**DATA MINING 2 – Analysis of factors that affect students’ academic performance**

2014210072 조은진

**1, 실험요약**

어떠한 이유들로 인해 학생의 academic performance가 좋지 않아서 course를 다시 수강하는 결과가 발생한다. evaluation tools을 사용하여 educational failure에 영향을 끼치는 요인들을 파악한다. 이런 요인들을 파악하고 예방책, 개선점을 찾으면 educational failure을 줄일 수 있다.

특히 수업이 진행될 때에 질문지를 작성하게 하여 course를 다시 수강할 확률이 높은 학생을 예측하여 그 학생이 course를 잘 마무리할 수 있도록 instructor는 개선점을 찾고 특별히 도와준다.

이는 학생의 시간적, 육체적, 정신적, 교육적 상황들을 개선시키는 데 큰 도움이 된다.

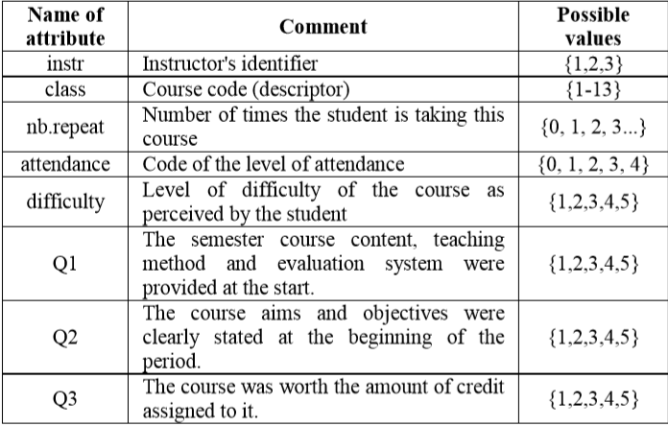
이를 위해 다양한 모델을 학습시켜보고 가장 성능이 좋은 모델을 채택하는 실험을 진행한다.

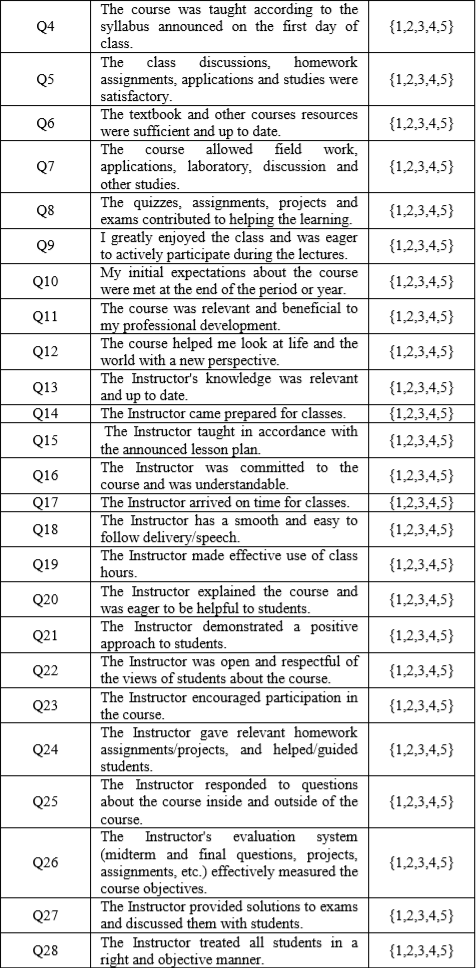
zeroR과 oneR을 baseline으로 사용하여 4개의 모델 (NaiveBayes, bayesNet, J48, Multilayer perceptron)을 cross – validation (10 folds)으로 비교분석한다. 모델을 비교분석하며 data set에 대한 분석도 함께 한다.

**2, 데이터 설명**

이 data set은 Gazi University in Ankara (Turkey) 의 학생들이 평가한 5820개의 평가 점수로 이루어져있다.

Attribute는 33개인데, 학생이 수강한 course에 관련된 atrribute는 nb.repeat, attendance, difficulty, Q1 to G12 attributes 이고 course instructor에 관련된 attribute는 Q13 to Q28 이다.





