Πρόβλεψη χρηματιστηριακών μεγεθών με τεχνικές εξόρυξης δεδομένων.

Αθανάσιος Μαζαράκης

Επιβλέπων: Νικόλαος Σαμαράς, Επίκουρος

Πανεπιστήμιο Μακεδονίας, Τμ. Εφαρμοσμένης Πληροφορικής

Θεσσαλονίκη, Ιούλιος 2007

Πρόβλεψη χρηματιστηριακών μεγεθών με τεχνικές εξόρυξης δεδομένων.

ΣΚΟΠΟΣ:

Η κατάλληλη επιλογή και επεξεργασία των χρηματιστηριακών δεδομένων προκειμένου να πετύχουμε την βέλτιστη εξαγωγή συμπερασμάτων - κανόνων πρόβλεψης με τεχνικές εξόρυξης δεδομένων.

Είδος Δεδομένων

Χρησιμοποιήθηκαν πραγματικά δεδομένα σε ημερήσια βάση, των παρακάτω μετοχών του χρηματιστηρίου αξιών Αθηνών.

1. ΆLFA (2/1/1986 - 5/1/2007)
2. ATE (19/1/2001 - 5/1/2007)
3. BIOXK (11/4/1997 - 5/1/2007)
4. DEH (15/2/2001 - 5/1/2007)
5. EEEK (15/7/1991 - 5/1/2007)
6. ELPE (30/6/1998 - 5/1/2007)
7. ELTEX (20/4/1994 - 5/1/2007)
8. EMP (5/9/1989 - 5/1/2007)
9. ETE (2/1/1986 - 5/1/2007)
10. EYROB (26/11/1987 - 5/1/2007)
11. FOLI (29/10/1997 - 5/1/2007)
12. INLOT (4/11/1999 - 5/1/2007)
13. KOSMO (6/10/2000 - 5/1/2007)
14. KYPR (9/11/2000 - 5/1/2007)
15. MOH (15/2/2001 - 5/1/2007)
16. OPAP (15/2/2001 - 5/1/2007)
17. OTE (19/4/1996 - 5/1/2007)
18. PEIR (19/4/1990 - 5/1/2007)
19. TITK (1/7/1987 - 5/1/2007)
20. TT (2/6/2006 - 5/1/2007)

Μορφή Δεδομένων

```
<TICKER>,<DTYYYYMMDD>,<OPEN>,<HIGH>,<LOW>,<CLOSE>,<VOL>
ATE,20010119,5.3512,5.3512,5.1235,5.2374,2222943
ATE,20010122,5.2374,5.2374,5.1349,5.2032,2132022
ATE,20010123,5.2032,5.2374,5.1008,5.2260,589111
ATE,20010124,5.1349,5.2260,5.1349,5.1805,250359
ATE,20010125,5.1805,5.1918,5.1235,5.1805,410096
ATE,20010126,5.1805,5.2032,5.1349,5.1805,171892
ATE,20010129,4.5884,5.1691,4.5884,5.1463,210379
ATE,20010130,5.1235,5.1691,5.1121,5.1577,320843
ATE,20010131,5.1577,5.1805,5.1463,5.1691,297225
ATE,20010201,5.1805,5.1805,5.1235,5.1463,190204
ATE,20010202,5.1691,5.1691,4.7933,5.1463,211565
ATE,20010205,5.1349,5.1349,5.0666,5.1235,183406
ATE,20010206,5.1235,5.1235,4.9527,5.1121,4265870
ATE,20010207,5.1121,5.1349,5.0552,5.0780,237633
ATE,20010208,5.0780,5.0780,5.0438,5.0666,207296
ATE,20010209,5.0666,5.1235,5.0552,5.0666,156478
ATE,20010212,5.0438,5.1008,5.0211,5.0552,188342
ATE,20010213,5.0552,5.0780,5.0097,5.0666,199795
ATE,20010214,5.0666,5.0666,4.9072,5.0666,113775
ATE,20010215,5.0666,5.0780,5.0438,5.0666,120256
ATE,20010216,5.0666,5.0780,4.9414,5.0552,193577
ATE,20010219,4.9641,5.0666,4.9527,5.0438,1400063
ATE,20010220,5.0324,5.0666,4.9527,5.0324,165243
ATE,20010221,5.0324,5.0324,4.8503,4.9300,87689
ATE,20010222,4.6339,4.9186,4.6339,4.7706,124525
```

Θέματα Τεχνικής Ανάλυσης - Δείκτες

- **Stochastic Oscillator** (%K-period,%D-slowing)

  =%(Κλείσιμο ημέρας- ελάχιστο περιόδου)/(μέγιστο περιόδου – ελάχιστο περιόδου)

- **MACD** (Moving Average Convergence/Divergence)

  = EKM(Close,12)-EKM(Close,26)

- **RSI** (Relative Strength Index – 14 Days)

  =100[100/(1+U/D)], U=Average of upward price change

  D=Average of downward price change

- **Momentum**

  =(τιμή κλεισίματος/τιμή κλεισίματος 12 ημερών προγενέστερα)*100

Σημείωση :ΕΚΜ σημαίνει ότι οι τιμές συμμετέχουν εκθετικά για τον υπολογισμό του μέσου όρου καθώς πλησιάζουμε τη τρέχουσα τιμή υπολογισμού.

