Lab04

martedì 22 novembre 2022 11:41

ES3_Data...

Data Science e Tecnologie per le Basi di Dati

Esercitazione #3 – Data mining

Obiettivo

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

Dataset

Il dataset denominato Utenti (Utenti.xls, scaricabile dalla pagina del corso all'indirizzo http://dbdmg.polito.it/wordpress/teaching/data-science-e-tecnologie-per-basi-dati/laboratori, raccoglie dati anagrafici e lavorativi relativi agli utenti americani di un'azienda. Gli utenti sono classificati come "basic" o "premium" in base alla tipologia di servizi richiesti. Ciascun record del dataset fa riferimento ad un utente diverso. Il dataset contiene circa 42.000 utenti differenti e ad ognuno di essi sono riportati alcuni dati personali relativi all'utente (ad es., età, sesso, settore lavorativo principale) e la classe a lui assegnata ("basic" o "premium"). L'attributo relativo alla classe dell'utente, 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] Workclass
[3] FinWgt
[4] Education record
[5] Education-num
[6] Marital status
[7] Occupation
[8] Relationship
[9] Race
[10] Sex
[11] Capital Gain
[12] Capital loss
[13] Hours per week
[14] Native country
[15] **Class (attributo di classe)**

Contesto di analisi

Gli analisti della compagnia vogliono predire la classe di un nuovo utente sulla base delle caratteristiche degli utenti che hanno sottoscritto questa richiesta. A tale scopo, gli analisti decidono di utilizzare tre differenti algoritmi di classificazione: un albero di decisione (Decision Tree), un classificatore Bayesiano (Naive Bayes), e un classificatore di tipo distance based (K-NN). Il dataset Utenti è utilizzato per la generazione dei modelli di classificazione e per la valutazione delle loro performance.

Finalità dell'esercitazione

Lo scopo dell'esercitazione è generare e analizzare diversi modelli di classificazione e validarne le performance sul dataset Utenti mediante l'ausilio del tool Rapid Miner. All'interno di Rapid Miner saranno generati diversi process. Per valutare le performance dei classificatori saranno testate diverse configurazioni e i rispettivi risultati saranno confrontati tra loro. Per la validazione delle performance sarà applicato un processo di Top 10-fold Stratified Cross-Validation. I risultati ottenuti saranno analizzati al fine di capire l'impatto dei principali parametri di input di ciascun algoritmo sulle performance di classificazione.

Domande

Rispondere alle seguenti domande:

1. Generare un albero di decisione usando l'intero dataset per il training e la configurazione di default per l'algoritmo Decision Tree.
   - (a) Quale attributo è considerato dall'algoritmo il più selettivo al fine di predire la classe di una nuova data 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.
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.
3. Cosa accade se modifichiamo l'attributo di classe da "Service Class" a "Native Country"? Rispondere nuovamente alla domanda (1) in questo nuovo scenario.
4. Considerando il nuovo attributo "Service Class" come attributo di classe e applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del minimal gain e del maximal depth sull'accuratezza media ottenuta dal Decision Tree? Confrontare le matrici di confusione ottenute usando diverse configurazioni per i parametri sopra citati (mantenere la configurazione di default per tutti gli altri parametri).
5. Considerando il classificatore K-Nearest Neighbor (K-NN) e applicando un 10-fold Stratified Cross-Validation, qual è l'effetto nel parametro K sulle performance del classificatore? Confrontare 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 classificando questo dataset anagrafico?
6. Analizzare la matrice di correlazione per valutare la correlazione tra coppie di attributi del dataset. Alla luce dei risultati ottenuti, l'ipotesi d'indipendenza Naïve risulta valida per il dataset Utenti?

Esercitazione

Installazione e configurazione del programma

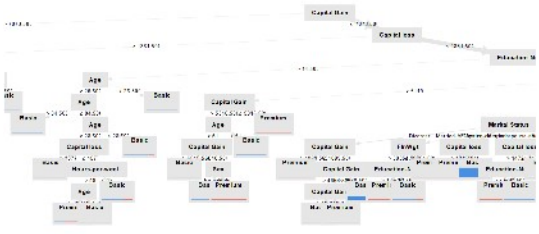
Lanciare RapidMiner in ambiente Windows

Generazione e analisi del processo

- Creare un nuovo processo in Rapid Miner.

