IL DATA MINING: INTRODUZIONE E CLASSIFICAZIONE DELLE TECNICHE

Slide 2 – Sommario

Benvenuti!

In questa lezione vedremo inizialmente che cosa si intende per Data Mining.

Presenteremo quindi una classificazione delle principali tecniche di Data Mining evidenziandone le caratteristiche fondamentali.

Cominciamo…

Slide 3 – Definizione di Data Mining

Con Data Mining si intende un processo iterativo per analizzare grandi quantità di dati al fine di estrarre informazione strategica che sia:

- accurata
- utile a fini decisionali
- precedentemente sconosciuta

Le informazioni strategiche estratte rappresentano quindi nuova conoscenza.

Slide 4 – Definizione di Data Mining (segue)

I processi di Data Mining si basano su metodi di **apprendimento induttivo**.

Questi metodi permettono di trarre conclusioni da un insieme di dati e di generalizzare le conclusioni individuate ad altri dati inizialmente non noti, nel modo più accurato possibile.

Gli obiettivi dei metodi di Data Mining sono due:

- alcuni metodi, detti metodi **descrittivi o interpretativi**, si pongono l'obiettivo di descrivere i dati, rappresentandone in modo efficace le regolarità.
- altri, detti metodi **predittivi,** predicono il valore che una variabile assumerà in futuro o stimano la probabilità di un evento futuro.

I dati a cui vengono applicati i metodi di Data Mining sono generalmente memorizzati in un Data Warehouse e gli strumenti di Data Mining sono offerti come strumenti di front-end dal sistema.

Slide 5 – Confronto tra gli strumenti di front-end in un Data Warehouse

A questo punto è importante confrontare gli obiettivi dei diversi strumenti di front-end disponibili in un Data Warehouse.

- Come abbiamo visto nella lezione 3, gli strumenti OLAP permettono l'estrazione di dati aggregati a partire da informazioni di dettaglio; questi dati aggregati rappresentano nuova conoscenza. Si tratta di strumenti utili, per esempio, per analizzare le vendite rispetto alle categorie di prodotto, al variare del mese di vendita.

- Gli strumenti statistici permettono, invece, di verificare sui dati alcune ipotesi formulate validando l'informazione aggregata. Per esempio, permettono di analizzare la varianza del ricavo di vendita rispetto alla categoria del prodotto.
- Infine, gli strumenti di Data Mining permettono, come abbiamo detto, di individuare pattern e ricorrenze nei dati di dettaglio, generando, così, conoscenza in termini di modelli e di pattern.

Ad esempio, permettono di caratterizzare i prodotti con caratteristiche simili rispetto alla vendita.

Slide 6 – Origini del Data Mining

Il Data Mining è una disciplina che si ispira a idee e tecniche derivate da diversi settori, tra cui: Machine learning, Statistical learning, Pattern recognition, Sistemi di gestione dati.

Queste tecniche vengono, però, contestualizzate in un processo analitico complesso, volto al supporto decisionale.

Slide 7 – Processo di Data Mining

Questa slide esemplifica l'intero processo di Data Mining, assumendo che i dati da analizzare siano memorizzati in un Data Mart.

Per prima cosa è necessario eseguire un'analisi esplorativa, tramite metodi grafici o statistici per determinare le caratteristiche degli attributi del Data Mart.

Quindi, viene valutata la rilevanza degli attributi rispetto agli obiettivi dell'analisi e vengono selezionati solo gli attributi di interesse.

Infine, le tecniche di Data Mining vengono applicate per generare modelli e pattern di conoscenza, utili a fini predittivi e interpretativi.

Si osservi che il processo in ogni momento può richiedere una revisione di quanto svolto nei passi precedenti.

Slide 8 – Esempi di strumenti di Data Mining

Esistono molteplici strumenti di Data Mining, alcuni commerciali, come SAS, altri disponibili con licenza gratuita o open source, come Weka.

Slide 9 – Terminologia

Prima di classificare le principali tecniche di Data Mining, è opportuno introdurre un po' di terminologia.

L'input di un metodo di Data Mining è un dataset di individui o istanze. Se il dataset corrisponde a un Data Mart, gli individui sono le righe delle tabelle.

Gli attributi delle istanze rappresentano le variabili chiamate anche **feature** o **elementi**.

