

마케팅 데이터를 대상으로 중요 통계 예측 기법의 정확성에 대한 비교 연구

조민호*

A Comparative Study on the Accuracy of Important Statistical Prediction Techniques for Marketing Data

Min-Ho Cho*

요 약

미래를 예측하는 기법은 통계에 기반을 둔 것과 딥러닝에 기반을 둔 기술로 분류할 수 있다. 그중 통계에 기반을 둔 것이 간단하고 정확성이 높아서 많이 사용된다. 하지만 실무자들은 많은 분석기법의 올바른 사용에 어려움이 많다. 이번 연구에서는 마케팅에 관련된 데이터에 다항로지스틱회귀, 의사결정나무, 랜덤포레스트, 서포트벡터머신, 베이زي안 추론을 적용하여 예측의 정확성을 비교하였다. 동일한 마케팅 데이터를 대상으로 하였고, R을 활용하여 분석을 진행하였다. 마케팅 분야의 데이터 특성을 반영한 다양한 기법의 예측 결과가 실무자들에게 좋은 참고가 될 것으로 생각한다.

ABSTRACT

Techniques for predicting the future can be categorized into statistics-based and deep-run-based techniques. Among them, statistic-based techniques are widely used because simple and highly accurate. However, working-level officials have difficulty using many analytical techniques correctly. In this study, we compared the accuracy of prediction by applying multinomial logistic regression, decision tree, random forest, support vector machine, and Bayesian inference to marketing related data. The same marketing data was used, and analysis was conducted by using R. The prediction results of various techniques reflecting the data characteristics of the marketing field will be a good reference for practitioners.

키워드

Statistical Forecasting, R, Regression, Random Forest, Decision Tree, Support Vector Machine
통계적 예측, R, 회귀, 랜덤 포레스트, 의사 결정 나무, 서포트 벡터 머신

1. 서 론

인공지능 연구와 함께 1990년 중반 이후에 가장 많은 발전을 이룬 분야가 통계를 기반으로 하는 예측분야라고 할 수 있다[1-2]. 회귀분석과 로지스틱 회귀,

다항 로지스틱 회귀를 시작으로 의사결정나무, 앙상블 개념을 적용한 랜덤포레스트 그리고 서포트벡터머신과 최근의 베이زي안추론까지 다양한 예측기법이 개발되어 활용되고 있다[3-5], [7-10].

이러한 통계적 예측 기법은 딥러닝을 포함한 인공

* 교신저자 : 중원대학교 컴퓨터공학과
• 접수 일 : 2019. 07. 25
• 수정완료일 : 2019. 08. 04
• 게재확정일 : 2019. 08. 15

• Received : Jul. 25, 2019, Revised : Aug. 04, 2019, Accepted : Aug. 15, 2019
• Corresponding Author : Min-Ho Cho
Dept. Computer System Engineering, JungWon University,
Email : chominhokr@jwu.ac.kr

지능의 기법보다 빠르고 간단하게 좋은 결과를 얻을 수 있다는 장점이 있다[6]. 하지만, 이것을 사용하고자 하는 사람에게는 많은 분석 기법의 종류가 큰 부담이 된다. 그래서, 어떤 기법이 어떤 환경에서 잘 적용되는 지에 대한 가이드가 필요하다.

이번 연구는 마케팅 분야의 데이터를 대상으로 통계적 예측기법 중에서 중요한 것 5개를 선별하여 동일한 환경에서 적용하여 보았다[11-12]. 이를 통하여 마케팅 분야에 대한 데이터에서 어떤 기법이 어떤 결과를 보이는 지에 대하여 비교할 수 있는 기본적인 자료를 제공하고자 한다. 이번 연구에서 비교할 예측 기법은 “다항로지스틱회귀”, “의사결정나무”, “랜덤포레스트”, “서포트벡터머신” 그리고 “베이지안 추론”이다.