1.

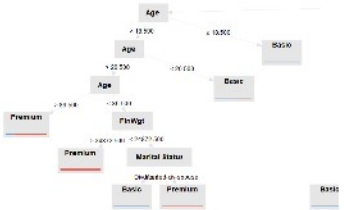
(a) L'attributo più selettivo è 'Capital Gain'



(b) Altezza dell'albero di decisione generato: 9 (considerando come altezza la lunghezza massima di un percorso che collega la radice ad una foglia)

(c) Un partizionamento puro è uno split sui valori di un attributo tale per cui i record corrispondenti appartengono tutti alla medesima classe

Per esempio, nella figura sottostante è possibile vedere che i valori dell'attributo Age sono splittati in due gruppi: >20.500, ≤20.500. Mentre la prima partizione è impura, perché copre record etichettati sia con la classe 'Basic' che 'Premium', la seconda è pura perché tutte le relative istanze appartengono alla classe 'Basic'.



2.

Il parametro "maximal depth" permette di specificare l'altezza massima dell'albero di decisione generato. Usando la configurazione di default (con maximal depth = 20) l'altezza

Installazione e configurazione del programma

- Lanciare RapidMiner in ambiente Windows

Generazione e analisi del processo

- Creare un nuovo processo in Rapid Miner.
- Comporre il flusso del processo di data mining da eseguire selezionando e trascinando gli operatori disponibili sul menu a sinistra all'interno della finestra relativa al processo principale.

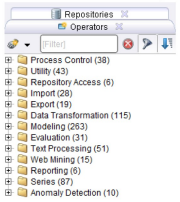


Figura 1. Operatori.

- Per gestire l'esecuzione del processo usare i pulsanti Start/Stop/Pause. Per visualizzare I risultati del processo, cambiare la prospettiva di visualizzazione da *Design* a *Results*.

Figura 2. Pulsanti di esecuzione / modifica della prospettiva di visualizzazione.

- Visionare il contenuto del dataset Utenti, disponibile in formato Excel (.xls).
- Importare i dati d'ingresso all'interno del processo di Data Mining principale mediante l'uso dell'operatore "Read Excel". Per importare i dati correttamente usare il *Data Import Wizard* configurando l'operatore come segue:
  - Selezionare il file sorgente desiderato (Step 1).
  - Selezionare tutte le celle del foglio di lavoro in cui sono contenuti i dati (Step 2).
  - Annotare la prima riga come quella contenente i nomi degli attributi (etichetta "name"), mantenendo non etichettate ("-") le righe dei dati allo Step 3.
  - Collegare il blocco di data import con il data source. Identificare il ruolo dell'attributo "Service class" come "label", ovvero "etichetta di classe" (Step 4).
- Includere, in coda al processo di mining, l'operatore relativo al classificatore "Decision Tree". Il processo costruito finora sarà analogo al seguente:

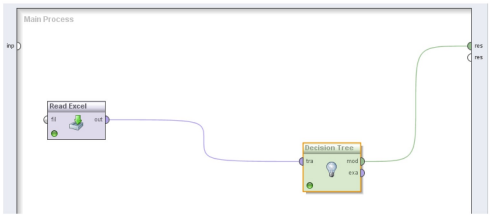


Figura 3. Processo di classificazione – Albero di decisione

- Eseguire il processo e analizzare l'albero di decisione generato mediante la Results perspective.
- Cambiare la configurazione dei parametri d'ingresso del classificatore cliccando sull'operatore Decision Tree e modificando le impostazioni relative nel menu posto sulla destra all'interno della Design perspective. In particolare, modificare i valori di maximal depth e di minimal gain, mantentendo inalterati tutti gli altri parametri, per analizzare il loro effetto sulle caratteristiche principali del modello di classificazione generato.
- Cliccare sull'operatore "Read Excel" al fine di poter modificare le impostazioni relative. Cambiare l'attributo di classe all'interno tra le "Data set metadata information" da "Service Class" a "Native Country" (altrimenti, rieseguire l'intero processo di data import mediante il wizard selezionando il nuovo attributo di classe allo Step 4).
- Rieseguire il processo per generare un nuovo albero di decisione.
- Modificare il flusso del processo principale per poter eseguire una 10-fold Stratified Cross-Validation. A tale scopo, come primo passo includere il blocco "Validation" al posto di Decision Tree nel processo principale.

