_i^{bag}} = \begin{cases} 1\,, \text{if } \widehat{p_i^{bag}} > 0.5 \\ 0\,, \text{if } \widehat{p_i^{bag}} < 0.5 \end{cases}$$
$$\widehat{y_i^{bag}} \sim i.i.d\ Bernoulli(0.5) \ \text{if } \widehat{p_i^{bag}} = 0.5$$
$$i = 1, 2, \cdots, s$$

만약 $\widehat{p_i^{bag}} = 0.5$인 경우에는 최종 예측값 $\widehat{y_i^{bag}}$을 임의로 1과 0 중 하나를 선택해서 대체한다. 단, 선택하는 확률은 동일하게 각각 $\frac{1}{2}$로 정하도록 한다. 이를 검증자료의 모든 관측치에 대해 적용한다.

그리고 $\widehat{y^{*}_{bag}}$을 얻는 또 다른 방법으로써 다중투표(majority vote)의 방법을 생각해 볼 수도 있으며 보통 앙상블 기법을 분류문제에 적용하는 경우에 이 방법이 주로 사용되는 것으로 알려져 있다. 다음과 같이 각각의 부트스트랩 훈련자료를 통해서 얻을 수 있는 검증자료에 대한 로짓(logit)추정량 $\widehat{\Theta^{b*}_{bag}}$을 구한다.

$$K^{b^{*}T}=\begin{pmatrix} k_1^{b^{*}T} \\ k_2^{b^{*}T} \\ \vdots \\ k_s^{b^{*}T} \end{pmatrix}$$

$$\widehat{\Theta_{bag}^{b^{*}}}=K^{b^{*}T}\hat{d^{b}}=\left(\widehat{\theta_1^{b^{*}}},\widehat{\theta_2^{b^{*}}},\cdots,\widehat{\theta_s^{b^{*}}}\right)^{T}$$

$$\widehat{\theta_i^{b^{*}}}=\log\left(\frac{\hat{p}\left(k_i^{b^{*}};\widehat{d^{b}}\right)}{1-\hat{p}\left(k_i^{b^{*}};\widehat{d^{b}}\right)}\right)=k_i^{b^{*}T}\hat{d^{b}}$$

$$k_i^{b^{*}}=\left(k_{i,1}^{b^{*}},k_{i,2}^{b^{*}},\cdots,k_{i,n}^{b^{*}}\right)^{T}$$

$$i=1,2,\cdots,s$$

이를 이용하여 다음과 같이 각각의 부트스트랩 훈련자료를 통해서 얻을 수 있는 검증자료에 대한 확률(probability)추정량 $\widehat{p_{bag}^{b^{*}}}$을 계산한다.

$$\hat{p}(k_i^{b^{*}};\widehat{d^{b}})=\frac{\exp(\widehat{\theta_i^{b^{*}}})}{1+\exp(\widehat{\theta_i^{b^{*}}})}=\frac{\exp(k_i^{b^{*}T}\widehat{d^{b}})}{1+\exp(k_i^{b^{*}T}\widehat{d^{b}})}\quad,\quad i=1,2,\cdots,s$$

$$\widehat{p_{bag}^{b^{*}}}=\left(\hat{p}(k_1^{b^{*}};\widehat{d^{b}}),\hat{p}(k_2^{b^{*}};\widehat{d^{b}}),\cdots,\hat{p}(k_s^{b^{*}};\widehat{d^{b}})\right)^{T}$$

위의 결과를 통해 b번째 부트스트랩 훈련자료에 의해서 얻어지는 검증자료의 반응변수에 대한 예측값

$\widehat{y_{bag}^{b^{*}}}=\left(\widehat{y_{1,bag}^{b^{*}}},\widehat{y_{2,bag}^{b^{*}}},\cdots,\widehat{y_{s,bag}^{b^{*}}}\right)^{T}$는 다음과 같이 판정하게 된다.

$$\widehat{y_{i,bag}^{b^{*}}}=\begin{cases}1\,,\,\text{if}\ \hat{p}(k_i^{b^{*}};\widehat{d^{b}})\geq 0.5\\ 0\,,\,\text{if}\ \hat{p}(k_i^{b^{*}};\widehat{d^{b}})<0.5\end{cases}\quad,\quad i=1,2,\cdots,s$$

이러한 과정을 B번 반복하여 $(b=1,2,\cdots,B)$ B개의 $\widehat{y_{bag}^{b*}}$을 얻는다. 이를 이용하여 다음과 같은 최종적인 배깅 기법을 이용한 예측값 $\widehat{y_{bag}^{*}}$을 얻을 수 있다.

$$\widehat{y_{bag}^{*}}=\left(\widehat{y_1^{bag}},\widehat{y_2^{bag}},\cdots,\widehat{y_s^{bag}}\right)^T$$

$$\widehat{y_i^{bag}}=\begin{cases}1\,,\,\text{if}\ \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\widehat{y_{i,bag}^{b*}}>0.5\\ 0\,,\,\text{if}\ \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\widehat{y_{i,bag}^{b*}}<0.5\end{cases}$$

$$\widehat{y_i^{bag}}\ \sim\ i.i.d\ Bernoulli(0.5)\ \ \text{if}\ \frac{1}{B}\sum_{b=1}^{B}\widehat{y_{i,bag}^{b*}}=0.5$$

$$i=1,2,\cdots,s$$

다시 말해서 각각의 검증자료의 관측치에 대한 B개의 부트스트랩 예측값들을 수집하여 그 값들 중 1이 0보다 많은 경우에는 최종 예측값을 1로, 1이 0보다 적은 경우에는 최종 예측값을 0으로 판정한다. 만약 1과 0의 개수가 같게 나온 관측치에 대해서는 최종 예측값 $\widehat{y_i^{bag}}$을 임의로 1과 0 중 하나를 선택해서 대체한다. 단, 선택하는 확률은 동일하게 각각 $\frac{1}{2}$로 정하도록 한다. 이를 검증자료의 모든 관측치에 대해 적용한다.

이러한 과정을 통해서 얻어진 예측값 $\widehat{y_{bag}^{*}}=\left(\widehat{y_1^{bag}},\widehat{y_2^{bag}},\cdots,\widehat{y_s^{bag}}\right)^T$와 검증자료의 반응변수 $y^{test}=\left(y_1^{test},y_2^{test},\cdots,y_s^{test}\right)^T$를 비교하여

다음과 같은 형태의 최종적인 검증 오분류율(test misclassification rate) $\widehat{\pi^{*}_{bag}}$을 계산한다.

$$\widehat{\pi^{*}_{bag}} = \frac{1}{s}\sum_{i=1}^{s} I\left(y_i^{test} \neq \widehat{y_i^{bag}}\right)$$

4.2 랜덤포레스트 기법을 이용한 커널 능형 로지스틱 회귀분류법

배깅 기법의 경우와 마찬가지로 관측치의 개수가 n이고 설명변수 공간이 p차원인 훈련자료 X가 있다고 하자. 이 자료에서 부트스트랩을 통해 다음과 같은 관측치들을 포함하고 있는 b번째 표본 X^b를 추출하도록 한다.

$(b = 1, 2, \cdots, B)$

$$(y_1^b, x_1^b), (y_2^b, x_2^b), \cdots, (y_n^b, x_n^b)$$
$$y_i^b \in \{0,1\}\ ,\ i = 1, \cdots, n$$

여기에서 랜덤포레스트 기법을 적용하는 경우에는 훈련자료 X에 대해 부트스트랩을 실행할 때 p개의 모든 설명변수들을 다 사용하는 것이 아니라 이들 중 m개만을 선택해서 부트스트랩 훈련자료 X^b에 포함시켜야 한다는 점에 주의할 필요가 있다. 그러므로 설명변수의 공간은 m차원으로 줄어들게 된다. 본 연구에서는 $m \approx \sqrt{p}$를 적용하였으며 이 값이 자연수의 형태로 나타나지 않

을 경우에는 소수점 이하 반올림을 사용하였다. 그리고 각각의 부트스트랩 훈련자료에 포함시킬 m개의 설명변수들은 부트스트랩 훈련자료를 추출할 때마다 다르게 선택해야 한다는 점도 유념해야 할 것이다.

최종적인 랜덤포레스트 기법을 이용한 검증자료의 반응변수에 대한 로짓추정량, 확률추정량, 예측값, 그리고 이에 따른 최종적인 검증 오분류율을 구하는 방법은 일부의 설명변수들을 선택해서 사용한다는 것만 차이가 있고 나머지는 배깅 기법과 동일하다. 따라서 이에 대한 자세한 설명은 생략하도록 한다. 다만 배깅 기법을 통해서 얻은 결과와 구별하기 위해서 이를 $\widehat{\Theta}_{rf}^{*}$, $\widehat{p}_{rf}^{*}$, $\widehat{y}_{rf}^{*}$, 그리고 $\widehat{\pi}_{rf}^{*}$으로 표기하도록 하겠다. 앙상블 기법을 이용한 커널 능형 로지스틱 회귀분류법이 일반적인 방법론들과 비교했을 때 더 우수한 성능을 보인다는 사실을 입증하기 위해서 실시한 모의실험과 실증분석에 대한 과정과 결과는 5장을 통해서 언급하도록 하겠다.

5 모의실험 및 실증분석

본 장에서는 커널 능형 로지스틱 회귀분류법에 관하여 기존에 활용되고 있는 2가지 방법론과 본 논문에서 제안하는 앙상블 기법을 이용하는 6가지 방법론 총 8가지 방법론들을 모의실험과 실증분석을 통해서 비교, 분석한다. 모의실험은 임의로 훈련자료(training data)와 검증자료(test data)를 각각 생성해서 실시하였으며 실증분석에서는 실제적인 데이터에서 일부를 무작위로 뽑아서 훈련자료를 만들고 나머지를 검증자료로 삼아서 분석하였다. 모든 모의실험과 실증분석은 패키지 R을 이용하여 실시하였다.

5.1 모의실험

모의실험에서는 임의로 생성한 훈련자료(training data)와 검증자료(test data)에 대해서 8가지 방법론들을 비교, 분석하였다. 사용하는 8가지 방법론들은 다음과 같다.

- KRLC : kernel ridge logistic regression classification
- KRLCS : kernel ridge logistic regression classification using sub-sampling

- KRLCB1 : kernel ridge logistic regression classification using bagging with logit estimate
- KRLCB2 : kernel ridge logistic regression classification using bagging with probability estimate
- KRLCB3 : kernel ridge logistic regression classification using bagging with majority vote

- KRLCR1 : kernel ridge logistic regression classification using random forests with logit estimate
- KRLCR2 : kernel ridge logistic regression classification using random forests with probability estimate
- KRLCR3 : kernel ridge logistic regression classification using random forests with majority vote

이들 중 KRLCS에서 사용되는 sub-sampling 이라는 방법론은 Huh (2015) 에 의하여 제안되었으며 이에 대해서는 모의실험 step을 설명하면서 간단히 언급하도록 하겠다.

