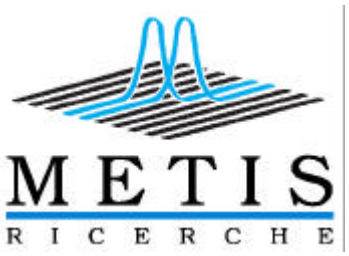

DALL'AUTOMAZIONE DELLE PROCEDURE ALL'AUTOMAZIONE DI PROCESSO

Flavio Bonifacio

Relazione al convegno

Data Mining e Metodologia della Ricerca Sociale: la Creazione di Valore Aggiunto per l'Utente, Torino, Aula Magna del Rettorato dell'Università, 20 Febbraio 2004

Introduzione

L'obiettivo della relazione è quello di attirare l'attenzione sulle peculiarità del Data Mining come strumento importante dell'Analisi Dati. Quest'ultima viene vista come il campo specifico di applicazione di professionalità nuove, che uniscono competenze diverse: la statistica, l'informatica, la matematica, la filosofia. La specificità di questa nuova professione viene individuata proprio nella sua doppia trasversalità, tra i domini e tra le professionalità.
Nella relazione si illustrano brevemente alcuni antefatti dell'attuale situazione e si focalizza l'attenzione su tre aspetti intorno a cui si costruisce l'evoluzione e il campo di applicazione del data Mining: l'evoluzione delle interfacce, l'evoluzione degli strumenti di analisi statistica, l'evoluzione dell'organizzazione del lavoro e delle competenze professionali.
Si definisce in primo luogo che cosa si intende per automazione delle procedure: come si passa dalla programmazione della statistica al suo utilizzo con linguaggi evoluti.
In secondo luogo si mostrano esempi di come l'evoluzione delle interfacce grafiche (Graphical User Interface) abbia influito sullo sviluppo del software medesimo e come le competenze richieste si siano modificate.
In terzo luogo si costruiscono esempi di analisi statistica e di Data Mining al fine di rendere evidenti le peculiarità dei due percorsi che vengono infine enucleate. Vengono altresì elencate le tecniche del Data Mining.
Infine viene riformulata la definizione del Data Mining come insieme complesso di procedure statistiche, utilizzabili in modo semiautomatico con interfacce grafiche su grandi masse di dati. Vengono poste in rilievo le caratteristiche che fanno diventare il Data Mining uno strumento forte di dislocazioni professionali alternative su percorsi organizzativi non consueti.

Premessa: interfacce e quadri di manovra

L'interfaccia è, nel campo che qui interessa, tutto ciò che sta in mezzo e serve da interprete tra il linguaggio della macchina, nella fattispecie il calcolatore, e il linguaggio umano. Molto sinteticamente è interfaccia tutto ciò che apre e chiude circuiti partendo da una sequenza di comandi codificati. La prima interfaccia fu il linguaggio macchina, seguita poi dall'assembler, dai linguaggi evoluti (Fortran, Cobol, PL/1,BASIC, C), dai linguaggi specializzati per macro funzione (la gestione dei dati, la statistica), dai linguaggi ad interfaccia grafica (GUI, Graphical User Interface). Sia seguendo l'ordine temporale, ovvero il periodo di implementazione, che l'ordine spaziale, ovvero la prossimità al linguaggio della macchina, le interfacce si sono indicate rispettivamente come interfacce di I,II,III,IV e V generazione. Il termine generazione si riferisce anche al fatto che ciascuna generazione successiva eredita i risultati di "comunicazione" raggiunti dalle precedenti. Nel senso che l'assembler usa il linguaggio macchina, il Fortran usa l'Assembler, ecc. Per fare un esempio che ci riguarda, il SAS fino ad un certo punto usava PL/1 (il ";" come terminatore di statement è rimasto nel linguaggio nativo SAS come testimone di quest' era) fino a quando è stato riscritto,15 anni fa circa, interamente in C. In tutte le generazioni poi c'è stata una spiccata tendenza all'autocannibalismo. Questa tendenza autofagica è stata particolarmente forte per certi linguaggi: alcuni di essi, come il C, avevano un set minimo, una ventina forse, di istruzioni elementari. Tutte le altre venivano costruite incapsulando le istruzioni elementari in funzioni, che costituivano poi il corpo vivente del linguaggio. L'autofagia ha raggiunto il suo massimo con la programmazione ad oggetti che ha fatto della riusabilità dei suoi elementi uno dei dogmi. *L'Object Oriented Programmation* si è introdotta tra la IV e la V generazione dei linguaggi permettendo di fatto il passaggio alle GUI che, come vedremo, costituiscono l'elemento portante dell'evoluzione moderna delle interfacce. Infine Internet e il WorldWideWeb: con il passaggio ad Internet e la separazione client/server si assiste ad una proliferazione di strumenti: da un lato quelli dedicati alla gestione del client, come HTML (*Hypertext Markup Language*) e ASP (*Active Server Pages*) che con JAVASCRIPT permette il controllo del dialogo client; dall'altro quelli dedicate alla gestione del Server, come Vbscript, Visual BASIC o Visual C++ e SQL/server che permettono l'accesso ai dati. Ma qui la sequenzialità si interrompe: ciascuno strumento segue la propria evoluzione rendendo difficile anche soltanto raccontare il seguito della storia. Ciò che comunque tranquillizza noi, formati tra la III e IV generazione, è che dietro la sigla più recente e alla moda o dietro la più bella icona c'è sempre una "IF". Istruzione che abbiamo particolarmente

