Clustering delle serie storiche economiche: Applicazioni e questioni computazionali

Sergio M. Focardi
The Intertek Group
94, rue de Javel - F-75015 Paris
Tel: +33 1/45 75 51 74
e-mail: interteksf@aol.com
www.theintertekgroup.com

CAPI 2001
Milano
16-17 Ottobre 2001

1 Summary

La ricerca di insiemi di serie storiche caratterizzate da pattern comuni è un importante problema statistico che si presenta in molti campi applicativi che vanno dall'economia e la finanza alle comunicazioni ed alla biomatematica. Facendo appello ad una nozione intuitiva di clustering di serie storiche, vengono presentate evidenze di clustering in problemi quali clustering di serie economiche per la definizione delle regioni economiche degli USA, clustering di comportamenti legati al credito per la creazioni di rating, clustering di serie di prezzi di azioni per la definizione di strategie d'investimento. La determinazione di cluster richiede la definizione di un criterio di similarità, o distanza, fra serie. Viene discussa la problematica della definizione di una metrica fra serie e vengono presentati alcuni criteri di distanza fra serie storiche utilizzati in pratica e proposti nella letteratura, collocando il clustering in un insieme coerente di metodologie statistiche per l'analisi delle serie storiche. Vengono poi discussi i problemi computazionali associati al clustering delle serie storiche. Vengono illustrati dati comparativi relativi al carico computazionale del clustering di vari insiemi di serie storiche, mostrando che la ricerca di cluster di serie finanziarie lunghe porta a problemi computazionalmente molto onerosi.

Key words: clustering, cointegrazione, correlazione, distanza, serie storiche, similarità.

2 Introduzione

La finanza è un utilizzatore relativamente recente di calcolo ad alte prestazioni. Grandi modelli macroeconomici erano in uso già a partire dagli anni 60 e, negli anni 80, il premio Nobel Kenneth Wilson riconosceva nella simulazione di sistemi economici una delle Grand Challenges computazionali poi inserite nei programmi del Department of Energy (DOE) del governo americano. In pratica, tuttavia, le aziende finanziarie divennero utilizzatori di calcolo ad alte prestazioni solo a partire dalla fine degli anni 80 quando alcune delle maggiori aziende acquisirono supercomputer della classe dei Cray o macchine parallele del tipo Thinking Machines.

Questo interesse per il supercalcolo coincise con la diffusione dei derivati in ambito finanziario e con l'arrivo del data mining in ambito marketing. Aziende quali Merrill Lynch e Prudential Securities, infatti, usavano i supercalcolatori per la valorizzazione dei derivati mentre l'American Express utilizzava Thinking Machines per scoprire pattern comportamentali nell'utilizzo delle carte di credito.

La rapidissima crescita della potenza elaborativa disponibile su computer desktop ha reso molti di questi problemi risolubili con questo tipo di macchine. L'interesse della comunità finanziaria verso il supercalcolo è oggi determinato sia da problemi di valorizzazione dei derivati e misura dei rischi che debbano essere risolti sotto severi vincoli di tempo, sia da problemi che coinvolgano un universo di titoli o contratti molto grande. Alcuni di questi problemi rimangono ai confini delle possibilità tecnologiche del calcolo. Problemi di clustering di serie storiche del tipo di quelli che andiamo a descrivere, se eseguiti su grandi universi di titoli e quindi di serie storiche, sono computazionalmente molto impegnativi e possono eccedere le capacità di computer desk-top.

Un'azienda finanziaria che voglia applicare tecniche di data mining per scoprire pattern e similarità nelle serie storiche di un insieme superiore ai diecimila titoli si trova ad affrontare problemi computazionali che richiedono un'attenta pianificazione delle risorse computazionali. Dal punto di vista della pianificazione delle operazioni e dei relativi strumenti analitici di un'azienda finanziaria, è importante valutare i vantaggi concorrenziali che l'analisi di similarità può offrire a fronte dei costi relativi. Vorremmo offrire prima un'idea generale delle motivazioni dell'analisi di similarità e presentare alcuni esempi che aiuteranno a situare tale tipo di analisi. Passeremo poi a discutere le tecnologie e delineare le questioni aperte.

Intutivamente l'analisi di similarità porta a riconoscere gruppi di serie storiche che hanno comportamenti simili. La similarità delle serie storiche segnala relazioni di causalità (oppure similarità strutturale) dei processi che le generano e permette di fare previsioni più accurate. Il raggruppamento per similarità è un concetto forte che conduce a relazioni stabili fra processi. La ricerca di similarità è perciò

la ricerca di un nucleo centrale di relazioni stabili e poco influenzate dal rumore.