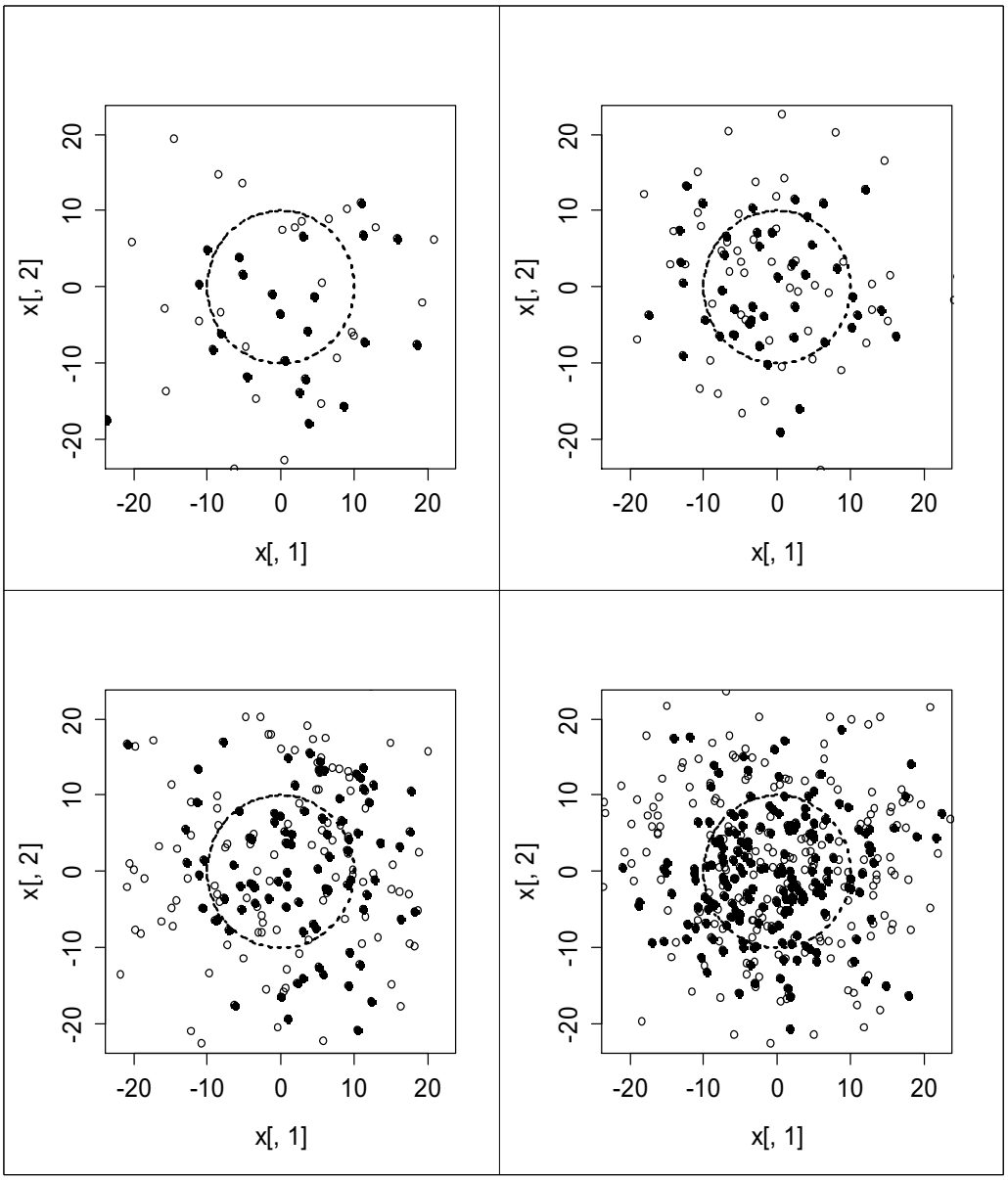
이러한 8가지 방법론들에 대해서 각각 모의실험을 실시하였다. 모의실험에 사용된 임의의 훈련자료(training data)와 검증자료(test data)에 대한 설명변수의 개수는 $p = 3, 5, 10, 20, 40$ 으로 설정하였고 이들 중 훈련자료에 대한 관측치의 개수는 $n = 50, 100, 200, 400$ 으로 설정해서 총 20가지의 상황을 가정하

였다. 그리고 검증자료에 대해서는 관측치의 개수를 $s=1000$ 으로 고정하였다. 훈련자료와 검증자료에 포함된 반응변수 y_i는 조건부 베르누이분포 $y_i|x_i \sim Bernoulli(p_i^*)$ 를 이용하여 독립적으로 생성하였다. 이 때 확률 p_i^*는 다음과 같이 지수형태(exponential form)로 설정하여 강한 비선형 관계(nonlinear relationship)를 만족하도록 하였다. 단, 설명변수의 개수가 $p=3,5$인 경우에는 $p/2 \rightarrow 2$ 로 설정하였다.

$$p_i^* = \exp\left\{-\frac{1}{(p/2)\times 10^2}\left(\sum_{j=1}^{p/2} x_{ij}^2\right)\right\},\ x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$$
$$i=1,2,\cdots,n(s) \quad , \quad j=1,2,\cdots,p$$

설명변수 x_{ij}는 정규분포 $N(0,10^2)$ 에서 독립적으로 추출하였다. 이러한 방법으로 모의실험 데이터를 생성하게 되면 1번째부터 $p/2$번째까지의 설명변수들만 반응변수에 영향력을 행사하게 된다. 그러므로 설명변수의 개수가 많아질수록 반응변수에 영향을 주지 않는 설명변수들이 늘어나게 된다. 그리고 반응변수의 분포형태는 그림 3과 같이 나타나게 된다.

<그림 3 : 모의실험 데이터의 분포형태>



위의 그림은 p차원 공간에서 정의된 모의실험 데이터를 1번째 설명변수와 2번째 설명변수로 이루어진 2차원 평면에 투영

(projection)시켜 바라봤을 때 나타나는 결과이며 검은색으로 표시된 점은 반응변수의 값이 1인 경우이고 흰색으로 표시된 점은 반응변수의 값이 0인 경우이다. 즉, 반지름의 길이가 10인 원의 외부에는 흰색 점이, 내부에는 검은색 점이 존재하는 것을 기대하고 실험자료를 생성하여 모의실험을 실시하였다.

모의실험은 다음과 같은 step을 통해서 진행한다.

* KRLC

Step1) 훈련자료와 검증자료를 각각 생성한다.

Step2) 5-fold CV(cross-validation)를 통해서 최적의 능형모수(ridge parameter) λ값을 선정한다.

Step3) Step2)에서 선정된 최적의 λ값과 훈련자료를 이용하여 $\hat{d}$ 값을 구하고 검증자료에 대해서 test MR(misclassification rate)을 계산한다.

-> Step1)부터 Step3)까지의 과정을 1000번 반복해서 총 1000개의 test MR의 수치들을 모은다.

* KRLCS (Sub-sampling)

Step1) 훈련자료와 검증자료를 각각 생성한다.

Step2) 훈련자료 중 70%를 임의로 선택해서 새로운 훈련자료를 만들고 나머지 30%를 새로운 평가자료로 사용하여 5-fold CV를 통해서 최적의 λ값을 구하고 test MR을

계산한다. 이 과정을 500번 반복하여 test MR의 값이 가장 작은 경우에 선택된 새로운 훈련자료와 이 훈련자료를 통해서 찾은 최적의 λ값을 최종적인 평가기준으로 선택한다.

Step3) Step2)에서 선정한 새로운 훈련자료와 λ값을 이용해서 $\hat{d}$ 값을 구하고 검증자료에 대해서 test MR을 계산한다.

-> Step1)부터 Step3)까지의 과정을 1000번 반복해서 총 1000개의 test MR의 수치들을 모은다.

* KRLCB1, KRLCB2, KRLCB3 (Bagging)

Step1) 훈련자료와 검증자료를 각각 생성한다.

Step2) 훈련자료에서 관측치의 개수가 같은 50개의 부트스트랩 표본들을 복원추출을 통해서 만들고 새로운 훈련자료로 삼는다. 각각의 표본에 대한 새로운 평가자료는 복원추출 시 뽑히지 않은 번호에 해당하는 관측치로 정한다.

Step3) Step2)에서 생성한 각 표본에 대해서 λ값마다 각각 $\hat{d}$값을 구하고 test MR을 계산한다.

Step4) Step3)에서 계산한 test MR의 결과를 각 λ값에 따라 정리하고 평균내서 그 결과가 가장 작은 경우의 λ값을 최종결정한다. 이에 해당하는 λ값이 여러 가지인 경우에는 그 중 가장 큰 값을 선택한다.

Step5) Step4)에서 결정한 λ값을 토대로 서로 다른 500개의 부

트스트랩 훈련자료들과 평가대상인 검증자료를 이용하여 최종적인 배깅 추정량(bagging estimator) $\widehat{\Theta^{*}_{bag}}$, $\widehat{p^{*}_{bag}}$과 이에 따른 예측값 $\widehat{y^{*}_{bag}}$을 결정한다. 이를 바탕으로 검증자료에 대한 test MR $\widehat{\pi^{*}_{bag}}$을 계산한다.

-> Step1)부터 Step5)까지의 과정을 1000번 반복해서 총 1000개의 test MR의 수치들을 모은다.

* KRLCR1, KRLCR2, KRLCR3 (Random forests)

Step1) 훈련자료와 검증자료를 각각 생성한다.

Step2) 훈련자료에서 관측치의 개수가 같은 50개의 부트스트랩 표본들을 복원추출을 통해서 만들고 새로운 훈련자료로 삼는다. 이 때 각각의 표본에 대해서 $m(\approx \sqrt{p})$개의 설명변수들을 임의로 추출해서 포함시킨다. 설명변수들의 종류는 표본마다 다르게 정한다. 각각의 표본에 대한 새로운 평가자료는 복원추출 시 뽑히지 않은 번호에 해당하는 관측치로 정한다. 물론 새로운 평가자료에 포함시킬 설명변수들의 종류는 대응되는 새로운 훈련자료의 경우와 같게 정하면 될 것이다.

Step3) Step2)에서 생성한 각 표본에 대해서 λ값마다 각각 $\hat{d}$값을 구하고 test MR을 계산한다.

Step4) Step3)에서 계산한 test MR의 결과를 각 λ값에 따라 정

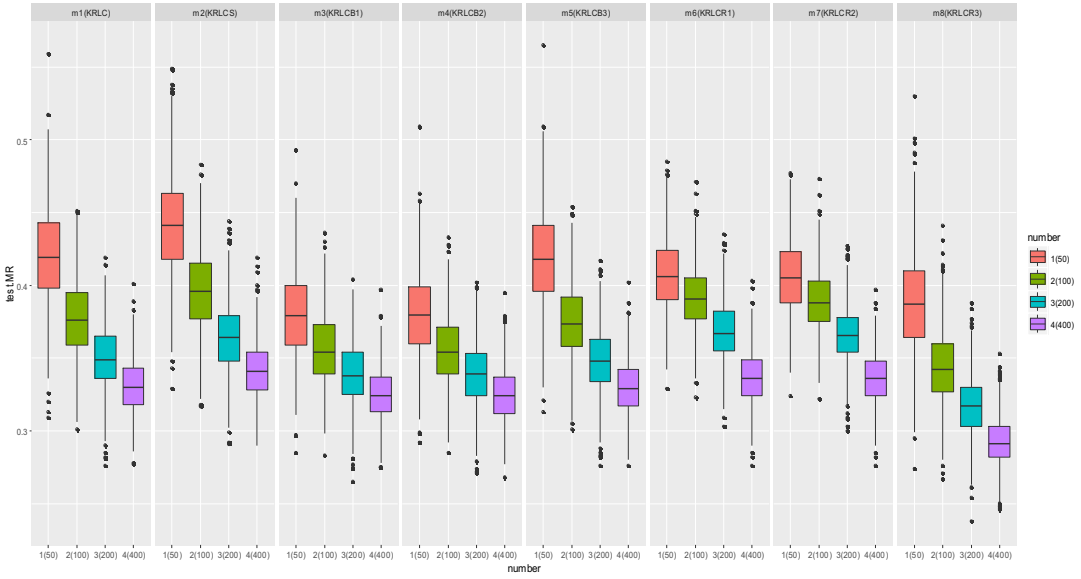
리하고 평균내서 그 결과가 가장 작은 경우의 λ값을 최종결정한다. 이에 해당하는 λ값이 여러 가지인 경우에는 그 중 가장 큰 값을 선택한다.

