UNIVERSITÀ DEGLI STUDI ROMA TRE

FACOLTÀ DI SCIENZE M.F.N.

Sintesi della Tesi di Laurea in Matematica

di

Maria Antonietta Galante

Applicazioni di Data Mining Logico a dati finanziari per la determinazione del Timing. Una verifica dell'ipotesi di efficienza dei mercati.

Relatore

Dott. Luca Torosantucci

Relatore interno

Prof. Francesco De Notaristefani

ANNO ACCADEMICO 2003 - 2004

Luglio 2004

Classificazione AMS : 68N17, 91B28

Parole Chiave : Learning Logic, Efficient Market Theory.

Sintesi

L'obiettivo della tesi è realizzazione di un test per la verifica dell'ipotesi di efficienza debole dei mercati finanziari. Il test è stato impostato sulla ricerca di una strategia di investimento che fosse in grado di battere il mercato; la strategia si basata sulle informazionioni relative all'andamento passato dei prezzi, in particolare su alcuni degli strumenti propri dall'analisi tecnica.
A tal fine si è condotta un'analisi su dati finanziari, forniti sotto forma di quotazioni storiche (cioè, aperture, minimi, massimi, chiusure e numero di azioni scambiate), con lo scopo di selezionare quei casi in cui sia possibile ottenere determinati rendimenti in determinati orizzonti temporali.
Data l'abbondanza di variabili che possono essere ricavate attraverso le informazioni utilizzate, per analizzare i dati a disposizione abbiamo utilizzato metodi di apprendimento logico usate nelle tecniche di data mining. I casi da selezionare, dunque, vengono riconosciuti attraverso formule logiche. A questo punto, le regole base della logica proposizionale consentono di sfruttare tali formule per la determinazione di vere e proprie strategie di investimento attivo il cui scopo sarà, come detto, quello di ottenere rendimenti superiori al mercato finanziario.

Secondo la definizione di Fama (1970), *un mercato si definisce efficiente quando i prezzi dei titoli riflettono in maniera pronta e corretta tutte le informazioni disponibili.* Questo significa che non è possibile implementare una strategia di investimento che sia in grado di battere il mercato. Le condizioni sufficienti perché un mercato risulti efficiente sono particolarmente restrittive;

in particolare si richiede che[1]:

1. Tutti gli investitori agiscano in modo razionale cercando il massimo profitto e non siano collegati tra loro;

2. Il set di informazioni sia sempre disponibile e gratuito per tutti gli operatori di mercato;

3. Gli investitori abbiano aspettative omogenee e concordino sull'influenza che le singole informazioni hanno sulle prospettive di crescita, e quindi sui prezzi attuali, dei singoli titoli;

4. Non esistano costi di transazione ed imposte.

Appare chiaro come tutto questo sia irreale, lo stesso Fama sottolinea come tali condizioni siano sufficienti ma non necessarie per la definizione di efficienza. Un mercato, dunque, può considerarsi efficiente quando:

- Un adeguato numero di partecipanti abbia accesso alle informazioni e queste non siano monopolio di pochi;

- Il disaccordo sull'influenza delle informazioni disponibili sul prezzo e il rendimento futuro (eterogeneità delle aspettative) non consenta a taluni investitori di battere costantemente il mercato;

- L'esistenza di costi di transazione e tasse non sia tale da scoraggiare gli scambi.

In base alla categoria di informazioni disponibile si definiscono tre forme di efficienza dei mercati [29]:

[1]Caparrelli, Fama (1970).

L'efficienza debole (Weak Hypothesis): *il prezzo attuale di un titolo incorpora tutte le informazioni contenute nelle serie storiche dei prezzi passati e più in generale da tutte le notizie che possono essere tratte dall'andamento passato del titolo.* In altre parole, analizzando l'andamento passato di un titolo o costruendo una qualsiasi strategia su di esso basata, non è possibile ottenere "costantemente" rendimenti superiori a quelli di mercato.

L'efficienza semi-forte (Semi-strong Hypothesis): *il prezzo attuale di un titolo incorpora tutte le informazioni pubblicamente disponibili.* Secondo quanto affermato dall'efficienza semi-forte, dunque, l'analisi di bilancio non fornisce suggerimenti operativi che permettano "costantemente" rendimenti superiori al mercato. Secondo questa forma di efficienza, quindi, tutti i titoli sono quotati ad un valore congruo alla situazione economica e finanziaria dell'azienda e alle sue prospettive di crescita.