Le variabili possono essere di due tipi:

- Sono **categoriche** se assumono un numero finito di valori distinti, rappresentano, cioè, una proprietà qualitativa. Le variabili categoriche, a loro volta, possono essere variabili booleane, ordinali, se i valori della variabile si possono ordinare, o nominali, se i valori non sono ordinabili in modo intuitivo.

- Sono **numeriche** se assumono un insieme finito o infinito di valori numerici. Le variabili numeriche si suddividono, a loro volta, in variabili discrete e continue.

Slide 10 – Classificazioni tecniche di Data Mining

Possiamo adesso passare a classificare le tecniche di Mining.

La prima grande suddivisione distingue tra **tecniche supervisionate** e **tecniche non supervisionate**.

Le prime sono guidate da un attributo target.

L'obiettivo di queste tecniche è spiegare i valori dell'attributo target, che rappresenta una variabile dipendente, rispetto ai valori di un insieme di variabili predittive, che rappresentano variabili indipendenti.

I principali approcci di tipo supervisionato sono la **classificazione** e la **regressione**.

Le tecniche non supervisionate, invece, non sono guidate da un attributo target e quindi non distinguono tra variabili dipendenti e indipendenti.

I principali approcci di tipo non supervisionato sono il **clustering** e il **pattern discovery**.

Cerchiamo adesso di capire qual è l'obiettivo delle tecniche citate, partendo dalle tecniche supervisionate.

Slide 11 – Classificazione

Iniziamo dalla classificazione.

In questo caso, l'attributo target è **categorico** e associa ogni individuo a una classe (un gruppo).

Il dataset di input viene chiamato **training set**.

La **classificazione** si pone un duplice obiettivo.

Il primo obiettivo è di tipo i**nterpretativo** e consiste nel creare un modello che spieghi le relazioni tra l'attributo target e gli altri attributi nel training set. Per validare il modello, cioè per determinarne l'accuratezza, viene utilizzato un dataset di test.

Il secondo obiettivo è di tipo **predittivo** e consiste nell'assegnare una classe agli individui di un insieme precedentemente sconosciuto, utilizzando il modello creato in precedenza.

Slide 12 – Classificazione – applicazioni

I contesti nei quali le tecniche di classificazione trovano impiego sono molteplici.

Tra gli esempi di possibili applicazioni ricordiamo i seguenti:

- Classificare i clienti di una agenzia assicurativa rispetto al rischio
- Classificare le transazioni con carta di credito come legittime o fraudolente
- Classificare una cellula tumorale come benigna o maligna
- Classificare le notizie riportate in un sito web rispetto al dominio di riferimento.

Slide 13 – Classificazione – esempio

Vediamo adesso un esempio concreto.

Supponiamo di considerare un dataset di training in cui ogni istanza rappresenta il tipo di auto posseduta da un individuo con una certa età.

L'attributo target è rappresentato dalla classe di rischio associata all'età e al tipo di auto. Le possibili classi rappresentano un rischio basso e un rischio alto.

Durante la fase di creazione del modello, il training set viene analizzato con l'obiettivo di estrarre regole che determinino il valore della variabile dipendente "Classe di rischio" rispetto al valore delle variabili indipendenti, "Età" e "Tipo auto".

Queste regole vengono spesso rappresentate come alberi, chiamati **alberi di decisione**.

Slide 14 – Classificazione – esempio (segue)

Durante la fase di validazione del modello, il classificatore viene utilizzato su un insieme di test per valutarne l'accuratezza.

Le regole vengono applicate a ogni istanza per dedurre un valore per la variabile target e verificarne la correttezza.

Nell'esempio, l'istanza in rosso non viene correttamente classificata.

Slide 15 – Classificazione – esempio (segue)

Infine, nella fase di previsione, il modello viene applicato a un dataset di istanze per dedurre un valore per l'attributo target.

Nell'esempio, all'istanza con età 35 anni e tipo auto "utilitaria" viene associata una classe di rischio bassa.

Slide 16 – Classificazione – algoritmi

Esistono moltissimi algoritmi di classificazione, implementati dai più noti strumenti di Data Mining. Tra questi ricordiamo: algoritmi per generare alberi di decisione, algoritmi Bayesiani, algoritmi di regressione logistica, reti neurali e support vector machine.

Slide 17 – Regressione

Passiamo adesso ad un altro tipo di tecnica di Data Mining supervisionata, la **regressione**.

