Il data mining

Introduzione

Viviamo in un mondo in cui qualsiasi cosa, in qualsiasi momento, viene convertita in dato: ogni condivisione, like, telefonata, tweet, click del mouse, genera molti e molti dati.

I dati possono avere qualsiasi natura: numeri, caratteri, testo, pagine web, immagini, video, e costituiscono l'aspetto centrale di qualsiasi attività in ricerca, negli affari, nell'industria, nel marketing, nelle telecomunicazioni, in finanza e in molti altri campi. I dati sono così importanti da essere considerati degli asset cioè patrimonio aziendale: in un mondo fortemente competitivo come il nostro, disporre di informazioni costantemente aggiornate, in linea con quanto scoperto nei dati, può contribuire efficacemente ai processi decisionali e nello sviluppo di nuove strategie di crescita.

La quantità di dati ottenuta e memorizzata aumenta in modo esponenziale: l'enorme quantità dei dati da processare ha portato alla necessità di sviluppare metodi automatici ed è una delle ragioni per le quali il data mining è essenziale.

Il **data mining** è il processo che permette di estrarre automaticamente informazioni utili da grandi volumi di dati.

Aiuta le aziende a identificare fatti, tendenze, modelli, relazioni ed eccezioni importanti che normalmente non vengono notati o sono nascosti. Le tecniche di data mining sono applicate in una vasta gamma di settori tra cui sanità, assicurazioni, finanza, vendita al dettaglio, produzione e così via.

Task

I dati di origine e i contesti applicativi possono essere

estremamente diversi ma le finalità principali per i quali si utilizza il data mining sono **la descrizione e la previsione.** Il data mining predittivo usa alcune variabili dei dati analizzati per predire valori sconosciuti e futuri di altre variabili di interesse, mentre il data mining descrittivo si focalizza sul trovare schemi interpretabili che forniscono informazioni sulla struttura dei dati. Sebbene i confini tra predizione e descrizione non sono netti (alcuni dei modelli predittivi possono essere descrittivi, nella misura in cui possono essere comprensibili e viceversa), la distinzione è utile per comprendere l'obiettivo generale di scoperta di conoscenza perseguito. Gli obiettivi di predizione e descrizione possono essere raggiunti usando diversi metodi *(task)* di data mining illustrati di seguito.

La **classificazione** è utile per apprendere un modello che mappa (classifica) un elemento, in base alle sue caratteristiche, (un dato osservato) in una dell'insieme di classi predefinite. La classificazione può essere usata, ad es., in ambito bancario per la concessione o la negazione di mutui o prestiti: gli utenti sono suddivisi nelle classi "affidabile"/"non affidabile" in base alle caratteristiche patrimoniali e finanziarie possedute.

La **regressione** è l'apprendimento di un modello predittivo che mappa un elemento in una variabile a valori reali. Questo metodo è usato, ad es., nel predire la domanda dei consumatori per un nuovo prodotto come funzione dell'investimento fatto per pubblicizzare il prodotto stesso, nello stimare la probabilità che un paziente possa sopravvivere dati i risultati di un insieme di test diagnostici, etc.

Il **clustering** è usato per identificare oggetti simili l'uno all'altro; il grado di similitudine tra oggetti diversi può essere decisa in base alle caratteristiche possedute dai dati. L'effetto del clustering è quello di suddividere i dati in un insieme finito di raggruppamenti che ne rappresenta il modello. La similarità può essere calcolata basandosi su

fattori quali, ad es., il comportamento esibito negli acquisti, la reattività a certe azioni, etc. in modo da individuare sottopopolazioni omogenee tra i consumatori. Le informazioni ottenute dal raggruppamento sono utili per capire meglio consumatori e fornire quindi servizi maggiormente personalizzati.

La **riepilogazione** include metodi che individuano una descrizione compatta dei dati analizzati o di un loro sottoinsieme. Vengono così fornite informazioni sintetiche sui dati. Le tecniche di riepilogazione sono spesso applicate all'analisi dei dati interattiva e esplorativa e nella generazione automatica dei report. Ad es. tali informazioni sono utili per l'analisi dettagliata dei comportamenti dei consumatori; gli acquisti fatti da un cliente possono essere riassunti in valori totali: quanti prodotti sono stati acquistati, quanto si è speso, le offerte usate, etc.