cara forse perché ricorda, a noi formati tra la III e la IV generazione, il titolo di un film del 1968 con Malcom Mc Dowell, per quella generazione, indimenticato interprete di Arancia Meccanica (1971) di Stanley Kubrick.
Per quel che qui interessa è importante notare ancora un punto: l'evoluzione delle interfacce non è soltanto una questione che ha riguardato gli addetti ai lavori. Essa ha soprattutto riguardato la possibilità di costruire applicazioni di tipo verticale, concernenti determinate funzioni aziendali, la contabilità, la produzione, l'analisi dati, più usabili e più facilmente apprendibili. Insomma la costruzione dei moderni quadri di manovra per la realizzazione di attività lavorative. Notiamo in passando che sono proprio questi gli aspetti che coinvolgono maggiormente la professionalità degli utilizzatori e la loro formazione.
Per illustrare l'evoluzione delle interfacce da questo punto di vista e le conseguenze di tale evoluzione sulla conoscenza e sulla formazione prendiamo ad esempio un campo che nulla ha a che vedere con la statistica: la musica.
In particolare osserviamo un possibile percorso di composizione che fa uso del computer.
1.Per prima cosa l'artista compositore proverà le sue idee su uno strumento.
2.Successivamente potrà scrivere la frase musicale che vi corrisponde.
3.Dopo aver deciso quali sono le parti, darà copia del documento a strumentisti che si preoccuperanno di eseguirle.
4.Raggiunto un numero sufficiente di prove i musicisti registreranno, producendo ciascuno una traccia registrata su CD.
5.Le tracce verranno rielaborate e mixate per produrre la registrazione finale.
In questo processo si sottolinea la mistura di processi automatici e non automatici e la potenza degli strumenti automatici che abbreviano notevolmente le fasi di lavoro strettamente tecnico. Questa potenza è resa usabile con interfacce veramente efficienti, che riducono di molto i tempi di apprendimento e perfezionano le abilità necessarie. Nel percorso disegnato le professionalità sono abbastanza ben individuabili e presenti, nonostante la presenza di notevoli risorse automatiche. Queste risorse o macchine chiudono sempre il feedback di funzionamento con l'utilizzatore. C'è cioè sempre un manovratore che sa cosa vuole e conosce il modo di farlo eseguire dalla macchina, attraverso la conoscenza delle opportune interfacce: compositore, esecutori, fonico.
Un processo ancora più automatizzato potrebbe eliminare qualcuno nel processo: forse gli esecutori? Probabilmente i primi candidati ad uscire sarebbero proprio loro. Se si ritorna al punto 2, possiamo immaginare una macchina che esegua i pezzi da sé. Tutto questo evidentemente

esiste già in qualche forma. Ovviamente è un discorso molto astratto ma non assurdo. Quel che posiamo dire è che nessuno vorrà seguire un concerto in cui l'esecutore è la macchina e basta. Ed anche se lo volesse i risultati che conseguirà saranno modesti rispetto ad assistere ad un concerto di professori d'orchestra. Ma in linea di principio quella soluzione, la soluzione della macchina esecutrice è plausibile. Sarebbe comunque interessante sapere che cos'è che ci attrae nell'esecuzione dei musicisti, visto che a tempo ci va molto meglio la macchina del musicista. Proprio il fatto che sbagliano, forse?
Chi è inalienabile è il compositore, colui che manovra con l'hardware più immediato e con l'interfaccia più flessibile: il cervello e le idee. Questo credo valga anche nel caso della musica elettronica, in cui si può supporre che la macchina abbia anche una parte creativa. Lasciamo per il momento da parte la musica e veniamo a noi. Riprenderemo la musica a tempo e luogo.

