Lezione 7 modulo 2

Descriviamo in questo modulo alcune tecniche di Data Mining e parleremo di regole associative, vorremmo estrarre delle regole dai nostri dati di classificazione e di clustering. Produrremo le tecniche illustrandole attraverso degli esempi, ma non ci concentreremo sull'aspetto più relativo alla ingegneria del software e lo sviluppo di algoritmi e di sistemi software tipici per l'esecuzione di queste tecniche di Data Mining. Cominciamo con il considerare le regole associative. Noi vogliamo estrarre delle regole del tipo se A, si verifica B, quindi avremo quello che viene chiamato un antecedente e un conseguente. Sostanzialmente vorremmo avere delle regole del tipo, ad esempio, se stiamo parlando di acquisti e quindi vogliamo analizzare questi dati, delle regole del tipo che se le persone comprano i cereali, allora comprano anche il latte, quindi delle regolarità che si trovano all'interno dei nostri dati grezzi che in questo caso saranno le transazioni raccolte nel punto di vendita. Quindi questo è il nostro obiettivo nell'usare la regola. Quello che è importante in questi casi è valutare il risultato ottenuto. Nel caso delle regole associative, la valutazione verrà effettuata utilizzando due tipi di parametri: uno è il calcolo di quello che è il supporto. Il supporto sarà il calcolo di quanto la regola effettivamente è verificata all'interno dell'insieme delle informazioni considerate. Quindi la probabilità di trovare insieme A e B. L'altra cosa è quella che viene chiamata la confidenza, che ci dice effettivamente quello che si può tradurre in una probabilità condizionata, qual è la probabilità di trovare B se ho trovato A, la probabilità di B dato A, quindi la probabilità di trovare insieme A e B dato A, quindi fratto la probabilità di A. Invece ci saranno due parametri che verranno valutati perché ovviamente io potrei estrarre una regola ma su un supporto basso. Cosa vuol dire questo? Ad esempio, se ho un insieme di 100 tuple all'interno di una tabella relazionale, se solo una tupla mi dice che avviene questo fatto, il supporto sarà molto basso perché sarà uno su 100, perché solo quella volta ho trovato queste informazioni. Siccome sto cercando delle regolarità, in genere io voglio avere un certo numero di tuple che mi dicono effettivamente che si è verificato questo. Per la confidenza, voglio non solo estrarre la regola che vale per alcune tuple, ma voglio sapere quante volte dato un valore si verifica l'altro. Vediamo questo con un esempio: noi potremmo ritrovarci delle tabelle con delle informazioni. Queste sono delle informazioni relative a delle partite, quindi l'obiettivo di questa tabella è registrare le condizioni in cui si è giocato o meno in una partita e teniamo presente che abbiamo 14 tuple perché questo è quello che ci servirà all'interno dei nostri conteggi. Vediamo in questa tabella diversi attributi: l'aspetto relativo alla condizione atmosferica, la temperatura tasso di umidità, la ventilazione e il gioco. Quando abbiamo un'estrazione di regole in modo non supervisionato, noi sostanzialmente andiamo a cercare delle regolarità all'interno di questi dati che sostanzialmente corrispondono a delle correlazioni che ho fra i valori degli attributi presenti nella tabella. Proviamo ad esaminare alcune situazioni che possono essere interessanti per quanto riguarda la relazione fra temperatura e tasso di umidità. Quindi in genere cercherò di vedere se ci sono delle relazioni fra tipologie di variabili e quindi potrò considerare le varie coppie di variabili e in questo caso prendiamo in esame temperatura e umidità e in particolare una regola che potremmo estrarre è questa, che se la temperatura è uguale a freddo noi avremo che il tasso di umidità è normale. Come possiamo valutare una regola di questo tipo? Abbiamo detto che vogliamo calcolare il supporto e la confidenza. Il supporto abbiamo detto che lo valutiamo contando il numero di tuple sul totale che ci danno questa informazione. Nel nostro caso il supporto sarà di 4/14 