La **modellazione delle dipendenze (dependency modelling)**, consiste nel trovare un modello che descriva significative dipendenze tra variabili. I modelli di dipendenza esistono a due livelli: quello strutturale specifica (spesso in forma grafica) quali variabili sono localmente dipendenti tra loro, e il livello quantitativo del modello specifica le forze delle dipendenze usando qualche scala numerica.

Il task di **individuazione di cambiamenti e deviazioni (change detection and deviation)** permette di scoprire dei cambiamenti più significativi nei dati rispetto a misurazioni fatte precedentemente o valori normati. Le tecniche utilizzate in questo task sono utili per il riconoscimento di frodi e, in generale, per individuare anomalie ed outlier in qualsiasi ambito.

I metodi

Esiste una grande varietà di metodi che trovano origine nelle discipline da cui il data mining ha origine: machine learning,

statistica e teoria delle basi di dati; di seguito sono illustrati i metodi basilari.

Alberi di decisione e regole di associazione trovano applicazione sia per il data mining descrittivo che per il data mining predittivo per task di classificazione.

Gli alberi di decisione sono usati per fare classificazione e regressione. Gli algoritmi che costruiscono alberi di decisione lavorano in modalità top down e seguono un approccio dividi-e-conquista: ad ogni stadio selezionano un attributo rispetto al quale creare una nuova suddivisione. Per processare un elemento, si parte dal nodo radice dell'albero; ogni nodo interno rappresenta un test sul valore di uno o più attributi e, in base al risultato del test, ci si sposta a uno dei figli del nodo. Il procedimento si ferma quando si raggiunge un nodo foglia cioè un nodo che non rappresenta nessun test. Il contenuto del nodo foglia raggiunto rappresenta l'etichetta della classe cui appartiene l'elemento analizzato correntemente o il valore ipotizzato nei task di regressione.

Una regola di associazione è una espressione della forma "se C allora P" dove C prende il nome di condizione della regola e P è la previsione. Le regole di associazione sono usate per fare classificazione, clustering e riepilogazione.

Oltre che nella classificazione, le artificial neural network (ANN) e support vector machine (SVM) vengono usate per i task di regressione, clustering e di individuazione di anomalie. La loro struttura è data da funzioni matematiche la cui struttura è tipicamente di difficile comprensione: sono quindi maggiormente adatte al data mining predittivo.

Una rete neurale artificiale si ispira alle reti neurali biologiche ed è costituita da un gruppo di nodi (interconnessioni di informazioni) costituiti da neuroni artificiali. E' un sistema adattivo che cambia la propria

struttura in base alle informazioni esterne o interne che l'attraversano.

Dall'analisi di un insieme di dati di training, una SVM costruisce un iperpiano o un insieme di iperpiani in uno spazio a molte o infinite dimensioni e individua l'iperpiano di appartenenza del dato osservato o del dato nuovo e/o non ancora osservato. Va sottolineato che, nonostante il problema originale sia definito in uno spazio finito, la soluzione è spesso costituita da insiemi di punti separabili solo tramite funzioni non lineari. Per questo motivo le SVM sono strumenti molto performanti.

Va infine citata la famiglia dei metodi *instance-based*. Gli algoritmi di questa famiglia sono usati per classificazione, clustering e per l'indivuazione di anomalie. In questi metodi il modello è costruito utilizzando *tout-court* un insieme di esempi fortemente rappresentativi presenti nel data set di training. La previsione su nuovi dati viene fatta calcolando la similarità che essi hanno con i dati appartenenti già al modello. Per il calcolo del valore di (dis-)similarità si utilizza una metrica matematica. Un semplice esempio è l'algoritmo di K-Nearest Neighbors (*KNN*) che prevede due passi: durante il primo passo, si calcola la distanza euclidea tra un nuovo campione *y* e ciascun campione *x* utilizzato nel training; durante il secondo passo, identifichiamo i *K* campioni che sono i più vicini a *y* e scegliamo la classe di maggioranza tra questi *K* campioni; la classe scelta è la classe di appartenenza di *y*.

