




Università degli Studi di Genova



Dipartimento di Ingegneria  
Biofisica ed Elettronica

	POS. N. 1 Avv. 4/2007
	Riferimento: Misura 3.7 sottomisura d) "Diffusione e trasferimento dell'innovazione" Docup Obiettivo 2 (2000-2006)

***Studio di fattibilità per la realizzazione di un sistema intelligente  
per la classificazione di basi dati eterogenee  
mediante l'utilizzo di algoritmi di clustering e machine learning***

## **Executive Summary**

Genova, 30 giugno 2008

## Introduzione

Questo documento riassume sinteticamente l'attività svolta nell'ambito del progetto sull'uso di tecniche innovative di clustering per data mining entro grandi basi di dati in applicazioni di informatica industriale. L'insieme di dati di riferimento prevede una cardinalità tipica di  $10^3 - 10^5$  elementi nel data base.

Obiettivi dello studio sono lo sviluppo di algoritmi di raggruppamento ('clustering') e rappresentazione aggregata di basi di dati secondo criteri dipendenti dal dominio, con successiva analisi dei risultati del clustering stesso ed interpretazione adattiva dello scenario osservato. Il principale vincolo realizzativo nello sviluppo della tecnologia consiste nella generale applicabilità, a contesti potenzialmente eterogenei, dei metodi implementati. Per verificare le fasi di realizzazione il progetto fa riferimento ad applicazioni in ambito 1) e-government – Sistemi Informativi Lavoro SIL (raggruppamento di *curricula* e profili professionali in grandi basi di dati di Centri Per l'Impiego) e 2) studio di banche dati in ambito di informatica medica (clustering di articoli scientifici).

## Metodologia scientifica

Il principio scientifico ispiratore nella realizzazione del sistema di clustering consiste nel disaccoppiare il contesto applicativo specifico, ovvero le informazioni relative ai dati empirici contenuti nel database, rispetto al 'motore' di calcolo che implementa le effettive operazioni di raggruppamento e successiva analisi. Tale risultato è conseguito attraverso due impostazioni fondamentali: 1) la definizione del concetto di 'documento', che accorpa tutte le informazioni relative ad una entità elementare contenuta nella base dati, e che definisce la granularità con cui si osserva il fenomeno; 2) la definizione di una 'metrica' di dominio che misura il grado di similarità fra due documenti omogenei.

La prima impostazione consente di personalizzare la tecnologia attraverso una interfaccia di acquisizione, che introduce all'interno del sistema un generico set di informazioni pertinenti, convertendole in un documento aggregato compatibile con la tecnologia di *clustering*. La seconda impostazione agisce solo su tali documenti e richiede semplicemente che venga definito un criterio oggettivo di misura di similarità.

Un importante accorgimento tecnico è l'uso di tecniche kernel-based nel computo delle similarità: le metriche oggettive sopra definite comportano necessariamente un problema di normalizzazione, che nello studio qui sviluppato è affrontato attraverso la definizione di un *kernel* apposito. I metodi kernel-based sono all'avanguardia dello stato dell'arte nella gestione di dati complessi; sono particolarmente efficaci quando non si ha disponibile a priori una tecnica esplicita di rappresentazione delle informazioni. Un kernel mappa i dati empirici in uno spazio (di Hilbert) di dimensione arbitraria, il cui unico requisito è la definizione di una similarità fra elementi attraverso un prodotto scalare. Nello studio in questione, un kernel di tipo Radial-Basis-Function (RBF) integra la definizione di metrica (specifica del dominio applicativo) con una implicita operazione di normalizzazione, e rende i risultati compatibili con l'algoritmo di raggruppamento indipendentemente dal dominio stesso. L'insieme delle similarità, trasformate dal kernel rbf, costituisce una matrice di distanze e prodotti scalari a supporto della successiva operazione di raggruppamento.