3 Qualche esempio di clustering delle serie storiche economiche

Le relazioni di similarità fra serie economiche sono analizzate nella pratica corrente della ricerca economica e finanziaria. Per illustrare questo punto, vorrei fare tre esempi pratici di clustering di serie storiche in ambito economico e finanziario.

- Il primo esempio riguarda l'applicazione delle tecniche di clustering a livello macroeconomico su grandi aggregati. Si tratta della suddivisione del territorio degli Stati Uniti in regioni economiche effettuata dalla Federal Reserve Bank of Philadelphia. Naturalmente gli stessi concetti e le stesse tecniche sono applicabili a qualunque altra nazione o regione geografica. Vi sono due divisioni in regioni degli Stati Uniti in uso corrente. La prima, adottata dal Bureau of the Census, comprende 9 regioni e risale al 1910. Quella più usata dagli economisti è la divisione in 8 regioni effettuata dal Bureau of Economic Analysis (BEA) nel 1950. Nel 1950, tuttavia, un comitato interagenzia guidato dal Department of Commerce raccomandò di riconsiderare la suddivisione regionale degli Stati Uniti, basandola su criteri economici e statistici. Riprendendo tali raccomandazioni, la Federal Reserve Bank of Philadelphia si è posta il problema di verificare la fondatezza economica della suddivisione regionale BEA e di proporne eventualmente una differente basata su criteri economici.

  A tal fine, si è scelto un insieme di variabili economiche che caratterizzano gli stati. Le variabili rappresentano quantità quali il reddito pro-capite, l'occupazione in settori non agricoli, il prodotto lordo dello stato. Le variabili sono successivamente aggregate in un indice rappresentativo di ogni stato. Si è scelto di fissare la granularità dell'analisi a livello stato perché a questo livello sono disponibili variabili con cadenza mensile. L'analisi è stata limitata ai 48 stati che sono geograficamente contigui. Il territorio occupato da questi stati è perciò rappresentato da insiemi di serie storiche mensili, ed in particolare da 48 serie mensili degli indici economici di ciascun stato.

  La Federal Reserve Bank of Philadelphia ha applicato tecniche di clustering a queste serie storiche per verificare quali gruppi di stati abbiano comportamenti simili e possano formare una regione. Il risultato dell'analisi di similarità (Crone, 1999) ha portato a definire 6 regioni contigue su 43 stati mentre altri 5 stati rimangono isolati o costituiscono gruppi troppo piccoli per formare una regione. Le regioni trovate mostrano un livello di omogeneità superiore a quello dei raggruppamenti tradizionali e rispettano le raccomandazioni del comitato interagenzia del Department of Commerce.

- Il secondo esempio riguarda applicazioni di misura del rischio di credito che coinvolgono un grande numero di contratti individuali. Al fine di misurare il rischio di credito è importante determinare quantità quali la probabilità individuale d'insolvenza e le correlazioni fra comportamenti creditizi. Queste valutazioni devono essere fatte su aziende che sono, in linea di principio, tutte diverse fra di loro.

  Se si disponesse di una teoria causale del comportamento dell'azienda, si potrebbero valutare le probabilità individuali d'insolvenza e ricostruire i comportamenti aggregati. La teoria stessa potrebbe essere validata su qualsiasi aggregato. Tuttavia non si dispone di una tale teoria e si possono solo fare misure statistiche in aggregato su popolazioni di aziende che riteniamo simili. Diventa pertanto importante determinare comportamenti aziendali simili, che permettano di raggruppare le aziende in segmenti omogenei sotto il profilo della probabilità d'insolvenza o del livello di correlazione.

  Per conto di un a banca italiana di media dimensione, abbiamo effettuato diverse esperienze di clustering di comportamenti creditizi su universi di contratti dell'ordine di 10.000 contratti. Nelle esperienze fatte, sono state usate variabili economiche quali gli indici di bilancio e variabili legate al comportamento dell'azienda nei confronti della banca che eroga il credito. Queste variabili sono ricavate dai rapporti del cliente con la banca. Potrebbero anche essere usate variabili qualitative ottenute codificando i rapporti d'agenzia.