L'automazione della procedura

Per procedura intendiamo qui in modo circoscritto la procedura statistica. In questa accezione il calcolo della varianza è una procedura statistica. Per automazione intendiamo, genericamente, il procedimento di calcolo eseguito sul *computer* atto a rendere automatico l'ottenimento dei risultati. Intorno alla fine degli anni 60 e all'inizio degli anni 70, periodo corrispondente allo sviluppo dei linguaggi di terza generazione, l'automazione veniva ottenuta con la scrittura di programmi. Il programma legge i record contenenti le osservazioni relative alla variabile oggetto di analisi, effettua i calcoli opportuni e scrive sull'output previsto il risultato richiesto. Successivamente, intorno alla metà degli anni 70, la necessità di conoscere un linguaggio di programmazione di terza generazione viene superata. Nell' ambito della statistica, e per il calcolo della varianza sono necessarie soltanto tre istruzioni che rendono oltretutto inutili le conoscenze della tecnica di programmazione. In questo periodo comincia a maturare una riflessione sui ruoli e sulle competenze nell'ambito della ricerca sociale. Cito da un seminario di formazione tenuto a Torino nel 1975, *Analisi quantitativa nella ricerca sociale*: "*E' ferma convinzione dei promotori che una volta fatta la scelta di impegnarsi in lavori di ricerca che necessitano del ricorso all'elaboratore, non si possa assolutamente delegare tale fase ad un tecnico. Alla base dell'ipotesi della delega vi è una concezione riduttiva del processo di analisi dei dati e una non esatta valutazione della gestione dell'interfaccia ricerca-elaborazione dati.Considerazioni di metodo ed organizzative richiederanno sempre di più che il ricercatore usi in prima persona dell'elaboratore*". Obiettivi che respirano l'aria dell'epoca. Certamente sono osservazioni che già allora valevano solo per il mondo della ricerca nell'Università. Soltanto all'Università ci si poteva proporre un obiettivo di quel genere. Ma anche per l'Università quell'obiettivo non poté essere perseguito: si formarono infatti professionalità diverse, in qualche modo a cavallo tra ricerca (o dominio dei contenuti) ed elaborazione statistica dei dati, che essendo in mezzo al guado, non riuscirono ad essere né l'una né l'altra cosa. Quella professionalità si riassunse in tre negazioni: né sociologi (o psicologi, o economisti), né informatici, né statistici. Ma quelle professionalità c'erano poi davvero? Come vedremo questa domanda, cambiate le cose da cambiare riferendole in generale alle organizzazioni complesse ed al mondo delle imprese, in qualche modo è viva ancora oggi. Anche se possiamo dire, almeno da un punto di vista antropologico, che esiste.

All'inizio degli anni 80, con l'avvento dell'informatica distribuita e dei personal computers, cominciano ad affacciarsi interfacce più evolute e soprattutto più user friendly, come allora si comincia a dire. Si va da semplici menu guidati, che dominano la scena fino alla fine degli anni 80,

a più sofisticate interfacce a icone agli inizi degli anni 90, con l'avvento e il definitivo affermarsi di Windows come sistema operativo (la versione 3.0 è del 90). Ulteriore impulso alla semplificazione dell'interfaccia viene dal WWW e da internet. Siamo sempre comunque sul piano dell'automazione della singola procedura statistica. Gli sviluppi sino a qui si limitano a facilitare l'esecuzione dei calcoli, ma non toccano ancora l'organizzazione del processo di produzione del dato. Una ricaduta importante comincia ad esserci sul contenuto del lavoro di quel professionista che non c'è: gli sembra che il suo compito gli sia reso più semplice. Non solo non deve più scrivere programmi: non ha nemmeno più bisogno di scrivere le poche istruzioni per SAS. A questo proposito, prima di compiere l'ultimo passo, bisogna osservare che l'attività di semplificazione del lavoro dovuta all'evolversi dell'interfaccia ha uno svantaggio: è difficile memorizzare le sequenze di click. Dal punto di vista dell'elaborazione statistica dei dati è qualche volta, cioè quasi sempre, importante garantire la ripetibilità delle operazioni. Per questi casi continua a convenire anche in quest'epoca scrivere programmi, sia pure semplificati nel linguaggio del package di IV generazione. Qui interviene l'ultima modificazione che consiste nell'automatizzare l'intero processo: non solo la singola procedura può essere automatizzata, ma un'intera sequenza delle medesime e, cosa ben più importante, l'intera sequenza può essere conservata e documentata in modo da non risolversi in una sequenza di click più o meno intelligenti, ma volatili. Il Data Mining, insieme con altri strumenti dell'ultima evoluzione informatica, può essere almeno in parte visto come protagonista di questo tipo di automatizzazione. Prima di valutarne le modalità e le conseguenze vediamo un po' più da vicino in che consiste il Data Mining nel campo più generale dell'Analisi Dati, partendo proprio dall'attività di elaborazione e di analisi che le necessità di conoscenza pongono quotidianamente all'ordine del giorno. Queste attività saranno provvisoriamente denominate "attività del miner".