L'efficienza forte (Strong Hypothesis): Secondo quanto suggerito prima da Roberts e poi dallo stesso Fama, l'efficienza forte afferma che *non esiste alcun gruppo di investitori avente "accesso monopolistico" a informazioni rilevanti.* In altre parole, pur esistendo gruppi di investitori che hanno informazioni privilegiate o riservate, tali informazioni non sono in grado di assicurare con il tempo guadagni superiori alla media.

Durante gli anni sessanta e settanta la quasi totalità della letteratura scientifica era concorde con l'ipotesi di efficienza del mercato. A partire

dagli anni ottanta, però, cominciarono ad apparire lavori che segnalarono forti anomalie rispetto a tutte e tre le forme di efficienza. Attualmente si ritiene che i mercati siano sostanzialmente efficienti e che le opportunità di guadagno superiori al mercato siano rare e di difficile sfruttamento, a volte per la rapidità con cui vengono assorbite, in parte per la difficoltà di individuarle.

L'impossibilità di ottenere rendimenti superiori al mercato spinge gli investitori fautori dell'efficienza a una strategia di investimento che miri a selezionare un portafoglio rappresentativo del mercato nel quale si intende investire e attenderne passivamente l'evoluzione. La strategia così implementata prende anche il nome di *buy and hold* e si parla in tal caso di *strategia d'investimento passivo*.

Il mondo operativo e una parte del mondo accademico non concordano con la teoria di Fama, e ritengono invece possibile implementare strategie alternative in grado di fornire rendimenti mediamente superiori al mercato. Tra questi i sostenitori dell'analisi tecnica, o traders, che ritengono sia possibile sfruttare l'andamento passato dei prezzi per implementare le strategie vincenti e migliori del mercato in termini di rendimenti; il loro obiettivo è quello di riuscire a determinare il *timing*, ovvero i momenti migliori per acquistare e vendere.

Verificare la validità dell'analisi tecnica significherebbe provare l'inefficienza del mercato in forma debole. L'anilisi tecnica si basa su un elevato numero di strumenti, ognuno di loro e qualsiasi loro combinazione rappresenta una strategia. Ogni traders opera sul mercato attraverso una di tali combinazioni e soprattutto di volta in volta può optare per una diversa.

Impostare un test di efficienza dedole dei mercati sulle strategie dell'analisi tecnica è molto difficile perchè, come detto, sono caratterizzate da molti strumenti e per di più variabili nel tempo. Le verifiche condotte in tal senso non possono ritenersi sufficienti perchè utilizzano un solo strumento per volta oltre tutto costante nel tempo.
L'unico modo per poter veramente testare la validità dell'analisi tecnica è riuscire a mettere insieme tutti gli strumenti disponibili, e ancor di più riuscire ad individuare di volta in volta le migliori. É proprio questo l'obiettivo che abbiamo raggiunto con il Data Mining.

Il Data Mining (DM) è un processo che impiega una o più tecniche di apprendimento computerizzate per analizzare automaticamente ed estrarre le conoscenze da grandi quantità di dati [30].
Le tecniche di DM si possono dividere in due grandi categorie: le tecniche di *apprendiemnto supervisionato* e le tecniche di *apprendimento non supervisionato* (o *clustering*). Le tecniche di apprendimento supervisionato si basano sulla costruzione di meccanismi che siano in grado di estrapolare in modo automatico delle regole a partire da un *insieme di esempi*; è possibile distinguerle ulteriormente in *stima*, *previsione* e *classificazione*. Le tecniche di *clusterizzazione*, invece, costruiscono modelli da dati senza classi predefinite. Tali tecniche si basano tutte su metodi statistici, informatici e di ottimizzazione; le prime hanno trovato un'importante applicazione nei problemi di marketing e vengono, ad esempio, utilizzate come supporto in alcune applicazioni di diagnosi medica. Nel nostro caso si è lavorato su tecniche di classificazione.

L'essenza delle tecniche di DM supervisionato, di cui fa parte la classifi-

cazione, può essere sintetizzata dai seguenti punti:

- si considera un numero finito di oggetti misurabili attraverso un numero finito di variaili; ogni oggetto può essere *rappresentato* da un vettore $x = (x_1, x_2, \ldots, x_m)$;
- si stabilisce, in base al problema considerato, la *caratteristica di interesse* y degli oggetti;
- l'obiettivo è la determinazione della *relazione* F tale che $y \cong F(x)$

F può avere una qualsiasi forma funzionale, o essere rappresentata da una formula logica, o da un algoritmo. La scelta della forma delle relazioni è un passo fondamentale e viene presa in base al tipo di problema considerato: ci sono dei casi in cui la forma non fa differenza purchè la relazione trovata funzioni; a volte, però, può essere opportuno scegliere una particolare forma funzionale adatta al tipo di applicazione o coerente con le conoscenze a priori del problema.
Altro passo importante è la scelta del *metodo* che si vuole utilizzare per trovare la relazione F, infatti, possono esserci modelli che sullo stesso problema funzionano meglio di altri.