Il processo

Seguendo una abitudine consolidatasi nell'ultimo decennio, abbiamo definito il data mining come il processo di estrazione di informazioni utili da grandi quantità di dati. In realtà il data mining è solo la fase centrale di tale processo il cui nome corretto è Knowledge Discovery in Databases (KDD). [riferimento Fayyad].

Il Knowledge Discovery in Databases (KDD) è costituito da 3 grandi fasi:

- la fase di data preprocessing;
- la fase di data processing cioè il data mining;
- la fase di data postprocessing.

tutte ugualmente rilevanti.

La fase di preprocessing è cruciale per la qualità della conoscenza estratta nella fase successiva: una preparazione dei dati effettuata superficialmente porta a risultati poco utili.

I dati grezzi (originali) vengono prodotti nativamente in formati idonei all'elaborazione e analisi nel contesto di riferimento; ad es. i dati di una rete di computer sono generati per essere immediatamente utilizzabili dai computer della rete stessa per le loro comunicazioni o tutt'al più per segnalare dei malfunzionamenti, ma non per individuare automaticamente comportamenti malevoli ed intrusioni.

Il preprocessing si occupa di preparare i dati originali per essere processati automaticamente nella fase di data mining e consiste di diversi passi. E' essenziale avere una esatta comprensione del dominio applicativo poiché questo fornisce informazioni utili a comprendere i dati grezzi posseduti e le loro correlazioni. Ciò permette di capire le necessità informative del fruitore finale del processo e di identificare la finalità principale richiesta ovverosia se attuare data mining predittivo o descrittivo o entrambi. La finalità determina il formato dei dati utile alla elaborazione: l'obiettivo è produrre un *dataset target* cioè un insieme di dati ottenuti dai dati originari ma idoneo alla fase successiva.

In generale i dati hanno bisogno di essere bonificati (*data cleaning*) perché possono presentare errori derivanti anche dalle misurazioni e dalle osservazioni fatte oltreché da

operazioni ordinarie su di essi; questo significa procedere a correzioni e/o rimozioni di ridondanze/o integrazioni di dati incompleti e/o obsoleti.

Successivamente i dati vengono "codificati" generando gli attributi di interesse. Le scelte da operare sono molteplici: a volte non vanno usati tutti gli attributi dei dati originali, a volte i dati vanno integrati con altri provenienti da altre fonti, a volte i dati devono subire una trasformazione intermedia (ad es. da dato continuo a dato discreto o da dato categorico a dato numerico).

In questo passo si è guidati dalla finalità richiesta e dal task su cui concentrarsi per i passi successivi: classificazione, regressione, clustering e così via. Infine si sceglie l'algoritmo da usare nella fase di processing.

Se la fase di preprocessing influisce sulla qualità delle informazioni estratte, la fase di postprocessing ci fornisce gli strumenti per poter analizzare la conoscenza ottenuta. L'informazione che otteniamo è data dall'output degli algoritmi di data mining usati nella fase centrale; è quindi più corretto parlare di "*nuggets of knowledge*" (letteralmente *pepite di conoscenza*) piuttosto che di "conoscenza". I dataset analizzati sono molto grandi e l'algoritmo di data mining usato può generare una grande quantità di *nuggets of knowledge* non necessariamente collegati all'aspetto di interesse. Il filtraggio della conoscenza non pertinente è il primo passo della fase di postprocessing. Si procede quindi ad interpretare e spiegare la conoscenza ottenuta per renderla comprensibile nel caso il fruitore finale sia una persona oppure per trasformarla in un formato adatto se il fruitore finale è un dispositivo. Quando la conoscenza è destinata alle persone, nel processo vengono utilizzate tecniche di visualizzazione.

La conoscenza ottenuta è sottoposta ad una fase di *valutazione*. In genere si usano dei metodi che, incrociando

opportunamente gli elementi del data set costruito, misurano la probabilità di errore presente nei modelli estratti.