Attività del Miner

Se immaginiamo la possibilità di un artista di grande sensibilità musicale, ma che non conosca la tecnica di scrivere musica abbiamo i riferimenti necessari per costruire analogie illustrative con il mondo del Data Mining e farci tornar utile l'esempio musicale. Nel Data Mining la musica è la statistica, l'esperto del dominio è l' artista che può non conoscere la tecnica statistica (e quasi sempre così capita). L'esperto del dominio, il marketing manager per esempio, ha idee più o meno precise di che cosa vuole. Si rivolge al Miner in quanto il Miner ha la capacità di tradurre in modelli le sue idee. Esattamente come il musicista ha la capacità di tradurre in armonie, cioè in modelli precostituiti, le idee dell'artista insipiente. Quest'opera di traduzione sarà più o meno vicina a ciò che il marketing manager avrebbe voluto esprimere. Non tutte le sensazioni e le impressioni tratte dall'esperienza del dominio potranno essere tradotte in statement del modello (allo stesso modo in cui l'esecuzione del pezzo nell'ambito musicale sarà vicino all'ispirazione dell'artista, ma non la ripeterà esattamente). Dopo di che il Miner prova i suoi modelli, compiendo operazioni statistiche di vario genere che come risultato forniranno una descrizione della realtà. L'aderenza del modello alla realtà sarà in genere questionabile, cioè sarà più o meno vicina a ciò che capita. Evidentemente il Miner fornirà misure di questa approssimazione. In genere gran parte del suo lavoro è proprio dedicata a capire di quanto i suoi modelli si allontaneranno dalla realtà, di quanto sbagliano, ovvero, dall'altro punto di vista, quanto i suoi modelli sono affidabili.
Di questo modo di procedere del Miner fornirò due esempi.

Primo esempio: è molto o è poco?

Primo esempio: un problema all'apparenza banale è quello di sapere se una certa quantità è molto o è poco. Prendiamo il caso di un'indagine sui giochi olimpici che abbiamo di recente condotto per la Città di Torino. Nel sondaggio, condotto per telefono a 900 torinesi, ad un certo punto si domandava se l'intervistato era a conoscenza di alcune opere in corso di realizzazione per i giochi del 2006. Questo per capire fino a qual punto era arrivata la circolazione dell'informazione sui giochi. La domanda, a risposte precodificate, era: "Delle seguenti opere, delle quali si è parlato, sa quali verranno effettivamente realizzate?". Dal momento che in tutti i sondaggi occorre superare una predisposizione psicologica che consiste nel cercare di compiacere l'interlocutore e di far, per così dire "bella figura" e per valutare l'entità di tale disturbo, nella batteria delle risposte possibili è stata introdotta una modalità civetta: il Palazzetto dello Sport nel quartiere Pozzo Strada che nessuno si è mai neanche sognato di costruire. Le risposte sono state: non so 52,3% , Sì 29,2%, No 18,3%. Per la Ristrutturazione di Palazzo a Vela, che è invece una delle attività previste, le risposte sono state: non so 19%, Sì 75,4%, No 5,6%. Già così la soluzione del problema consistente nel capire se il 75% è molto od è poco e se, soprattutto, il dato riflette un dato di conoscenza reale e non indotto è risolto: nel secondo caso ci sono meno "non so" e più "sì", come ci saremmo aspettati. Ma un Miner non si contenterà. Ci potrà essere qualcuno che non riterrà sufficienti le differenze riscontrate nelle distribuzioni e dirà per esempio che il 75% degli interlocutori che affermano che un'opera prevista sarà effettivamente realizzata, contro un 30% che lo afferma per un'opera non prevista è troppo poco. Adducendo per esempio a motivo che per valutare positivamente una campagna d'informazione occorre che nessuno pronostichi la realizzazione di un'opera non prevista e tutti la pronostichino per un'opera prevista. A questo punto il Miner interverrà contrapponendo la sua soluzione: in realtà il problema consiste nel capire se gli intervistati rispondono a caso oppure no. Si assume che la risposta data a caso abbia la distribuzione delle risposte alla domanda civetta. Dunque, dice il Miner, assumiamo la distribuzione delle risposte alla domanda civetta casuale e la denominiamo "teorica".Consideriamo poi la differenza tra questa distribuzione e quella della risposte alla domanda relativa ad opere realmente previste che denominiamo "osservata". Se la differenza è significativa in senso statistico, stabilendo per esempio una soglia di probabilità d'errore al 5%, allora rifiutiamo l'ipotesi che le risposte siano date a caso, altrimenti non la rifiutiamo. Tutto questo verrà fatto attraverso il calcolo di una grandezza denominata di $?^2$ (una specie di somma delle differenze algebriche tra frequenze osservate e frequenze