Chiariti i punti essenziali delle tecniche di apprendimento supervisionato in genere, si mostra ora in cosa consiste la *classificazione*.
Dato un insieme T (*training-set*) di oggetti appartenenti a due classi distinte A e B, scopo della classificazione è stabilire una regola che associ gli oggetti alla propria classe d'appartenenza. Tale regola detta *classificatore* viene riapplicata all'insieme T per testarne la *correttezza* ovvero per verificare che riesca effettivamente a separare gli oggetti nelle due classi prestabilite; una

volta testato il grado di affidabilità del classificatore questo viene utilizzato per classificare nuovi insiemi di oggetti (*testing-set*).
Il classificatore non è altro che la relazione F introdotta sopra e la caratteristica di interesse y serve a distinguere le due classi A e B. Da ciò è necessario stabilire uno *spazio di rappresentazione* per gli oggetti da classificare. Le possibilità sono diverse e dipendono dal problema di apprendimento in questione, ad esempio gli attributi dei vettori di rappresentazione possono avere valore *reale*, *nominale* o *logico*.

Vediamo in che modo il metodo è stato adattato al nostro caso.
L'idea è quella di riuscire a riconoscere, nell'ambito delle operazioni finanziarie, gli investimenti in grado di restituire un determinato guadagno in un determinato orizzonte temporale.

Per far questo abbiamo rappresentato ogni possibilità d'investimento, cioè ogni giorno con i suoi relativi prezzi di, chiusura, apertura, minimo e massimo, in vettori di m componenti a valore $\{0, \pm 1\}$ ovvero i *vettori di rappresentazione* x. Ogni variabile si basa sulla segnalazione di una delle strategie della analisi tecnica implementate, che in totale sono 149. La variabile x_i riferita alla i-esima strategia vale 1 se la strategia restituisce un segnale d'acquisto, -1 se restituisce un segnale di vendita e 0 se non restituisce segnali.

La caratteristica di interesse y è stata fissata in modo tale da assumere valore 1 per il relativo dato se questo verifica un determinato rendimento in un determinato numero di giorni e -1 altrimenti.

Abbiamo scelto di rappresentare la relazione F, tale che $y \cong F(x)$, attraverso *formule logiche*: questo, come vedremo più avanti, ne faciliterà l'utilizzo e la comprensione.

Il sistema di riconoscimento automatico realizzato in questa lavoro si basa sul metodo di apprendimento per problemi espressi in domini logici descritto da Felici e Truemper in un articolo del 2001 [12].

Il metodo opera su vettori $r \in \{0, \pm 1\}^n$ denominati *records* ad ognuno dei quali è associata una variabile logica che può assumere valore *vero* o *falso*. La presenza di un 1 in un record significa che una certa variabile logica, w ha valore *vero*; mentre -1 significa che w ha valore *falso*. La presenza dello 0 indica che il valore della variabile w non si conosce o non è significativa.

Il metodo ricava dei vettori $s \in \{0, \pm 1\}^n$ che possono essere usati per calcolare il valore *vero* o *falso* associato ad ogni record r. Sostanzialmente i vettori s separano i records che hanno valore *vero* (ad esempio, la classe A) dai records che hanno valore *falso* (ad esempio, la classe B), per questo motivo vengono chiamati *vettori di separazione*. La totalità dei vettori di separazione costituisce un *insieme di separazione*. Da ogni insieme di separazione è possibile derivare una formula logica che utilizza i valori $\{0, \pm 1\}$ delle componenti dei records r per calcolarne il valore *vero* o *falso*.

Gli insiemi di separazione sono determinati attraverso un algoritmo iterativo. In ogni iterazione vengono risolti due problemi di minimizzazione logica (MINSAT) per ottenere un vettore di separazione. Le soluzioni dei problemi MINSAT vengono fornite da un sitema di programmazione logica avanzato chiamato *Leibniz System* [37]. I problemi SAT in generale, appartengono alla classe dei cosiddétti $\mathcal{NP}$-completi; in questa classe rientrano tutti i problemi che non sono risolvibili in un tempo di ordine polinomiale e per questo motivo rappresentano un caso molto studiato nella teoria della complessità computazionale. Il Leibniz System, però, consente di risolvere in modo effi-

ciente e con prestazioni garantite i problemi di tipo SAT.
Per la correttezza del metodo è importante che le due classi di records A e B siano *separabili*, ovvero che non esista un records che appartenga sia ad una classe che all'altra o che esistano due records rappresentati dallo stesso vettore e tali che uno appartenga ad A ed uno appartenga a B.