Επεξεργασία των Δεδομένων (1)

- Υπολογίσαμε ίδιου μεγέθους διανύσματα ή καλύτερα σειρές με τις αρχικές, για κάθε έναν από τους προαναφερθέντες δείκτες σε σχέση πάντα με την πραγματική τιμή κλεισίματος.

- Υπολογίσαμε το συντελεστή Cross correlation μεταξύ της ημερήσιας τιμής κλεισίματος κάθε μετοχής και των δεικτών :

  - RSI
  - Momentum
  - MACD
  - Stochastic Oscillator

Επεξεργασία των Δεδομένων (2)

- Cross Correlation για καθυστέρηση d

$$r = \frac{\sum_i [\,(x(i) - mx) * (y(i-d) - my)\,]}{\sqrt{\sum_i (x(i) - mx)^2}\sqrt{\sum_i (y(i-d) - my)^2}}$$

- Cross Correlation για καθυστερήσεις d=0,1,2,…,N-1

$$r(d) = \frac{\sum_i [\,(x(i) - mx) * (y(i-d) - my)\,]}{\sqrt{\sum_i (x(i) - mx)^2}\sqrt{\sum_i (y(i-d) - my)^2}}$$

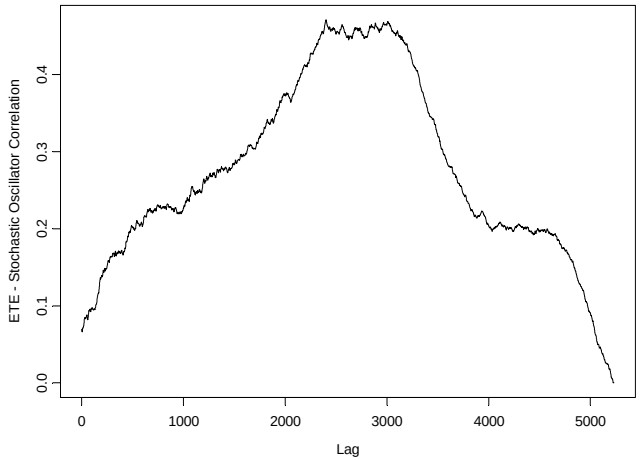
Υλοποίηση Συνάρτησης Cross Correlation

```
void correlation (double vect1[], double vect2[], double *vect3){

	double numerator[N],sy[N],denom[N];

	for(int k=0; k<N; k++)
	numerator[k]=sy[k]=denom[k]=0;
	double sx;
	for (int d=0; d<K; d++){
		sx=0;//gia kathe d tha prepei na mhdenizoyme to sx
		for(int i=0; i<K; i++){
			if((i-d)>0){
			    numerator[d]=numerator[d]+((vect1[i]-mean(vect1))*(vect2[i-d]-mean(vect2)));
			    sy[d]=sy[d]+((vect2[i-d]-mean(vect2))*(vect2[i-d]-mean(vect2)));
			}
			else{
			     numerator[d]=numerator[d]+((vect1[i]-mean(vect1))*(-mean(vect2)));
			     sy[d]=sy[d]+((-mean(vect2))*(-mean(vect2)));
			}
		    sx=sx+((vect1[i]-mean(vect1))*(vect1[i]-mean(vect1)));
		}
		denom[d]=sqrt((sx*sy[d]));
		*vect3=numerator[d]/denom[d];
		vect3++;
	}
}
```

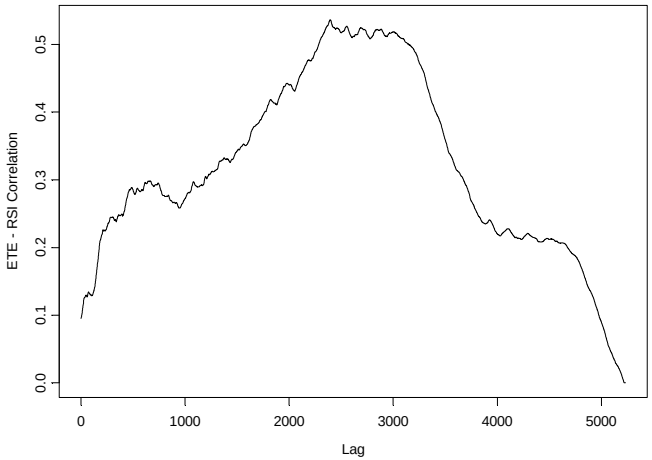
Αποτελέσματα Cross Correlation (1)

Μετοχή ETE (2/1/1986 - 5/1/2007)