teoriche), che confrontata con la funzione di densità della sua distribuzione di probabilità, ci darà la risposta. Nel caso presente in cui le distribuzioni da confrontare, considerando solo le risposte affermative, sono 70 e 30 contro 25 e 75, il test di $?^2$ sulle proporzioni porge un valore di 108 ed una probabilità di 0,0001, inferiore alla soglia critica scelta al 0,05 (5%). Ragione per la quale decidiamo di considerare le risposte alla domanda relativa alle future opere non casuale. Dovute dunque ad una reale conoscenza. Il pregio maggiore di questo modo di ragionare è la sua incontrovertibilità, in quanto si basa su metodologie scientifiche consolidate. Esprimere critiche a risultati così ottenuti è molto più difficile e impegnativo (anche se non impossibile) che non esprimere disaccordo con affermazioni di buon senso. Tutto ciò è tanto più vero quanto più le frequenze teoriche e quelle osservate sono vicine e prendere una decisione di buon senso si fa difficile ed opinabile.
Un'ultima osservazione. In questi come in altri casi ha molta importanza in modo in cui si pone la domanda nel questionario. Il modo in cui la domanda viene posta risponde a due criteri: facilità di comprensione da parte di chi l'ascolta e possibilità di classificazione per una rapida presentazione dei risultati. Solitamente si usa questa metodologia a risposta chiusa, se l'interesse precipuo è rivolto alla quantificazione, mentre si preferiscono risposte aperte se l'interesse è rivolto piuttosto a far emergere problemi o a capire l'effettivo campo di variazione delle risposte. Come mostrerà l'intervento di Gianluca Bo sulle tecniche di *text mining* questo modo di procedere si rivela interessante in certi contesti ed applicazioni, per esempio nel campo della *Customer Satisfaction*.

Secondo esempio: fenomeni rari

Esistono fenomeni naturali o sociali che normalmente si ripetono con frequenze analoghe, la cui distribuzione è approssimabile da una legge esponenziale. Sono di questo tipo la popolazione delle città, il numero di addetti delle imprese, i redditi delle persone, la frequenza delle parole, la magnitudine dei terremoti, la gravità degli incidenti, la durata in vita di un cliente, il tempo speso su internet e così via. La caratteristica di questo tipo di distribuzioni, pur con diverse approssimazioni, implica che le frequenze delle osservazioni per valori bassi siano estremamente comuni, mentre le frequenze per valori alti siano molto rare. Per esempio in provincia di Torino a fronte di 38.000 imprese del settore privato con più di 2 addetti, solo 29 ne hanno più di 1.000. Tra più di 8.000 comuni italiani solo 3 hanno più di 1.000.000 di abitanti, tra le migliaia di incidenti stradali solo 1, per fortuna, ha avuto le dimensioni dell'incidente nel tunnel del Monte Bianco.
Dal nome di chi le ha studiate, tali distribuzioni prendono nomi diversi: Zipf, Bendford, Mandelbrot, Pareto. Esse sono, per i valori estremi, quello che è la curva di Gauss per le distribuzioni dei valori normali.
Su quest'ultima funzione di distribuzione illustro una metodologia automatica per l'individuazione di valori estremi. L'esempio è costruito sulla popolazione dei comuni Italiani e automaticamente individua le maggiori città. Noi l'abbiamo utilizzata per l'individuazione delle agenzie di una compagnia di assicurazione che hanno valori estremi sugli incassi annuali nei vari rami.
Mentre il precedente esempio ci illustrava un'attività di Analisi Dati avente per obiettivo la determinazione di un criterio oggettivo di valutazione, quest'ultima attività ci ha fatto scoprire l'utilizzo di un metaprogramma (in questo caso SAS) che metabolizza sue proprie istruzioni per automatizzare completamente un percorso di ricerca. Noi non conosciamo fin dall'inizio come stanno le cose, cioè quale forma ha la distribuzione di cui stiamo trattando. Possiamo sospettarlo, ma non esserne certi. Se per la distribuzione delle città possiamo immaginarlo prima di vedere i risultati, non altrettanto possiamo fare per la distribuzione delle agenzie di una compagnia di assicurazione. Le statistiche che pubblichiamo alla fine del percorso hanno questo compito: valutare se i dati di partenza sono congrui con l'ipotesi del modello, che cioè la distribuzione di cui stiamo trattando sia almeno approssimativamente, se non strettamente, Paretiana. Se così è l'incremento di R^2 rispetto all'ipotesi di linearità sarà significativo.