Gli insiemi di separazione generati dal meccanismo di apprendimento su dati logici descritto sopra, non sono altro che i classificatori utilizzati per riconoscere i giorni favorevoli all'investimento. Ogni insieme di separazione è formato da un numero finito k di vettori di separazione. Data la natura booleana delle sue componenti, ogni vettore di separazione $s = (s_1, s_2, \ldots, s_m)$ può essere trasformato in una *clausola logica* $(s_1 \wedge s_2 \wedge \ldots \wedge s_m)$ e, di conseguenza, ogni insieme di separazione in una *formula logica in forma disgiuntiva normale* (DNF) $s^{(1)} \vee s^{(2)} \vee \ldots \vee s^{(k)}$.
Siccome ogni variabile s_i rappresenta il risultato di una ben determinata strategia operativa, è possibile trasformare una *formula classificatrice* in *strategie classificatrici*.

Il sistema sviluppato in questa tesi è stato sperimentato secondo due modalità:

- I test che chiameremo di *tipo classico* sono quelli che valutano le percentuali di riconoscimento di ognuna delle formule classificatrici. Il procedimento adottato è il seguente: si considera un insieme di dati finanziari sotto forma di quotazioni storiche e lo si divide in altri due insiemi, il *training-set* ed il *testing-set*; gli elementi dei due insiemi e i due insiemi stessi devono rispettare l'ordinamento cronologico originale; sul training-set viene applicato il meccanismo d'apprendimento

che genera le formule logiche classificatrici; tali formule vengono sfruttate per il riconoscimento degli elementi, tra classe A e classe B, del testing-set, in base al numero di classificazioni corrette si calcola la percentuale di riconoscimento di ogni formula.

- Il secondo tipo di test condotti mirano in maniera ancora più evidente alla verifica d'efficienza debole. Sulla base delle regole di classificazione restituite dal metodo viene realizzata una strategia d'investimento attivo finale.

I nostri risultati sono stati i seguenti. Sul primo tipo di test sono stati condotti degli esperimenti sia sull'indice S&P500, che rappresenta il 90% dei titoli quotati del NYSE, che sul titolo Fiat; abbiamo voluto valutare i due casi differenti, indice e singolo titolo, e soprattutto due tipi di andamento diverso; nello stesso periodo, infatti, la correlazione tra le due serie di quotazioni storiche analizzate è pari a -0.36; si sono osservate in media percentuali di riconoscimento del 70% in entrambi i casi. Secondo la definizione di efficienza debole che sostanzialmente afferma che non è possibile valutare il futuro dell'andamento dei prezzi da quello passato, tale percentuale dovrebbe oscillare intorno 50%.

La strategia implementata per il secondo tipo di test è stata messa a confronto con la strategia di investimento passivo del *buy and hold*, su periodi di circa sei mesi, sia sull'indice S&P500 che sul titolo Fiat. I risultati ottenuti hanno fatto registrate performance superiori al buy and hold di circa il 5%, costi di transazione inclusi, sia in caso di crescita che di perdita.

In sostanza i risultati ottenuti con il metodo sviluppato in questa tesi sembrerebbero mostrare delle anomalie nell'efficienza debole dei mercati: impie-

gando un sistema esteso per la rappresentazione del trend di un titolo e un algoritmo sofisticato di data mining, è possibile ottenere in diversi casi un comportamento migliore del mercato, anche se il differenziale non è molto alto. Ovviamente, tale affermazione richiede un lavoro di estensione sperimentale voluminoso prima che possa essere considerata sufficientemente solida; in tal senso, il lavoro svolto intende essere un primo passo verso lo studio di nuovi metodi di analisi del trend e di definizione di strategie attive.

Bibliografia

[1] Berry M.J. Linoff, 2001, *Data Mining*, Apogeo, Milano.

[2] Brealey R.A., Myers S.C., 1993, *Principi di Finanza Aziendale*, McGraw-Hill, Milano.

[3] Boros E., Ibaraki T., Makino K., 1999, Logical analysis of binary data with missing bits *Artificial Intelligence* 107, 219-263.

[4] Boros E., Ibaraki T., Kogan A., Mayoraz E., Muchnik I., 1996, An implementation of logical analysis of data *RUTCOR Research Report* 29-96, Rutgers University, NJ.