quindi circa il 30 per 100 delle tuple che ci danno queste informazioni. Quanto siamo certi di queste informazioni, date le informazioni di partenza di questo Risultato? Questo ci viene dato dalla confidenza: nel nostro caso tutte le tuple che hanno una temperatura freddo, vedete non ci sono altre tuple all'interno di questa colonna con il valore freddo, hanno un tasso di umidità che viene associato al valore normale, quindi in questo caso la confidenza è del 100 per 100. Quindi abbiamo che, dato un certo valore di temperatura uguale a freddo, noi abbiamo un tasso di umidità che è normale nel 100 per 100 dei casi. Quindi abbiamo un'estrazione di regole che ci dicono dai dati a disposizione che si verifica una certa regolarità dei dati. Abbiamo visto quindi che questa è una tecnica di tipo non supervisionato, non ho definito una variabile target ma la sto estraendo guardando i dati e valutando ovviamente il supporto e la confidenza per decidere se è un valore che considereremo

accettabile e avremo una tecnica di tipo descrittivo perché andiamo a descrivere, a partire dai dati disponibili, quali sono le regolarità che possono essere identificate. Passiamo adesso ad un altro tipo di tecnica, alla classificazione. La classificazione assegna oggetti a classi. In questo caso, noi vogliamo dire che un certo oggetto appartiene a una certa classe, quindi di fatto definisco una variabile target che mi indica qual è la classe in cui è collegato l'oggetto. Tipicamente abbiamo casi di tipo supervisionato per fare questo tipo di operazione. Avremo a disposizione per ogni oggetto vari aspetti che caratterizzano l'oggetto, varie informazioni sull'oggetto. Quello che ci interessa capire è quali sono le informazioni utili per poter classificare automaticamente un oggetto in una certa classe. Quindi, a partire da un nostro training set, noi avremo un algoritmo che ci consentirà di assegnare automaticamente un oggetto ad una classe. Abbiamo detto che un aspetto importante per noi è quello della valutazione. In questo caso avremo due parametri di valutazione che considereremo che vengono chiamati Precision e Recall. In realtà esistono molti parametri di valutazione che possono essere considerati sulle classificazioni, ma questi sono i due principali che vengono utilizzati in letteratura. Cominciamo a dare una loro definizione esaminando una figura che vediamo qua. Noi, quando andiamo a valutare una classificazione, avremo degli elementi che supponiamo che sia una classificazione di elementi rilevanti o non rilevanti, quindi di tipo binario e quindi questo poi si può estendere anche ad altre classificazioni quindi avremo gli elementi classificati in modo giusto oppure sbagliato. Gli elementi che sono elementi rilevanti sono quelli che vogliamo che il nostro algoritmo classifichi effettivamente come rilevanti. Gli altri saranno gli altri elementi che vengono chiamati qua elementi negativi. Quindi sostanzialmente avremo gli elementi rilevanti che chiameremo poco positivi e gli elementi non rilevanti vengono indicati come elementi negativi. L'algoritmo che cosa farà? Cercherà di classificare ciascun elemento e nella classificazione, l'algoritmo potrà effettivamente trovare quelli che vengono chiamati Veri Positivi e quindi TP sta per True Positive e i Falsi Negativi saranno quelli in cui l'algoritmo dirà che non sono rilevanti, ma invece sono rilevanti. Quindi falsi negativi perché l'algoritmo dice no, mentre avrebbe dovuto dire di sì. Se guardiamo l'altro insieme, quello dei negativi, ovviamente può succedere la situazione simmetrica. L'algoritmo classifica un elemento che in effetti è negativo come positivo e questi sono chiamati Falsi Positivi, oppure effettivamente l'algoritmo funziona correttamente e quindi i negativi vengono effettivamente classificati come negativi, quindi abbiamo il True Negative, in questo caso. Abbiamo detto che vogliamo dare due parametri di valutazione: uno è la precisione, quanto effettivamente il nostro algoritmo mi classifica correttamente