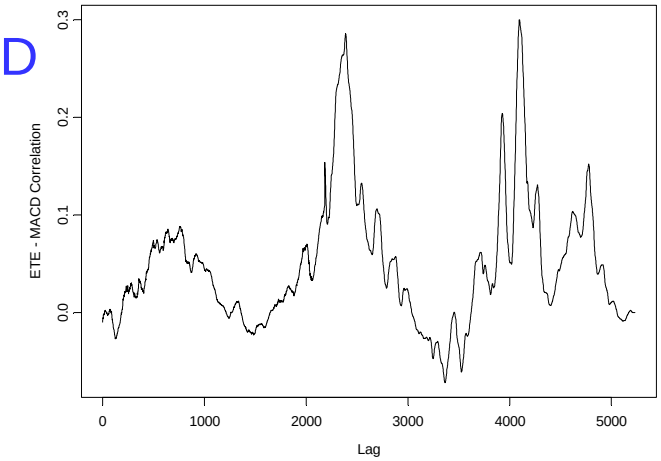
Stochastic Oscillator →



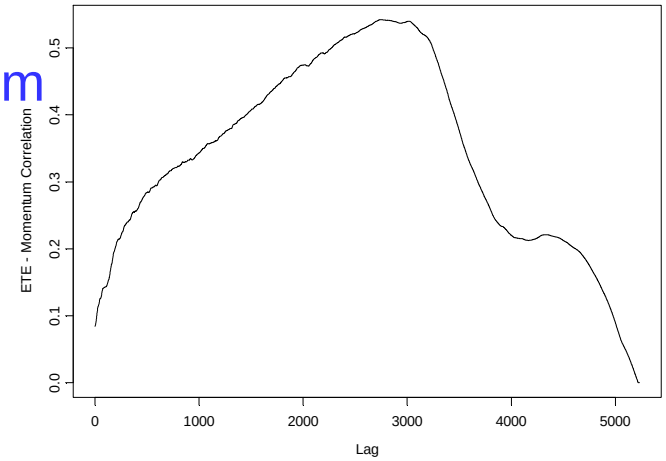
RSI →



MACD →



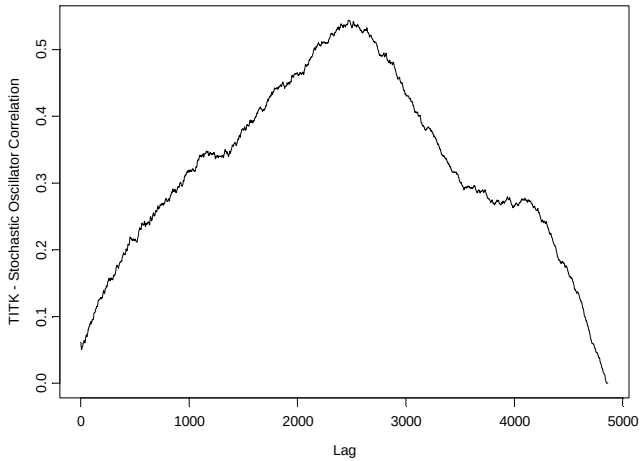
Momentum →



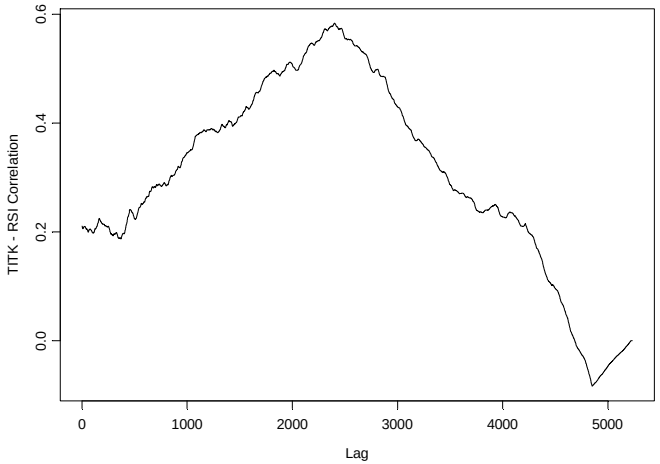
Αποτελέσματα Cross Correlation (2)

Μετοχή TITK (1/7/1987 - 5/1/2007)