Step5) Step4)에서 결정한 λ값을 토대로 서로 다른 500개의 부트스트랩 훈련자료들과 평가대상인 검증자료를 이용하여 최종적인 랜덤포레스트 추정량(random forests estimator) $\widehat{\Theta_{rf}^{*}}$, $\widehat{p_{rf}^{*}}$과 이에 따른 예측값 $\widehat{y_{rf}^{*}}$을 결정한다. 이 때 각각의 부트스트랩 훈련자료에 대해서 $m(\approx \sqrt{p})$개의 설명변수들을 임의로 추출해서 포함시킨다. 설명변수들의 종류는 부트스트랩 훈련자료마다 다르게 정한다. 물론 평가대상인 검증자료에 포함시킬 설명변수들의 종류는 부트스트랩 훈련자료들의 기준을 따라서 정하면 될 것이다. 이를 바탕으로 검증자료에 대한 test MR $\widehat{\pi_{rf}^{*}}$을 계산한다.

-> Step1)부터 Step5)까지의 과정을 1000번 반복해서 총 1000개의 test MR의 수치들을 모은다.

위에서 설명한 과정대로 각각의 방법론에 대해 총 20가지의 상황을 가정하여 컴퓨터 모의실험을 실시하고 이를 통해서 얻은 검증 오분류율에 대한 결과를 그림 4~8과 표 1~5로 정리하였다. 평균값과 분산이 작을수록 이에 해당하는 방법론의 분류정확도와 신뢰성이 더 높다는 것을 의미한다. 따라서 이러한 결과를 제시하는 방법론의 성능이 더 좋다고 말할 수 있다. 분류정확도는 상자그림의 높이, 신뢰성은 상자그림의 크기를 통해서 판단할 수 있다. 또한 검증 오분류율에 대한 전체적인 분포의 특징도 상자그림을 통해서 시각적으로 확인해 볼 수 있다. 이를 통해서 어떠한 방법론의 성능이 더 좋은지 판단할 수 있겠다. 상자그림과 표는 모두 설명변수의 개수인 p를 기준으로 정리되었으며 추가적으로 표의 경우에는 각 상황마다 가장 좋은 성능을 보인 방법론에 대해 해당하는 수치를 강조해서 표시하였다.

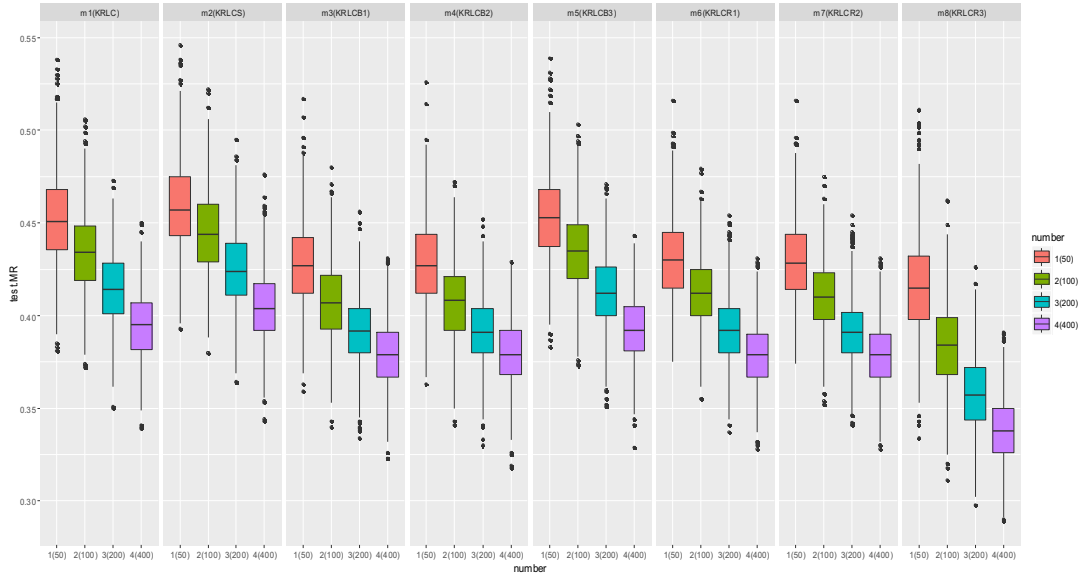
<그림 4 : $p=3$인 경우의 모의실험 결과에 대한 상자그림>



<표 1 : $p=3$인 경우의 모의실험 결과에 대한 표>

		KRLC	KRLCS	KRLCB 1	KRLCB 2	KRLCB 3	KRLCR 1	KRLCR 2	KRLCR 3
$n=50$	mean	0.4199	0.4419	**0.3800**	**0.3800**	0.4186	0.4067	0.4056	0.3881
	sd	0.0331	0.0343	**0.0293**	**0.0293**	0.0319	0.0245	0.0247	0.0343
$n=100$	mean	0.3774	0.3967	0.3556	0.3554	0.3754	0.3912	0.3889	**0.3438**
	sd	0.0258	0.0276	0.0244	0.0243	0.0253	0.0215	0.0215	**0.0250**
$n=200$	mean	0.3500	0.3638	0.3390	0.3388	0.3487	0.3681	0.3661	**0.3170**
	sd	0.0217	0.0234	0.0213	0.0212	0.0216	0.0198	0.0193	**0.0204**
$n=400$	mean	0.3308	0.3414	0.3250	0.3249	0.3298	0.3367	0.3362	**0.2925**
	sd	0.0179	0.0191	0.0177	0.0178	0.0178	0.0182	0.0178	**0.0169**

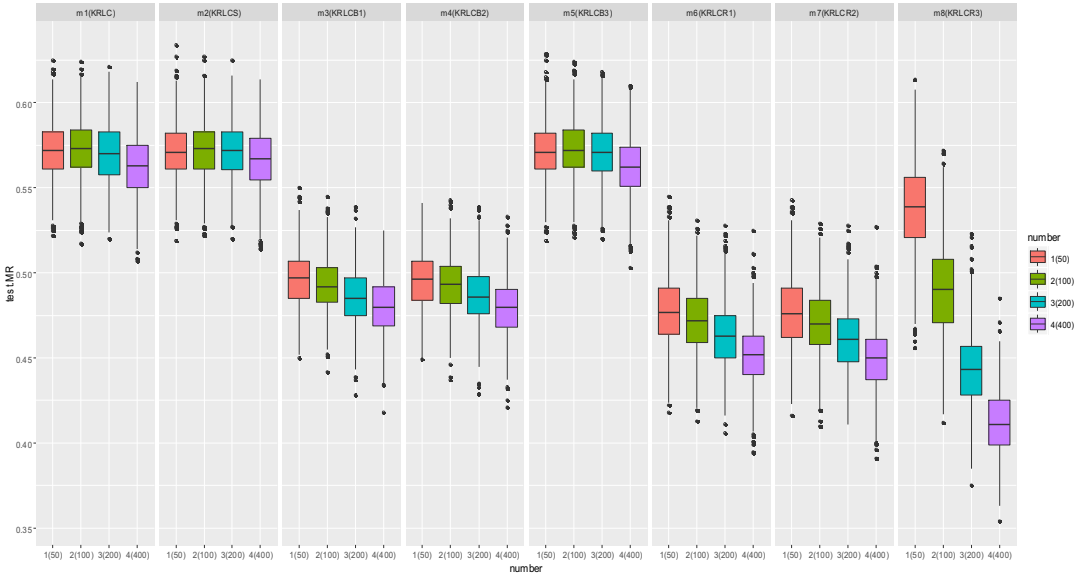
<그림 5 : $p=5$인 경우의 모의실험 결과에 대한 상자그림>



<표 2 : $p=5$인 경우의 모의실험 결과에 대한 표>

		KRLC	KRLCS	KRLCB 1	KRLCB 2	KRLCB 3	KRLCR 1	KRLCR 2	KRLCR 3
$n=50$	mean	0.4517	0.4588	0.4278	0.4282	0.4530	0.4308	0.4293	**0.4159**
	sd	0.0240	0.0243	0.0233	0.0235	0.0234	0.0221	0.0221	**0.0266**
$n=100$	mean	0.4342	0.4450	0.4074	0.4071	0.4349	0.4122	0.4101	**0.3838**
	sd	0.0221	0.0232	0.0212	0.0213	0.0219	0.0193	0.0191	**0.0226**
$n=200$	mean	0.4142	0.4250	0.3923	0.3919	0.4130	0.3924	0.3915	**0.3576**
	sd	0.0194	0.0211	0.0184	0.0182	0.0190	0.0182	0.0180	**0.0210**
$n=400$	mean	0.3949	0.4047	0.3797	0.3797	0.3929	0.3781	0.3780	**0.3382**
	sd	0.0182	0.0187	0.0178	0.0178	0.0178	0.0172	0.0173	**0.0177**