- 이 Data Set을 선택한 이유 : 대학생으로서 students’ academic performance에 관련된 데이터를 분석하면 분석결과를 해석하는 것이 더 쉽고 정확도와 예측도가 높을 것 같아서 선택하게 되었다.

**3, 실험 내용**

**3.1 Grouping of attribute values**

이 실험의 목적에 따라 nb.repeat attribute가 class attribute로 선택되고

numerical 값들은 학생이 한번 넘게 수강하면(2이상) Yes, 그렇지 않으면

No로 대체한다. 또한 instr attribute, class attribute는 nomial 하지만, numeric 값들로 되어있어서

데이터 마이닝 알고리즘에서 정확성이 보장되지 않을 수 있고 더 명확한 실험을 위해

instr attribute의 값 1,2,3을 각각 A,B,C로 대체한다. class attribute의 값 1부터 13도 각각

A 부터 M으로 대체한다.

마지막으로 attendance, difficulty , Q1 to Q28 attribute의 값들 중 1,2 -> low / 3 -> middle/ 4,5-> high로 대체한다. (attendance의 0값은 제외한다)

**3.2 select attributes**

predictive power가 낮은 attribute들을 제거하여 효율성과 일반성을 높인다.

- Cfs picked class attendance

- InfoGainAttributeEval randked class attendance instr difficulty q14 q25 q2 q21 q15 q17 q13 q19 q20 q22 q4 q23 q28 q3 q18 ….

- SymmetricalUncertAttributeEval ranked class instr attendance difficulty q14 q25 q2 q21 q15 q17 q13 q19 q20 q22 q4 q23 q28 q3 q18….

=> most important attributes : class attendance instr difficulty q14 q25 q2 q21 q15 q17 q13 q19

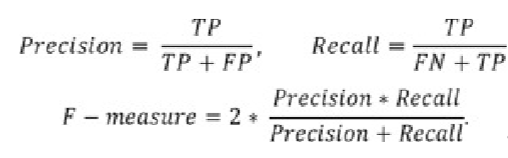
학생이 수강한 course에 관련된 질문들보다 course instructor에 관련된 질문들이 더 높은 순위로 매겨지는 현상을 발견할 수 있다. 이를 통해 course보다는 instructor의 영향력이 훨씬 더 크다는 사실을 알 수 있고 다음 질문지를 만들 때도 instructor에 대한 더 많고 세부적인 질문들을 집어넣는 것이 평가와 예측에 도움이 된다는 것을 생각해볼 수 있다.

**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.

**3.4 overfitting 방지**

Data set의 No case가 4909개 / Yes case가 911개이다. data set의 데이터가 매우 No 쪽으로 치우쳐 있는 No에 과도하게 적합된 특징을 학습하게 되어 overfitting현상이 발생한다. 이렇게 overfitting된 데이터는 일반성을 신뢰할 수 없기 때문에 새로운 데이터에 대한 예측 모형으로 사용할 수 없다. overfitting을 방지하기 위해 다음과 같은 방법을 사용한다.

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

2, No cases 중 Yes cases와 동등한 수인 911만 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 zeroR과 oneR 모두 84% 정도이다.

하지만 data set의 데이터가 매우 No 쪽으로 치우쳐 있다는 이유 때문에

두 모델은 data set의 모든 데이터를 no로 구분하며 True Positive Rate가 0이다.

따라서 baseline으로의 사용을 고려해보아야한다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 911만 랜덤하게 뽑아내어 학습에 사용하면 Correctly Classified Instances는 zeroR은 52%, oneR은 92% 정도이다.

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

**2) Naïve Bayes**

이 모델은 Correctly Classified Instances는 79% 정도, Yes class 에 대한 True Positive Rate(TPR)는 0.290, F-Measure는 0.303이다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 911만 랜덤하게 뽑아내어 학습에 사용하면 다음과 같은 변화가 있다.

Correctly Classified Instances는 92% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.905, F-Measure는 0.924 이다.

**2) bayesNet**

이 모델은 Correctly Classified Instances는 79% 정도, Yes class 에 대한 True Positive Rate(TPR)는 0.288, F-Measure는 0.301이다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 911만 랜덤하게 뽑아내어 학습에 사용하면 다음과 같은 변화가 있다.