gli elementi positivi. E quindi vado a prendere quelli che ha classificato positivamente e lo valuto rispetto a tutti quelli che ha classificato positivamente, quindi, se vogliamo vedere, l'area indicata qua in rosso. Quindi questo vuol dire quanto è preciso l'algoritmo quando mi dice che un elemento effettivamente è positivo. Ovviamente io voglio avere una buona precisione dell'algoritmo, quindi eliminare i falsi positivi quando possibile. D'altra parte quello che può succedere se io cerco di eliminare i falsi positivi è che l'algoritmo trovi anche pochi veri positivi, nel senso che uno degli obiettivi è che effettivamente il mio algoritmo trovi quanti più positivi possibili. Quindi l'altro parametro che devo utilizzare è il Recall, che mi dirà: di tutti gli elementi positivi, quanti ne ha trovati il nostro algoritmo? Quindi, avrò i veri positivi, diviso questa volta per l'insieme di tutti i positivi che sono stati valutati come positivi all'interno del nostro sistema. Quindi abbiamo due parametri che ci dicono che vogliamo una precisione nella valutazione, ma anche vogliamo che effettivamente il nostro algoritmo classifichi una buona percentuale degli elementi in modo corretto fra quelli che sono gli elementi da riconoscere all'interno di una certa classe. Questo è il risultato, quindi vogliamo valutare se l'algoritmo alla fine ha dato il risultato che possiamo considerare un buon risultato. Come fa a classificare un certo elemento a partire da un insieme di informazioni e ottenere una classificazione? Ci sono tante tecniche, noi vedremo quella degli alberi di decisione. In particolare, un albero di decisione, lo vediamo su un esempio, ha questa forma. Avremo un insieme di variabili. In questo caso, noi vogliamo andare a considerare delle persone, quindi queste variabili saranno associate a delle persone, e come variabile avremo, ad esempio, l'età della persona, la zona in cui abita e se è laureata oppure no. Abbiamo detto che in questi casi abbiamo anche un target. Nel nostro caso, il nostro target verrà indicato con i valori si e no, supponiamo che sia l'acquisto di un televisore ad alte prestazioni. Quindi sostanzialmente, acquisto, e questo acquisto potrà

assumere questi due valori, sì o no. Quindi la nostra predizione, in questo caso, sarà: date certe informazioni sulle persone, riesco a predire correttamente se acquisterà o meno un certo televisore di un certo tipo? Questo verrà ottenuto costruendo degli alberi. Gli alberi utilizzeranno le variabili per discriminare la popolazione e quindi, ad esempio per l'età, noi avremo dei valori, supponiamo che l'età di chi va acquistare un televisore vada dai 18 ad un valore che è superiore ai 60 anni, ad esempio, noi andiamo a decidere che avremo tre range che sono interessanti per discriminarli, per quelli sotto i 30 anni, quelli fra i 31 e i 60 e quelli superiori ai 60 anni. Sotto i 30 anni, essendo televisori costosi, supponiamo che non venga effettuato l'acquisto, ovviamente poi dovremo valutare questo risultato, in una popolazione reale poi ci saranno alcuni che potranno effettivamente comunque acquistare questo televisore anche sotto i 30 anni. Quindi, quello che andremo a valutare poi nell'algoritmo è la precisione di questa decisione. Vediamo che qua si ferma l'albero. L'albero considera una variabile, vedremo che ci possono essere diversi tipi di alberi, in questo caso ci fermiamo perché sostanzialmente sotto i 30 anni viene ritenuto sufficiente guardare l'età e non occorre vedere altre informazioni per discriminare in modo accettabile. Se andiamo su una fascia intermedia, un'altra discriminante invece che troviamo qua è la laurea, quindi è un secondo livello dell'albero, la domanda successiva che ci facciamo quando classifichiamo è se la persona è laureata oppure no. Se è laureata, è abbastanza probabile che compri il televisore, altrimenti no. Vediamo che allo stesso livello dell'albero, invece, se l'età è superiore ai 60 anni, andremo prima vedere qual è la zona in cui abita la persona, qua