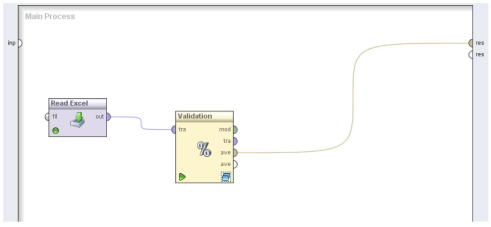


Figura 4. 10-Fold Cross-Validation.

Come passo successivo, fare doppio click sull'operatore "Validation" e creare un processo innestato analogo a quello sotto riportato:

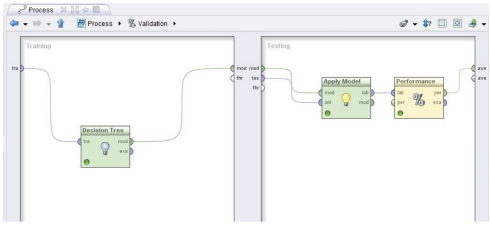


Figure 5. Validation subprocess.

- Tornare al Results perspective e analizzare la matrice di confusione generata dal processo di

2. Il parametro "maximal depth" permette di specificare l'altezza massima dell'albero di decisione generato. Usando la configurazione di default (con maximal depth = 20) l'altezza dell'albero risulta essere 15 e quindi il processo ricorsivo di learning dell'albero viene completato. Al contrario, settando un valore di maximal depth inferior a 15 (ad es. 5) la ricorsione viene interrotta e quindi la qualità del modello generato potenzialmente decresce. Il parametro "minimal gain" permette di scegliere se splittare ulteriormente un nodo dell'albero oppure no. In particolare, un nodo viene splittato se il suo gain è superiore alla soglia minima (minimal gain). Valori elevati di minimal gain producono un numero limitato di partizionamenti e, di conseguenza, alberi di decisione più piccoli. Valori troppo elevati di minimal gain (ad es., 0.9) impediscono completamente lo split dei valori degli attributi e quindi l'albero risultante conterrà un singolo nodo. Dato che il valore di default di minimal gain è moderatamente basso (ad es., 0.1), esso produce generalmente uno splitting degli attributi abbastanza fitto.

3. Impostando come attributo di classe "Native Country" l'analista può predire la nazionalità dell'utente che sottomette una nuova richiesta di servizio sulla base delle richieste passate e delle caratteristiche degli utenti che le hanno sottomesso.

   (a) L'attributo più selettivo è 'Education-num'
   (b) Altezza dell'albero: 8
   (c) Esempio di partizionamento puro

4. Matrici di confusione

accuracy: 85.09%

	true Basic	true Premium	class precision
pred. Basic	6991	1031	87.15%
pred. Premium	425	1321	75.66%
class recall	94.27%	56.16%	

Maximal depth: 20, Minimal gain:0.01

accuracy: 82.82%

	true Basic	true Premium	class precision
pred. Basic	7323	1585	82.21%
pred. Premium	93	767	89.19%
class recall	98.75%	32.61%	

Maximal depth: 10, Minimal gain: 0.01

accuracy: 82.37%

	true Basic	true Premium	class precision
pred. Basic	7294	1600	82.01%
pred. Premium	122	752	86.04%
class recall	98.35%	31.97%	

Maximal depth: 5, Minimal gain: 0.01

accuracy: 85.37%

	true Basic	true Premium	class precision
pred. Basic	7029	1042	87.09%
pred. Premium	387	1310	77.20%
class recall	94.78%	55.70%	

Maximal depth:20, Minimal gain: 0.05

accuracy: 82.31%

	true Basic	true Premium	class precision
pred. Basic	7324	1636	81.74%
pred. Premium	92	716	88.61%
class recall	98.76%	30.44%	