Correctly Classified Instances는 92% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.905, F-Measure는 0.924 이다.

**3) J48**

이 모델은 Correctly Classified Instances는 84% 정도이지만 Yes class에 대한 True Positive Rate(TPR)는 0, F-Measure는 0이다. data set의 데이터가 매우 No 쪽으로 치우쳐 있는 심각한 overfitting으로 인해 Number of Leaves와 Size of the tree가 1이고 data set의 모든 데이터를 no로 구분하는 큰 문제점이 있다.

Overfitting을 방지하기 위해 No cases 중 Yes cases와 동등한 수인 911개만 랜덤하게 뽑아내어 학습에 사용하면 다음과 같은 변화가 있다.

Correctly Classified Instances는 94% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.889, F-Measure는 0.941이다.

이 때의 decision tree를 보면, instr와 class attribute에 의해서만 yes, no가 결정되는 구조이므로 instr와 class attribute를 제거하고 실행시켜본다. 이는 다음학기에 instructor와 class가 바뀔 때에도 질문들에 근거하여 어느 정도 예측을 할 수 있게 하기 위해서이기도 하다.

이 decision tree를 보면, Attendance가 가장 중요한 요소임을 알 수 있다. Attendance가 high, middle이면 그 다음으로 Q2(The course aims and objectives were clearly stated at the beginning of the period)가 중요한 요소인 반면, low이면 Q25(The Instructor responded to questions about the course inside and outside of the course.)가 그 다음으로 중요한 요소이다. Attendance가 0이면 difficulty가 그 다음으로 중요한 요소이다.

그러므로 학생들의 attendance를 높여주는 것이 가장 중요하다. 출석 점수 비중을 높이거나 꾸준히 출석을 격려하고 그에 따른 보상을 주는 등 다양한 방법을 시도해볼 수 있다.

Attendance를 높이는 것이 어렵다면, attendance level에 따라 다른 전략을 취하는 것이 좋다.

1. Attendance level이 high, middle인 학생은 학기 초에 course aims and objectives가 명확히 제시되었을 때 의욕을 가지고 course를 잘 마무리한다. 특히 attendance와 Q25가 모두 high인 학생들은 모두 course를 한번에 수강하는 것을 성공하였다. 이런 학생들을 위해 course aims and objectives를 학기 초에 명확히 제시해주는 것이 필요하다.
2. Attendance level이 low일 때, Q25가 middle, low면 repeat이 yes이다. Attendance가 낮은데 Instructor가 질문에 응답을 잘 해주지 않으면 학생들은 course를 성공적으로 수강하는 데 큰 어려움을 겪고 다시 course를 수강하게 됨을 예측해볼 수 있다. 이런 학생들을 위해 instructor가 질문에 응답해주는 것이 매우 중요하다. 특히 Attendance level이 낮은 학생들에게 특별히 질의응답시간을 마련해주거나 메일로 질문에 대해 상세한 답을 해주는 등 다양한 방법들을 시도해볼 수 있다.
3. Attendance level이 0일 때 difficulty가 high면 repeat은 yes 이다. Attendance가 매우 낮은 학생들은 진도를 따라가고 수업을 이해하는 것에 매우 큰 어려움이 있어서 difficulty가 높으면 한번에 수강하는 것이 쉽지 않다. 특별한 사정으로 attendance가 낮은 학생들에게 보충을 해주거나 강의를 녹화하여 인터넷에 올려주어서 difficulty가 높아도 course를 잘 따라갈 수 있도록 도와주는 것이 좋다.

**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 24개 만들어준다. Correctly Classified Instances는 93% 정도이고, Yes class 에 대한 True Positive Rate(TPR)는 0.913, F-Measure는 0.933이다.

