**DATA MINING 1 – Bank Marketing**

2014210072 조은진

**1, 실험요약**

Classficiation의 목표는 고객이 term deposit을 구독할 것인가를 예측하는 것이다. 데이터 마이닝은 상품을 구매할 가능성이 높은 고객들을 구별하게 도와줘서, 회사는 그 고객들을 타겟으로 하여 캠패인을 진행할 수 있다. 이는 마케팅 비용과 노력을 크게 줄여주고 캠패인의 방향성을 결정지어 준다. 이를 위해 다양한 모델을 학습시켜보고 가장 성능이 좋은 모델을 채택하는 실험을 진행한다.

Portuguese banking institution 마케팅 캠페인에 관해 전체 data set과 랜덤으로 10%를 선택한 data set이 주어졌는데, 후자를 test data set으로 사용한다.

zeroR과 oneR을 baseline으로 사용하여 4개의 모델 (NaiveBayes, bayesNet, J48, Multilayer perceptron)을 cross – validation (10 folds)으로 비교분석한다. 모델을 비교분석하며 data set에 대한 분석도 함께 한다.

**2, 데이터 설명**

- Number of Instances: 45211

- Number of Attributes: 16 + output attribute.

- Input variables:

# bank client data:

1 - age (numeric)

2 – job: type of job (categorical)

"admin.","unknown","unemployed","management","housemaid","entrepreneur","student",

"blue-collar","self-employed","retired","technician","services"

3 - marital : marital status (categorical)

"married","divorced","single"; note: "divorced" means divorced or widowed

4 - education (categorical)

"unknown","secondary","primary","tertiary"

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical)

"unknown","telephone","cellular"

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical)

"jan", "feb", "mar", ..., "nov", "dec"

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical)

“unknown","other","failure","success"

Output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: "yes","no")

+ Missing Attribute Values: None

- 이 Data Set을 선택한 이유 :

마케팅에 빅데이터 분석이 사용되는 많은 사례들을 보며 직접 마케팅 관련된 Data Set을 분석해보고 싶었다.

**3, 실험 내용**

**3.1 Initial test (10 folds cross – validation)**

다음은 선택한 각 알고리즘에 대한 정확도이다.

1) zeroR

Correctly Classified Instances 88.3015 %

2)oneR

Correctly Classified Instances 88.5625 %

3) NaiveBayes

Correctly Classified Instances 88.0073 %

4) bayesNet

Correctly Classified Instances 88.3767 %

5) J48

Correctly Classified Instances 90.3187 %

6) Multilayer perceptron (hidden layer = 10 일 때)

Correctly Classified Instances 89.9095 %

**3.2 select attributes**

predictive power가 낮은 attribute들을 제거하여 효율성과 일반성을 높인다.

- Cfs picked 3 7 8 12 16

- InfoGainAttributeEval randked 12,16,14,11,9,15,1,7,2,6,10,13,8,4,3,5

- SymmetricalUncertAttributeEval ranked 12,16,14,15,9,11,7,1,6,2,8,13,10,4,3,5

=> most important attributes : 1,7,9,11,12,14,15,16

위의 attribute들만 적용했을 때, 선택한 각 알고리즘에 대한 정확도이다.

1) zeroR

Correctly Classified Instances 88.3015 %

2) oneR

Correctly Classified Instances 88.5625 %

3) NaiveBayes

Correctly Classified Instances 88.8235 %

4) bayesNet

Correctly Classified Instances 88.474 %

5) J48

Correctly Classified Instances 90.2656 %

6) Multilayer perceptron (hidden layer = 10 일 때)

Correctly Classified Instances 90.2479 %

또한 ‘bank client data’에 관련된 attribute는 8개 중 2개,

‘related with the last contact of the current campaign’에 관련된 attribute는 4개 중 3개,

‘other attributes’에 관련된 attribute는 4개 중 3개가 most important attributes 로 선택된 것을 주목해 볼 수 있다.

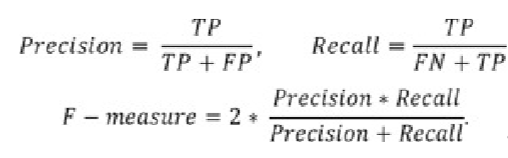
이를 통해 yes, no의 구분에 bank client의 정보가 큰 영향력을 끼치지 않음을 알 수 있다. 따라서 고객의 구독 여부에 대한 예측 시, 선택된 age와 housing을 제외하고 여러 고객정보들을 확인하는 것에 크게 시간과 비용을 투자하지 않아도 된다.

