Business Intelligence

Esercitazione di laboratorio N. 4

L'obiettivo dell'esercitazione è:

- **utilizzare il software Rapid Miner per preparare i dati relativi ad una campagna promozionale e i dati testuali relativi ad un argomento specifico per analisi successive**

Dati strutturati

Il dataset denominato UsersSmall (UsersSmall.xls) è disponibile sul sito del corso (http://dbdmg.polito.it/wordpress/teaching/business-intelligence/). Esso raccoglie dati anagrafici e lavorativi relativi a circa 300 persone contattate da un'azienda per proporgli l'iscrizione ad un loro servizio. Per tali utenti è noto se, dopo essere stati contattati, si sono iscritti al servizio proposto oppure no (valore del campo Response).

La lista completa degli attributi del dataset a disposizione (UsersSmall.xls) è riportata di seguito.

(1) Age
(2) Workclass
(3) Education record
(4) Marital status
(5) Occupation
(6) Relationship
(7) Race
(8) Sex
(9) Hours per week
(10) Native country
(11) Response.

Dati testuali

Il dataset denominati ObamaNews (ObamaNews.zip) è disponibile sul sito del corso (http://dbdmg.polito.it/wordpress/teaching/business-intelligence/). Esso contiene una collezione di news scaricate mediante il servizio Google News. La collezione rappresenta l'insieme delle prime 10 news (pagine contenenti notizie) restituite da Google News a fronte della specifica della parola chiave *Obama*.

Preparazione dei dati strutturati

Obiettivo 1 – Import dei dati

Importare il dataset UsersSmall in Rapid Miner (operatore *Read Excel*).

Configurare i parametri di import attraverso la procedura guidata (*Configuration Wizard*).

Analizzare la semantica degli attributi e il loro ruolo a seconda degli obiettivi dell'analisi svolta.

Obiettivo 2 – Gestione dei dati mancanti

Verificare la presenza di eventuali dati mancanti e gestirli con opportuni passi di trasformazione (operatori *Decleare Missing values* e *Replace Missing Values*).

Obiettivo 3 – Discretizzazione

Verificare la presenza di attributi continui nei dati di origine.

Discutere l'eventuale necessità di applicare un processo preliminare di discretizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati.

Applicare diverse tecniche di discretizzazione (operatori *Discretize by binning*, *Discretize by frequency*, *Discretize by size*, *Discretize by entropy*) e confrontare i risultati.

Obiettivo 4 – Normalizzazione

Discutere l'eventuale necessità di applicare un processo preliminare di normalizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati (operatore *Normalize*).

Obiettivo 5 – Analisi delle correlazioni

Analizzare la correlazione tra coppie di attributi (operatore *Correlation Matrix*).

Obiettivo 6 – Feature selection

Selezionare gli attributi d'interesse in funzione delle obiettivi delle analisi successive (operatore *Select attributes*).

Preparazione dei dati testuali

Obiettivo 1 – Import dei dati

Importare il dataset ObamaNews in Rapid Miner (operatore *Process Documents From Files*).

Obiettivo 2 – Generazione dei token

Dividere il testo in parole (operatore *Tokenize*) e uniformare maiuscole/minuscole (operatore *Transform cases*).

Obiettivo 3 – Stopword e stemming

Applicare algoritmi di stemming (operatore *Stemming (Snowball)*) e di eliminazione dello stopword sui dati di origine. Per le stopword usare l'operatore *Filter Stopwords* (Dictionary) e specificare come file contenente le stopword il file *stopwordsItalian.txt* disponibile sul sito del corso (Impostare il tipo di encoding a UTF-8 per una corretta interpretazione del contenuto del file delle stopword).

Obiettivo 4 – Matrice delle occorrenze e del tf-idf

Generare le matrici delle occorrenze e del tf-idf dei termini singoli. Confrontare le matrici ottenute.