Il Data Mining e la statistica

Non è chiarissimo quale sia la differenza. Cercherò di rintracciare qualche linea di demarcazione seguendo Berry e Linoff[1] sull'argomento:
"La statistica non si riferisce soltanto a tecniche analitiche che processano dati, ma anche allo sviluppo di buoni esperimenti per la raccolta dei medesimi. Un grande esempio di un esperimento accuratamente progettato fu quello di Mendel per descrivere l'eridatarietà. Avendo coltivato e osservato migliaia di piante su molte generazioni, egli fu capace di inferirne l'esistenza dei geni. Egli condusse il lavoro in mancanza di ogni forma di automatizzazione. Fortunatamente, essendo monaco, aveva tempo da spendere per raccogliere e analizzare i suoi dati a mano. Spese la vita per questo.
Oggi la statistica ha qualche vantaggio che le tecniche di data Mining non offrono. Ci sono legioni di persone altamente qualificate che hanno studiato la statistica e le sue applicazioni si può dire in ogni area. Il software statistico è disponibile, lavora su molte piattaforme, anche sulle piattaforme parallele altamente performanti. Le tecniche stesse sono state estensivamente analizzate, cosicché gli statistici possono spiegare con grande dovizia di complessa matematica cosa funziona e cosa non funziona. E, in qualche caso, gli statistici producono risultati altrettanto buoni delle tecniche descritte successivamente (Data Mining specifiche?, *nda*).
Qualche volta, ma non sempre. C'è infine spazio per le tecniche di Data Mining. In breve, la statistica è molto utile, ma non risolve ogni problema del Data Mining. Campionare per ridurre la dimensione di un Data Set può causare la perdita di importanti sottoinsiemi di dati, e con ciò perdere importanti opportunità. A causa dell'enfasi sulle funzioni continue di forma conosciuta, le regressioni non sono generalizzabili così come lo sono le tecniche di Data Mining. La complessità computazionale dell'impostazione statistica non sembra adeguarsi bene a dati di grandi dimensioni".

Sembrerebbe possibile arguire che la differenzia stia 1) nel fatto che il Data Mining è soprattutto dedicato ad analizzare dati già esistenti in formato elettronico, 2) e nel fatto che è particolarmente adatto ad indagare grandi masse di dati. Ci sono però altri elementi che vengono qua e là indicati per differenziare i due ambiti: anche se si afferma che il data Mining è adatto ad una procedura top-down o hypotheses testing gli esempi che di solito vengono portati sembrano orientare tali tecniche di più sul versante dell'esplorazione delle relazioni tra i dati. Soprattutto il

[1] Michael J.A. Berry Gordon Linoff, *Data Mining Techniques*, Wiley, 1997, pp. 114-115

Data Mining viene indicato in congiunzione con un'area tematica precisa, quella della Business Intelligence ed in una sua particolare area: quella dedicata a migliorare il mercato delle imprese, le vendite, il supporto ai clienti ed al CRM in generale.

Le tecniche del Data Mining

Come precedentemente anticipato le tecniche usate si riferiscono e sono adatte alla soluzione di problemi legati alle analisi di mercato.
La Basket analysis che studia la composizione delle borsa della spesa, l'analisi di Cluster che raggruppa records simili, la Link analysis che analizza i legami tra utenti o clienti utilizzando la teoria dei grafi, gli alberi decisionali usati per scopi di classificazione, le reti neuronali che costruiscono attraverso meccanismi di apprendimento testati su dati *training* criteri generali di associazione o di previsione da usare su altri dati, ed altri ancora. Anche se meno univocamente riferibili al Data Mining le tecniche di regressione ai minimi quadrati ordinari o la regressione logistica non vengono disdegnati dai software che implementano soluzioni di Data Mining.
Ciò che qualifica maggiormente le tecniche di Data Mining è comunque la prevalenza del risultato sul metodo utilizzato. Nell'ambito del Data Mining si sottolinea maggiormente l'utilizzo ibrido di tecniche diverse in modo da poter giungere ad un confronto (*assessement*) dei risultati. La resa migliore di un modello in termini di corrette previsioni può essere utile, si dice,cioè portatrice di notevoli guadagni, anche se non statisticamente significativa. In una campagna postale rivolta a 50.000.000 di persone anche un ritorno dell' 1% è significativo dal punto di vista del business poiché sono almeno 500.000 prodotti venduti. Un ritorno di un decimo maggiore, cioè dell'1,1% vorrebbe comunque dire una vendita superiore per almeno 5.000 pezzi. Differenze percentuali piccole fanno differenze grandi in valore assoluto se i numeri sono grandi. Ma, verrebbe da dire, se i numeri sono grandi anche i test statistici fanno in fretta a diagnosticare differenze significative!