Infine i *nuggets of knowledge* vengono integrati: ciò significa combinarli in modo da ottenere informazioni cioè nuovi concetti, schemi, tendenze, outlier o relazioni presenti ma nascosti in grandi volumi di dati. [fayyad, et altri.]

Cross-industry standard process for data mining – CRISP-DM

Il processo di Knowledge Discovery prevede in totale 9 passi suddivisi nelle 3 grandi fasi illustrate. E' nato in ambito accademico e necessita di forti competenze nell'ambito del data mining.

L'adozione di tale metodologia al di fuori della realtà accademica ha reso necessario rivederne la struttura. Nel 1996 un consorzio di grandi imprese avviò i lavori per la definizione di uno standard industriale per la modellazione del processo che prese il nome di *Cross-industry standard process for data mining – CRISP-DM;* nel 1999 venne pubblicata la sua prima versione e nel 2006 vennero iniziati i lavori per il *CRISP-DM ver.2.0* ma la seconda versione non venne mai pubblicata.

Malgrado ciò è una metodologia molto valida, ed è utilizzata dalle aziende nei progetti di data mining; il CRISP-DM parte dal presupposto che il processo deve essere usato da persone con elevata conoscenza del business da analizzare ma con poca o nessuna competenza nel data mining; in particolare gli obiettivi principali previsti dallo standard sono:

- assicurare la qualità dei risultati dei progetti di data mining;
- ridurre le competenze richieste per la conduzione del progetto;
- tesaurizzare l'esperienza per il riuso;

- generalizzare la metodologia;
- robustezza rispetto ai cambiamenti di contesto;
- indipendenza dagli strumenti e dalle tecniche usate;
- compatibilità rispetto agli strumenti usati.

La metodologia prevede l'applicazione di 6 fasi che possono essere iterate al fine di migliorare il modello, descrittivo o predittivo, generato: business understanding → data understanding → data preparation –> modeling → evaluation → deployment.

Dopo aver definito gli obiettivi di progetto e i requisiti di business, questi vengono trasformati nel problema da risolvere. Vengono quindi raccolti i dati grezzi ritenuti interessanti e rappresentati tramite campi e attributi; questo passo permette di studiare i dati e di individuare sottoinsiemi di dati interessanti su cui formulare ipotesi di schemi da individuare e/o verificare. Nella fase di preparazione, i dati sono trasformati per costruire l'insieme finale dei dati su cui verranno condotte le osservazioni; in genere in questa fase si ha già un'idea del modello che si vuole ricavare. Si costruisce quindi il modello su un sottoinsieme di dati ritenuto significativo e quindi lo si valuta su parte o tutto il rimanente dell'insieme di dati. Se il modello costruito dimostra di avere un'alta qualità, viene convertito in strategia di business e utilizzato nelle varie fasi del processo di business aziendale. Il modello è tanto migliore quanto più si adatta agli obiettivi di progetto e ai requisiti di business fissati inizialmente.

L'utilità di CRISP-DM sta nella flessibilità del suo processo di analisi dei dati che assicura alta qualità dell'output ottenuto indipendentemente dalla variabilità nella tipologia dei dati trattati e dagli obiettivi da raggiungere e nel dotare gli analisti di un framework unificato nell'approccio, supportato da documentazione e linee guida di esperti.

Sfide

Il data mining è uno strumento essenziale per le aziende ma presenta degli aspetti sfidanti che è utile segnalare. L'analisi che segue non è esaustiva ma sono elencati i problemi più comuni che i professionisti del data mining si trovano ad affrontare.

Sono necessari classificatori che trattino efficacemente dati con centinaia di milioni o miliardi di attributi per applicazioni quali il text mining o in analisi chimiche per l'individuazione di droghe; tali problemi non solo presentano dati con un numero abnorme di attributi ma richiedono l'analisi anche delle interazioni tra gli attributi.