la zona potrà essere centro oppure periferia e nel caso della periferia decideremo che è no e invece nel caso del centro, andremo a vedere di nuovo se è laureato oppure no e quindi andremo a discriminare nella nostra popolazione sulla base anche di questo terzo parametro. Cosa vediamo in questo caso? Lo abbiamo visto con un esempio, cosa dovremo decidere con un algoritmo? Innanzitutto dovremmo decidere qual è la variabile da scegliere, vediamo che all'inizio abbiamo scelto l'età, ovviamente avremmo potuto cominciare dalla laurea oppure dalla zona. Questo verrà scelto sulla base di un parametro che è l'entropia, sceglieremo quella entropia che ci consentirà di distinguere meglio fra i diversi gruppi di popolazione, quindi di discriminare maggiormente quella che è la variabile target, quindi l'acquisto. Vediamo che per ogni ramo dell'albero, successivamente andiamo a scegliere, non per quel livello dell'albero, ma per ciascun ramo dell'albero qual è la variabile successiva da considerare. Dobbiamo anche decidere, abbiamo visto, quando fermarci, perché abbiamo visto che qua ci siamo fermati qui, qua ad un secondo livello e qui siamo arrivati ad un terzo livello e quindi dovremmo avere un criterio per fermarsi nella costruzione dell'albero. Questo potrà essere predefinito, quindi potremmo decidere di avere alberi a due livelli, a tre livelli in modo predefinito oppure possiamo valutare se effettivamente esistono delle variabili che consentono di scardinare sufficientemente la popolazione rimanente, perché qua, quando saremo a questo livello, noi avremo selezionato dal nostro insieme di persone quelle sopra ai 60 anni che vivono in centro, quindi andremo a valutare se c'è un numero sufficiente di persone da considerare e se all'interno di questo una variabile ulteriore ci può aiutare a discriminare per fare la scelta successiva oppure è sufficiente fermarsi a questo livello. Quindi, gli alberi vengono costruiti come modelli di situazioni che potranno essere utilizzati in modo predittivo. Quindi, quando mi arriverà un'altra persona, io andrò a utilizzare quest'albero per discriminare sulle caratteristiche della persona e se mi arriva una persona di 40 anni dovrò anche chiedere se è laureato e a quel punto il mio predittore mi dirà se è probabile o meno che compri il televisore. Ovviamente questo tipo di algoritmi può essere utilizzato, ad esempio, per supportare delle offerte commerciali che vengono fatte alle persone, cercando di classificarle in gruppi in modo da offrire, ad esempio, il prodotto a quelli che hanno più probabilità, perché ci sono certe categorie, di acquistare questo prodotto. Come abbiamo accennato prima, possono esserci varie tipologie di alberi. Innanzitutto possiamo avere degli alberi di tipo generale, come abbiamo visto prima, oppure di tipo binario, in cui voglio sempre dividere la mia popolazione in due sottoclassi diverse ad ogni passaggio. Potremmo avere degli alberi di tipo univariato, come quello che abbiamo visto prima, consideriamo una variabile alla volta, oppure combinazioni di variabili quindi abbiamo alberi di tipo multivariato. Quindi considero più attributi contemporaneamente nel momento in cui scelgo un percorso dell'albero. L'ultima tecnica che andiamo a considerare è quella del clustering: quale obiettivo? Quello di creare dei gruppi, di elementi, a

partire da un insieme di dati di addestramento, che abbiano caratteristiche simili. Ovviamente questa funzione di somiglianza dipenderà da qualcosa e quindi sarà necessario definire il modo di valutare una distanza fra due elementi, quindi io cercherò di mettere all'interno di un gruppo di elementi che sono più vicini fra di loro, a partire da una distanza che è calcolata a partire da caratteristiche della popolazione, quindi da informazioni che ho. Il clustering è una tecnica non supervisionata, quindi non definirò dei gruppi a priori, ma sarà l'algoritmo a trovarmi dei gruppi di persone o di elementi che sto considerando. Ad esempio, qua possiamo vedere una visualizzazione, spesso