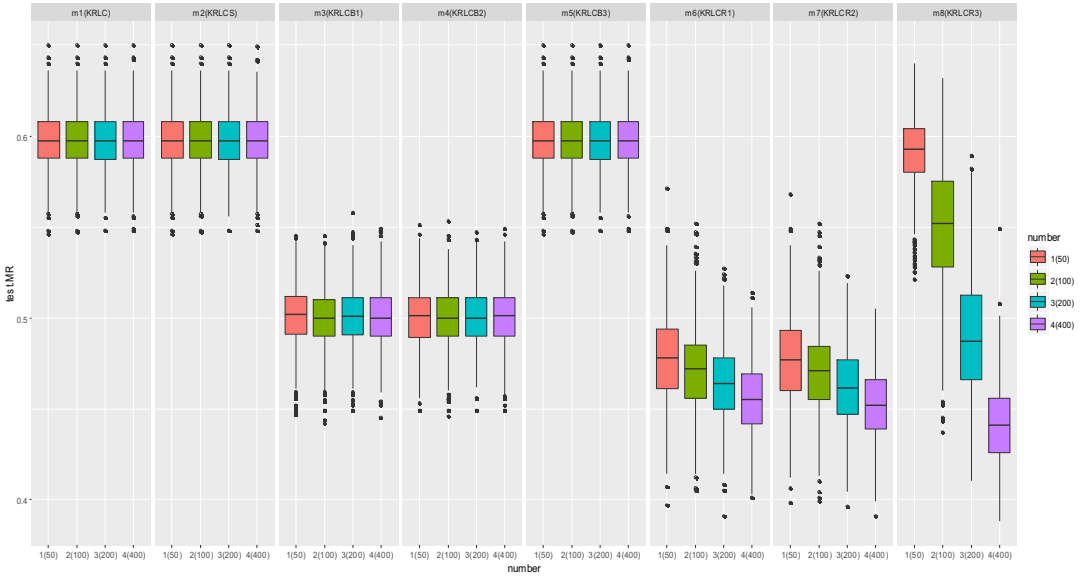
<그림 6 : $p=10$인 경우의 모의실험 결과에 대한 상자그림>



<표 3 : $p=10$인 경우의 모의실험 결과에 대한 표>

		KRLC	KRLCS	KRLCB 1	KRLCB 2	KRLCB 3	KRLCR 1	KRLCR 2	KRLCR 3
$n=50$	mean	0.5717	0.5712	0.4962	0.4959	0.5714	0.4771	**0.4765**	0.5381
	sd	0.0162	0.0163	0.0160	0.0161	0.0162	0.0203	**0.0204**	0.0252
$n=100$	mean	0.5727	0.5727	0.4926	0.4928	0.5727	0.4723	**0.4711**	0.4906
	sd	0.0166	0.0165	0.0158	0.0159	0.0164	0.0186	**0.0189**	0.0272
$n=200$	mean	0.5704	0.5721	0.4858	0.4865	0.5708	0.4627	0.4610	**0.4439**
	sd	0.0168	0.0170	0.0162	0.0165	0.0167	0.0185	0.0188	**0.0221**
$n=400$	mean	0.5627	0.5670	0.4799	0.4791	0.5624	0.4514	0.4491	**0.4119**
	sd	0.0178	0.0172	0.0165	0.0165	0.0173	0.0181	0.0187	**0.0186**

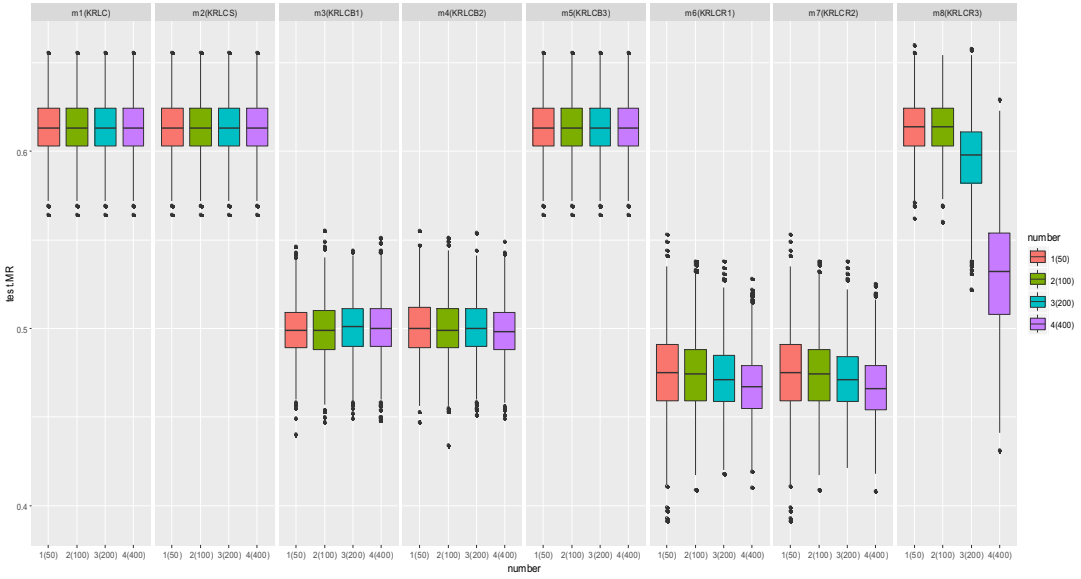
<그림 7 : $p=20$인 경우의 모의실험 결과에 대한 상자그림>



<표 4 : $p=20$인 경우의 모의실험 결과에 대한 표>

		KRLC	KRLCS	KRLCB 1	KRLCB 2	KRLCB 3	KRLCR 1	KRLCR 2	KRLCR 3
$n=50$	mean	0.5974	0.5974	0.5009	0.5002	0.5974	0.4775	**0.4771**	0.5915
	sd	0.0148	0.0148	0.0158	0.0163	0.0148	0.0239	**0.0241**	0.0190
$n=100$	mean	0.5975	0.5975	0.5000	0.5004	0.5975	0.4712	**0.4701**	0.5501
	sd	0.0148	0.0148	0.0158	0.0158	0.0148	0.0220	**0.0225**	0.0330
$n=200$	mean	0.5975	0.5975	0.5004	0.5003	0.5975	0.4640	**0.4621**	0.4896
	sd	0.0148	0.0147	0.0160	0.0153	0.0148	0.0209	**0.0214**	0.0328
$n=400$	mean	0.5976	0.5976	0.5006	0.5000	0.5978	0.4549	0.4522	**0.4414**
	sd	0.0148	0.0148	0.0161	0.0159	0.0148	0.0194	0.0200	**0.0220**

<그림 8 : $p=40$인 경우의 모의실험 결과에 대한 상자그림>



<표 5 : $p=40$인 경우의 모의실험 결과에 대한 표>

		KRLC	KRLCS	KRLCB 1	KRLCB 2	KRLCB 3	KRLCR 1	KRLCR 2	KRLCR 3
$n=50$	mean	0.6134	0.6134	0.4991	0.5004	0.6134	**0.4753**	**0.4753**	0.6136
	sd	0.0157	0.0157	0.0158	0.0166	0.0157	**0.0239**	**0.0239**	0.0158
$n=100$	mean	0.6134	0.6134	0.4992	0.4994	0.6134	**0.4739**	**0.4739**	0.6140
	sd	0.0157	0.0157	0.0163	0.0164	0.0157	**0.0213**	**0.0213**	0.0152
$n=200$	mean	0.6134	0.6134	0.5003	0.5001	0.6134	0.4716	**0.4715**	0.5958
	sd	0.0157	0.0157	0.0160	0.0155	0.0157	0.0195	**0.0195**	0.0218
$n=400$	mean	0.6134	0.6134	0.5000	0.4985	0.6134	0.4672	**0.4666**	0.5305
	sd	0.0157	0.0157	0.0159	0.0161	0.0157	0.0185	**0.0187**	0.0320

위의 5가지 상자그림과 표를 통해서 볼 때 설명변수의 개수 p와는 상관없이 대체적으로 훈련자료에 대한 관측치의 개수 n(number)이 커질수록 8가지 방법론들 모두 검증 오분류율(test misclassification rate)의 평균값이 낮아짐을 확인할 수 있다. 즉, 훈련자료에 대한 관측치의 개수가 크다는 것은 그만큼 사전정보를 많이 가지고 있다는 의미이고 이에 따른 분류의 정확도는 높아진다고 말할 수 있는 것이다. 그리고 전체적으로 살펴보면 랜덤포레스트(random forests) 기법을 적용했을 경우에 가장 우수한 결과를 얻게 됨을 확인할 수 있을 것이다. 3장을 통해 앙상블 기법(ensemble method)을 소개하면서 언급했듯이 랜덤포레스트 기법에 의해 추정량을 구하게 되면 분산을 크게 줄일 수 있을 뿐만 아니라 부트스트랩(bootstrap)을 실행 시 설명변수들 중 일부만을 선택해서 사용하게 되므로 뽑히는 표본들 간의 연관성도 줄일 수 있게 된다. 상자그림과 표를 통해서 비교해보면 랜덤포레스트 기법에 의해 산출된 검증 오분류율의 평균값이 다른 방법론들에 의해 얻어진 검증 오분류율의 평균값들보다 대체적으로 더 작다는 것을 확인할 수 있을 것이다. 즉, 다른 방법론들에 비해 상자그림이 더 밑으로 내려가는 경향을 나타내는 것을 볼 때 분류정확도 측면에서 랜덤포레스트 기법을 적용한 방법론의 성능이 더 우수하다고 말할 수 있다. 또한 배깅(bagging) 기법에 의한 결과도 기존에 사용되고 있는 방법론들에 의한 결과들과 비교해보면 대체적으로 상자그림의 높이가 더 낮게 나타나고 있으므로 분류정확도 측

면에서 더 바람직하다는 것을 알 수 있다. 결과적으로 이 모의실험을 통해서 본 논문에서 제안하고자 하는 커널 능형 로지스틱 회귀분류법(kernel ridge logistic regression classification)에 앙상블 기법을 적용하는 방법론이 기존에 사용되고 있는 일반적인 방법론들에 비해 더 우수한 성능을 보임을 확인할 수 있었다.

5.2 실증분석

실증분석에서는 총 11개의 실제적인 데이터들을 이용하여 8가지 방법론들(KRLC, KRLCS, KRLCB1, KRLCB2, KRLCB3, KRLCR1, KRLCR2, KRLCR3)의 성능을 비교, 분석하였다. 모든 데이터들은 UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html)에서 얻었다. 그리고 각 데이터 안에 있는 관측치들 중 50%$\left(0.5 = \frac{n}{N}\right)$, 즉, 절반을 무작위로 추출하여 훈련자료(training data)를 구성하고 추출되지 않은 남은 관측치들을 검증자료(test data)로 사용하였다. 모의실험과 동일한 step으로 각각의 방법론마다 100번의 반복을 실시하였으며 반복 시마다 추출하는 훈련자료는 모두 다르게 하였다. 관측치들 중 결측치가 존재하는 경우에는 이를 제거하였으며 문자형이나 범주형인 설명변수가 있는 경우에도 이를 제거하고 분석을 실시하였다. 분석에 사용된 데이터들과 관련된 내용은 아래의 표

6을 통해서 정리하도록 하겠다.