**3.3 모델평가기준**

1, Correctly Classified Instances

전체적인 정확도이며 높을수록 좋다

2, F - Measure



TP  True Positive (Actually “yes” and classified as “yes”) – We want a high TP rate

FP  False Positive (Actually “no” but classified as “yes”)

FN False Negative (Actually “yes” but classified as “no”)

F – measure가 1에 가까울수록 좋은 모델이다.

3, Validation using test data set

Test data set에 대한 model의 performance를 측정한다. 측정 기준은 TP rate of class ”yes” in the test set으로 한다. ( how many of the actual “yes” cases were classified as “yes” by the classifier.)

**3.4 overfitting 방지**

Data set의 No case가 39922개 / Yes case가 5289개이다. data set의 데이터가 매우 No 쪽으로 치우쳐 있는 No에 과도하게 적합된 특징을 학습하게 되어 overfitting현상이 발생한다. 이렇게 overfiting된 데이터는 일반성을 신뢰할 수 없기 때문에 새로운 데이터에 대한 예측 모형으로 사용할 수 없다. overfiting을 방지하기 위해 다음과 같은 방법을 사용한다.

1, data set과 주어진 test data set은 모두 no cases가 현저하게 많은 data set이므로, TP rate 값을 가장 우선순위에 두어 평가한다.

2, No cases 중 Yes cases와 동등한 수인 5289만 랜덤하게 뽑아내어, no cases와 yes cases의 개수가 동일한 dataset을 새로 만든 후, 학습에 사용한다.

이 방법은 계산속도도 높여준다.

**3.5 Models**

**1)** **zeroR , oneR**

Correctly Classified Instances는 zeroR 88%, oneR 88% 정도이다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 5289만 랜덤하게 뽑아내어 학습에 사용하면 Correctly Classified Instances는 zeroR은 50%, oneR은 94% 정도이다.

이를 baseline으로 사용하였기 때문에, 밑의 모델들은 위의 퍼센트보다 낮아질 수 없다.

**2) Naïve Bayes**

이 모델은 Correctly Classified Instances는 89% 정도, Yes class 에 대한 True Positive Rate(TPR)는 0.459, F-Measure는 0.490 이다. Test data set을 적용했을 때, TP rate of class yes 는 0.455 이다.

Correctly Classified Instancesss는 높은 편이지만, Yes class 에 대한 True Positive Rate가 낮은 것을 고려해야한다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 5289만 랜덤하게 뽑아내어 학습에 사용하면 다음과 같은 변화가 있다.

Correctly Classified Instances는 94% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.915, F-Measure는 0.93 이다. Test data set을 적용했을 때, TP rate of class yes는 0.902이다.

**2) bayesNet**

이 모델은 Correctly Classified Instances는 89% 정도, Yes class 에 대한 True Positive Rate(TPR)는 0.459, F-Measure는 0.490 이다. Test data set을 적용했을 때, TP rate of class yes 는 0.455 이다.

Correctly Classified Instancesss는 높은 편이지만, Yes class 에 대한 True Positive Rate가 낮은 것을 고려해야한다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 5289만 랜덤하게 뽑아내어 학습에 사용하면 다음과 같은 변화가 있다.

Correctly Classified Instances는 96% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.94, F-Measure는 0.956 이다. Test data set을 적용했을 때, TP rate of class yes는 0.927이다.

**3) J48**

이 모델은 Correctly Classified Instances는 90% 정도, Yes class 에 대한 True Positive Rate(TPR)는 0.463, F-Measure는 0.527이다. Test data set을 적용했을 때, TP rate of class yes 는 0.455 이다.

Correctly Classified Instancesss는 높은 편이지만, Yes class 에 대한 True Positive Rate가 낮은 것을 고려해야한다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 5289개만 랜덤하게 뽑아내어 학습에 사용하면 다음과 같은 변화가 있다.

Correctly Classified Instances는 95% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.955, F-Measure는 0.959이다. Test data set을 적용했을 때, TP rate of class yes는 0.94이다.

