Business Intelligence per i Big Data

Quaderno #1 – Classificazione

Obiettivo

Applicare algoritmi di data mining per la classificazione al fine di analizzare dati reali mediante l'utilizzo dell'applicazione RapidMiner.

Dataset

Il dataset Breast (Breast.xls, downloadable at http://dbdmg.polito.it/wordpress/teaching/businessintelligence/) raccoglie dati medici relativi a pazienti che hanno contratto il cancro al seno. Ciascun record del dataset corrisponde ad un paziente differente e contiene un insieme di caratteristiche relative al paziente, alla malattia e alla terapia seguita (ad es., l'età del paziente, la dimensione del tumore). A seconda se il tumore risulta essere un evento ricorrente o meno nella vita del paziente, ciascun record è inoltre etichettato come "Recurrence events" o "No-recurrence events". Quest'ultimo attributo, usato come attributo di classe durante l'esercitazione, è riportato come ultimo attributo di ciascun record.

La lista completa degli attributi del dataset da analizzare è riportata di seguito.

(1) Age
(2) Menopause
(3) Tumor-size
(4) Inv-nodes
(5) Node-caps
(6) Deg-malig
(7) Breast
(8) Breast-quad
(9) Irradiat
(10) **class (attributo di classe)**

Contesto di analisi

Degli esperti di oncologia sono interessati a predire la proprietà di ricorrenza o meno del tumore al seno sulla base delle caratteristiche del paziente, della malattia e della terapia seguita. A questo scopo, usano tre differenti algoritmi di classificazione: un albero di decisione (Decision Tree), un classificatore Bayesiano (Naïve Bayes), e un classificatore distance-based (K-NN). Il dataset Breast è utilizzato per la generazione dei modelli di classificazione e per la validazione delle loro performance.

Domande

Rispondere alle seguenti domande:

1. Generare un albero di decisione con l'algoritmo Decision Tree usando l'intero dataset per il training, settando il minimal gain a 0.01 e mantenendo la configurazione di default per gli altri parametri. (a) Quale attributo è considerato dall'algoritmo il più selettivo al fine di predire la classe di un nuovo dato di test? (b) Qual è l'altezza dell'albero di decisione generato? (c) Trovare un esempio di partizionamento puro all'interno dell'albero di decisione generato e riportare un screenshot che mostri l'esempio trovato.
2. Analizzare l'impatto del minimal gain (considerando il gain ratio come criterio di splitting) e del maximal depth sulle caratteristiche dell'albero di decisione generato dall'intero dataset (mantenendo la configurazione di default per gli altri parametri di configurazione). Riportare almeno 5 screenshot differenti che mostrino gli alberi di decisione (o porzioni di essi) generati con differenti configurazioni.
3. Applicando un 10-fold Stratified (paramentro Sampling type dell'operatore X-Validation) CrossValidation (X-Validation), qual è l'effetto del minimal gain e del maximal depth sull'accuratezza media ottenuta da Decision Tree? Riportare almeno 5 sceenshot che mostrino le matrici di confusione ottenute usando diverse configurazioni per i parametri sopra citati (considerare *almeno* le 5 configurazioni usate per rispondere alla domanda 2). Mantenere la configurazione di default per tutti gli altri parametri.
4. Considerando il classificatore K-Nearest Neighbor (K-NN) e applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del parametro K sull'accuratezza media del classificatore? Riportare almeno 5 sceenshot che mostrino le matrici di confusione ottenute usando diversi valori di K. Applicare un 10-fold Stratified Cross-Validation con il classificatore Naïve Bayes. K-NN ottiene mediamente prestazioni superiori o inferiori a Naïve Bayes classifier sul dataset analizzato? Riportare uno screenshot che mostri la matrice di confusione ottenuta con Naive Bayes sul dataset analizzato.
5. Analizzare la matrice di correlazione per valutare la correlazione tra coppie di attributi del dataset. Riportare uno screenshot che mostri la matrice di correlazione ottenuta. Alla luce dei risultati ottenuti, l'ipotesi d'indipendenza Naïve risulta valida per il dataset Breast? Qual è la coppia di attributi maggiormente correlati?

Assignment

Riportare sul quaderno:

- **Screenshot dei processi di analisi generati con RapidMiner;**
- **Commento di 1-2 pagine per descrivere i risultati dell'analisi includendo le risposte alle domande precedentemente riportate.**