<표 6 : 실증분석에 사용된 데이터>

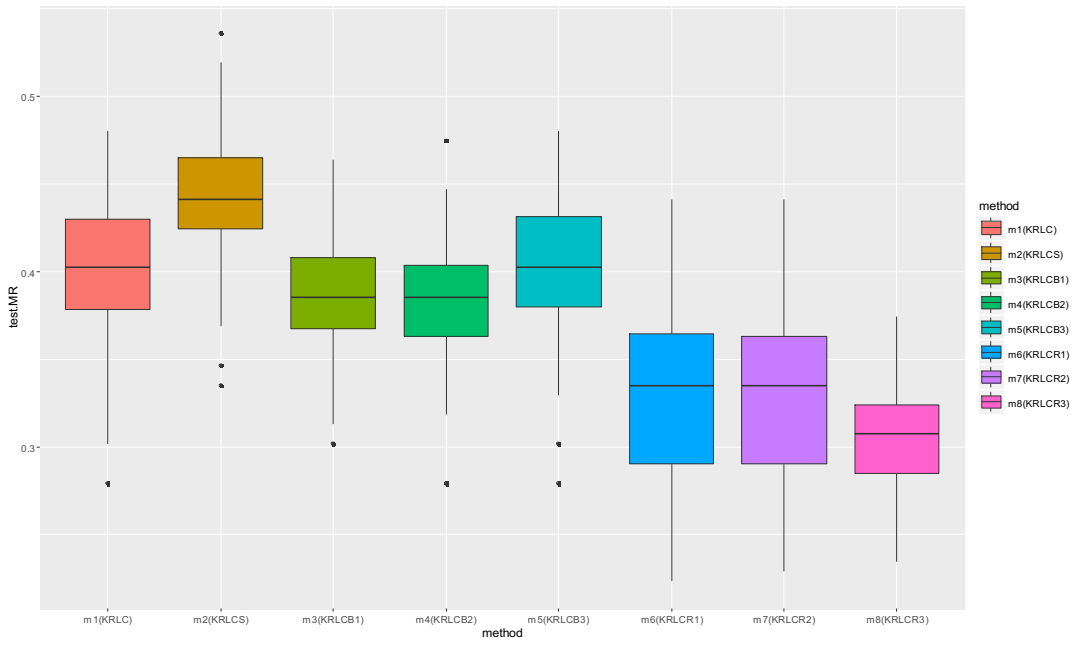
데이터 이름	관측치의 개수 N	반응변수의 값이 1인 관측치의 개수 n^*	설명변수의 개수 p	비고
Cylinder Bands	512 -> 358	133	38 -> 23	2개의 범주를 가지는 반응변수 문자형 설명변수 15개 제거 연속형 설명변수 23개 범주형 설명변수 0개 결측치 154개 제거
Forest Type Mapping	326	136	27	4개의 범주를 가지는 반응변수 -> 2개의 범주로 조정 문자형 설명변수 0개 연속형 설명변수 27개 범주형 설명변수 0개 결측치 없음
Dow Jones Index	750 -> 720	330	15 -> 13	2개의 범주를 가지는 반응변수 문자형 설명변수 2개 제거 연속형 설명변수 13개 범주형 설명변수 0개 결측치 30개 제거
Haberman's Survival	306	225	3	2개의 범주를 가지는 반응변수 문자형 설명변수 0개 연속형 설명변수 3개 범주형 설명변수 0개 결측치 없음
Ionosphere	351	225	34	2개의 범주를 가지는 반응변수 문자형 설명변수 0개 연속형 설명변수 34개 범주형 설명변수 0개 결측치 없음

Pima Indians Diabetes	768	268	8	2개의 범주를 가지는 반응변수 문자형 설명변수 0개 연속형 설명변수 8개 범주형 설명변수 0개 결측치 없음
Statlog (Heart)	270	183	13 -> 5	2개의 범주를 가지는 반응변수 문자형 설명변수 0개 연속형 설명변수 5개 범주형 설명변수 8개 제거 결측치 없음
Blood Transfusio -n Service Center	748	178	4	2개의 범주를 가지는 반응변수 문자형 설명변수 0개 연속형 설명변수 4개 범주형 설명변수 0개 결측치 없음
Breast Tissue	106	54	9	6개의 범주를 가지는 반응변수 -> 2개의 범주로 조정 문자형 설명변수 0개 연속형 설명변수 9개 범주형 설명변수 0개 결측치 없음
Urban Land Cover	168	76	147	9개의 범주를 가지는 반응변수 -> 2개의 범주로 조정 문자형 설명변수 0개 연속형 설명변수 147개 범주형 설명변수 0개 결측치 없음
Statlog (Australian Credit Approval)	690 -> 502	355	14 -> 5	2개의 범주를 가지는 반응변수 문자형 설명변수 0개 연속형 설명변수 5개 범주형 설명변수 9개 제거 결측치 188개 제거

#1) Cylinder Bands Data

이 데이터는 공장에서 만들어진 cylinder와 관련한 정보를 담고 있다. 설명변수들은 각각의 cylinder들에 대한 크기와 생산할 때 사용된 재료의 양, 소요시간, 온도 등과 관련한 정보를 포함하고 있다. 이러한 요소들을 통해 만들어진 cylinder의 type이 band인지 no band인지를 판단하고자 하는 것이 목적이라고 할 수 있겠다. 분석하기 전 데이터에 포함된 154개의 결측치와 15개의 문자형 설명변수는 제거하였다.

<그림 9 : Cylinder Bands Data 실증분석 상자그림>



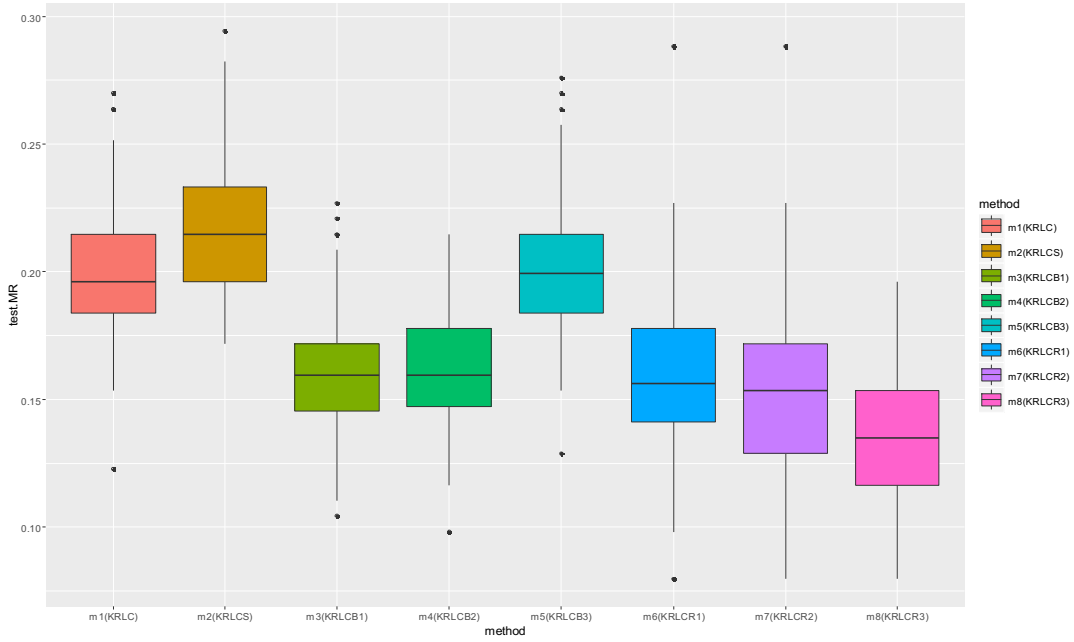
실증분석 결과 앙상블 기법을 적용하는 방법론이 기존에 사용되고 있는 방법론들보다 대체적으로 분류정확도 측면에서 더 좋은

성능을 보이고 있음을 확인할 수 있었다. 그리고 상자그림을 통해서 랜덤포레스트 기법을 적용하는 방법론의 성능이 다른 방법론들과 비교했을 때 가장 우수하다는 사실도 파악할 수 있었다. 이를 통해 본 논문에서 제안하고자 하는 방법론이 일반적인 방법론들에 비해 더 좋은 성능을 보인다고 판단할 수 있었다.

#2) Forest Type Mapping Data

이 데이터는 일본 내에 있는 여러 숲 지형의 특성을 분석하기 위해서 만들어졌다. 설명변수들은 모두 각 숲 지형의 지질에 포함된 성분의 수치, 풍화정도 등과 관련된 정보를 담고 있다. 분석을 위해 4개의 범주로 구성되어 있는 반응변수를 2개의 범주로 재조정하였다. 재조정은 숲이 형성될 수 있는 지역과 그렇지 않은 지역을 기준으로 실시하였다.

<그림 10 : Forest Type Mapping Data 실증분석 상자그림>

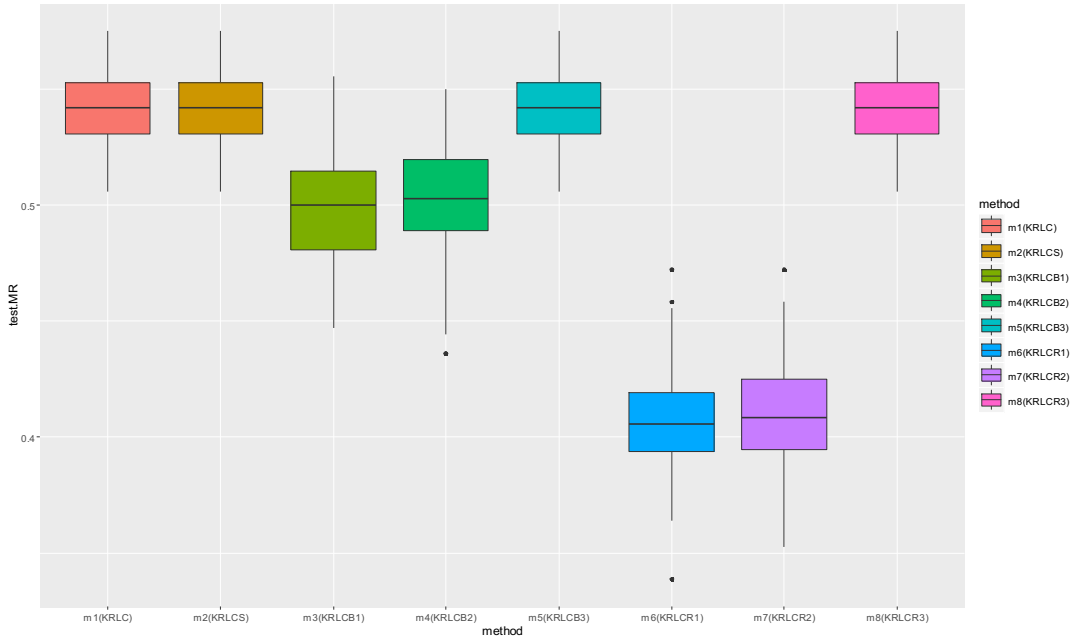


실증분석 결과 앙상블 기법을 적용하는 방법론이 일반적인 방법론들에 비해 더 좋은 성능을 보이고 있음을 확인할 수 있었다. 그리고 상자그림을 통해서 배깅 기법을 적용했을 때 신뢰성 측면에서 가장 바람직한 결과를 얻을 수 있고 랜덤포레스트 기법을 적용했을 때 분류정확도 측면에서 가장 바람직한 결과를 얻을 수 있다는 사실도 파악할 수 있었다. 이를 통해 본 논문에서 제안하고자 하는 앙상블 기법을 적용하는 방법론이 기존에 사용되고 있는 방법론들에 비해 더 좋은 성능을 보인다고 판단할 수 있었다.

#3) Dow Jones Index Data

이 데이터는 경제학 분야에서 유명하다고 알려진 Dow Jones Index와 관련한 정보를 담고 있다. 설명변수들은 일주일동안의 주식거래량, 근무자들의 근무시간 등을 포함하고 있으며 분석목적은 주식의 분기(상반기, 하반기)를 구분하는 것이라고 할 수 있겠다. 분석하기 전 30개의 결측치와 2개의 문자형 설명변수는 제거하였다.