II. 사용할 데이터 분석

이번 연구를 위하여 사용할 데이터는 사용자가 스마트폰을 구매할 때 영향을 미치는 요인에 대한 것이다. 즉, 삼성, LG, 애플, 기타의 4가지 스마트폰을 사용자가 선택하는 데 중요한 영향을 미친 요인이 무엇인가에 대한 설문 조사이다. 20~25세 사이의 연령을 대상으로 조사하였고 설문 응답자는 215명이었다. 스마트폰 구매에 영향을 미치는 요인으로는 11가지가 고려되었다

- x1 : 화면의 색감 만족도
- x2 : 가격대 성능 만족도
- x3 : 사용의 용이성(화면의 구성분야)
- x4 : 카메라의 성능
- x5 : 폰의 크기와 자재의 느낌
- x6 : 어플리케이션의 다양성
- x7 : 폰의 하드웨어적인 디자인 만족도
- x8 : 사후 지원의 만족도
- x9 : 제조사의 이미지
- x10 : 어플리케이션 수행 성능 만족도
- x11 : 종합적 디자인 만족도

11가지 요인을 대상으로 소비자가 선택하는 스마트폰은 0:삼성폰, 1:LG폰, 2:애플폰, 3:기타폰으로 구분하여 설문지가 완성되었다. 데이터에 대한 상세한 사항은 그림 1에 정리되어 있다.

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Result	
5	1	5	5	5	4	4	2	5	5	5	0	
5	5	3	3	3	3	3	3	4	4	5	2	
5	5	5	3	4	4	4	3	4	4	3	0	
5	3	5	2	4	5	2	4	3	4	4	5	2
4	3	4	4	4	5	5	5	3	4	4	4	0
5	3	5	3	5	5	5	5	1	5	5	5	2
4	3	4	3	3	3	3	3	4	3	2	3	1
4	3	4	4	4	3	4	4	4	4	5	5	2
3	4	4	4	4	4	3	3	3	3	3	3	0
4	4	4	3	4	3	3	3	4	4	4	1	
4	4	4	4	3	4	4	4	4	4	4	3	0
5	5	5	5	5	5	5	5	5	5	5	5	0

그림 1. 설문조사 데이터 모습(일부)
Fig. 1 Part of survey data

예측기법의 적용을 위해서는 데이터 중에서 Result 부분을 Factor형으로 변경하여야 하므로, 간단한 작업을 거쳐 데이터를 분석이 가능한 형태로 변형하여 준비하였다.

III. 통계적 예측기법의 적용

많은 인공지능 기법과 동일하게 통계적 예측기법도 지도학습의 개념이 적용된 것이므로, 분석을 함에 있어서 학습용과 테스트용 데이터를 구분하는 것이 바람직하지만, 마케팅 데이터가 일반적으로 크지 않아서 분리하기 어렵다는 상황을 고려하여, 별도로 분리하지 않고 동일한 데이터를 이용하여 학습과 테스트를 수행하는 방향으로 진행하였다. 이번 연구의 목표가 예측기법의 비교이므로 동일한 환경을 준다는 점에서 모든 기법의 적용에 동일하게 적용하였다

3.1 다항 로지스틱회귀

준비된 데이터의 형태가 Result에서 4가지의 모양을 가지게 됨으로 다항로지스틱 회귀를 적용하여 분석을 수행하였다. 분석한 절차는 그림 2에 정리되어 있다.

분석은 R을 이용하였고, 상세 절차는 그림 2의 내용을 통하여 확인할 수 있다. 이후의 모든 예측방법들도 동일한 환경에서 R을 이용하여 수행하였다

그림 2의 결과를 살펴보면 학습이 이루어진 후 수행된 예측의 정확도는 95%정도임을 알 수 있다. 간단하지만 높은 정확도를 보이고 있음을 확인할 수 있다

```
[다항 로지스틱 회귀 적용]
> library(nnet)

> Model <- multinom(Result ~.,data=phone)
# weights: 52 (36 variable)
initial value 267.554812
iter 10 value 77.502861
iter 20 value 23.411242
iter 30 value 11.790402
iter 40 value 11.754122
final value 11.754073
converged

> predicted <- predict(Model, newdata=phone)
> sum(predicted == phone$Result) /NROW(predicted)
[1] 0.9585492 // 예측의 정확도
>
```

그림 2. 다항회귀 분석 결과
Fig. 2 Result of multinomial regression analysis

3.2 의사결정나무

의사결정나무는 CART(Classification and Regression Tree)와 C4.5, C5.0 그리고 CHAID(Chi-squared Automatic Interaction Detection) 등 다양한 알고리즘이 데이터의 형태와 분석 대상에 따라 나뉘어서 사용되고 있다. 이번 연구에서 사용하는 데이터의 모습에 따라 적용할 수 있는 CART알고리즘과 그것을 개선한 조건부 추론나무를 적용하여 분석해 보았다.