  Per una banca di medie dimensioni, in genere si trova un numero di cluster significativi dell'ordine di 6-8 cluster in aggiunta a cluster satelliti che sono troppo piccoli e debbono essere aggregati manualmente. L'analisi in cluster trova segmenti più o meno omogenei in funzione della clientela della banca e della sua presenza regionale. Va osservato che per una banca media, in genere si trova che il rischio è molto concentrato in piccoli segmenti formati da grosse esposizioni mentre la clientela media presenta rischi molto bassi e livelli medi di insolvenza bassi. Questo comportamento è fortemente determinato dalla distribuzione della dimensione dei contratti che, in genere, segue molto da vicino la legge di Zipf, cioè la dimensione delle esposizioni è approssimativamente inversamente proporzionale al suo rango.

- Un terzo esempio di analisi in cluster è costituito dall'analisi dei titoli azionari in settori, stili e temi d'investimento. Questa analisi è importante ai fini della gestione del rischio, sia a livello banche ed assicurazioni, sia a livello gestione degli investimenti. In tutti i processi di gestione del rischio, infatti, è importante diversificare le assunzioni di rischio in modo da eliminare i rischi inutili e sfruttare le opportunità di hedge naturali. Più in generale, il processo di gestione dei rischi è un processo di ottimizzazione che dipende in

modo critico dalle correlazioni fra processi. La segmentazione del mercato in settori corrisponde ad un primo livello di analisi delle correlazioni.

Tradizionalmente i mercati azionari sono stati segmentati in aree geografiche e settori merceologici, oppure usando indicatori quali la capitalizzazione od il tasso di crescita aziendale. Varie ragioni, soprattutto legate a fenomeni di globalizzazione e di formazione di grandi conglomerati, rendono oggi tali segmentazioni poco efficaci. Si è ricercato pertanto un nuovo modo di segmentare i mercati finanziari basato sull'analisi delle correlazioni empiriche fra titoli.

Queste tecniche stanno entrando nella pratica corrente, ed esistono ormai molti esempi di banche ed aziende finanziarie che utilizzano tecniche di clustering più o meno sofisticate per segmentare i mercati azionari.

Vediamo perciò che il clustering delle serie finanziarie è un processo di raggruppamento e discretizzazione delle correlazioni (coarse graining) che soddisfa alcune esigenze fondamentali tra cui:

- identifcare aggregati a cui corrispondano scelte politiche od economiche, come nel caso della segmentazione territoriale in regioni;
- identificare aggregati a cui corrispondano particolari caratteristiche, come nel caso dei rating di credito;
- ridurre la dimensionalità dei problemi per renderli trattabili sotto il profilo matematico e statistico, come nel caso della determinazione dei temi d'investimento;
- separare il rumore dall'informazione costruendo correlazioni stabili, come nel caso della gestione del rischio.

Notiamo che il problema di definire la similarità fra serie si trova anche nel contesto dei database e data-warehouse. Infatti è stato posto da tempo il problema di fare ricerche per similarità in una base dati di oggetti ad alta dimensionalità (Agrawal, Faloutsos e Sami, 1993). Questo è un problema classico di pattern recognition che ha condotto a metodologie applicative nuove in domini quali la biomatematica, la sismologia, le telecomunicazioni ed i problemi di intelligence quali il riconoscimento di esplosioni nucleari a partire da pattern sismologici.

4 Similarità e distanza

Passiamo ora ad illustrare in modo sintetico le tecnologie e le problematiche che si incontrano nell'analisi di similarità. La clusterizzazione per similarità delle serie storiche poggia su due pilastri: 1) la definizione di una distanza tra serie e 2) i

metodi di clustering. Infatti clusterizzare significa trovare insiemi di elementi simili, dove la similitudine è quantificata in un concetto di distanza fra serie. Mentre efficaci procedure di clustering sono ben note e sono implementate in software commercialmente disponibili, quali MathWorks, SAS e SPSS, la definizione di concetti di distanza tra serie è forse l'elemento più nuovo su cui si concentra una buona parte della ricerca.