[5] Breiman, Friedman, Olshen, Stone, 1984, *Classification & Regression Trees*,Pacific Grove, Wadsworth.

[6] Caparrelli F., 1995, *Il Mercato Azionario*, McGraw-Hill, Milano.

[7] Curram S.P., Mingers J., 1994, "Neural Networks, Decision Tree Induction and Disciminant Analysis: an Empirical Comparison", *J.Op.Res.Soc.*, vol.45, n.4, 440-450.

[8] Di Lorenzo R., 1999, *Guadagnare in borsa con l'Analisi Tecnica* Voll.1-3, Ed. Sole24ORE.

[9] Elton-Gruber, 1995, *Modern Portfolio Theory and Investment Analysis*, John & Wiley & Sons Inc.

[10] Evans, Archer, "Diversification and the Reduction of Dispersion: an Empirical Analysis" - *Journal of Finance*, XXIII, n° 5 (Dec.1968), pp. 761-767.

[11] Felici G., 1995, *Il Problema di riconoscimento automatico: proprietá ed algoritmi di soluzione*, Tesi di Dottorato, Biblioteca Nazionale di Roma, Italy.

[12] Felici G., Truemper K., 2001, "A MINSAT Approach for Learning in Logic Domains", *INFORMS Journal on Computing*, Vol.13, No.3, pp.1-17

[13] Felici G., Truemper K., 1997, *Learning Logic*, IASI Technical Report 450.

[14] Fornasini A., 1991, *Analisi Tecnica e Fondamentale di borsa*, ETAS.

[15] Francis K., 1990, *Investment Analysis and Management*, SH Saw and Lim, Longmans.

[16] Fuller R.J., Farrell J.L., 1993, *Analisi degli Investimenti Finanziari*, McGraw-Hill, Milano.

[17] Graham B., 1972, *The Intelligent Investor*, Harper & Row.

[18] Hastie T., Tibshirani R., Friedman J., 2001 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

[19] Haugen, 1999, *Modern Investment Theory*, Prentice Hall.

[20] Hooker J., Chandru V. 1999 *Optimization Methods for Logical Inference*, Wiley-Interscience.

[21] Jonhson, Shannon, "A Note of Diversification and the Reduction of Dispersion", *Journal of Financial Economics*, 1, n° 4, (dec. 1974), pp. 365-372.

[22] Kantardzic M., 2003, *Data Mining: Concepts, Models, Methods, and Algorithms*, IEEE Press, Piscataway - NJ.

[23] Leibniz System, 1996, Version 4.3 Leibniz Company, Plano, TX.

[24] Leibniz System Software, http://www.leibnizsystems.com/.

[25] Lorie, Hamilton, 1973, *The Stock Market: Theories and Evidence*, Irwin.

[26] Malkiel B.G., 1981, A Random Walk down Wall Street, Down Jones Irwin.

[27] Pinches G.E., "The Random Walk Hypothesis and Technical Analysis", *Financial Analysts Journal*, Marzo-Aprile 1970.

[28] Pring M.J. 2003 *Analisi tecnica dei mercati finanziari*, McGraw-Hill.

[29] H. Roberts: "Statistical Versus Clinical Prediction of the Stock Markets" - Studio inedito presentato al *Seminar on the Analysis of Security Prices*, University of Chicago, Maggio 1967.

[30] Roiger R.J., Geats M.W., 2004 *Introduzione al DATA MINING*, McGraw-Hill.

[31] Samuelson, "Challenge to Judgement", *Journal of Portafoglio Management*, n° 1, 1974.

[32] Samuelson, "The Judgement of Economic Science on Rational Portfolio Management: Indexing, Timing, and Long Horizon Effect", *Journal of Portafolio Management*, n° 1, 1974.

[33] Samuelson, "Asset Allocation Could Be Dangerous to Your Health", *Journal of Portafolio Management*, n° 3, 1990.

[34] Tinic, West, 1979, *Investing in Securities: An Efficient Market Approach*, Addison-Wesley.

[35] Tobin J. 1984 *On the Efficiency of the Financial System*, Lloyds Bank Review.

[36] Torosantucci L., 2002, *Guadagnare in Borsa con il Metodo di Benjamin Graham*, Experta, Forlì.

[37] Truemper K. 1998 *Effective Logic Computation*, Wiley-Interscience.

[38] Vaciago G., Verga G., 1995, *Efficienza e Stabilità dei Mercati Finanziari*, Il Mulino, Bologna.

[39] Whitmore, "Diversification and the Reduction of Dispersion: a Note" - *Journal of Financial and Quantitative Analysis*, V, n° 2 (May 1970), pp. 263-264.