그림3은 R 패키지의 CART 알고리즘을 적용하여 분석한 결과이다. 92%의 정확도를 보이고 있다

```
[ 의사결정나무 중 CART 알고리즘 ]

> library(rpart)
> rpartTree <- rpart(myF, data=phone)
> rpartForecast <- predict(rpartTree, newdata=phone, type="class")
> sum(rpartForecast == phone$Result) / NROW(rpartForecast)
[1] 0.9170984 // 예측의 정확도
```

그림 3. CART 알고리즘 적용 결과
Fig. 3 Result of CART algorithm

그림4는 CART알고리즘이 가지는 과적합문제와 다양한 값으로 분할 가능한 변수가 다른 변수에 비해 선호되는 현상을 제거한 조건부추론나무의 분석 결과이다.

```
[ 의사결정나무 중 조건부 추론나무(Conditional Inference Tree) ]

> install.packages("party")
> library(party)

> myF <- Result ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10
> ctreeResult <- ctree(myF, data=phone)
> forecast <- predict(ctreeResult, data=phone)
> sum(forecast == phone$Result) / NROW(forecast)
[1] 0.9170984 // 예측의 정확도
>
```

그림 4. 조건부추론나무 적용 결과
Fig. 4 Result of conditional inference tree

의사결정나무를 적용한 분석 결과는 92%정도의 정확성을 보이고 있으며, CART와 조건부추론나무가 동일한 정확도를 나타내고 있다. 이는 소규모의 데이터를 기반으로 하는 경우에 CART와 조건부추론나무가 거의 동일한 결과를 보이고 있다고 볼 수 있으며, 데이터가 작으면 회귀분석이 예측을 함에 있어서 보다 나은 결과를 보여준다고 볼 수 있다. 이제 의사결정나무를 개선한 방법을 테스트해 볼 차례이다

3.3 랜덤포레스트

통계적 예측기법에서 부각되는 기법중의 하나가 랜덤포레스트이다. 랜덤 포레스트는 배깅(Bagging), 부스팅(Boosting)과 함께 앙상블기법을 적용한 대표적인 분석기법이다. 앙상블은 주어진 자료로부터 예측모형을 여러 개의 만들고 이것을 결합하여 최종적인 예측모형을 만드는 방법을 통칭하는 개념이다. 실무에서 앙상블개념을 적용한 예측방법은 좋은 결과를 보이는 것으로 알려져 있다.그림5는 랜덤포레스트기법을 적용하여 분석한 결과이다

```
[ 의사결정나무 중 RandomForest 기법의 사용 ]

> install.packages("randomForest")

> library(randomForest)
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.
> randomF <- randomForest(myF, data=phone, ntree=100, proximity=T)
> randomForecast <- predict(randomF, newdata=phone)
> sum(randomForecast == phone$Result) /NROW(randomForecast)
[1] 0.9585492 // 예측의 정확도
>
```

그림 5. 랜덤포레스트 분석 결과
Fig. 5 Result of random forest analysis

그림5에서 확인할 수 있는 것과 같이 랜덤포레스트는 의사결정나무를 보완하여 개발된 분석 기법이므로 보다 나은 예측정확도를 보여주고 있다. 특히 크지 않은 데이터의 특성을 고려할 때, 잘 작동하는 예측기법이라고 판단할 수 있다.

3.4 서포트벡터머신

다음으로 딥러닝이 나오기 전에 개발되어 좋은 반응을 얻은 서포트벡터머신을 적용해 보겠다. 서포트벡터머신은 서포트벡터분류기를 확장하여 비선형 클래스 경계를 수용할 수 있도록 개발한 분류방법이다. 즉, 선형분류기를 비선형 구조로 변경하여 데이터를 분류하는 개념이다. 이 개념을 적용한 대표적인 경우가 커널방법이며, 커널의 차원을 높임으로써 좀 더 다양한 경계를 만들 수 있는데 이것을 다항식 커널이라고 한다. 서포트벡터머신은 다항식 커널의 개념이 적용된 분류 방법으로 좋은 예측 결과를 보여준다.