J48 모델이 위의 두 모델보다 성능이 더 좋다는 것을 알 수 있다.

또한 decision tree를 통해, duration이 yes case와 no case가 나눠지는 것에 가장 영향을 많이 끼치는 요소임을 확인할 수 있다.

**4) Multilayer perceptron**

위의 두 모델과 같이 Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 5289개만 랜덤하게 뽑아내어 학습에 사용한다. 더불어 이 모델에서 성능을 결정짓는 매우 중요한 요소는 Hidden node의 개수이다. Hidden node가 너무 많으면 지금 있는 training data들에 대해서만 너무 세분화되어 일반성을 띄지 못하는 overfitting 문제가 발생한다. 반면 hidden node를 너무 적게 사용하면 underfitting 문제가 발생한다.

weka에서 자동으로 hidden node를 생성해주는 설정을 하면, weka는 hidden node를 13개 만들어준다. Correctly Classified Instances는 96% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.954, F-Measure는 0.956이다. Test data set을 적용했을 때, TP rate of class yes는 0.956이다.

Hidden node 개수를 3,5,10,18 로도 설정 해보고 가장 적합한 hidden node 개수를 선택한다.

3 – TPR : 0.957/ TEST TPR : 0.942

5 - TPR : 0.976/ TEST TPR : 0.965

10 - TPR : 0.953 / TEST TPR : 0.952

13 - TPR : 0.954 / TEST TPR : 0.956

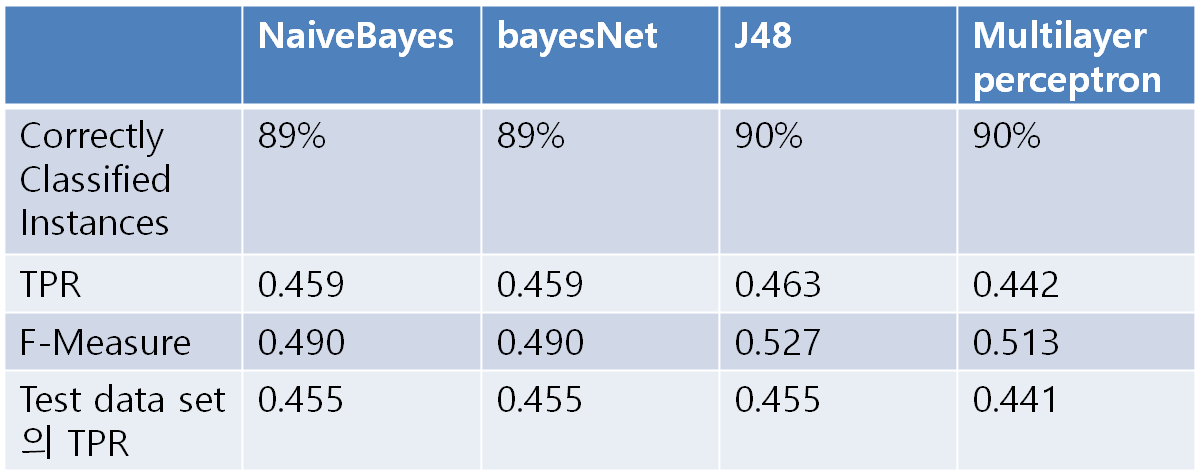
18 - TPR : 0.957 TEST TPR: 0.952

TPR 과 TEST TPR 이 가장 높은 5를 hindden node 개수로 설정한다.

이 때, Correctly Classified Instances 96% 정도이고 , F-Measure는 0.963이다.