In questo caso, l'attributo target è continuo, ma l'approccio è simile a quello discusso per la classificazione:

- da un punto di vista **interpretativo**, la regressione permette di descrivere la dipendenza tra la variabile dipendente e le variabili indipendenti, partendo dalle istanze del training set, attraverso una funzione. Anche in questo caso, viene utilizzato un test set per validare il modello, cioè determinarne l'accuratezza.

- da un punto di vista predittivo, permette, poi, di predire il valore della variabile target per ogni istanza di un insieme precedentemente sconosciuto, utilizzando la funzione individuata.

Slide 18 – Regressione – applicazioni

Anche la regressione trova impiego in moltissimi contesti.

Permette, ad esempio, di:

- Predire il totale delle vendite di un nuovo prodotto sulla base della spesa per la relativa campagna pubblicitaria
- Predire la velocità del vento in funzione di temperature, umidità e pressione
- Predire gli anni di vita attesa di una persona rispetto al peso.

Vediamo adesso un esempio concreto.

Slide 19 – Regressione – esempio

Consideriamo come variabile dipendente gli anni di vita attesa per una persona e come variabile indipendente il suo peso.

La regressione permette di individuare la funzione che, partendo da un training set in cui ogni istanza registri gli anni di vita attesa e il peso delle persone, permetta di calcolare gli anni a partire dal peso.

Un esempio di funzione potrebbe calcolare gli anni di vita attesa moltiplicando il peso per 0.8 e sottraendo il risultato da 124.

La funzione ottenuta può, poi, essere utilizzata per predire gli anni di vita attesa di individui non inclusi nel training set.

Ad esempio, per un individuo che pesa 50kg, gli anni di vita attesa stimati saranno 84.

Slide 20 – Regressione – algoritmi

Gli algoritmi di regressione differiscono per molteplici fattori, tra cui ricordiamo il numero di variabili indipendenti considerate e il tipo di funzione appresa.

Quando la funzione è lineare rispetto a una singola variabile indipendente si parla di algoritmi di regressione lineare semplice.

Gli algoritmi di regressione lineare multipla assumono, invece, che esista più di una variabile indipendente.

Slide 21 – Clustering

Passiamo adesso ad analizzare le principali tecniche di Mining non supervisionato.

La prima tecnica considerata è il **clustering**.

In questo caso l'obiettivo è **interpretativo** e consiste nell'individuare i sottogruppi, chiamati **cluster**, di una popolazione contenenti individui simili, quindi omogenei, rispetto alle variabili che li descrivono.

Nell'individuare questi cluster si cerca di minimizzare la distanza tra individui appartenenti allo stesso cluster e massimizzare la distanza tra individui appartenenti a cluster diversi.

Le tecniche di clustering si basano quindi su una **funzione di distanza**.

La scelta della funzione di distanza dipende dal tipo di variabili considerate: variabili continue o categoriche porteranno, quindi, a utilizzare funzioni di distanza differenti.

Poiché il clustering genera sottoinsiemi di dati simili, si presta ad essere utilizzato come passo iniziale prima di applicare altre tecniche di Mining su ogni cluster.

Slide 22 – Clustering – applicazioni

Le tecniche di clustering sono molto utilizzate.

Tra i possibili impieghi ricordiamo i seguenti:

- Clusterizzare i clienti di una azienda rispetto alle informazioni personali e alle caratteristiche di acquisto in modo da inviare agli individui di ogni cluster informazioni pubblicitarie mirate
- Clusterizzare un insieme di documenti rispetto alle parole chiave che contengono in modo da suddividerli in gruppi di affinità tematica
- Clusterizzare i clienti di una agenzia assicurativa rispetto alle richieste di rimborso presentate
- Clusterizzare gli appartamenti in vendita in una certa zona rispetto alle loro caratteristiche.

Slide 23 – Clustering – esempio

La figura che vi mostro adesso illustra l'effetto dell'applicazione di una tecnica di clustering.

Utilizzando un approccio iterativo, il dataset iniziale viene suddiviso in sottoinsiemi — in questo caso particolare disgiunti — di istanze simili.

Il processo viene ripetuto fino a generare un numero di cluster predefinito: 3 nell'esempio in figura.

Slide 24 – Clustering – algoritmi

Gli algoritmi di clustering possono essere suddivisi in tre categorie principali.