Maximal depth:20, Minimal gain:0.1

accuracy: 80.20%

	true Basic	true Premium	class precision
pred. Basic	7410	1928	79.35%
pred. Premium	6	424	98.60%
class recall	99.92%	18.03%	

Maximal depth:20, Minimal gain: 0.2

5. **accuracy: 72.27%**

	true Basic	true Premium	class precision
pred. Basic	5971	1264	82.53%
pred. Premium	1445	1088	42.95%
class recall	80.52%	46.26%	

K-NN. Matrice di confusione. K=1

accuracy: 75.50%

	true Basic	true Premium	class precision
pred. Basic	6513	1490	81.38%
pred. Premium	903	862	48.84%
class recall	87.82%	36.65%	

K-NN. Matrice di confusione. K=3

accuracy: 76.50%

	true Basic	true Premium	class precision
pred. Basic	6742	1621	80.62%
pred. Premium	674	731	52.03%
class recall	90.91%	31.08%	

K-NN. Matrice di confusione. K=5

accuracy: 78.85%

	true Basic	true Premium	class precision
pred. Basic	7101	1751	80.22%

Figure 5. Validation subprocess.

- Tornare al Results perspective e analizzare la matrice di confusione generata dal processo di validazione.
- Disabilitare temporaneamente l'operatore Decision Tree (cliccando col tasto destro sull'operatore e eliminando il segno di spunta a lato di "Enable Operator"). Sostituire l'operatore Decision Tree con Naïve Bayes prima e con K-NN successivamente.
- Confrontare le performance di K-NN e Naïve Bayes, in termini di accuratezza media, precisione e richiamo, analizzando le rispettive matrici di confusione. Per il classificatore K-NN, variare i valori del parametro K usando il menu sul lato destro nella Design perspective.
- Per analizzare la matrice di correlazione associata al dataset in esame tornare al processo principale (click sul pulsante "Process"), disabilitare temporaneamente l'operatore Validation (cliccando col tasto destro sull'operatore e eliminando il segno di spunta a lato di "Enable Operator"), inserire l'operatore "Correlation Matrix" in coda al processo e visualizzare la rispettiva matrice collegando il plug-in del blocco denominato "mat" al plug-in "Result" sulla destra della finestra del processo principale. Il processo così generato sarà analogo al seguente:

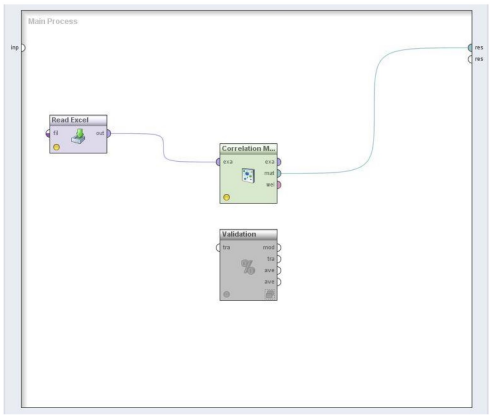


Figure 6. Matrice di correlazione tra attributi

Tornando al Results perspective, per ordinare le correlazioni trovate tra coppie di attributi in ordine decrescente selezionare la "Pairwise Table view" e cliccare sul campo "Correlation" della tabella visualizzata.

K-NN. Matrice di confusione. K=5

accuracy: 78.85%

	true Basic	true Premium	class precision
pred. Basic	7101	1751	80.22%
pred. Premium	315	601	65.61%
class recall	95.75%	25.55%	

K-NN. Matrice di confusione. K=10

accuracy: 79.93%

	true Basic	true Premium	class precision
pred. Basic	7268	1812	80.04%
pred. Premium	148	540	78.49%
class recall	98.00%	22.96%	

K-NN. Matrice di confusione. K=15

accuracy: 79.98%

	true Basic	true Premium	class precision
pred. Basic	7306	1846	79.83%
pred. Premium	110	506	82.14%
class recall	98.52%	21.51%	

K-NN. Matrice di confusione. K=20

accuracy: 83.23%

	true Basic	true Premium	class precision
pred. Basic	6924	1146	85.80%
pred. Premium	492	1206	71.02%
class recall	93.37%	51.28%	

Naïve-Bayes

Come mostrato nelle figure, Naïve Bayes ottiene un'accuratezza media più elevata di K-NN (83.23% contro79.98%) sul dataset analizzato.