<그림 11 : Dow Jones Index Data 실증분석 상자그림>



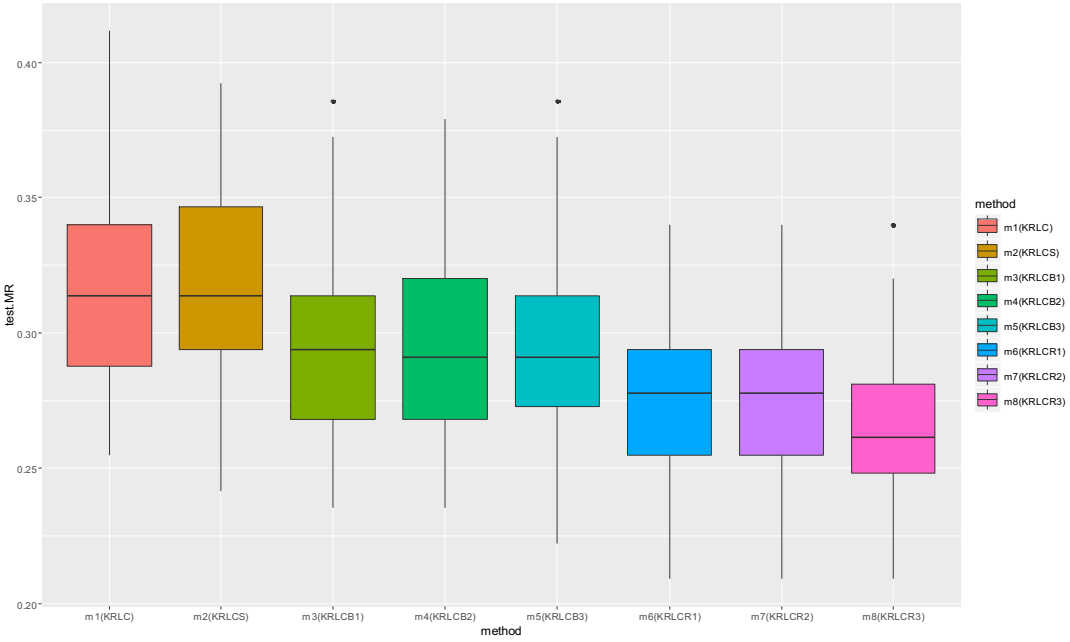
실증분석 결과 앙상블 기법을 적용하는 3가지 방법론들 중 다중투표를 이용하는 방법론은 기존에 사용되고 있는 방법론들에 비해 성능의 향상이 나타나지 않았다. 하지만 로짓추정량 및 확률추정

량과 관련된 앙상블 기법을 적용하는 방법론의 경우에는 일반적인 방법론들보다 분류정확도 측면에서 확실히 좋은 성능을 보이고 있음을 확인할 수 있었다. 이를 통해 앙상블 기법을 적용하는 방법론이 기존에 사용되고 있는 방법론들보다 대체적으로 더 효율성이 높다고 판단할 수 있었다.

#4) Haberman’s Survival Data

이 데이터는 1958년부터 1970년 사이에 University of Chicago's Billings Hospital에서 유방암(breast cancer) 수술을 받은 환자들과 관련된 정보를 담고 있다. 설명변수들은 환자의 나이, 수술을 받은 기간, 양성 반응 정도를 나타내고 있으며 반응변수는 5년 이내에 환자가 사망했는지의 여부와 관련되어 있다.

<그림 12 : Haberman's Survival Data 실증분석 상자그림>

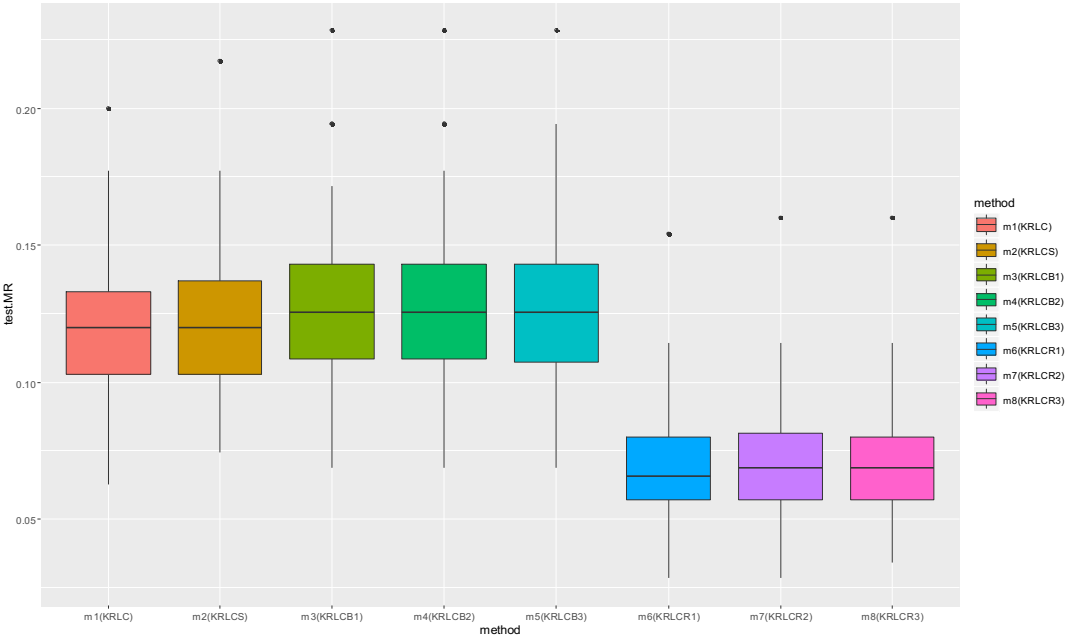


실증분석 결과 앙상블 기법을 적용하는 방법론이 일반적인 방법론들에 비해 전체적으로 분류정확도와 신뢰성 측면에서 더 좋은 성능을 보이고 있음을 확인할 수 있었다. 이를 통해 본 논문에서 제안하고자 하는 방법론이 기존에 사용되고 있는 방법론들보다 대체적으로 더 효율성이 높다고 판단할 수 있었다. 그리고 랜덤포레스트 기법을 적용하는 방법론이 다른 방법론들과 비교했을 때 가장 우수한 성능을 보인다는 사실도 위의 상자그림을 통해서 추가적으로 파악할 수 있었다.

#5) Ionosphere Data

이 데이터는 1989년 Goose Bay system에 의해서 수집되었으며 지구의 전리층(ionosphere)을 분석하여 얻어진 정보를 담고 있다. 설명변수들은 모두 전리층에 있는 자유전자(free electrons)에 대한 특성, 이동속도 등을 수치화하여 얻어졌다. 최종목표는 자유전자의 전리층 경로 이탈여부를 구분해내는 것이다.

<그림 13 : Ionosphere Data 실증분석 상자그림>



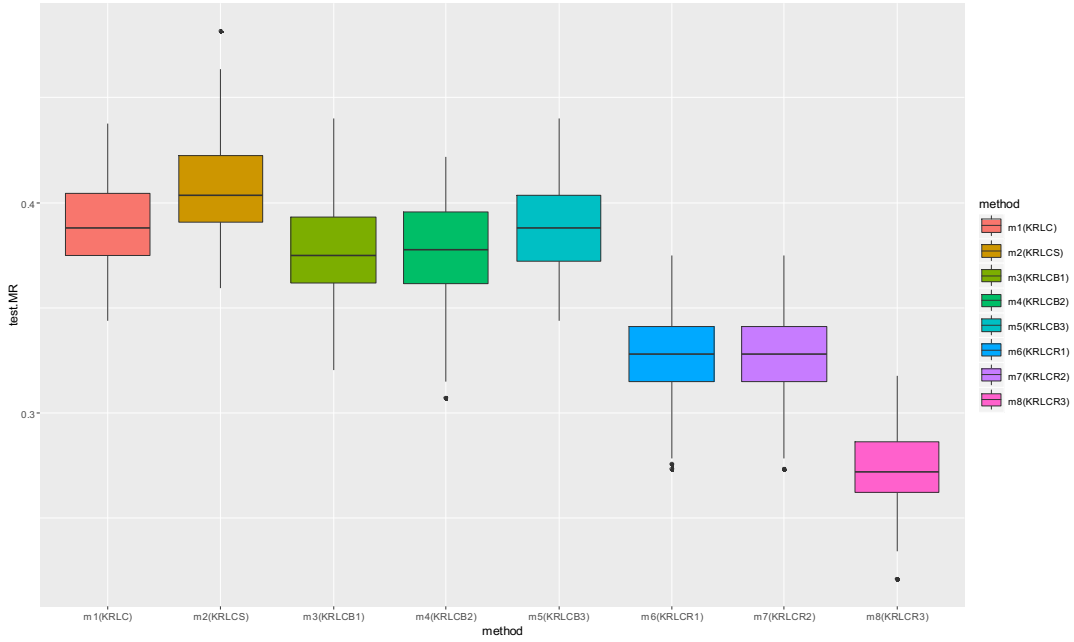
실증분석 결과 다른 방법론들의 결과와 비교했을 때 랜덤포레스트 기법을 적용하는 방법론에 의해서 산출된 검증 오분류율의 평균값과 분산이 가장 작게 나온다는 것을 쉽게 파악할 수 있었다. 이를 통해 랜덤포레스트 기법을 적용하는 방법론이 다른 방법론들

에 비해 분류정확도와 신뢰성 측면에서 가장 우수한 성능을 보임을 확인할 수 있었다. 다만 배깅 기법을 적용하는 방법론의 경우에는 일반적인 방법론들에 비해 성능의 향상이 나타나지 않았다.

#6) Pima Indians Diabetes Data

이 데이터는 Pima Indian 유전자를 가진 21세 이상의 여성 환자와 관련한 정보를 담고 있다. 설명변수들은 환자의 키, 몸무게, 혈압, 그리고 임신 경험 여부 등과 관련되어 있다. 그리고 반응변수는 신장(kidney) 질환의 유무를 나타내고 있다.