In molte applicazioni i dati sono conservati in database e analizzati off line ma, sempre più spesso, c'è la necessità di fare analisi in tempo reale di flussi di dati (*data stream*) come, ad es., nell'individuazione di intrusioni e/o anomalie su reti ad alta velocità o in osservazioni provenienti da satelliti o da reti di sensori del mondo dell'IoT.

Se la necessità di analizzare i dati con un ritardo il più possibile ridotto rende il task critico, la dinamicità dei dati complica ulteriormente il lavoro. La dinamicità implica che il dato è variabile ovverosia che le sue proprietà possono cambiare mentre lo si sta osservando. Ciò significa che è lo stesso modello descrittivo o predittivo corrente che va a sua volta modificato.

Un problema particolarmente critico si ha nell'analisi delle serie temporali (es. di tali serie sono le osservazioni di serie finanziarie o serie sismiche). Le loro osservazioni sono prese in istanti di tempo discreti, distanziati l'uno dall'altro e spesso sono affette da rumore che gli algoritmi di data mining trattano con difficoltà. I task predittivi possono essere meno accurati in presenza di rumore nei dati originari.

La parte maggiore della conoscenza di un'azienda è complessa e conservata in dati non relazionali quali testi, immagini, multimedia e web, tipicamente dati non strutturati e quindi non facili da rappresentare in una forma adatta ad un database relazionale e altrettanto non facili da elaborare come nell'elaborazione di riepiloghi di documenti oppure conoscenza temporale e spaziale sulle operazioni svolte dalle persone e censite in pagine web e log wireless.

Altra forma di complessità proviene dai dati non indipendenti e identicamente distribuiti. Questo problema può presentarsi quando i dati sono generati da relazioni multiple cioè tabelle diverse di un database. Nella maggior parte dei domini, i dati analizzati non sono indipendenti e non sono tutti dello stesso tipo; rientrano in questa casistica pagine web linkate tra loro, social network, reti cellulari, etc.

Infine non vanno dimenticati gli aspetti legati alla protezione della privacy degli utenti cui appartengono i dati analizzati. Spesso i dati vengono analizzati dopo essere stati intenzionalmente modificati per proteggere l'identità degli utenti. Queste modifiche possono distorcere, in modo non prevedibile, la conoscenza contenuta nei dati originali. Di conseguenza, una delle sfide si affrontano è lo sviluppo di misure per valutare l'integrità della conoscenza del dataset e l'integrità di schemi individuali nascosti nei dati.

Le applicazioni del data mining

Oggigiorno il data mining è usato moltissimo dalle aziende con una forte attenzione al consumatore – grande distribuzione, assicurazioni e banche, servizi di comunicazione, marketing. Le aziende determinano prezzi, preferenze del consumatore e le politiche di marketing per influenzare vendite, soddisfare il cliente e aumentare i profitti. La *market basket analysis* permette di comprendere le esigenze del cliente e il rivenditore può modificare i prodotti offerti nel punto vendita. Il data mining aiuta a classificare i clienti in

segmenti distinti (*customer segmentation*) e può suggerire come adattare le offerte in base alle esigenze dei clienti migliorando le tecniche di *CRM*.

Il rilevamento delle frodi *(fraud detection)* è un grosso problema per le società finanziarie e assicurative: il data mining viene utilizzato per identificare le transazioni che hanno maggiori probabilità di essere fraudolente. Il data mining può aiutare a migliorare il rilevamento delle intrusioni *(intrusion detection)* grazie alle tecniche di rilevamento delle anomalie perché supporta un analista di rete nel distinguere un'attività diversa dall'attività di rete quotidiana comune. Le tecniche di *lie detection,* basate sul text mining, vengono utilizzate dalle forze dell'ordine per indagare sui crimini e monitorare la comunicazione di sospetti terroristi.

L'elevato volume di set di dati sulla criminalità (*criminal investigation*) e anche la complessità delle relazioni tra questi dati hanno reso la criminologia un campo appropriato per l'applicazione delle tecniche di data mining.