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

10 – Correctly Classified Instances : 93%정도 / TPR : 0.910 / F-Measure: 0.931

15 - Correctly Classified Instances : 92%정도 / TPR : 0.914 / F-Measure: 0.930

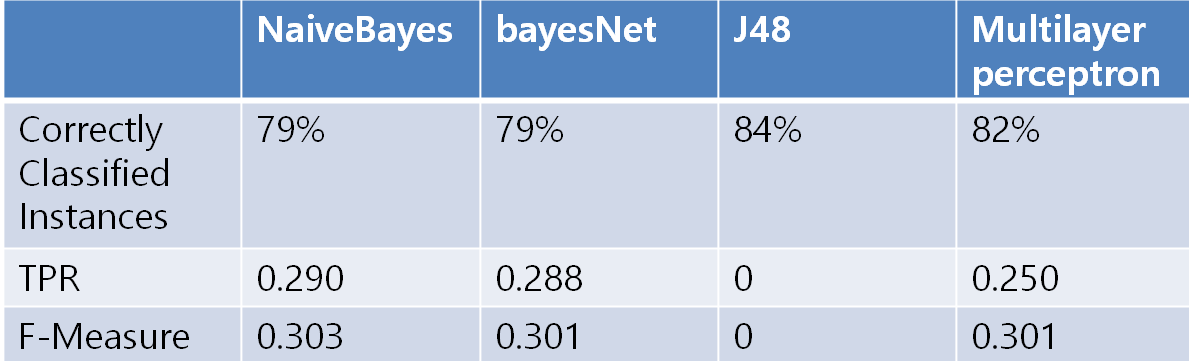
20 - Correctly Classified Instances : 93%정도 / TPR : 0.914 / F-Measure: 0.929

25 - Correctly Classified Instances : 93%정도 / TPR : 0.913 / F-Measure: 0.929

큰 차이는 없지만, hidden node가 24일 때 가장 성능이 좋으므로 24를 선택한다.

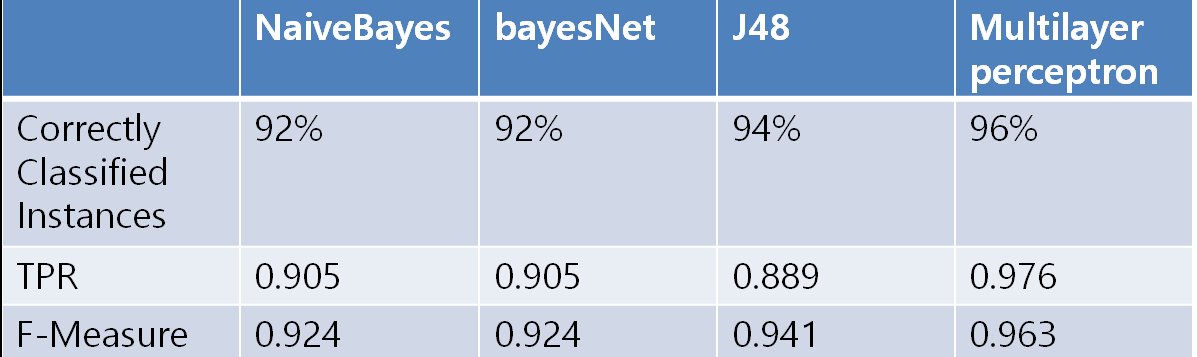
**3.6 Model comparison**

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



no case에 치중된 data set이므로 TPR을 가장 우선으로 고려해야 한다. NaiveBayes가 가장 좋은 모델이다.

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



Multilayer perceptron(hidden node = 24)이 가장 분류를 정확하게 하는 모델임을 알 수 있다. 또한 TPR , F – Measure도 높다.

**4, 결론**

No case와 yes case의 개수가 같은 data set을 사용하여 Multilayer perceptron(hidden node = 24)를 선택하는 것이 좋다. 하지만 Multilayer perceptron은 성능이 좋은 반면, decision tree처럼 쉽게 이해하고 설명하기 어렵다. 성능이 크게 차이 나지 않는다면, 이해하기 쉬운 J48을 사용하는 것도 좋은 선택이다.

또한 위에 언급한 것처럼 데이터 마이닝으로 educational failure에 영향을 끼치는 요인들을 파악하여 미리 개선하고 course를 다시 수강할 확률이 높은 학생을 예측하여 그 학생이 course를 한 번에 잘 마무리할 수 있도록 맞춤형 개선점을 제공한다.

- 참고 문현 :

Analysis of factors that affect students’ academic performance - Data Mining Approach

(Nahid Hajizadeh, Marzieh Ahmadzadeh)