

## Semestrální práce - report

### Bankroty

/Y336VD/

Tomáš Dlouhý

## Zadání

Použijte metodu k nejbližších sousedů nebo metodu rozhodovacích stromů ke klasifikaci cílového atributu zvoleného datasetu. Zvolte si jednu třídu jako cílovou - pozitivní. Naleznete nejlepší klasifikátor, který má false positive rate  $FPr < 0.3$ . Spočtete přesnost (accuracy) a true positive rate tohoto klasifikátoru.

## Data

K tvorbě semestrální práce byl zvolen dataset „Bankrot“, který obsahuje 44 ekonomických atributů v součtu 2050 firem. Některé firmy přežijí a jiné zbankrotují. Mezi nedostatky tohoto datasetu patří neexistence popisků, čímž je znemožněno získat relevanci těchto dat. Data je též nutné před zahájením práce normalizovat.

## Zpracování dat

Ke zpracování dat byl použit program OpenOffice Calc, pomocí kterého byly data normalizována na interval  $<0, 1>$  čímž byla zajištěna stejná váha atributů pro klasifikaci. Dataset byl následně rozdělen na dvě stejné části trénovací a testovací aby byl zajištěna dostatečná kvalita modelu.

Následně jsem musel nalézt vhodné  $k$ , které jsem získal pomocí iterace z intervalu  $<1; 30>$ , kdy jsem se zaměřil na jejich přesnost s pomocí parametru 2. Jako nejlepší  $k$  se ukázalo  $k=10$  s přesností  $accuracy=0,9844$ .

Pak jsem pro získání vzorců použil funkci programu RapidMiner BinomialClassification Performance a dosadil do vzorce:  $FPr = \frac{FP}{FP + TN}$  a výsledek byl 0,0315. Dále jsem spočítal  $Tpr = \frac{TP}{TP + FN} = 0,989$  a  $Acc = \frac{TP + TN}{TP + TN + FP + FN} = 0,9805$ .

## Závěr

Nejllepší klasifikátor byl s parametrem  $K=10$ , pro něj je  $Fpr = 0,0315$ ,  $Acc=0,9805$  a  $Tpr=0,38$

## Použitá literatura

[1] Stránky předmětu Y336VD - <http://ida.felk.cvut.cz/moodle>

## Přílohy