Dato un insieme di serie storiche, la **funzione distanza** fra di esse, nella sua definizione più generale, è una funzione non negativa definita su ogni coppia di serie e tale che . Un elevato livello di similarità fra due serie è caratterizzato da un piccolo valore della loro distanza. La distanza fra serie può essere definita in modo da originare una **metrica**. In tal caso, la funzione distanza $d(a, b)$ fra le serie a,b rispetta le condizioni della metrica:

$$d(a, b) > 0, \text{if} a \neq b \tag{1}$$

$$d(a, a) = 0, \forall a \tag{2}$$

$$d(a, b) = d(b, a), \forall a, b \tag{3}$$

$$d(a, c) \leq d(a, b) + d(b, c), \forall a, b, c (\text{diseguaglianza triangolare.} \tag{4}$$

Si chiama **quasi-metrica** una funzione distanza che soddisfi le condizioni precedenti eccetto la diseguaglianza triangolare. Si chiama **ultrametrica** ogni funzione distanza che soddisfi le quattro condizioni precedenti più la relazione:

$$d(a, c) \leq \max(d(a, b), (d(b, c)) \tag{5}$$

Ogni funzione distanza implementa un concetto di similarità. Non esiste una distanza ottimale che implementi il "vero" concetto di similarità; ogni distanza serve ad un obiettivo specifico. La scelta della funzione distanza da utilizzare dipende perciò dal problema che si vuole risolvere. Differenti nozioni di distanza richiedono l'applicazione di differenti tecniche di segmentazione e clustering.

5 Distanza, correlazione e cointegrazione

Il problema di stabilire relazioni di somiglianza fra serie non è nuovo in statistica. Infatti, il concetto classico di **correlazione** e di **regressione** è stato introdotto per caratterizzare processi che hanno un andamento simile. Dato un insieme di serie storiche, la matrice di correlazione o di varianza-covarianza misura, infatti, il grado di similarità fra le varie serie. Si noti, tuttavia, che i coefficienti di correlazione non costituiscono una distanza perché possono assumere valori compresi fra -1 e +1.

Diverse motivazioni hanno spinto ad andare oltre il concetto di correlazione. Una prima motivazione è l'applicabilità del concetto di correlazione. Infatti, si possono trovare similitudini fra pattern o spezzoni di serie che sarebbero statisticamente non correlati. Ne abbiamo visto un esempio nella ricerca di pattern di comportamento creditizio che conducono al fallimento. Si possono trovare percorsi simili verso il fallimento anche tra aziende lontane o nel tempo o nello spazio, anche se non vi è alcuna correlazione diretta tra tali percorsi. I concetti di similarità si applicano perciò a contesti più generali della correlazione.

Una seconda, forte, motivazione che ha indotto a superare il concetto di correlazione è la fondamentale instabilità che può manifestarsi nelle relazioni di correlazione. Se si considera un insieme abbastanza ampio di titoli, ad esempio i titoli appartenenti all'indice Standard & Poor 500 (S&P 500), si trova che la matrice di **varianza-covarianza** fra questi titoli non è stabile a meno di valutare tale matrice su periodi estremamente lunghi. E' stato osservato empiricamente (Laloux, Cizeau, Bouchaud e Potters, 1999, Plerou, Gopikrishnan, Rosenow, Amaral e Stanley, 1999) che se il numero dei titoli è dello stesso ordine di grandezza del numero dei punti su cui si valutano le correlazioni, la matrice delle correlazioni non è stabile ma presenta fluttuazioni casuali.

Considerare periodi molto lunghi per valutare le correlazioni non è realistico dal punto di vista economico perché si valuterebbero correlazioni fra entità aziendali che nel frattempo hanno, in genere, subito importanti cambiamenti. Limitando le finestre temporali a periodi dell'ordine di 1000 giorni, la matrice di correlazione su grandi aggregati è instabile.

La rumorosità delle matrici di varianza-covarianza è solo un aspetto del problema della stabilità delle correlazioni. Infatti, il problema della debolezza intrinseca delle correlazioni era noto in statistica da parecchio tempo, anche prima dell'osservazione che le matrici di varianza-covarianza sono molto rumorose. La possibilità di osservare correlazioni spurie era stata osservata da Yule nel 1926 e Granger e Newbold (1974) avevano mostrato che è possibile trovare correlazioni spurie in fenomeni perfettamente scorrelati, anche se i campioni osservati sono molto grandi. Infatti vi sono casi in cui i metodi correnti di determinazione delle correlazioni convergono anche se le serie sono indipendenti. Le correlazioni trovate sono completamente spurie.

Questa osservazione è una delle motivazioni che hanno spinto all'introduzione delle tecniche di **cointegrazione**. Due o più processi sono cointegrati se esiste una loro combinazione lineare che sia stazionaria. Processi cointegrati rimangono pertanto vicini a meno di trasformazioni lineari. Le loro correlazioni sono stabili. Dato un insieme di serie storiche empiriche, è possibile eseguire test di cointegrazione che determinano se è ragionevole considerare le serie generate da processi cointegrati e determinare i processi stessi. Ad esempio, dato un insieme

di serie storiche si possono stimare modelli a correzione d'errore (ECM) che producono processi cointegrati.