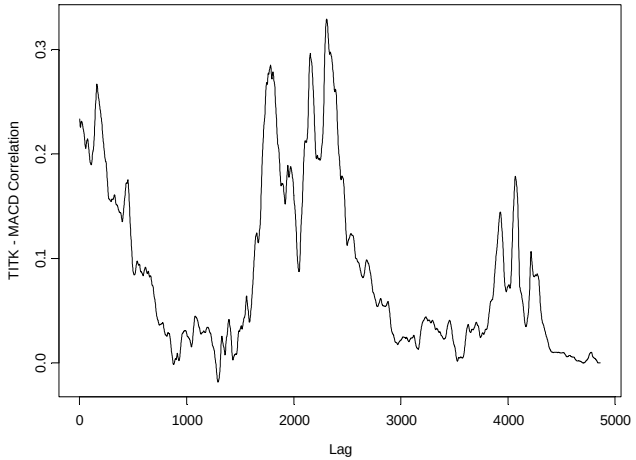
Stochastic Oscillator →



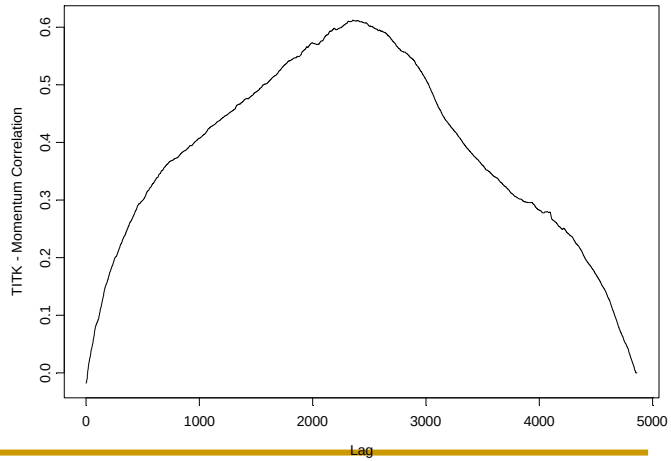
RSI →



MACD →



Momentum →



Συμπεράσματα από το Cross Correlation.

- Η διακύμανση της τιμής των μετοχών είναι ένα πολυπαραγοντικό γεγονός, για το λόγο αυτό τα αποτελέσματα της ετεροσυσχέτισης μπορούν να ερμηνευτούν ως εξής:

Correlation	Negative	Positive
Small	−0.29 to −0.10	0.10 to 0.29
Medium	−0.49 to −0.30	0.30 to 0.49
Large	−1.00 to −0.50	0.50 to 1.00

- Τι μας προσφέρει ο συντελεστής ετεροσυσχέτισης;
  - Εκτίμηση – αξιολόγηση των δεικτών για την χρηματιστηριακή αγορά - επενδυτική διαδικασία.
  - Δυνατότητα βραχυπρόθεσμης – μακροπρόθεσμης έμμεσης πρόβλεψης της τάσης της αγοράς - μετοχών.

Unsupervised Learning - Clustering

- Η μη καθοδηγούμενη εκμάθηση δεν οδηγεί σε συγκεκριμένα αποτελέσματα, απλά χρειάζεται κανείς να εξάγει από μόνος του τη γνώση και όχι να την έχει προσδιορίσει αρχικά ως ζητούμενο.

- Εφαρμόσαμε τον αλγόριθμο K-Means σε δύο ομάδες δεδομένων:

  1. Όλες τις μετοχές (20) για το ίδιο χρονικό διάστημα που καθορίζεται από τη μετοχή με τις λιγότερες εγγραφές.

  2. Όλες τις υπόλοιπες μετοχές (19) (πλην της ΤΤ) για μεγαλύτερο χρονικό διάστημα.

Clustering – Μορφή Δεδομένων.

1.Data set(20 μετοχών)

<TICKER>,<DTYYYYMMDD>,<OPEN>,<HIGH>,<LOW>,<CLOSE>,<VOL>
ΤΤ,20060602,12.5000,12.5000,12.5000,12.5000,0
ΤΤ,20060605,13.5000,14.0000,13.5000,13.9800,5680924
ΤΤ,20060606,13.9800,13.9800,13.5400,13.6000,2613329
ΤΤ,20060607,13.5000,13.9200,13.1000,13.7000,1753924
ΤΤ,20060608,13.1000,13.6000,13.1000,13.3800,652366
ΤΤ,20060609,13.6200,13.9000,13.5000,13.7000,611909
ΤΤ,20060613,13.5000,13.5000,13.3000,13.4600,445125
ΤΤ,20060614,13.5000,13.7000,13.4200,13.4800,266244
ΤΤ,20060615,13.5800,13.7600,13.5600,13.7200,421242
ΤΤ,20060616,13.9600,14.3200,13.8600,14.2000,1631619
ΠΕΙΡ,20060718,18.3200,18.5000,18.2000,18.4800,349455
ΠΕΙΡ,20060719,18.4200,18.9200,18.4200,18.9000,585642
ΠΕΙΡ,20060720,19.3000,19.3000,18.9800,19.1400,657298
ΠΕΙΡ,20060721,19.3400,19.6600,19.1600,19.2600,1092921
ΠΕΙΡ,20060724,19.3000,19.6800,19.2800,19.5600,299579
ΜΟΗ,20061229,19.7000,19.7200,19.5200,19.5200,77383
ΜΟΗ,20070102,19.5200,19.7800,19.5000,19.6400,98089
ΜΟΗ,20070103,19.3600,19.9800,19.3600,19.8600,267986
ΜΟΗ,20070104,19.7200,19.8000,19.5000,19.6000,443167
ΜΟΗ,20070105,19.6000,19.6600,19.4800,19.6000,261246
ΚΥΠΡ,20060602,7.1000,7.2600,7.0600,7.1000,977959
ΚΥΠΡ,20060605,7.0800,7.0800,6.9200,6.9400,621143
ΚΥΠΡ,20060606,6.7800,6.8800,6.5000,6.6600,1228292
ΚΥΠΡ,20060607,6.6000,6.7800,6.4000,6.7800,1419348
ΙΝΛΟΤ,20060804,21.7600,22.4600,21.6800,22.3800,163589
ΙΝΛΟΤ,20060807,22.3800,22.3800,21.8200,22.1000,69369
ΙΝΛΟΤ,20060808,22.1000,22.1800,21.8000,21.9400,151418
ΙΝΛΟΤ,20060809,22.0200,22.0800,21.5000,21.8800,40154