<그림 14 : Pima Indians Diabetes Data 실증분석 상자그림>

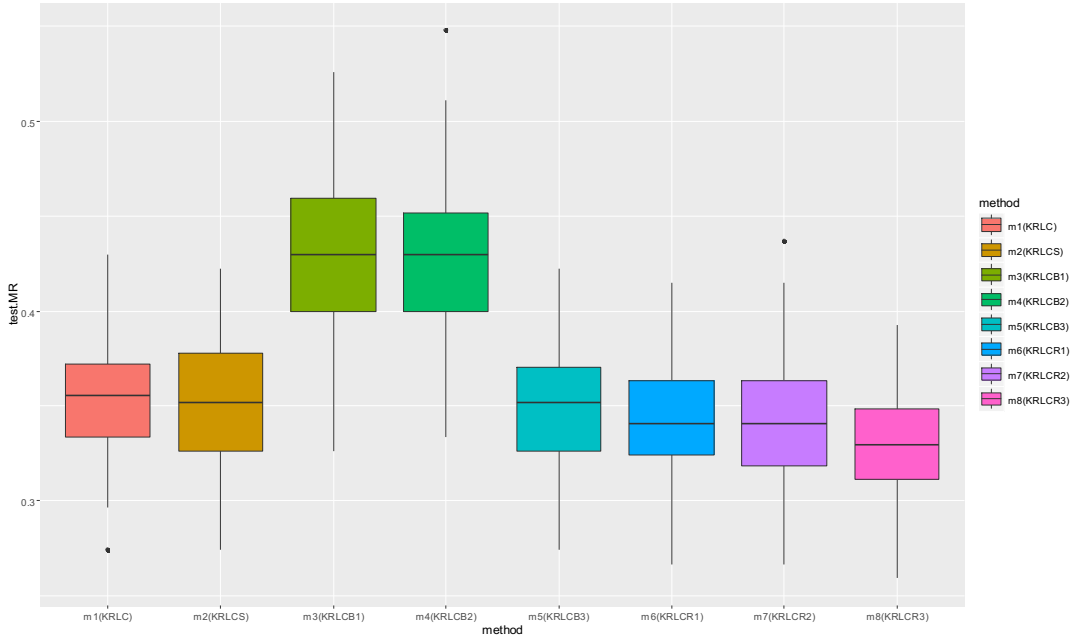


실증분석 결과 앙상블 기법을 적용하는 방법론이 기존에 사용되고 있는 방법론들에 비해 대체적으로 분류정확도 측면에서 더 좋은 성능을 보인다는 것을 확인할 수 있었다. 특히 랜덤포레스트 기법을 적용하는 방법론의 경우에는 다른 방법론들에 비해서 확실히 좋은 성능을 보인다는 사실을 위의 상자그림을 통해서 쉽게 확인할 수 있을 것이다. 결과적으로 본 논문을 통해 제시하고자 하는 앙상블 기법을 적용하는 방법론이 일반적인 방법론들보다 더 우수한 성능을 보임을 입증할 수 있었다.

#7) Statlog (Heart) Data

이 데이터는 환자들의 심장질환(heart disease) 유무를 구분하기 위해 만들어졌다. 언제, 어디에서 조사가 실시되었는지는 알려져 있지 않았다. 설명변수들은 나이, 혈압, 성별 등의 환자의 상태와 관련되어 있으며 반응변수는 심장질환 유무를 나타내고 있다. 분석하기 전 8개의 범주형 설명변수는 제거하였다.

<그림 15 : Statlog (Heart) Data 실증분석 상자그림>



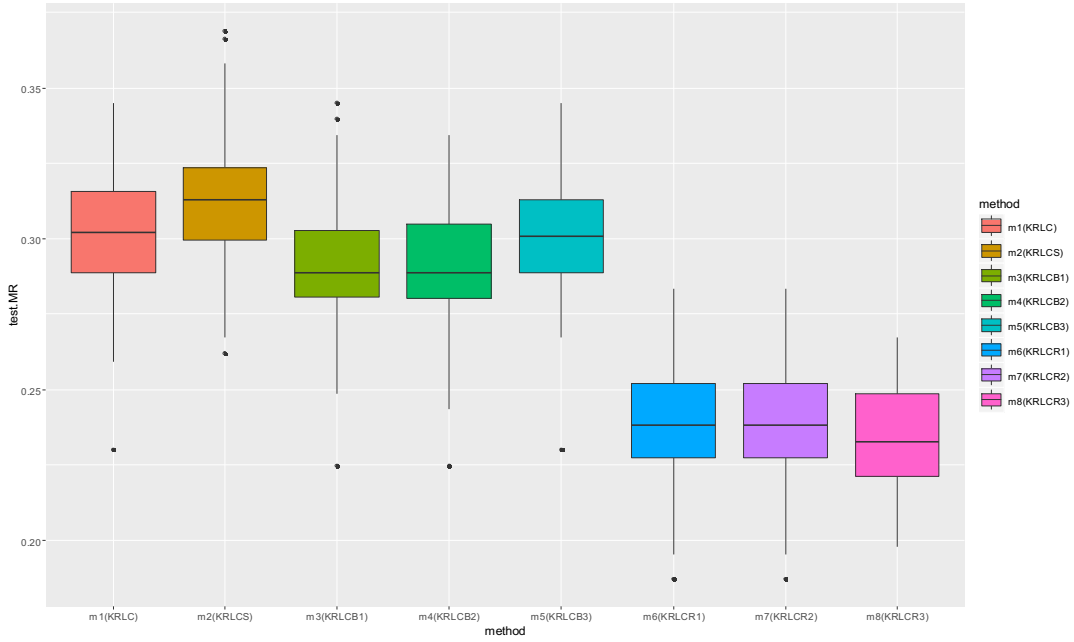
실증분석 결과 로짓추정량 및 확률추정량과 관련된 배깅 기법을 적용하는 방법론이 다른 방법론들에 비해 분류정확도와 신뢰성 측면에서 가장 좋지 않은 성능을 보였다. 하지만 랜덤포레스트 기법이나 다중투표와 관련된 배깅 기법을 적용하는 방법론의 경우에는

큰 차이를 보이지는 않았지만 기존에 사용되고 있는 방법론들에 비해 대체적으로 분류정확도 측면에서 더 좋은 성능을 나타냈다. 그러므로 전체적으로 볼 때 본 연구에서 제안하고자 하는 방법론이 일반적인 방법론들보다 더 성능이 좋다고 판단된다.

#8) Blood Transfusion Service Center Data

이 데이터는 2007년 3월에 Taiwan의 Hsin-Chu City 내에 있는 Blood Transfusion Service Center에서 방문자에 대해 실시한 헌혈 여부 조사를 통해 만들어졌다. 즉, 헌혈 여부를 구분해 내는 것이 목적이라고 할 수 있으며 설명변수들은 방문 횟수, 체류시간 등을 나타내고 있다.

<그림 16 : Blood Transfusion Service Center Data 실증분석 상자그림>

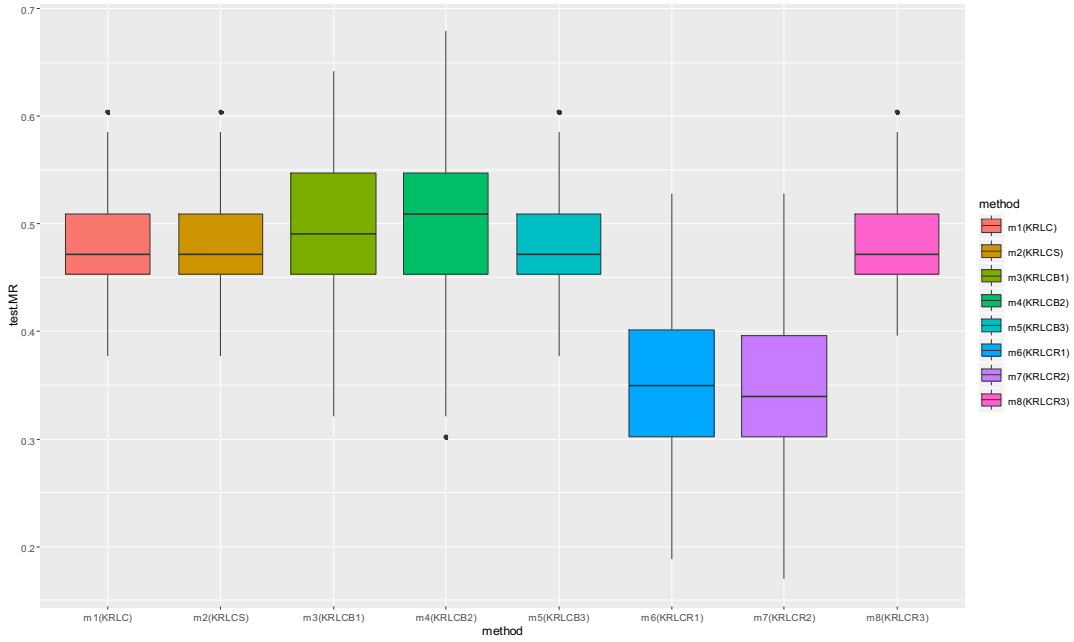


실증분석 결과 다른 방법론들의 결과와 비교했을 때 랜덤포레스트 기법을 적용하는 방법론에 의해서 산출된 검증 오분류율의 평균값과 분산이 가장 작게 나온다는 것을 쉽게 확인할 수 있었다. 그리고 배깅 기법을 적용하는 방법론의 경우에도 기존에 사용되고 있는 방법론들에 비하면 대체적으로 검증 오분류율의 평균값과 분산이 더 작게 나온다는 사실도 추가적으로 살펴볼 수 있었다. 결과적으로 본 논문에서 제안하고자 하는 앙상블 기법을 적용하는 방법론이 일반적인 방법론들보다 분류정확도와 신뢰성 측면에서 더 우수한 성능을 보인다는 것을 입증할 수 있었다.

#9) Breast Tissue Data

이 데이터는 제조된 breast tissue의 특성과 관련한 정보를 가지고 있다. 설명변수들은 자외선 주파수 등의 제조 시 사용한 재료들의 특성을 나타내고 있으며 반응변수는 총 6가지의 type들을 보여주고 있다. 분석을 위해 반응변수의 범주를 2가지로 재조정하였으며 이 때 서로 유사한 성질을 나타낸다고 여겨지는 type들을 하나의 group으로 묶는 방법을 적용하였다.

<그림 17 : Breast Tissue Data 실증분석 상자그림>



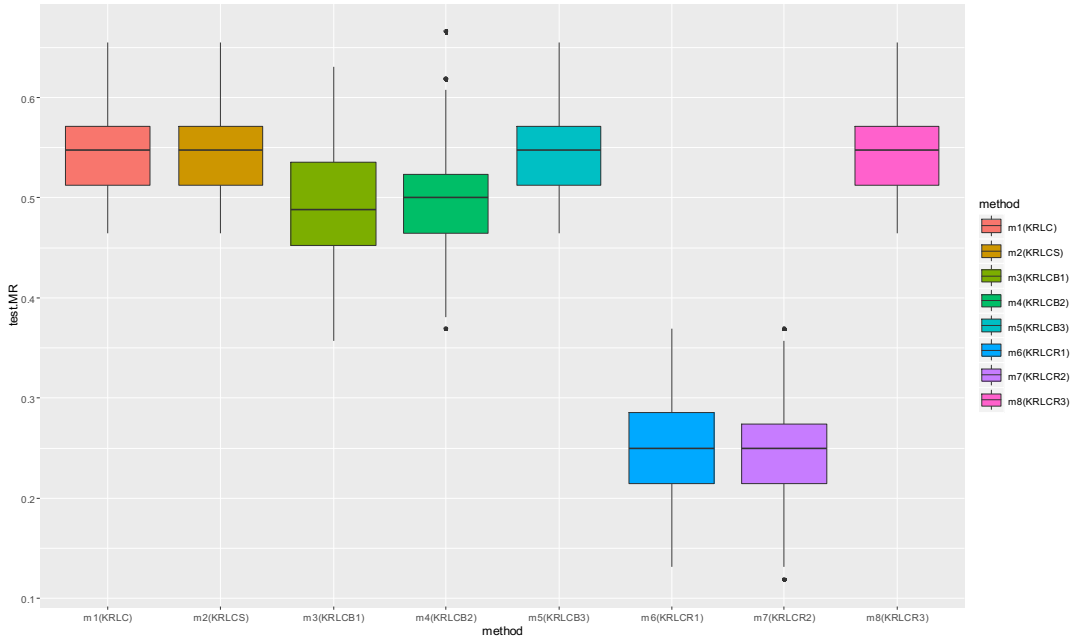
실증분석 결과 배깅 기법을 적용하는 방법론이 기존에 사용되고 있는 방법론들과 비교했을 때 확실히 좋다고 할 만한 결과를 내놓지는 못하였다. 하지만 랜덤포레스트 기법을 적용하는 방법론의

경우에는 다른 방법론들에 비해 분류정확도 측면에서 더 바람직한 결과를 보여주고 있음을 확인할 수 있었다. 그러므로 본 논문에서 제안하고자 하는 방법론이 일반적인 방법론들보다 더 성능이 좋다고 판단된다.