Non è agevole, tuttavia, adottare gli stessi metodi per analizzare grandi insiemi di serie empiriche in cui solo alcuni cluster sono eventualmente cointegrati. Per tornare ad un esempio finanziario, è evidente che su un grande aggregato come l'S&P 500 non tutti i titoli saranno cointegrati, anche se è possibile ipotizzare che esistano cluster di titoli fortemente correlati o anche cointegrati. Per risolvere questi problemi sono state proposte varie strategie che conducono tutte a segmentazioni del mercato in segmenti internamente correlati.

Alcuni ricercatori tra cui Laloux, Cizeau, Bouchaud e Potters (1999) e Ormerod e Mounfield (2000) hanno proposto di utilizzare la teoria delle **matrici casuali**, una teoria matematica sviluppata nell'ambito della meccanica quantistica negli anni 50. Se si considera la matrice di correlazione di serie storiche completamente indipendenti e casuali, si trova che gli autovalori di tale matrice hanno una distribuzione caratteristica che può essere calcolata teoricamente. E' stato perciò proposto di confrontare tale distribuzione teorica con la distribuzione degli autovalori di serie empiriche per discriminare il rumore dall'informazione nella matrice di correlazione. I risultati empirici mostrano che la distribuzione empirica degli autovalori della matrice di correlazione del S&P 500 è sorprendentemente vicina a quella di una matrice casuale.

Esistono tuttavia importanti deviazioni che segnalano l'esistenza di correlazioni vere e stabili al di sopra del rumore. Questi risultati possono essere interpretati con la presenza di cluster di titoli fortemente correlati al loro interno mentre il rumore maschera le eventuali correlazioni fra altri titoli. L'analisi basata sulle matrici casuali mostra perciò che esiste una struttura di correlazione su insiemi del tipo S&P 500 distribuita su cluster correlati (Ormerod e Mounfield, 2000).

E' naturale, allora, porre il problema di ricercare direttamente i cluster di titoli solidamente correlati. Come osservato in precedenza, i coefficienti di correlazione non costituiscono una distanza, in quanto possono assumere valori negativi. Al fine della ricerca di cluster è necessario introdurre una misura quantitativa della similarità fra serie che sia una distanza. Data una distanza, possono essere applicate procedure di clustering che conducono a determinare cluster di elementi vicini.

Si vede perciò che correlazione, cointegrazione e similarità sono nozioni che possono stare in varie relazioni. Similarità è un concetto più generale di correlazione. Può essere applicato a processi che abbiano una somiglianza strutturale opportunamente definita ma che non sono necessariamente correlati. Ad esempio, i prezzi di due azioni diverse possono seguire percorsi strutturalmente simili in momenti diversi e su intervalli temporali diversi senza essere necessariamente correlati.

Tuttavia, sotto opportune condizioni aggiuntive, si possono trovare cluster di processi simili che siano stabilmente correlati ed eventualmente cointegrati. Si noti che la matrice di similarità (o distanza) non è necessariamente più stabile della matrice di varianza-covarianza (o correlazione). La similarità, tuttavia, può essere definita in modo sufficientemente stringente per cui cluster di elementi simili mostrano correlazioni stabili.

6 Definizioni di similarità

Passiamo ora ad esaminare alcuni tra i più diffusi concetti di similarità che sono stati proposti. Date due o più serie prese negli stessi N intervalli temporali, la più semplice misura di similarità è la distanza L_p, considerando le serie come punti in uno spazio ad N dimensioni. L_1, talvolta chiamata **distanza Manhattan**, è definita come la somma (o la media) dei moduli delle differenze fra le serie, L_2 è la **distanza euclidea**, L_q è la **distanza di Minkowsky**, definita come $L_q = (\sum_{i=1}^{N} |a_i - b_i|^q)^{\frac{1}{q}}$, L_∞ è la massima distanza tra punti omologhi. Queste distanze sono tutte metriche in quanto rispettano le condizioni della metrica. Esse hanno il vantaggio di essere semplici ma sono piuttosto restrittive. Innanzitutto esse richiedono che le serie siano definite negli stessi punti. Inoltre sono sensibili ad eventuali outlier e disallineamenti. Ad esempio, se due serie prezzi con forti fluttuazioni sono molto simili nell'andamento ma leggermente disallineate, la loro distanza euclidea può risultare molto grande.