2.Data set (19 μετοχών)

<TICKER>,<DTYYYYMMDD>,<OPEN>,<HIGH>,<LOW>,<CLOSE>,<VOL>
ΔΕΗ,20011212,12.3200,12.4600,12.0000,12.0600,20905756
ΔΕΗ,20011213,12.0000,12.0000,11.6600,11.7400,1491111
ΔΕΗ,20011214,11.6000,11.9400,11.6000,11.8000,439880
ΔΕΗ,20011217,11.8000,12.0000,11.7600,11.8800,412250
ΔΕΗ,20011218,11.9400,12.3000,11.8800,12.2000,435262
ΔΕΗ,20011219,12.2000,12.3000,12.1600,12.1800,229340
ΔΕΗ,20011220,12.2000,12.2000,11.9000,12.0400,278890
ΔΕΗ,20011221,11.9400,12.1000,11.9000,12.0200,223050
ΒΙΟΧΚ,20020307,8.2400,8.3600,8.1800,8.2600,47087
ΒΙΟΧΚ,20020308,8.3800,8.3800,8.2000,8.2400,31030
ΒΙΟΧΚ,20020311,8.3000,8.3200,8.2000,8.2000,17270
ΒΙΟΧΚ,20020312,8.1000,8.2000,8.0600,8.0800,30397
ΒΙΟΧΚ,20020313,8.0800,8.1600,7.8800,7.9400,62063
ΒΙΟΧΚ,20020314,8.0200,8.0400,7.9400,8.0200,12475
ΒΙΟΧΚ,20020315,8.0400,8.1800,8.0200,8.1000,22516
ΒΙΟΧΚ,20020319,8.1000,8.1000,7.7200,7.7800,55398
ΒΙΟΧΚ,20020320,7.6000,7.9600,7.6000,7.7800,17636
ΒΙΟΧΚ,20020321,7.7800,7.9000,7.6800,7.8800,25440
ΑΛΦΑ,20070102,23.2400,23.4200,23.0400,23.3000,651023
ΑΛΦΑ,20070103,23.4800,24.1400,23.4800,23.9000,1327221
ΑΛΦΑ,20070104,23.9000,24.4000,23.6600,24.2000,1294802
ΑΛΦΑ,20070105,24.2000,24.7400,24.1400,24.6200,1662125
ΑΤΕ,20011212,4.5315,4.5429,4.5315,4.5315,75027
ΑΤΕ,20011213,4.5656,4.5656,4.5315,4.5429,79254
ΑΤΕ,20011214,4.5315,4.5542,4.5315,4.5429,24771
ΑΤΕ,20011217,4.5315,4.5429,4.5315,4.5315,18792
ΑΤΕ,20011218,4.5315,4.5429,4.5315,4.5315,32634
ΑΤΕ,20011219,4.5315,4.5429,4.5315,4.5315,46038

Clustering

- Προκειμένου να διαπιστώσουμε σχέσεις μεταξύ των μετοχών, εφαρμόσαμε τον KMeans αλγόριθμο επαναληπτικά, μέχρι να λάβουμε το καλύτερο αποτέλεσμα σχετικά με τον αριθμό των σχηματιζόμενων Clusters.

- Επιπλέον η πληροφορία που θέλαμε να εξάγουμε μεταξύ των μετοχών αφορούσε τα ακόλουθα χαρακτηριστικά των μετοχών:

  - Τιμή κλεισίματος - Χαμηλότερη τιμή (Close - Low)
  - Τιμή κλεισίματος - Υψηλότερη τιμή (Close - High)
  - Τιμή κλεισίματος - Όγκος συναλλαγών (Close - Volume)

Clustering – Ενδεικτικά Αποτελέσματα (1)