Il Data Mining e il Warehousing: il Data Mining di fronte alle esigenze dell'utilizzatore

Un problema importante dell'Analisi dei Dati in generale e del Data Mining in particolare è il reperimento del loro oggetto.
Vi è innanzitutto la necessità di una comprensione dei contenuti delle basi dati: molto spesso, anzi quasi sempre, le necessità di tipo organizzativo e gestionale per cui i dati sono stati raccolti non coincidono con gli scopi in cui altre funzioni aziendali li vorrebbero utilizzare. Ciò è vero sia nel marketing assicurativo come in altri campi: nella pubblica amministrazione o in ambito sanitario ad esempio.
Il percorso di reinterpretazione dei dati, che spesso implica una loro ricostruzione storica (chi lo ha reperito, come è stato raccolto, a quali necessità doveva servire in origine, che significato aveva), dovrebbe dar luogo ad una *definizione condivisa* nell'ambito del nuovo dominio di utilizzo e, possibilmente, anche nel vecchio. Anche se questo problema sarebbe, a rigor di logica, di pertinenza dell'area *warehousing* non è inopportuno sottolinearlo qui. Molto spesso infatti il warehouse si scontra con il problema definitorio di cui sopra data l'intrinseca necessità di operare astrazioni generalizzanti per arrivare a repository d'informazione general pourpose, polifunzionali. Cosa che ne ritarda l'operatività e lo fa essere quasi sempre "non ancora pronto". Per questa ragione il problema del Data Mining non è in principio un problema del Data Mining, ma un problema del data warehouse.

Un altro problema che si pone all'altro capo del percorso è la descrizione degli obiettivi. Il Data Mining ha compiti facilitati quando sa per che cosa scavare. Ma quest'attività non gli deve essere estranea o sopraggiungere dell'esterno: è parte integrante delle proprie funzioni. Pertanto, prima di essere un bagaglio di tecniche statistiche, il Mining deve consistere in attività concettuali, di affinamento e perfezionamento "di idee", che adotta metodiche interattive di lavoro. Il Miner *aiuta* a reperire i dati ed a strutturare un percorso logico finalizzato: egli sa però che entrambi esistono già in qualche forma nell'organizzazione che lo ospita. In una forma che bisognerà scoprire.
Quest'attività *maieutica* si esprime proprio in gruppi di lavoro votati al contenuto, per così dire, prima che alle tecniche. Oltre che Data Miner, l'esperto di tecniche dovrà diventare un *concepts mixer*, un frullatore di concetti.

In terzo luogo il Data Mining deve scoprire il proprio valore aggiunto: un' attività di Mining che inizi nel 2004 e termini nel 2008, per quanto interessante e foriera di buoni risultati, alla fine risulterà inutile.

Beninteso possono essere concepite ancor oggi attività di lunga durata. Il tempo dipende infine dalla dimensione dei problemi. Non è concepibile però che non siano previsti punti di emersione e di verifica intermedi. La misura dei risultati è infatti un terzo, importantissimo compito dell'attività di Data Mining. Le previsioni di vendita elaborate dall'ufficio del marketing devono poter trovare un riscontro nelle campagne di vendita sia a fini di validazione che di affinamento dei modelli. Tra l'altro è proprio attraverso l'analisi dei risultati delle diverse attività di marketing e di vendita che, in un contesto commerciale, il Data Mining misura sé stesso.