Inoltre, in molte applicazioni si vogliono considerare simili serie a meno di fattori di scala e di spostamenti del valor medio. Ad esempio, le serie che rappresentano i prezzi di due azioni differenti possono avere lo stesso andamento e le stesse fluttuazioni percentuali ma valori assoluti molto diversi. Perciò si può estendere la definizione precedente misurando la similarità fra serie che sono state assoggettate a trasformazioni, sia per cambiarne i valori medi sia per variarne la scala.

Questo concetto di similarità è ancora troppo restrittivo per molte applicazioni finanziarie. Ad esempio, nel caso della valutazione del rischio creditizio, aziende differenti possono andare verso l'insolvenza seguendo percorsi strutturalmente simili ma su un lasso di tempo differente. E' perciò importante identificare similarità anche fra serie che occupino intervalli di tempo differenti. A tal fine è stato introdotto il concetto di **time-warping**. Una serie è soggetta a time-warping se ai suoi elementi vengono aggiunti elementi contigui identici. Ad esempio, le serie $(2, 21, 3, 4)$ e $(2, 2, 1, 1, 3, 4, 4)$ sono identiche dopo un'operazione di time-warping.

Il time-warping consiste perciò nello stirare o comprimere localmente una serie. La distanza di time-warping è definita come la minore possibile distanza (ad esempio in senso L_p) dopo operazioni di time-warping. Un algoritmo di distanza time-warping allungherà o comprimerà localmente le serie fino a che la loro distanza non sia minima.

E' necessario, tuttavia, estendere ulteriormente il concetto di similarità per includere serie che sono simili solo in parte. Similarità di questo genere sono importanti quando si studiano, ad esempio, comportamenti particolari di titoli che abbiano seguito un pattern specifico in un certo intervallo. Sono state avanzate varie proposte che conducono a considerare la distanza fra due serie solo in un sottoinsieme di punti.

Restando nell'ambito delle distanze metriche, un concetto differente di distanza è stato proposto da Bollobas, Das, Gunnopulos e Mannila (1997). Questi autori definiscono innanzitutto un intervallo di tolleranza che prescrive se punti omologhi sono considerati simili o no. La distanza fra due serie è definita dal rapporto fra il numero di punti simili rispetto al numero di punti totale. Naturalmente si possono assoggettare le serie ad operazioni di cambiamento di scala prima di calcolare la distanza in questo senso.

E' stata avanzata l'ipotesi che la similarità delle serie finanziarie possa essere meglio descritta da una ultrametrica, introducendo così una fondamentale gerarchizzazione nella similarità (Bonanno, Vandewalle e Mantegna, 2000 e Mantegna, 1999). E' stato mostrato (Ormerod e Mounfield, 2000) come si possa costruire una ultrametrica fra serie in modo molto semplice a partire dai consueti coefficienti di correlazione fra serie, definendo la distanza come:

$$d(a,b) = \sqrt{2(1 - C(a,b)} \tag{6}$$

dove $C(a,b)$ è il consueto coefficiente di correlazione.

Le distanze fra serie definite nei precedenti paragrafi sono esempi di distanze geometriche calcolate direttamente sui punti delle serie, eventualmente dopo trasformazioni di scala o di base dei tempi. Sono stati proposti concetti di similarità definiti su spazi correlati formati da varie trasformate delle serie. La prima proposta di similarità fra serie, infatti, (Agrawal, Faloutsos e Swami, 1993) era basata sul proiettare le serie nello spazio delle Fast Fourier Transform (FFT). Proposte più recenti includono l'uso di wavelet (Huhtala, Karkkainen e Toivonen, 1999), approssimazioni lineari a pezzi (Keogh, Chakrabarti, Pazzani e Mehrotra, 2000) o l'estrazione di varie caratteristiche delle serie (Chang-Shing Perng, Wang, Zhang e Stott Parker, 2000).

Trasformazioni particolarmente importanti per la finanza e l'economia, portano dallo spazio delle serie allo spazio delle loro distribuzioni di probabilità. Le distanze precedentemente definite non dipendono direttamente dalla distribuzione statistica dei punti delle serie. Naturalmente tutte queste distanze possono essere interpretate statisticamente, ricavando appropriate statistiche eventualmente con metodi di simulazione. La loro motivazione, tuttavia, non è statistica ma è legata a vari concetti di somiglianza fra le forme delle serie, totali o parziali.