- Για το πρώτο data set (20 μετοχών) – close - Volume

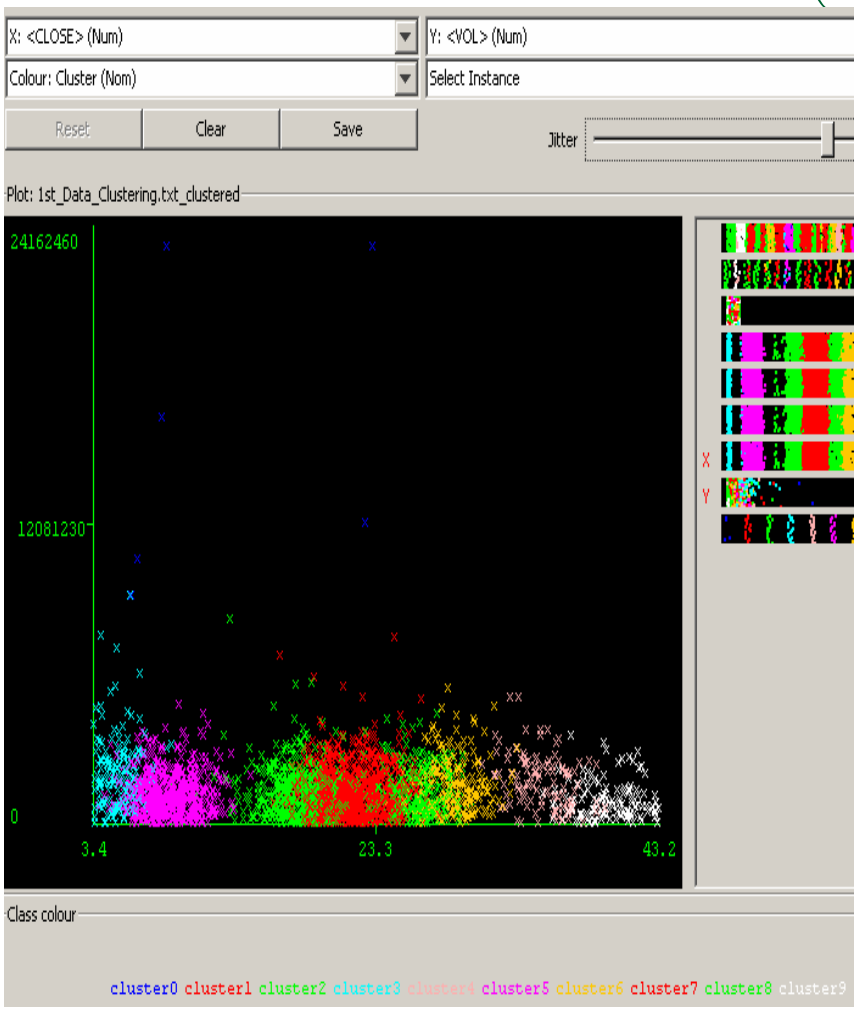
Within cluster sum of squared errors: 4.212002905474135

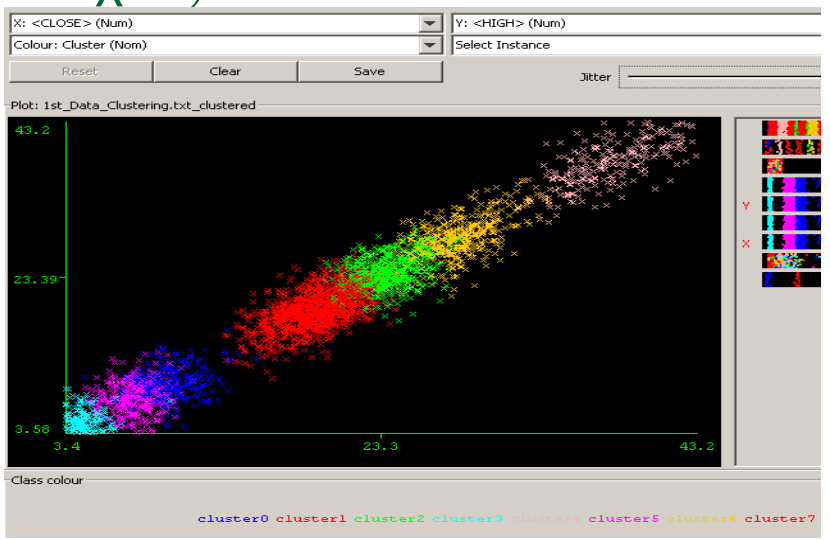
```
Cluster 0
        Mean/Mode:    12.8633 15273820.8333
        Std Devs:     7.6945  6532603.3128
Cluster 1
        Mean/Mode:    20.3792  471389.5895
        Std Devs:     0.7263   520483.5315
Cluster 2
        Mean/Mode:    26.035   511434.3591
        Std Devs:     0.8055   529390.3109
Cluster 3
        Mean/Mode:    4.0657  1373293.7405
        Std Devs:     0.6396  1069166.8506
Cluster 4
        Mean/Mode:    34.8865  1120937.9474
        Std Devs:     1.4582   749560.1374
Cluster 5
        Mean/Mode:    8.6974   435054.8603
        Std Devs:     1.2055   509315.4889
Cluster 6
        Mean/Mode:    28.6942  770674.4523
        Std Devs:     1.0232   654380.235
Cluster 7
        Mean/Mode:    23.0986  427017.3304
        Std Devs:     0.7452   561851.859
Cluster 8
        Mean/Mode:    17.7385  444704.5057
        Std Devs:     1.0751   511873.9783
Cluster 9
        Mean/Mode:    39.5173  109815.3561
        Std Devs:     1.6623   69565.6704

Clustered Instances

0      6 (  0%)
1    531 ( 18%)
2    259 (  9%)
3    158 (  5%)
4    114 (  4%)
```