L'automazione di processo

Il processo è definibile come una sequenza di operazioni: la definizione di obiettivi di marketing coerenti con l'informazione disponibile; l'organizzazione dell'informazione disponibile in archivi strutturati su livelli di analisi appropriati, l'applicazione di adeguate procedure per l'estrazione e il campionamento dei dati da analizzare secondo le opportune procedure statistiche, la valutazione ex-post dei risultati ottenuti da diversi modelli, la generalizzazione dei risultati ottenuti a tutto il portafoglio o comunque ad ampie popolazioni, la validazione dei modelli sul campo. Delle operazioni citate tutte le operazioni, dall'estrazione dell'informazione in poi, possono essere ingabbiate, da un punto di vista tecnico, in un unico progetto di Data Mining e questo progetto potrà essere riapplicato molte volte su input diversi, purchè i risultati delle analisi vengano considerati stabili, o per lo meno venga considerato valido il set di informazioni iniziale e, nel caso di modelli top-down, venga considerata valida l'identificazione del set delle variabili indipendenti. L'invarianza dei valori stimati dei coefficienti è considerata meno importante in quanto essi sono automaticamente ristimati tutte le volte che il processo viene attivato. Nell'ambito del Data Mining queste attività vengono rappresentate in grafi che il Miner connette secondo sequenze logiche prestabilite, in parte logicamente e in parte tecnicamente prescritte. Come ciò concretamente avvenga è per buona parte determinato dal software usato.
Ci domandiamo ora quali sono i mutamenti che l'automazione di processo determina all'interno dell'organizzazione del lavoro e delle competenze professionali. Riprendiamo l'affermazione intorno al ruolo dell'elaborazione dei dati nella ricerca citata all'inizio e giriamola nei termini di un'organizzazione aziendale:
"*E' ferma convinzione del marketing manager che una volta fatta la scelta del marketing analitico appoggiato su DataMart che necessitano di tecniche IT o ICT, non si possa assolutamente delegare la loro gestione al CED. Alla base dell'ipotesi della delega vi è una concezione riduttiva del processo di analisi dei dati e una non esatta valutazione della gestione dell'interfaccia user e data manager. Considerazioni di metodo ed organizzative richiederanno sempre di più che il marketing manager usi in prima persona le tecniche IT o ICT*".
Tutto stride in quest'affermazione, a partire dalla parola CED, Centro Elaborazione Dati. Ma cosa ci combina in questo discorso? Eppure è opportuno notare che l'oggetto di tutto l'intervento concerne l'Analisi Dati e dunque il CED, che tradizionalmente è depositario dei dati e delle metodologie che riguardano il loro trattamento. Può darsi che lo stridore sia dovuto all'impossibilità di separare il CED dalla sua funzione storica che è sempre stata quella di preparare dati per usi di tipo amministrativo.

Una prima conseguenza dell'applicazione estensiva di metodologie di Data Mining è dunque la trasformazione delle funzioni del CED: da centro di elaborazione dati specialmente dedicato agli usi di tipo amministrativo, a centro multipurpose per l'analisi dei dati che considera le competenze analitiche parte integrante dei propri savoir-faire. Oppure a centro aperto alle esigenze analitiche maturate dall'utenza e le cui competenze l'utenza stessa si incarica di formare o reperire sul mercato. Se venga scelta l'una o l'altra strada non ha molta importanza: sarà localmente deciso sulla base di considerazioni di opportunità.
Da questo punto di vista, anche se nell'affermazione citata stride anche la critica dell'ipotesi della delega, si sottolinea un punto importante: la necessità di ridefinire le modalità dell'interfaccia tra utilizzatore e gestore dei dati. Invece appare non condivisibile l'ultima affermazione, a meno che non venga inteso in luogo di marketing manager l'intera funzione di marketing di una grande azienda: in questo caso la specializzazione all'interno di uno o più esperti di analisi dati può essere decisiva per rendere più redditizio il rapporto con il CED da un lato, e con chi svolge attività in quell'ambito e che collabora o dovrà collaborare con l'azienda dall'altro.
Dunque sul piano dell'organizzazione del lavoro dal Data Mining viene una forte spinta ad una maggior integrazione di ambiti diversi: l'ambito dell'analisi con le implicazioni del dominio su cui si applica e che si tira dietro, e l'ambito dell'elaborazione dei dati. E quel professionista di cui si diceva né,né,né? E' chiaro a questo punto che il Miner è lui: colui che in realtà ha le competenze statistiche, informatiche e di contenuto necessarie per costruire il ponte tra le competenze di cui si diceva.
All'Università il compito di tracciare un percorso didattico che lo formi.

Conclusioni

Spesso lavorando all'analisi dei dati, ci viene spontanea una domanda: "Chissà perché, mentre ognuno di noi non si sognerebbe mai di far dipendere la realtà dai nostri desideri, quando si tratta dell'Analisi Dati ognuno di noi vorrebbe che essa raccontasse quello che noi vogliamo sentirci raccontare?". In Metis non ci siamo ancora dati una risposta: l'abbiamo però buttata in Filosofia e ci siamo dati allo studio di come ci si possa mettere alla ricerca della Verità. Qualche dubbio però ci è venuto: non vorremmo finire come *L' Uomo che guardava passare i treni* di cui Simenon ci racconta che finì in un manicomio e al visitatore scrisse, al fondo della storia: "la Verità non esiste".
La nostra idea è che, al di là delle nostre preferenze, l'Analisi Dati aiuti la conoscenza dei fenomeni e dunque in fondo ci aiuti a lavorare ed a vivere meglio.

Indice