Le banche generano ogni giorno una quantità enorme di nuove transazioni *(financial banking)*. Il data mining può contribuire a risolvere problemi di business nel settore bancario e finanziario trovando modelli, causalità e correlazioni nelle informazioni aziendali e nei prezzi di mercato che non sono immediatamente evidenti ai responsabili perché i dati sul volume sono troppo grandi o vengono generati troppo rapidamente per essere sottoposti a verifica da parte di esperti. I gestori possono trovare queste informazioni per segmentare, indirizzare, acquisire, conservare e mantenere i clienti.

Il data mining ha un grande potenziale per migliorare i sistemi sanitari *(healthcare)*. Utilizza dati e analisi per identificare le migliori pratiche che migliorano l'assistenza e riducono i costi. Vengono sviluppati processi che assicurano

che i pazienti ricevano cure adeguate nel posto giusto e al momento giusto. Il data mining può anche aiutare gli assicuratori sanitari a rilevare frodi e abusi.

Gli strumenti di data mining possono scoprire gli schemi nascosti nei processi produttivi (*manufacturing*); la conoscenza estratta è usata per prevedere il tempo di sviluppo di un prodotto, il costo e le dipendenze tra le attività.

Gli approcci basati sul data mining sono ideali per la bioinformatica (*bioinformatics)*, dal momento che sono ricchi di dati. Le applicazioni del data mining alla bioinformatica comprendono la ricerca genetica, l'inferenza della funzione proteica, la diagnosi della malattia, la prognosi della malattia, l'ottimizzazione del trattamento della malattia, etc.

C'è un nuovo campo emergente, chiamato *Educational Data Mining*, che riguarda lo sviluppo di metodi che scoprono la conoscenza da dati provenienti da ambienti educativi. Gli obiettivi dell'EDM sono la previsione del comportamento di apprendimento futuro degli studenti, lo studio degli effetti del supporto educativo e l'avanzamento delle conoscenze scientifiche sull'apprendimento. Il data mining può essere utilizzato da un'istituzione per prendere decisioni accurate e anche per prevedere i risultati dello studente. Analizzando i risultati ottenuti, l'istituzione può concentrarsi su cosa insegnare e come insegnare.

Conclusioni

Col presente articolo si è cercato di fornire i concetti teorici fondamentali e gli strumenti tecnici essenziali per iniziare a muoversi nel data mining.

Il mondo dell'analisi dei dati è in costante evoluzione e si è arricchito di nuove espressioni e tecnologie negli ultimi anni a dimostrazione del grande e rinnovato interesse nella

capacità di produrre conoscenza dai dati. La figura professionale dal data scientist è molto richiesta dalle aziende. I metodi hanno subìto e subiscono modifiche e miglioramenti spronati, in questo, da nuove esigenze di business e nuove forme di dati sempre più complesse.

Il processo di estrazione di conoscenza dai dati è comunque rimasto sostanzialmente invariato.

Bibliografia e sitografia

- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. *From Data Mining to Knowledge Discovery in Databases*. 1996.
- Kurgan, L., Musilek, P. *A survey of Knowledge Discovery and Data Mining process models*. The Knowledge Engineering Review.
- Klösgen, W., Zytkow, J.M. *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press. 2002.
- Witten, I.H., Frank, E. *Data Mining, Practical Machine Learning Tools and Techniques*. Morgan Kauffmann Publishers. 2005.
- https://www.kdnuggets.com/.

A cura di: **Costantina Caruso**



Costantina Caruso è un esperto di sicurezza informatica, sistemista e amministratrice di rete. E’ stata sia il Chief Information Officer che il Chief Security Officer dell’Università degli Studi di Bari Aldo Moro dal 2010 al 2013. Ha un PhD in data mining e sicurezza informatica. E’ membro del Laboratorio Nazionale di Sicurezza Informatica del CINI. Attualmente è il research manager del Dipartimento di Informatica di UNIBA. I suoi interessi principali sono: anomaly detection, fraud detection, open source intelligence, malware detection, secure code, secure by design, offensive and defensive techniques, data mining, big data analytics and security, cryptography, blockchain.

https://www.ictsecuritymagazine.com/articoli/il-data-mining/