서포트벡터머신은 커널방법을 사용하므로 코스트와 감마라는 두 개의 파라미터를 결정하는 과정이 필요하다. 이외에도 복잡한 과정이 필요하지만 R 패키지를 사용하면 계산과정에 대한 고민을 줄일 수 있다.

서포트벡터머신을 사용한 분석은 방법에 따라서 3가지로 분리할 수 있다. “radial 방법”, “linear 방법”, 그리고 “Polynomial방법”이다. 3가지 방법 모두가 주어진 환경에 따라 조금씩 다른 결과를 보여줄 수 있다.

그림6에서는 가장 기본적인 radial 방법을 사용한 서포트벡터머신 분석의 과정을 보여준다. 과정이 길고 보이는 데이터와 내용이 너무 많아서 일부를 삭제하고 우리가 관심을 가지는 예측의 정확성에 중점을 두고 정리하여 제시하였다

그림6의 결과를 살펴보면 적은양의 데이터를 대상으로 분석하였음에도 분류의 결과가 95%정도를 보이고 있다. 상당히 좋은 예측 결과이며 소규모의 데이터를 이용하는 마케팅 분야에서 서포트벡터머신은 잘 작동한다는 것을 알 수 있다

이제 관심을 갖는 것은 서포트벡터머신의 다른 방법들을 마케팅데이터에 적용하였을 때, 이들의 결과에 대한 것이다. 이 결과에 따라서 마케팅 관련 분석을 하는 경우에 서포트벡터머신의 어떤 방법을 써야 하는지에 대한 의사결정을 할 수 있게 될 것이다.

그림7에 “linear방법”에 대한 분석결과가 제시되어

있고 그림8에 “Polynomial방법”에 대한 분석결과가 제시되어 있다. 분석한 결과도 차이를 보이지 않으므로 소규모마케팅 데이터에서는 모델에 따른 차이가 없다고 볼 수 있다.

```
[ 서포트벡터머신 중 radial 방법 적용 ]
> phone$Result <- as.factor(phone$Result) // 데이터의 형식 변경
> library(caret) // 필요한 패키지의 호출
> library(e1071)
// 기본형(radial)에 대한 튜닝 수행
> result<-tune.svm(Result~.,data=phone, gamma=2^(-5:0), cost=2^(0:4), kernel="radial")
> result$best.parameters // 필요한 파라미터 값을 얻음
gamma cost
3 0.125 1
// 분석을 수행
> normal_svm <- svm(Result~., data=phone, gamma=0.125, cost=1, kernel="radial")
> summary(normal_svm) // 결과 보기
... 내용 생략
> normal_svm_predict <- predict(normal_svm, phone) //예측수행
> confusionMatrix(normal_svm_predict, phone$Result) //예측결과보기
Confusion Matrix and Statistics

          Reference
Prediction 0  1  2  3
0          105  8  0  0
1           0  16  0  0
2           0  0  56  0
3           0  0  0  8

Overall Statistics
Accuracy : 0.9585 // 예측의 정확성
95% CI : (0.92, 0.9819)
No Information Rate : 0.544
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.9292
McNemar's Test P-Value : NA

... 이하 내용 생략
```

그림 6. 방사형 방법 분석 결과

Fig. 6 Result of radial method Analysis

3.5 베이지안 방법론

베이지안 방법론은 베이즈 확률 이론을 적용한 예측모델로 전통 통계학의 빈도주의와 함께 현대 통계학의 중요한 축이다. 이번에는 베이지안 방법론을 이용해서 예측을 수행해 보자. 결과는 그림9에 있다.