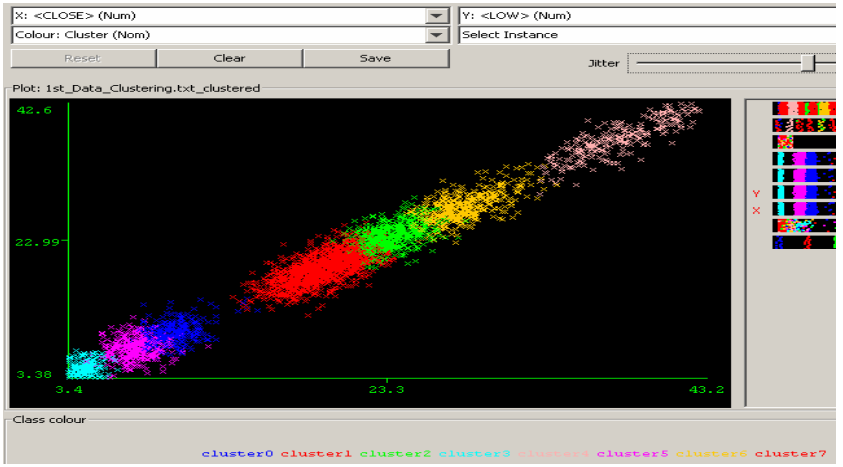
```
Classes to Clusters:

  0   1   2   3   4   5   6   7   8   9 <-- assigned to cluster
  0   0   0   0   0   1   0   0 150   0 | ΤΤ
  0   0   0   0  19   0   0   0   0 132 | ΤΙΤΚ
  1  69   4   0   0   0   0  46  31   0 | ΠΕΙΡ
  0  50   0   0   0   0   0  24  77   0 | ΟΤΕ
  0   0  43   0   0   0 106   2   0   0 | ΟΠΑΠ
  0 135   0   0   0   0   0   7   9   0 | ΜΟΗ
  4   0   0   7   0 140   0   0   0   0 | ΚΥΠΡ
  0  30   0   0   0   0   0  17 104   0 | ΚΟΣΜΟ
  0  67  18   0   0   0   2  58   6   0 | ΙΝΛΟΤ
  0  21  12   0   0   0  22  43  53   0 | ΦΟΛΙ
  0  17  51   0   0   0  15  68   0   0 | ΕΥΡΩΒ
  0   0   2   0  95   0  54   0   0   0 | ΕΤΕ
  0   1  33   0   0   0   7 110   0   0 | ΕΜΠ
  0   0   0   0   0 151   0   0   0   0 | ΕΛΤΕΧ
  0   0   0   0   0 151   0   0   0   0 | ΕΛΠΕ
  0   0  89   0   0   0  35  27   0   0 | ΕΕΕΚ
  0   0   0 151   0   0   0   0   0   0 | ΑΤΕ
  1  69   7   0   0   0   0  55  19   0 | ΑΛΦΑ
  0   0   0   0   0 151   0   0   0   0 | ΒΙΟΧΚ
  0  72   0   0   0   0   0   0  79   0 | ΔΕΗ

Cluster 0 <-- ΚΥΠΡ
Cluster 1 <-- ΜΟΗ
Cluster 2 <-- ΕΕΕΚ
Cluster 3 <-- ΑΤΕ
Cluster 4 <-- ΕΤΕ
Cluster 5 <-- ΕΛΤΕΧ
Cluster 6 <-- ΟΠΑΠ
Cluster 7 <-- ΕΜΠ
Cluster 8 <-- ΤΤ
Cluster 9 <-- ΤΙΤΚ

Incorrectly clustered instances : 1897.0      62.8146 %
```

Clustering – Ενδεικτικά Αποτελέσματα (2)

Data set (20 μετοχών)







Clustering - Συμπεράσματα

Στα μοντέλα που κατασκευάσαμε παρατηρούμε ότι:

- μεγάλος αριθμός στιγμιοτύπων δεν μπορούν να συσταδοποιηθούν σωστά πχ: Close - Volume

  Incorrectly clustered instances : 1897.0 62.8146 %

- Δεν μπορούμε να εξάγουμε σαφή πληροφορία για τη μεταξύ των μετοχών ενδεχόμενη σχέση.

Τα ανωτέρω συμβαίνουν διότι τα δεδομένα μας είναι ακουλουθιακά – χρονικά. Οπότε ο κλασσικός τρόπος συσταδοποίησης δεν ενδείκνυται.

Supervised Learning – Classification/Prediction

- Η διαδικασία της καθοδηγούμενης εκμάθησης είναι πιο σαφής εξ αρχής, διότι προϋποθέτει να έχουμε προσδιορίσει το τι θέλουμε να προβλέψουμε.
- Η πρόβλεψη (Prediction) αφορά μελλοντικές και όχι τρέχουσες τιμές όπως συμβαίνει στην κατηγοριοποίηση.
- Έτσι λοιπόν η προσπάθεια μας κατευθύνεται στην κατασκευή ενός μοντέλου το οποίο θα προβλέπει, σε ημερήσια βάση, το αν θα πρέπει να επενδύσουμε ή όχι στο χρηματιστήριο για συγκεκριμένη μετοχή.
- Ιδιαίτερη σημασία για το αποτέλεσμα έχουν:
  - *Το είδος – μορφή των δεδομένων που θα χρησιμοποιηθούν.*
  - *Η ποιότητα των δεδομένων.*
  - *Ο αλγόριθμος της καθοδηγούμενης εκμάθησης.*
  - *Η ερμηνεία των στατιστικών αποτελεσμάτων.*

Prediction – Επεξεργασία Δεδομένων(1)

- Ως δεδομένα χρησιμοποιήσαμε τις υπολογισμένες τιμές των δεικτών, σε ημερήσια βάση, που αναλύθηκαν παραπάνω, καθώς και τις πραγματικές τιμές κλεισίματος για κάθε μετοχή ξεχωριστά.