Gli algoritmi di **partizionamento** operano in modo analogo a quanto descritto nell'esempio precedente, producono quindi cluster che rappresentano un partizionamento del dataset iniziale e sono quindi disgiunti. L'algoritmo di partizionamento più noto è l'algoritmo chiamato K-MEANS.

Gli algoritmi **gerarchici**, invece, creano una decomposizione gerarchica del dataset che non rappresenta necessariamente una partizione del dataset iniziale.

Infine, gli algoritmi **basati sulla densità**, come DBSCAN, individuano i cluster cercando le aree dense di punti nel dataset.

Slide 25 – Pattern discovery

La seconda tecnica di Mining non supervisionato è nota come pattern discovery.

In questo caso l'obiettivo è estrarre dai dati pattern di conoscenza.

Questi pattern di conoscenza vengono frequentemente rappresentati come regole di associazione del tipo IF X THEN Y (o X → Y) dove X e Y rappresentano valori di variabili distinte.

Per esempio, la regola IF Età > 26 and TipoAuto ='sportiva' THEN Rischio = 'alto', già incontrata durante la presentazione delle tecniche di classificazione, dice che se in una istanza il valore per l'attributo età è superiore a 26 e il tipo dell'auto è 'sportiva', allora il rischio di incidente è alto.

Più in generale, se le istanze contengono un insieme di elementi T, presi da una certa collezione, la regola X → Y indica che, se T contiene X, allora molto probabilmente contiene anche Y.

Le tecniche di pattern discovery hanno un duplice obiettivo: permettono di descrivere gli individui nel dataset utilizzando le regole di associazione e di predire le caratteristiche di nuovi dati utilizzando le regole individuate.

Slide 26 – Pattern discovery – applicazioni

Anche le regole di associazione trovano molteplici ambiti di applicazione.

Ecco alcuni esempi:

- individuare i prodotti venduti insieme per migliorare le strategie di vendita;
- individuare quali libri vengono comprati da chi compra il libro "La Divina Commedia;"
- individuare se una richiesta di congedo per malattia è sempre collegata a una richiesta di visita a un medico nello stesso periodo;
- individuare quali pagine web vengono accedute frequentemente nell'ambito della stessa sessione.

Slide 27 – Pattern discovery – esempio

Un esempio molto noto è quello che individua regole di associazione a partire dalle informazioni relative a un insieme di transazioni di acquisto.

Ogni transazione di acquisto è rappresentata da una riga di una tabella in cui, oltre all'identificatore, troviamo un insieme di prodotti.

Le regole individuate a partire da questo dataset sono due. La prima indica che chi compra latte compra anche la coca cola; la seconda dice che, nelle transazioni di acquisto che includono pannolini e latte, è inclusa anche la birra.

Slide 28 – Pattern discovery – algoritmi

Per quanto riguarda gli algoritmi per l'estrazione di regole di associazioni, benché ne esistano molti, i principali sono varianti di un algoritmo molto noto chiamato algoritmo **APRIORI.**

Ricordiamo inoltre che le regole di associazione possono anche essere estratte dagli alberi di decisioni generati da task di classificazione.

Slide 29 – Qualità dei modelli e dei pattern individuati

Prima di concludere la lezione, è importante ricordare che tutti gli algoritmi di Mining finora citati portano all'individuazione di modelli o pattern di "buona qualità" cioè che rappresentano ragionevolmente bene il dataset iniziale.

La qualità della conoscenza estratta può essere quantificata con modalità che dipendono dal tipo di tecnica considerata.

Per quanto riguarda la classificazione e la regressione, le misure di qualità vengono definite a partire da una matrice di confusione.

Per quanto riguarda il clustering, le distanze inter e intra cluster rappresentano buoni indicatori della qualità del risultato.

Quando invece si considera una regola di associazione X →Y, gli indicatori di qualità principali sono il **supporto**, definito come la frequenza con cui X U Y appare tra tutte le istanze, e la **confidenza**, definita come la frequenza con cui Y appare nelle istanze che includono X.

Slide 30 – Riepilogo e conclusioni finali

Bene, siamo giunti alla fine di questa lezione.

Ti ricordo che abbiamo introdotto che cosa si intende per Data Mining, chiarendone il suo ruolo tra gli strumenti di front-end di un sistema di Data Warehousing.

Abbiamo quindi classificato le principali tecniche di Data Mining, fornendo esempi e descrivendone le caratteristiche principali.

Grazie per l'attenzione!