예측의 결과는 66%정도로 나쁘지 않지만 다른 기법들의 정확도에 비하면 저조한 편이다. 이러한 이유는 베이지안 방법론이 가지는 기본 원리 때문인 것으로 판단된다, 즉, 기존 자료의 결과를 피드백 받아서 반영한 다음 결과를 생성하는 방법으로 볼 때, 이번 연구에서 사용하는 데이터의 숫자가 너무 작아서 피드백이 제대로 이루어지지 않았기 때문인 것으로 판단된다. 그러므로 실무에서 베이지안방법론을 적용하고자 하는 경우에는 충분히 큰 데이터를 확보한 상태에서 사용하는 것이 좋은 결과를 얻는데 도움이 될 것이다. 일반적인 마케팅 데이터처럼 만들어지고, 수

집된 특정 부분에 대한 소규모의 데이터를 이용한 분석에는 베이지안 분석은 좋은 결과를 보이지 못할 것으로 판단된다.

```
[ 서포트벡터 머신 중 linear 방법을 적용한 경우 ]
.... 앞부분의 내용은 생략
> result1 <- tune.svm(Result~., data=phone, cost=2*(0.4), kernel="linear")
> result1$best.parameters // 파라미터를 구한다
cost
3 4
> normal_svm1 <- svm(Result~., data=phone, cost=4, kernel="linear") //분석
> normal_svm1_predict <- predict(normal_svm1, phone) // 예측
> confusionMatrix(normal_svm1_predict, phone$Result) // 예측결과보기
Confusion Matrix and Statistics

      Reference
Prediction 0 1 2 3
0 105 8 0 0
1 0 16 0 0
2 0 0 56 0
3 0 0 0 8

Overall Statistics
          Accuracy : 0.9585 // 예측의 정확도
          95% CI : (0.92, 0.9819)
No Information Rate : 0.544
P-Value [Acc > NIR] : < 2.2e-16
          Kappa : 0.9292
.... 이하의 내용 생략
```

그림 7. 선형 방법 분석 결과
Fig. 7 Result of linear method Analysis

```
[ 베이지안 방법론을 적용하여 예측한 경우 ]

> library(e1071)
> Bayes <- naiveBayes(Result~., data=phone)
> Bayes
... 상세한 내용 생략
> predicted <- predict(Bayes, phone, type="class") // 예측
> confusionMatrix(predicted, phone$Result) //예측의 정확도
Confusion Matrix and Statistics

      Reference
Prediction 0 1 2 3
0 64 0 0 0
1 0 16 0 0
2 0 0 40 0
3 41 8 16 8

Overall Statistics

          Accuracy : 0.6632 // 예측정확도
          95% CI : (0.5918, 0.7295)
No Information Rate : 0.544
P-Value [Acc > NIR] : 0.0005065

          Kappa : 0.5408
.... 이하의 내용 생략
```

그림 9. 베이지안방법론 분석 결과
Fig. 9 Result of Bayesian method Analysis

```
[ 서포트벡터 머신중 Polynomial 방법 적용 ]

... 앞부분의 내용은 생략
// 분석수행
> result2 <- tune.svm(Result~., data=phone, cost=2*(0.4), degree=2.4,
kernel="polynomial")
> result2$best.parameters //파라미터구하기
degree cost
8 3 4
> normal_svm2 <- svm(Result~., data=phone, cost=4, degree=3,
kernel="polynomial") // 분석
> normal_svm2_predict <- predict(normal_svm2, phone) // 예측
> confusionMatrix(normal_svm2_predict, phone$Result) // 예측결과보기
Confusion Matrix and Statistics

      Reference
Prediction 0 1 2 3
0 105 8 0 0
1 0 16 0 0
2 0 0 56 0
3 0 0 0 8

Overall Statistics
          Accuracy : 0.9585 //분석정확도
          95% CI : (0.92, 0.9819)
No Information Rate : 0.544
P-Value [Acc > NIR] : < 2.2e-16

          Kappa : 0.9292
.... 이하생략
```

그림 8. 다항 방법 분석 결과
Fig. 8 Result of Polynomial method Analysis

IV. 결론

통계적 예측기법은 딥러닝에 비하여 간단하며 높은 정확성을 보여주는 좋은 방법이다. 이번 연구를 통하여 현재까지 활용되고 있는 “다항로지스틱회귀”, “의사결정나무”, “랜덤포레스트”, “서포트벡터머신” 그리고 “베이지안 추론”을 마케팅 성격의 데이터에 적용했을 때 각 기법의 정확성을 비교하였다.

