RULE-BASED CLASSIFIER

Basato sul rappresentare la conoscenza nel formato IF-THEN.

Esempio

IF AGE = "youth" AND STUDENT = "yes" THEN BUYS_PC = "yes"

COVERAGE e ACCURACY

M_{COVERS} = numero di tuple che la regola riesce a coprire

$M_{CORRECT}$ = numero di tuple che la regola riesce a classificare correttamente

$$Coverage(R) = \frac{M_{COVERS}}{|D|}$$

$$accuracy(R) = \frac{M_{CORRECT}}{M_{COVERS}}$$

Se più di una regola viene interpellata durante la classificazione avviene un CONFLITTO.

RISOLUZIONE DEI CONFLITTI

1) Ordine di dimensione:
Prima le regole che comprendono più attributi.

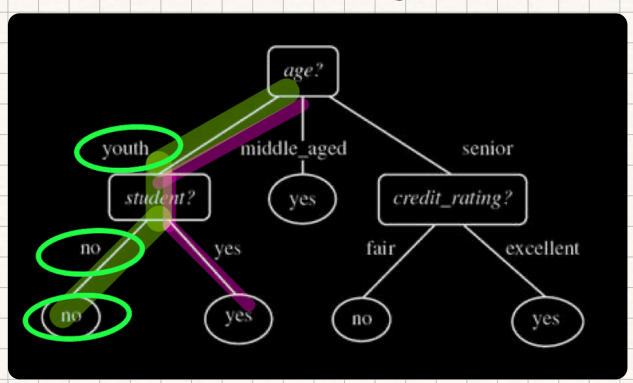
2) Ordinamento class-based:
Per esempio la regola che porta alla classe più prevalente ha la priorità.

3) Ordinamento rule-based
Semplicemente ordinando le regole secondo una qualche misura.

ESTRAZIONI DELLE REGOLE

DECISION TREE BASED

È possibile estrarre le regole da un decision tree, visitandolo dalla radice alle foglie, e ogni nodo di mezzo da vita a una regola



Ogni coppia attributo-valore di un path forma una congiunzione fra regole. Un rule-based creato dal DT è exclusive (non ci sono conflitti fra regole) e anche exhaustive (esiste una regola per ogni coppia attributo-valore: A_i, a_i)

Essendoci una regola per foglia, se il DT è in overfitting o/e presenta ripetizioni / replicazioni diventa difficile estrarre regole facilmente interpretabili

Pruning
Possiamo eliminare ogni condizione che non migliora l'accuratezza della regola.
Possiamo eliminare ogni regola che non migliora l'accuratezza totale del modello.
Attenzione: dopo questa procedura il risultato non è più esclusivo e esaustivo

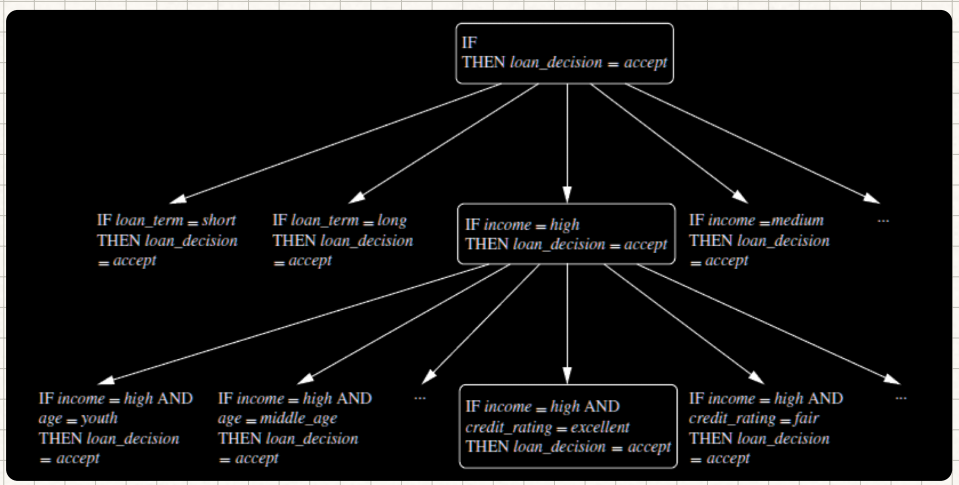
APPROCCIO EURISTICO

Questi algoritmi vengono definiti sequential covering algorithm, e hanno l'obiettivo di estrarre le regole direttamente dal set di training.
Le regole vengono estratte in modo sequenziale, ognuna per una data classe.

Passi:
1) Una regola per volta
2) Imparata una regola, rimuovere le tuple con la classe che la regola copre.
3) Ripetere finché non si verifica la stopping-cond.

Come costruire una regola?

Si parte dalla regola vuota, per poi considerare ogni combinazione dei valori degli attributi della tupla con classe C.



A questo punto bisogna valutare la qualità delle regole trovate e prendere la migliore.
Per valutare la regola non basta l'accuratezza ma bisogna considerare anche la copertura perché una regola che copre pochissimi casi può avere accuratezza massima.

FOIL-GAIN

È una misura di qualità della regola R rispetto a R':

$$\text{FOIL-GAIN} = pos' \cdot \left(\log_2 \frac{pos'}{pos'+neg'} - \log_2 \frac{pos}{pos+neg} \right)$$

pos e neg sono il numero di tuple positive e negative che la regola copre. (R)

pos' e neg' fanno riferimento alla nuova regola che si vuole confrontare con la vecchia. (R')

Questo tipo di approccio non usa un test-set e per questo la valutazione della regola è ottimistica.
Potrebbe essere necessario fare pruning sulla base di un set indipendente di dati. Per pruning si intende togliere congiunzioni alla regola.
Per valutare se potare o no si valuta:

$$\text{FOIL-PRUNE} = \frac{pos - neg}{pos + neg}$$

Il FOIL-PRUNE viene valutato per la regola R e per la sua versione potata R', se il FOIL-PRUN è maggiore per R' allora si procede col pruning.
Si eliminano una a una le congiunzioni finché non si migliora il FOIL-PRUNE.