Il metodo di riferimento per l'operazione di clustering è invece generale e prescinde dalla specifica applicazione. Il modello di apprendimento si appoggia all'algoritmo k-means, che rappresenta uno dei metodi più diffusi per il clustering in presenza di grandi masse di dati. La scelta del particolare algoritmo dipende dal fatto che altri modelli, pur sofisticati, mostrano criticità nella loro complessità computazionale o realizzativa, che ne rendono proibitivo l'impiego in presenza di

grandi masse di dati; lo stato dell'arte di letteratura mostra che il k-means, con opportune varianti di ottimizzazione, resta l'unica soluzione tecnica di interesse pratico.

La stessa versione di riferimento dell'algoritmo, in ogni caso, non ne consente un impiego efficiente entro problemi di data mining, soprattutto per le difficoltà di gestione di matrici di grandi dimensioni. A tal fine lo studio ha sviluppato una variante per grandi data set basata su tecniche di tipo 'decomposition', che sfruttano la località delle informazioni contenute nelle matrici di prodotti scalari. In particolare la tecnica realizzata sfrutta uno schema di decomposizione in cui il concetto di "località" è riferito agli elementi della matrice di kernel. Agire verso questa direzione ha consentito di gestire in modo modulare ed efficiente grandi matrici (di dimensione superiore a  $10^5$  righe/colonne) ove il paradigma classico di esecuzione del k-means avrebbe avuto notevoli limiti. Il risultato finale è che un calcolatore di fascia media è in grado di affrontare, con tempi di calcolo ragionevoli, problemi estremamente complessi di data mining come quelli proposti in questo contesto.

I risultati ottenuti sono del tutto generali in quanto la matrice di kernel è sempre, vista la simmetria, di taglia  $n \cdot (n+1)/2$  indipendentemente dal sottostante corpus documentale. Eventuali differenze nei tempi di esecuzione si possono avere cambiando la metrica quindi in fase di costruzione della matrice di kernel e non nel momento dell'esecuzione vera e propria del k-means. Questo aspetto in pratica significa che ove il vector space abbia poche features la metrica sarà più rapida (dati ben codificati e strutturati come SIL) diversamente nel caso del documento grezzo (fonte documentale di Aitek) il vector space sarà decisamente più ampio con conseguenze ovvie sui tempi di esecuzione. Si ricorda inoltre che il clustering è un tipico processo *batch*, il che significa che non è importante che esso abbia tempi di esecuzione ristretti, conta invece che estragga informazioni utili interrogabili a posteriori conclusa la fase di addestramento. Detto questo, la fase di raggruppamento non conclude l'esecuzione del motore, infatti è necessario un processo di estrazione sintetica dell'informazione. Questo consiste nell'analisi dei clusters ottenuti dall'algoritmo k-means. I gruppi ottenuti sono osservati e descritti sotto diverse prospettive:

- Nel caso SIL il processo di estrazione sintetica consiste nell'estrazione e relativo "etichettamento" dei cluster utilizzando le informazioni su: le principali qualifiche, la provincia di residenza e la cittadinanza. Questa scelta ha consentito un buon compromesso per avere una significativa descrizione dei gruppi senza creare descrittori, benché ricchi, troppo complessi ai fini dell'intelligibilità da parte dell'operatore. Si noti poi che questa operazione è stata facilitata nel contesto SIL dal buon condizionamento dei dati, ovvero a causa della loro struttura ben codificata.
- Nel caso di analisi di fonte documentale non sussiste l'ipotesi di buona strutturazione dei dati. Questo ha reso necessario un approccio diverso per l'estrazione concettuale da siffatti clusters. Nella fattispecie l'estrazione dell'informazione si basa sostanzialmente nell'analisi delle parole più frequenti, tecnica generalmente usata in problemi di text mining. In tal caso le informazioni estratte caratterizzano meglio di altri indicatori sintetici il sottostante insieme di documenti. Con tale tecnica sono stati trovati gruppi ben distinti che, nel contesto corrente, corrispondevano spesso a diagnostica, tecniche terapeutiche, studi clinici, malattie in genere ed altri campi di interesse medico.