대부분의 기법들이 지도학습의 성격을 가지므로 데이터가 많은 경우에는 좋은 결과를 보이게 된다. 하지만, 실무에서 분석을 위하여 일부러 수집한 데이터는 크기가 제한적이기 마련이다. 제한된 크기를 가지는 데이터를 대상으로 예측을 수행하는 경우에 예측의 정확성은 “서포트벡터머신”과 “랜덤포레스트”가 가장 좋은 성능을 보이고 있다. “의사결정나무”의 경우는 기법의 성격상 “서포트벡터머신”과 “랜덤포레스트”보다 좋은 성능을 보이지 못하고 있다. 특이한 점은 데이터가 크지 않은 환경에서는 “베이지안추론”이 좋은 결과를 보이지 못한다는 것이다.

추가로 “의사결정나무”기법의 알고리즘인 CART나 조건부추론나무는 예측에 있어서 정확성의 차이를 보이지 않았다. 이러한 점은 “서포트벡터머신”의 방사형(radial), 선형(linear), 다항(Polynomial) 커널 알고리

즘에서도 동일하였다. 데이터가 작은 경우에는 동일 분석 기법에 속한 알고리즘간에 정확도의 차이를 보이지 않는다는 점을 확인할 수 있다. 결론적으로 소규모 데이터를 사용하는 예측의 경우에는 예측기법에 따라 정확성의 차이가 명확해서 서포트벡터 머신이 가장 유용한 것으로 생각된다. 다만, 동일 기법에서의 알고리즘 차이는 크지 않은 것으로 판단된다.

References

- [1] Y.Guo, Y. Liu, A. Oerlenmans, S. Lao, S. Wu, and M. Lew, "Deep learning for visual understanding : A review", *Neurocomputing*, vol. 187, 2016, pp. 27-48.
- [2] S. Kotsiantis, "Decision trees:a recent overview", *Artificial Intelligence Review*, vol. 39, issue. 4, 2013, pp. 261-283.
- [3] J. Peng, K. Lee, and C. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", *The Journal of Education Research*, vol., 96. Sept. 2002, pp. 3-14..
- [4] A. El-Habil, "An Application on Multinomial Logistic Regression Model", *Pakistan journal of Statsitics and operation Research*, vol., 8. no. 2, Mar. 2012, pp. 271-291.
- [5] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", *Journal of Biomedical Informatics*, vol. 35, 2002, pp. 352-359.
- [6] P. Bossaerts and P. Hillion, "Implementation Statistical Criteria to Select Return Forecasting Models:What Do We learn?," *The Review of Financial Studies Summert*, vol. 12, no. 2, 1999, pp. 405-428.
- [7] J. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, vol., 9. June. 1999, pp. 293-300.
- [8] B. Gupta, A. Rawat, A. Jain, A. Arora and N.Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining," *International Journal of Computer Applications*, vol. 163, no. 8, Apr. 2017, pp. 15-19.
- [9] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, issue. 1, Oct. 2001, pp. 5-32.
- [10] S. Kim and K. Suk, "Effective real-time identification using Bayesian statistical method gaze Network," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 3, 2016, pp. 331-338.
- [11] Y. Bae and K. Park, "Performance Comparison of Machine Larning in the Various Kind of Prediction," *J. of the Korea Institute of lectronic Communication Science*, vol. 14, no. 1, 2019, pp. 169-178.
- [12] W. Hauser, "Marketing analytics:the evolution of marketing research in the twenty-first century," *Direct Marketing : An International Journal*, vol., 1, no.1, Apr. 2007, pp. 38-54.

저자 소개

조민호(Min-Ho CHO)



1989년 인하대학교 산업공학과 졸업(공학사)
HP Korea, Openwave, SK C&C 등 산업체 근무

2003년 숭실대학교 대학원 컴퓨터공학과 졸업(공학박사)
2012년~ 중원대학교 컴퓨터공학과 교수
※ 관심분야 : 소셜네트워크, 소프트웨어공학, 데이터 마이닝 및 통계분석