Se si lascia cadere la condizione di triangolarità e si fanno alcune ipotesi statistiche sulle serie, si possono adottare altri tipi di distanza che si sono rivelate vantaggiose in molte applicazioni. In particolare (Kakizawa, Shumway e Taniguchi, 1998), sono stati proposti concetti di similarità basati sulla **teoria dell'informazione**. E' stato proposto di utilizzare l'entropia di Kullback-Leibler, definita come

$$I(a,b) = E[\log(\frac{p(x)}{q(x))}] \tag{7}$$

o la misura di informazione di Chernoff definita come

$$B_\alpha = -\log E[(\frac{q(x)}{p(x)})^\alpha] \tag{8}$$

per definire la quasi-distanza fra due serie come: $J(a,b) = I(a,b)+I(b,a)$, oppure $J(a,b) = B_\alpha(a,b) + B_\alpha(b,a)$ rispettivamente.

Sempre nell'ambito di distanze definite in termini di concetti probabilistici, sono state proposte distanze basate sui modelli di Markov nascosti (**hidden Markov models** - HMM). In questo caso si cerca una rappresentazione delle serie in termini di catene di Markov con un numero di stati che deve essere determinato con criteri statistici (Xianping Ge e Padraic Smyth, 2000).

7 Tecnologie e problematiche di clustering

Le procedure di clustering basate su una metrica non sono, in generale, di tipo gerarchico. Questo significa che non è possibile ordinare la clusterizzazione in ordine crescente di numero di cluster in modo che cluster più piccoli siano sempre contenuti in cluster di livello superiore. In generale aumentando il numero dei cluster ci saranno cambiamenti globali di struttura. E' stato proposto, sia in ambito economico che biomatematico, di restringere l'analisi a cluster di tipo gerarchico.

Clusterizzazioni di tipo gerarchico conducono in modo naturale ad ultrametriche. Si può dimostrare infatti che ogni ultrametrica induce una clusterizzazione di tipo gerarchico. Per converso, ogni clusterizzazione di tipo gerarchico induce una ultrametrica. Data una matrice di distanze di,j che rispettino le condizioni dell'ultrametrica, esistono (e sono implementate in software commerciali) procedure di determinazione dei cluster basate sulla creazione di alberi minimi.

Sotto il profilo computazionale, le metodologie di clustering e di ricerca di similarità sono molto onerose soprattutto quando applicate a grandi insiemi di dati. Per dare un'idea della dimensione dei problemi, si consideri il calcolo della matrice di varianza-covarianza per un insieme di 500 titoli su una finestra temporale dell'ordine di 100 giorni. Si tratta di calcolare circa centomila numeri diversi,

dove ciascun calcolo richiede l'esecuzione di un numero di operazioni dell'ordine delle migliaia. Il calcolo richiede un numero di operazioni dell'ordine di 10 alla 9 ed è eseguibile giornalmente su desk-top dedicati.

Se i coefficienti di correlazione sono sostituiti da misure di similarità complesse, ad esempio considerando il time-warping, il numero di operazioni può risultare moltiplicato di vari ordini di grandezza (Das, Gunopulos e Mannila, 1997). A questi tempi vanno aggiunti i tempi per eseguire e far convergere le procedure di clustering. A livello insiemi del tipo S&P 500, i problemi computazionali presentati da metodologie di clustering complesse possono eccedere le capacità di sistemi desk-top. Se si considerano aggregati più vasti, dell'ordine delle migliaia di titoli, è richiesta una pianificazione accurata delle risorse computazionali ed eventualmente il ricorso a metodologie semplificate ed a postazioni di supercalcolo.

E' naturale chiedersi quale sia il vantaggio reale che ci si può aspettare dall'adozione di questo tipo di metodi. L'ampiezza delle caratterizzazioni di similarità porta naturalmente a chiedersi se si può stabilire quali di queste caratterizzazioni siano le migliori, e se rispondano ai criteri che si erano indicati inizialmente. Esistono alcuni studi di confronto fra metodologie di clustering di serie finanziarie, quali ad esempio lo studio effettuato da Gavrilov, Anguelov, Indyk e Motwani (2000).

La determinazione delle procedure migliori è un problema empirico che dipende dalla base dati considerata. Ad esempio, nel caso della determinazione dei rating di credito, l'utilizzabilità effettiva di procedure di clustering basate su dati storici e comportamentali dipende strettamente dalla disponibilità di dati e dal costo di fruizione dei dati teoricamente disponibili.

