1. 연구 내용:

**An empirical comparison of Performance of classification techniques using public and opened medical data with weka**

(weka와 공개 및 공개 의료 데이터를 사용하여, classification 기술의 성능(accracy)을

실험적으로 비교함)

1. 데이터 수정

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [5] | [6] | [20] | [22] | [24] | [25] | [26] | [27] | [28] | [29] | [31] | [33] |
|
|
| : 수정없음 | : class 수정 | : class 수정 | :dat  파일을 csv파일로 변경 후,  class 수정 | : missing value 수정 | :class수정 | : 수정 내용 없음 | :arff  파일을 csv파일로 변경 | :class를  0==P  1==M으로 수정 | : class 수정함 | : missing value가 많았음->0 | : 순서 나타내는 첫번째 열 제거, class변수이름->a, b로 변경 |
|
| 1. Nominal->numerical 수정내용 | | | | | | | | | | | |
| [5] | [6] | [20] | [22] | [24] | [25] | [26] | [27] | [28] | [29] | [31] | [33] |
|  |  |  |  |  | **BMI**  Thin==0  Moderate==1  Overweight==2 | **Sex**  Female==0  Male==1 | **V1**  DGN1==0  DGN2==1  DGN3==2  DGN4==3  DGN5==5  DGN6==5  DGN8==6 |  | **V2**  Female==0  Male==1 |  | **rbs**  abmormal==0  normal==1 |
|  |  |  |  |  | **Blood**  <140/90  ==0  >=140/90  ==1 | **History of high blood**  No==0  Yes==1 | **V4**  PRZ0==0  PRZ1==1  PRZ2==2 |  |  |  | **pc**  Abnormal==0  Normal==1 |
|  |  |  |  |  | **FBS**  <126==0  >=126==1 | **History of** **use drugs for high blood pressure**  항목|  No==0  Yes==1 | **V5**  F==0  T==1 |  |  |  | **pcc**  notpresent==0  present==1 |
|  |  |  |  |  | **IBS**  <200==0  >=200==1 | **History of diabetes**  No==0  Yes==1 | **V10**  OC11==0  OC12==1  OC13==2  OC14==3 |  |  |  | **ba**  notpresent==0  present==1 |
|  |  |  |  |  | **Plasma**  Low==0  High==1 | **Family history of diabetes**  No==0  Yes==1 |  |  |  |  | **htn**  no==0  yes==1 |
|  |  |  |  |  | **Class**  0==P  1==M | **History of aborted baby**  No==0  Yes==1 |  |  |  |  | **cad**  no==0  yes==1 |
|  |  |  |  |  |  | **History of pregnacy**  No==0  Yes==1 |  |  |  |  | **appet**  poor==0  good==1 |
|  |  |  |  |  |  | **result of high blood pressure screening**  negative==0  new patient==1  old patient==2 |  |  |  |  | **pe**  no==0  yes==1 |
|  |  |  |  |  |  |  |  |  |  |  | **ane**  no==0  yes==1 |
|  |  |  |  |  |  |  |  |  |  |  | **dm**  No==0  Yes==1 |

* + extra tree classifier 을 이용하기 위해

( 이 기법은 class이외의 데이터는 nominal 보다는 numerical로 바꾸어 돌려야 했다.)

1. 5X2CV 시행: 열 순서에서 첫번째부터 다섯 번째로 순서대로 데이터를 재정렬하여 5번시행

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 각 분류 기법에 대한5X2CV 시행시 규칙: | | | | | | | | | | | | |
| 1회) | V1 | V1 | V1 | V1 | V1 | Age | Sex | V1 | V9 | V1 | Age | age |
| 2회) | V3 | V2 | V2 | V2 | V3 | Gender | Age | V2 | V10 | V2 | Number of sexual partners | bp |
| 3회) | V4 | V3 | V3 | V3 | V4 | BMI | History of high blood | V3 | V11 | V3 | First sexual intercourse | sg |
| 4회) | V5 | V4 | V4 | V4 | V5 | Blood | History of use drugs for high blood pressure | V4 | V12 | V4 | Num of pregnancies | al |
| 5회) | V6 | V5 | V5 | V5 | V6 | Hyper | Systolic blood pressure | V5 | V13 | V5 | Smokes | su |

1. [27]데이터와 [31]데이터의 LDA와 MDA 빈칸 채우기

1) QDA

by fitting a multivariate Gaussian to each class and applying Bayes' rule.

2)FLDA

Builds Fisher's Linear Discriminant function.

The threshold is selected so that the separator is half-way between centroids.

The class must be binary and all other attributes must be numeric.

Missing values are not permitted.

Constant attributes are removed using RemoveUseless.

No standardization or normalization of attributes is performed.