In linea generale l'architettura sviluppata consente in modo modulare di costruire l'apposito analizzatore per ogni apposito dominio. Il disaccoppiamento fra analisi e raggruppamento ha consentito di creare una struttura facilmente adattabile ad altri domini applicativi secondo la logica per cui, dato un dominio applicativo, ne discende una ben codificata struttura di classi (concetti) tutte afferenti all'engine comune e alla stessa logica descrittiva.

## Metodologia applicativa

### *Dominio SIL*

La base di dati su cui è stata effettuata la validazione dell'engine di clustering è un estratto di  $10^5$  documenti provenienti da vari SIL principalmente della Calabria. ETT ha definito e fornito un formato testuale a cui appoggiarsi per il caricamento dei dati. Il formato è composto da tag che segnalano le sezioni e le sottosezioni del particolare CV in esame. Globalmente il formato è gerarchicamente organizzato in: sezioni, sottosezioni e campi, inoltre è dotato di terminatori ed è sempre indicato il numero di sottosezioni in modo da semplificare la procedura di lettura dei dati.

I dati forniti provengono in modo "grezzo" dal sottostante database SIL: ciò significa che nessuna operazione di filtraggio e pulizia dei dati è stata effettuata a monte del processo di raggruppamento. Questo aspetto ha costituito un ulteriore banco di prova per l'engine sviluppato in modo tale da sottolineare la capacità, non solo di raggruppamento di dati "puliti", ma anche di trovare gruppi di dati omogeneamente "rumorosi", garantendo così un ulteriore, e spesso inaspettato in senso positivo, discernimento fra dati completi, e dunque degni di analisi, e dati talmente rumorosi da non essere rilevanti.

### *Dominio documentale*

In questo secondo contesto applicativo si è passati all'analisi di una forma testuale destrutturata, ovvero gli abstract di articoli scientifici. In tal caso la metrica utilizzata è stata quella euclidea basata su un vector space. Nella fattispecie per aumentare la robustezza del clustering è stata effettuata la rimozione delle stopwords e lo stemming, ovvero rispettivamente la rimozione di token lessicali come articoli, preposizioni ecc.. e l'estrazione delle radici delle parole; per quest'ultimo passo in particolare è stato utilizzato l'algoritmo di Porter. In tal caso la sfida scientifica tecnica non consisteva tanto nella mole di dati ( $10^3$ ) tanto più che altro nella cura nelle operazioni di "pulizia" dei testi originali attraverso alcuni accorgimenti e nella fase di calcolo delle distanze. Qui infatti alcuni accorgimenti tecnici pensati ad hoc per dati documentali hanno consentito la creazione di un metodo efficiente di calcolo della metrica.

## Risultati conseguiti

Gli esperimenti in entrambi i dominini hanno seguito una struttura standard per validare e ispezionare nell'ordine: la struttura generale della soluzione, la metrica e la resa finale.

In entrambi i domini i risultati ottenuti hanno consentito di valutare come consistente il paradigma adottato sia in termini di raggruppamento in quanto tale dall'analisi delle curve dei costi, sia dalla reale fruibilità dei risultati ottenuti da un operatore umano. L'analisi finale proposta è fatta da non esperti di dominio: questo sottolinea ulteriormente come un operatore umano non analista del settore possa già trarre una notevole quantità di informazioni dal sistema con una conoscenza a priori pari a zero.

A maggior ragione, a nostro avviso, questo aspetto valida i risultati raggiunti in vista di analisi approfondite effettuate da analisti specifici. Per quanto concerne altri aspetti operativi gli interessati congiuntamente con gli implementatori hanno creato scenari di interesse commerciale per la tecnologia realizzata decisamente realistici e fondati da un punto di vista e produttivo e di qualità intrinseca e percepita dal cliente finale.