Un problema correntemente oggetto di ricerca riguarda le serie finanziarie. In particolare si vorrebbe capire se metodologie di clustering e di "coarse graining" dell'informazione conducano effettivamente a determinare relazioni di correlazione stabili e, più in generale, ad una efficace discriminazione fra informazione e rumore. E' importante sottolineare le difficoltà metodologiche a cui ogni attività di ricerca va incontro, difficoltà essenzialmente dovute alla paucità dei dati ed alla non stazionarietà dei fenomeni.

Da un punto di vista pratico, va sottolineato che si ricercano gli aspetti stabili dell'economia e del mercato. Queste stabilità possono condurre a notevoli miglioramenti nella gestione del rischio e nella gestione degli investimenti ma richiedono di essere compresi nella loro fondamentale dimensione economica.

References

[1] Agrawal, R., C. Faloutsos e A. N. Swami, "Efficient Similarity Search in Sequence Databases", *FODO*, 1993.

[2] Bollobas, B., G. Das, D. Gunopulos e H. Mannila, "Time-Series Similarity Problems and Well-Separated Geometric Sets", in *Proceedings of the Association for Computing Machinery Thirteenth Annual Symposium on Computational Geometry*, 454–476, 1997.

[3] Bonanno, G., N. Vandewalle e R. N. Mantegna, "Taxonomy of Stock Market Indices", *Physical Review E62, R7615-R7618*, 2000.

[4] Crone, T., "Using State Indexes to Define Economic Regions in the US", *Working Paper N. 99-19, Federal Reserve Bank of Philadelphia*, Novembre 1999.

[5] Das, G., D. Gunopulos e H Mannila. "Finding similar time series", in *Proceedings of The Fourth International Conference on Knowledge Discovery and Data Mining*, Agosto 1997.

[6] Gavrilov, M., D. Anguelov, P. Indyk e R. Motwani, "Mining the Stock Market: Which Measure is Best?" in *Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining*, 487–496, 2000.

[7] Granger, C.W.J. e P. Newbold, "Spurious Regressions in Economics", *Journal of Econometrics*, 2, 111-120, 1974.

[8] Huhtala, Y., J. Karkkainen e H. Toivonen, "Mining for Similarities in Aligned Time Series Using Wavelets", in *Data Mining and Knowledge Discovery: Theory, Tools and Technology, SPIE Proceedings Series Vol. 3695*, 150-160, Orlando, 1999.

[9] Keogh, E., K. Chakrabarti, M. Pazzani e S. Mehrotra, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases", *Journal of Knowledge and Information Systems*, 2000.

[10] Kakizawa, Y., R.H. Shumway e M. Taniguchi, "Discrimination and Clustering for Multivariate Time Series", *Journal of the American Statistical Association*, Marzo 1998.

[11] Laloux,, L., P. Cizeau, J.P. Bouchaud e M. Potters, "Noise Dressing of Financial Correlation Matrices", *Phys Rev Lett 83*, 1467, 1999.

[12] Laloux, L., P. Cizeau, J.P. Bouchaud e M. Potters, "Random Matrix Theory and Financial Correlations", *Dublin Conference 1999*, World Scientific Publishing Company.

[13] Mantegna, R.N., "Hierarchical Structure in Financial Markets", *Eur Phys J B11, 193*, 1999.

[14] Ormerod, P. e C. Mounfield, "Localised Structures in *Temporal Evolution of Asset Prices", New Approaches to Financial Economics, Santa Fe Conference*, Ottobre 2000.

[15] Chang-Shing Perng, H. Wang, S. R. Zhang e D. Stott Parker, "Landmarks: a New Model for Similarity-based Pattern Querying in Time Series Databases", *International Conference on Data Engineering*, San Diego, USA, 2000.

[16] Plerou, V., P. Gopikrishnan, B. Rosenow, L.A.N Amaral e H.E. Stanley, "Universal and Non-Universal Properties of Cross-correlations in Financial Time Series", *Phys Rev Lett 83, 1471*, 1999.

[17] Xianping Ge e Padraic Smyth, "Deformable Markov Templates for Time Series Pattern Matching", *Technical Report N. 00-10, University of California at Irvine*, March 2000.

[18] Yule, G.W., "Why Do We Sometimes Get Nonsense Correlations Between Time Series? A Study on Sampling and the Nature of Time Series", *Journal of the Royal Statistical Society, 89, 1-64*, 1926.