6.

Attributes	Workclas...	Workcl...	Workcl...	Workcl...	Workcl...	Workcl...	Workcl...	Workcl...	Workcl...	Educati...	Educati...	Educati...
Workclass = State-gov	1	-0.059	-0.309	-0.036	-0.053	-0.050	-0.038	-0.004	-0.003	0.024	-0.051	-0.028
Workclass = Self-emp-not-inc	-0.059	1	-0.441	-0.051	-0.076	-0.071	-0.055	-0.006	-0.004	-0.006	0.011	-0.019
Workclass = Private	-0.309	-0.441	1	-0.264	-0.398	-0.371	-0.286	-0.031	-0.022	-0.033	0.066	0.037
Workclass = Federal-gov	-0.036	-0.051	-0.264	1	-0.046	-0.043	-0.033	-0.004	-0.003	0.027	-0.018	-0.025
Workclass = Local-gov	-0.053	-0.076	-0.398	-0.046	1	-0.064	-0.049	-0.005	-0.004	0.045	-0.046	-0.027
Workclass = ?	-0.050	-0.071	-0.371	-0.043	-0.064	1	-0.046	-0.005	-0.004	-0.046	-0.017	0.037
Workclass = Self-emp-inc	-0.038	-0.055	-0.286	-0.033	-0.049	-0.046	1	-0.004	-0.003	0.041	-0.029	-0.024
Workclass = Without-pay	-0.004	-0.006	-0.031	-0.004	-0.005	-0.005	-0.004	1	-0.000	-0.009	0.014	-0.004
Workclass = Never-worked	-0.003	-0.004	-0.022	-0.003	-0.004	-0.004	-0.003	-0.000	1	-0.007	-0.006	0.008
Education = Bachelors	0.024	-0.006	-0.033	0.027	0.045	-0.046	0.041	-0.009	-0.007	1	-0.305	-0.086
Education = HS-grad	-0.051	0.011	0.066	-0.018	-0.046	-0.017	-0.029	0.014	-0.006	-0.306	1	-0.133
Education = 11th	-0.028	-0.019	0.037	-0.025	-0.027	0.037	-0.024	-0.004	0.008	-0.086	-0.133	1
Education = Masters	0.070	-0.005	-0.092	0.013	0.129	-0.029	0.015	-0.005	-0.003	-0.105	-0.163	-0.046

La figura mostra la matrice di correlazione ottenuta dal dataset analizzato. Essa riporta la correlazione mutua (e simmetrica) tra coppie di attributi. Per esempio, l'attributo "Age" risulta essere molto correlato con l'attributo "Marital status" Dato che sussistono correlazioni significative tra attributi, ad es., tra "Age" e "Marital Status" (correlazione = 0.316), tra "Sex" e "Relationship" (correlazione = 0.319), l'ipotesi Naïve risulta essere irrealistica per il dataset analizzato. Tuttavia, le performance di Naïve Bayes risultano essere mediamente buone.

First Attribute	Second Attribute	Correlation ↓
Occupation = Prof-specialty	Education-Num	0.419
Race = Asian-Pac-Islander	Native Country = Philipp...	0.408
Education = Masters	Education-Num	0.360
Marital Status = Divorced	Relationship = Unmarried	0.329
Relationship = Unmarried	Sex = Female	0.321
Relationship = Wife	Sex = Female	0.319
Relationship = Husband	Age	0.317
Marital Status = Married-civ-spouse	Age	0.316
Education = 5th-6th	Native Country = Mexico	0.310
Marital Status = Never-married	Relationship = Not-in-fa...	0.297
Marital Status = Widowed	Age	0.265
Education = Prof-school	Occupation = Prof-speci...	0.265
Occupation = Adm-clerical	Sex = Female	0.263