#10) Urban Land Cover Data

이 데이터는 2014년 일본에 있는 여러 지역들 중 하나인 하마야 정(Hamaya, Kanagawa)을 조사하면서 얻어졌다. 분석 목적은 이 지역 안에 있는 소지역들의 사용 목적을 구분하는 것이다. 설명변수들은 모두 각 소지역들의 색채, 크기, 밀도, 모양 등과 관련된 수치로 이루어져 있다. 분석을 위해 9개의 범주로 구성되어 있는 반응변수를 2개의 범주로 재조정하였다. 재조정은 인위적인 개발의 여부를 기준으로 삼아서 실시하였다.

<그림 18 : Urban Land Cover Data 실증분석 상자그림>

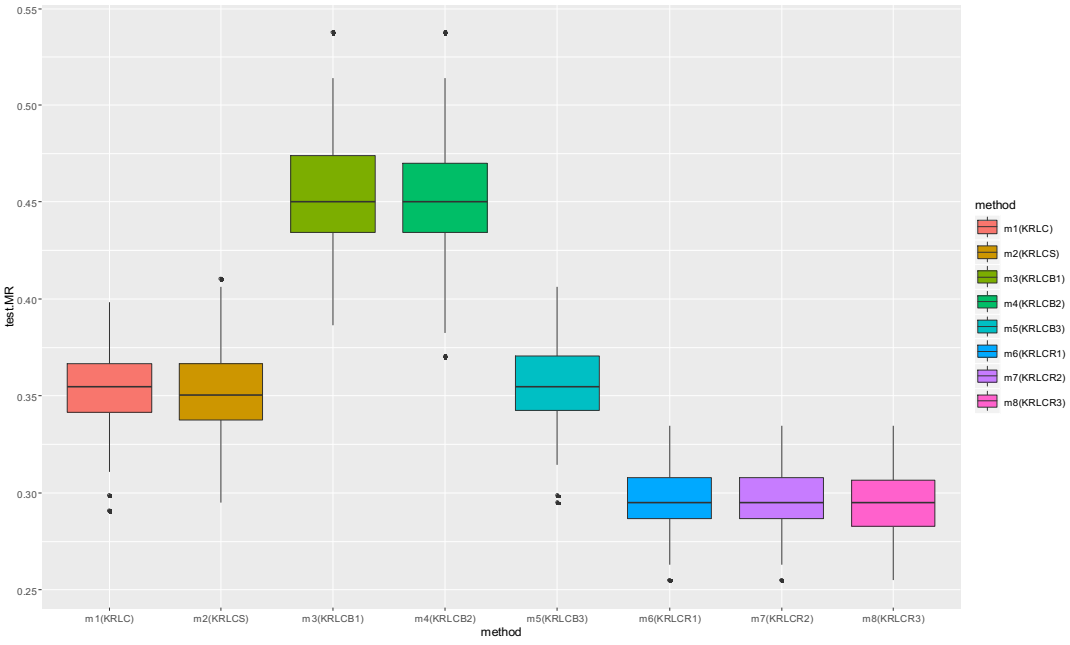


실증분석 결과 앙상블 기법을 적용하는 3가지 방법론들 중 로짓 추정량과 확률추정량을 이용하는 방법론이 일반적인 방법론들에 비해 확실히 좋은 성능을 보이고 있음을 확인할 수 있었다. 다중 투표와 관련된 앙상블 기법을 적용하는 방법론이 기존에 사용되고 있는 방법론들보다 더 좋은 성능을 보이지는 못했지만 전체적으로 살펴봤을 때 본 논문에서 제안하고자 하는 방법론의 성능이 일반적인 방법론들보다 더 우수하다고 판단된다.

#11) Statlog (Australian Credit Approval) Data

이 데이터는 호주에 있는 어느 한 기업의 신용카드를 사용하는 고객들에 대한 신상정보를 담고 있다. 목표는 고객들에 대한 신용도 평가라고 할 수 있으며 반응변수는 신용도가 좋은지, 그렇지 않은지에 대한 척도로 구성되어 있다. 분석하기 전 188개의 결측치와 9개의 범주형 설명변수는 제거하였다.

<그림 19 : Statlog (Australian Credit Approval) Data 실증분석 상자그림>



실증분석 결과 랜덤포레스트 기법을 적용하는 방법론의 성능이 다른 방법론들과 비교했을 때 분류정확도와 신뢰성 측면에서 가장 우수하다는 사실을 쉽게 확인할 수 있었다. 그러나 배깅 기법을

적용하는 방법론의 경우에는 일반적인 방법론들에 비해 성능의 향상이 나타나지 않았다. 하지만 전체적으로 살펴봤을 때 본 연구에서 제안하고자 하는 방법론의 성능이 기존에 사용되고 있는 방법론들에 비해 더 우수하다고 판단된다.

아래에 제시된 표 7은 실증분석에 대한 결과를 8가지 방법론들에 따라 나누어서 정리한 것이다. 각 데이터마다 가장 좋은 성능을 보인 방법론에 대해 해당하는 수치를 강조해서 표시하였다. 이를 통해 실증분석 결과에 대해서 전체적으로 살펴보면 앙상블 기법을 적용했을 경우가 그렇지 않은 경우에 비해 검증 오분류율의 평균값이 더 작게 산출된 것을 확인할 수 있을 것이다. 결과적으로 이 실증분석을 통해서 본 논문에서 제안하고자 하는 방법론이 일반적인 방법론들보다 자료에 대한 분류를 더 정확하게 수행한다는 것을 입증할 수 있었다.

<표 7 : 실증분석 결과에 대한 표>

		KRLC	KRLCS	KRLCB 1	KRLCB 2	KRLCB 3	KRLCR 1	KRLCR 2	KRLCR 3
Cylinder Bands	mean	0.4011	0.4445	0.3870	0.3825	0.4023	0.3275	0.3291	**0.3037**
	sd	0.0365	0.0342	0.0321	0.0344	0.0369	0.0482	0.0462	**0.0285**
Forest Type Mapping	mean	0.1987	0.2188	0.1599	0.1623	0.2019	0.1585	0.1552	**0.1354**
	sd	0.0266	0.0282	0.0233	0.0235	0.0268	0.0331	0.0342	**0.0248**
Dow Jones Index	mean	0.5423	0.5423	0.4998	0.5021	0.5423	**0.4067**	0.4092	0.5423
	sd	0.0164	0.0164	0.0238	0.0244	0.0164	**0.0214**	0.0209	0.0164
Haber -man's Survival	mean	0.3136	0.3196	0.2957	0.2956	0.2937	0.2751	0.2751	**0.2637**
	sd	0.0347	0.0352	0.0330	0.0327	0.0315	0.0274	0.0274	**0.0265**
Ionos -phere	mean	0.1187	0.1204	0.1267	0.1263	0.1254	**0.0683**	0.0705	0.0709
	sd	0.0258	0.0257	0.0272	0.0273	0.0271	**0.0211**	0.0205	0.0200
Pima Indians Diabetes	mean	0.3885	0.4070	0.3766	0.3774	0.3879	0.3269	0.3269	**0.2729**
	sd	0.0203	0.0240	0.0225	0.0239	0.0205	0.0227	0.0229	**0.0194**
Statlog (Heart)	mean	0.3535	0.3493	0.4299	0.4261	0.3484	0.3422	0.3424	**0.3279**
	sd	0.0309	0.0323	0.0400	0.0376	0.0300	0.0296	0.0314	**0.0272**
Blood Trans -fusion Service Center	mean	0.3016	0.3126	0.2912	0.2914	0.3009	0.2385	0.2385	**0.2341**
	sd	0.0198	0.0203	0.0206	0.0195	0.0195	0.0185	0.0185	**0.0179**
Breast Tissue	mean	0.4779	0.4789	0.4955	0.4983	0.4779	0.3542	**0.3508**	0.4804
	sd	0.0446	0.0453	0.0673	0.0707	0.0446	0.0726	**0.0741**	0.0439
Urban Land Cover	mean	0.5450	0.5450	0.4926	0.4987	0.5450	0.2499	**0.2427**	0.5450
	sd	0.0382	0.0382	0.0560	0.0553	0.0382	0.0506	**0.0471**	0.0382
Statlog (Austra -lian Credit Approval)	mean	0.3538	0.3514	0.4529	0.4514	0.3542	0.2985	0.2984	**0.2956**
	sd	0.0201	0.0224	0.0293	0.0300	0.0206	0.0178	0.0179	**0.0177**

6 결론

본 연구를 통해 분류문제에서 자주 사용되는 로지스틱 회귀분류법에 커널기법과 능형 회귀분석 방법, 그리고 앙상블 기법을 적용하는 방법론을 제안하였다. 서론에서도 언급했듯이 자료를 통해서 제대로 된 결과를 얻어야 하는 분석전문가의 입장에서 실제로 마주하게 되는 데이터들의 대다수는 분석이 용이하게 정형화되어있지 않은 것이 사실이다. 이러한 데이터들을 제대로 이해하고 분석하기 위해서는 비선형성과 다중공선성 등의 문제들을 미리 고려해야 한다. 이러한 점들을 해결하기 위해 데이터의 특성에 맞게 형태를 자동으로 변환해주는 커널트릭 기법을 적용하였다. 그리고 분석을 통해 얻어지는 추정량에 대한 정확성과 신뢰성을 높이기 위해 앙상블 기법 중 배깅 기법과 랜덤포레스트 기법을 추가적으로 적용하여 모의실험과 실증분석을 실시하였다. 그 결과 기존에 사용되고 있는 방법론들에 비해 본 연구에서 제안하고자 하는 방법론의 성능이 더 우수함을 입증할 수 있었다. 본 논문을 통해 제시한 방법론 이외에도 다른 방향으로 여러 가지 통계적인 방법론들을 적절하게 적용한다면 더 우수한 성능을 보이는 방법론을 추가적으로 제시할 수 있을 것이다. 이는 향후 연구로 남기도록 하겠다.

7 참고문헌

Hastie, T., Tibshirani, R., Friedman, J. (2011).
The Elements of Statistical Learning, 2nd Edition, Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2014).
An Introduction to Statistical Learning with Applications in R, Springer.

Huh, M. (2015).
Kernel-trick regression and classification.
Communications for Statistical Applications and Methods 22, 201-207.

Schölkopf, B. and Smola, A. J. (2002).
Learning with Kernels, MIT Press.

McCullagh, P. and Nelder, J. A. (1989).
Generalized Linear Models, 2nd Edition, Chapman & Hall.

한선우. (2016).
커널능형회귀분석에서 앙상블기법을 이용한 효율성 연구
(석사학위논문, 한국외국어대학교 대학원)