avremo n variabili, ma in questo caso possiamo considerare una visualizzazione su due variabili in cui abbiamo la fascia di reddito e il livello professionale. La fascia di reddito è bassa, media, alta, e il livello professionale è suddiviso in quattro categorie: professionale, semi-professionale, specializzato e non specializzato. Originariamente, noi posizioneremo questi valori che supponiamo, per come sono messi in questo esempio, possano essere anche graduati all'interno delle varie categorie, come punti su questo spazio. L'algoritmo di clusterizzazione, quello che cercherò di fare è trovare dei gruppi in modo che gli elementi all'interno di questi gruppi siano abbastanza vicini tra di loro. Questo lo può fare senza aver definito il numero dei gruppi da trovare a priori oppure potrò dire, ad esempio in questo caso, che voglio trovare cinque gruppi e quindi l'algoritmo cercherà di inserire i punti nei gruppi in modo che venga massimizzata la funzione di somiglianza all'interno dei vari gruppi complessivamente. Spesso noi daremo come parametro il numero dei gruppi e ci sono degli algoritmi, ad esempio, K-Means, che consentono di trovare gruppi di elementi dove il numero K è dato a priori. Anche per la clusterizzazione, noi possiamo avere varie tipologie di algoritmi che faranno questi raggruppamenti in modo diverso. Gli algoritmi potranno essere distinti in algoritmi in cui i gruppi hanno un'appartenenza esclusiva ad un unico gruppo, e quindi, come abbiamo visto prima, i gruppi sono separati e ogni elemento sta all'interno di gruppo oppure con sovrapposizione, in cui io potrò avere dei gruppi che condividono degli elementi. Un altro criterio di distinzione che posso avere è se, dato un insieme di elementi, li classifico tutti in un gruppo oppure li classifico solo parzialmente, perché io avrò dei casi in cui, ad esempio, potrei avere degli elementi organizzati in modo sparso per alcuni degli elementi e quindi definire dei gruppi in questi casi può essere anche a volte fuorviante. Quindi si definiranno algoritmi in cui vorrò dare questa classificazione in modo completo oppure parziale. Nel caso parziale, nel formare i gruppi non considererò gli outliers e quindi definirò dei gruppi solo per gli elementi che sono vicini fra di loro e gli altri non li andrò a considerare nella definizione dei gruppi. Vediamo adesso come possono essere usate queste tecniche di analisi dei dati, di data Mining, ad esempio in un sistema di tipo CRM in cui andiamo ad occuparci di clienti. Può essere interessante definire delle regole, perché noi potremmo identificare, ad esempio, quali sono i prodotti che vengono sempre comprati insieme. Questo ad esempio può guidare la disposizione dei prodotti all'interno degli scaffali. Un altro aspetto che può essere interessante è quello della classificazione, abbiamo detto, predittiva e quindi, ad esempio, può essere utile per campagne di vendita. Abbiamo già visto che all'interno dei sistemi CRM, una delle operazioni che vorremmo fare è selezionare quali possono essere i clienti da raggiungere all'interno di una campagna per un certo prodotto e quindi cercherò di contattare i clienti che saranno più probabilmente interessati all'acquisto di un certo prodotto. Un altro caso che possiamo trovare è quello del clustering e quindi, ad esempio, potrò classificare i miei clienti in categorie, ad esempio, Gold, Silver, Iron, categorie che mi dicono quali sono i clienti più interessati ai prodotti, più profittevoli per la compagnia, quelli intermedi e quelle che invece non risultano particolarmente per le loro caratteristiche interessanti per la compagnia e quindi ovviamente, quando ad esempio potremmo identificare le situazioni in cui il nostro cliente dà dei segnali di voler lasciare, ad esempio, i servizi di una compagnia di servizi, è chiaro che potrei decidere di comportarmi in modo diverso a seconda della classificazione del cliente e quindi prestare più attenzione ai clienti ritenuti più interessanti per una certa azienda.