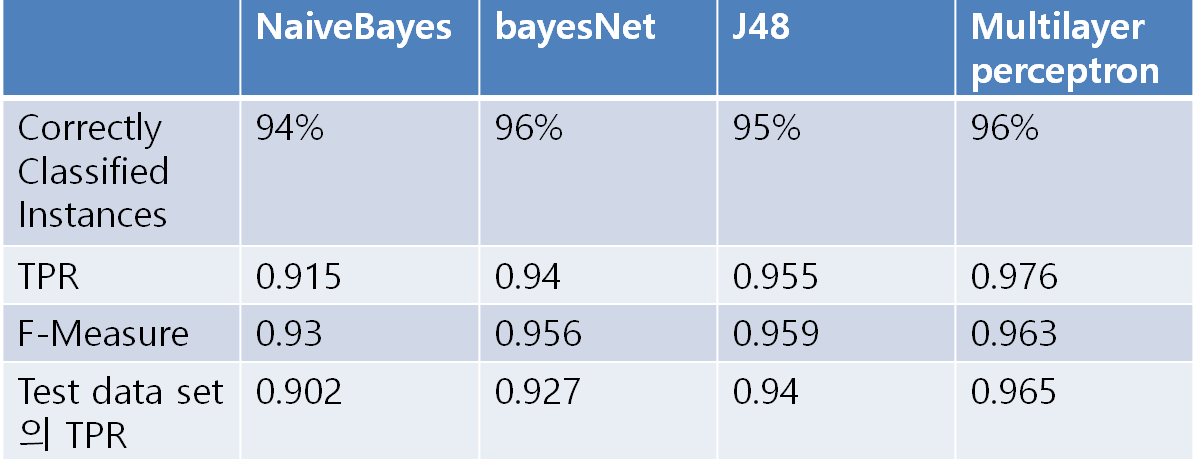
**3.6 Model comparison**

전체 data set을 사용했을 때 4개의 모델의 결과는 다음과 같다.



J48과 multilayer perceptron(hidden node = 5)의 correctly classified instance가 가장 높다. 하지만 위의 data set은 no case에 치중된 data set이므로 다른 요소들도 반드시 고려해야한다. J48이 TPR , F – Measure, test data set의 TPR도 가장 높음을 알 수 있다.

No case와 yes case의 개수가 같은 data set을 사용했을 때 4개의 모델의 결과는 다음과 같다.



bayesNet과 Multilayer perceptron(hidden node = 5)이 가장 분류를 정확하게 하는 모델임을 알 수 있다. J48도 correctly classified instance에서 큰 차이는 나지 않고 TPR , F – Measure, test data set의 TPR이 bayesNet에 비해 높다.

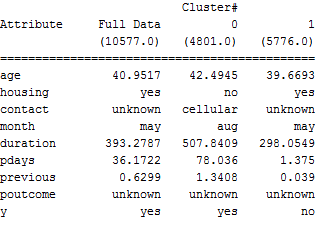
**3.7 Clustering**

클러스터링은 매우 informative and effective 하다. 클러스터링을 통해 상품을 구매할 가능성이 높은 고객들의 특성을 쉽게 알 수 있기 때문이다.

Simple K means 를 사용한다.

Number of cluster 의 개수를 2로 설정하면, 두 개의 클러스터 모두 output이 No이다. Data set의 data set의 데이터가 매우 No 쪽으로 치우쳐 있어서 발생하는 현상이다.

No cases 중 Yes cases와 동등한 수인 5289만 랜덤하게 뽑아내어 학습에 사용하면, 다음과 같은 결과를 얻는다.



Ouput이 yes인 클러스터와 no인 클러스터를 비교해보면, duration과 pdays에서 뚜렷한 차이를 보인다는 것을 알 수 있다.

Yes 클러스터의 duration이 훨씬 많은데, Duration이 많다는 것은 고객이 관심 있는 상품에 대해 상세하게 물어보기 때문임을 추측해 볼 수 있다.

이전에 접촉이 없던 고객일 경우, Pdays(number of days that passed by after the client was last contacted from a previous campaign)는 -1로 기록된다. 이로 인해 No 클러스터의 pdays가 현저하게 낮음을 알 수 있다.

**4, 결론**

상품을 구매할 가능성이 높은 고객을 분류하는 것에 J48을 사용하는 것이 좋다. No case와 yes case의 개수가 같은 data set을 사용한다면, Multilayer perceptron(hidden node = 5)모델도 좋은 선택이다. 하지만 Multilayer perceptron은 성능이 좋은 반면, decision tree처럼 쉽게 이해하고 설명하기 어렵기 때문에 클러스터링을 함께 사용하여 상품을 구매할 가능성이 높은 고객들의 특징들을 파악하는 것이 큰 도움이 된다.

- 참고 문현 :

Data Mining Application in Direct Marketing: Identifying Hot Prospects for Banking Product(Sagarika Prusty)