Impact Studies and Sensitivity Analysis in Medical Data Mining

with ROC-based Genetic Learning

Michèle Sebag, Jérôme Azé, Noël Lucas

LRI, CNRS UMR 8623, Université Paris-Sud Orsay
Yet another learning criterion
Bradley 97, Provost et al. 98

Usual predictive accuracy \( \frac{a+d}{a+b+c+d} \)

\( \hat{t}_c \cdot t_c \) positive negative
positive a b
negative c d

**Drawbacks**
- skewed distributions (positive << negative)
- misclassification costs

Receiver Operating Characteristics (ROC) Curve:
True Positive Rate vs False Positive Rate

![ROC Curve Graph](image-url)
Optimizing the Area Under the ROC Curve
Mozer et al. NIPS 01, Flach et al., ICML 02

- Search space: Linear hypotheses
- Baseline: SVMs
- Criterion:
  \[ h \rightarrow \text{Order on examples} \]
  \[ \text{Quality}(h) = \text{Sum ranks positive examples} \]

NP-complete optimization \( \rightarrow \) Evolutionary Computation

Evaluation on Irvine repository
Artificial Evolution VI, 2004

Compared to linear SVMs: similar performances
smaller computational cost

Application to Medical Data Mining
Identify Atherosis Risk

PKDD Challenge 2002, 2003
Atherosclerosis

Experimental setting 2/3 training, 1/3 test

On each training set, 21 independent runs
Display the median ROC curve

ROGER: ROC-based Genetic LearneR vs SVM Torch
Influence Analysis - The tobacco factor

Procedure

A = { 100 non smokers }
B = { 100 heaviest smokers }
Sort individuals in A (in B) by increasing risk
Plot (i, risk(i))
Influence Analysis - The alcool factor

Procedure

A = { 100 light drinkers }
B = { 100 heavy drinkers }
Sort individuals in A (in B) by increasing risk
Plot (i, risk(i))

no light drinkers in the db...
Sensitivity Analysis - For free

21 runs, 21 solutions, 21 curves: 

\[(i, \text{weight(attribute}_i))\]
Conclusions - Perspectives

Present

• Good predictive performances
• Affordable complexity
• UNDERSTANDABLE RESULTS

Using Vision to Think, Card et al. 2001

Future

• Extend to kernel spaces
• Use for constructive induction