(Μετοχή ΕΤΕ) →

Date	MACD	Stochastic Oscillator	Relative Strength Index	Momentum	Close	Invest?
11/4/1997	0.2295	38.0581	64.8692	105.5125	8.5137	no
14/4/1997	0.2066	32.0777	56.9036	103.5556	8.3558	no
15/4/1997	0.1857	35.3858	56.5626	100.1704	8.3486	no
16/4/1997	0.1784	53.6239	61.7326	100.6799	8.4993	yes
17/4/1997	0.1729	69.3844	62.6445	102.1489	8.528	yes
18/4/1997	0.1693	86.7334	63.8032	101.5294	8.5638	yes
21/4/1997	0.1667	91.5665	64.7505	101.9579	8.5926	yes
22/4/1997	0.1639	91.4862	65.2369	101.1803	8.6069	yes
23/4/1997	0.1614	93.9445	65.9967	99.9166	8.6284	yes
24/4/1997	0.1631	96.2531	68.4747	101.1686	8.7002	yes
29/4/1997	0.1668	94.9356	70.3329	101.9191	8.7575	yes
30/4/1997	0.1679	90.6954	70.3329	102.6057	8.7575	yes
2/5/1997	0.1749	91.8339	73.7078	104.1275	8.8651	yes
5/5/1997	0.1833	88.965	75.5066	106.8671	8.9296	yes
6/5/1997	0.1905	91.1332	76.4715	107.3893	8.9655	yes
7/5/1997	0.1962	92.101	77.2458	105.824	8.9943	yes
8/5/1997	0.2006	92.9237	78.0195	105.8032	9.0229	yes
9/5/1997	0.2304	95.0749	85.2147	109.7994	9.403	yes
12/5/1997	0.2641	92.3216	87.2498	111.4343	9.5751	yes
13/5/1997	0.2860	89.0285	85.6708	111.0005	9.5537	no
14/5/1997	0.3107	88.1681	87.3261	112.3858	9.6971	yes
15/5/1997	0.3332	90.4637	88.207	112.448	9.7832	yes
16/5/1997	0.3481	94.7374	88.3518	111.8755	9.7975	yes
19/5/1997	0.3732	94.6519	90.391	114.4962	10.027	yes
20/5/1997	0.3742	68.7472	77.9876	110.9226	9.8334	no
21/5/1997	0.3869	62.8339	81.092	112.5302	10.049	yes
22/5/1997	0.4296	68.5298	85.989	117.6008	10.544	yes
23/5/1997	0.4746	89.4485	87.5042	119.6169	10.759	yes
27/5/1997	0.5072	94.6639	87.7417	119.6345	10.795	yes
28/5/1997	0.5277	91.0361	87.7913	114.874	10.802	yes
29/5/1997	0.5266	81.4385	80.3735	111.2375	10.651	no
30/5/1997	0.5053	62.5044	71.9477	109.4592	10.457	no
2/6/1997	0.4974	43.4879	74.7934	109.838	10.651	yes

Prediction – Επεξεργασία Δεδομένων(2)

- Επιπλέον παρήγαμε και ένα «attribute» (categorical) για τις κλάσεις ETE,TITK όπως φαίνεται στη διπλανή εικονά –Invest? (yes / no).

(Μετοχή TITK)→

Date	Relative Strength Index	Stochastic Oscillator	Momentum	MACD	Close	Invest?
16/9/1987	56.1973	55.5556	93.5484	0.0055	0.29	yes
17/9/1987	49.9965	64.7059	94.8276	0.0043	0.275	no
18/9/1987	48.0916	50.0001	96.4286	0.003	0.27	no
21/9/1987	50.1375	28.5715	94.8276	0.0023	0.275	yes
22/9/1987	57.3746	50	98.3333	0.0033	0.295	yes
23/9/1987	58.9776	70.5882	101.6949	0.0044	0.3	yes
24/9/1987	63.4218	86.9565	106.7797	0.0063	0.315	yes
25/9/1987	68.3459	87.5	113.5593	0.0092	0.335	yes
28/9/1987	72.3538	86.3636	129.0909	0.0129	0.355	yes
30/9/1987	56.85	66	114.5454	0.0127	0.315	no
1/10/1987	52.3223	39.2157	111.1111	0.0112	0.3	no
2/10/1987	57.2152	17.3913	112.2807	0.0115	0.32	yes
5/10/1987	55.6768	15.5555	108.6207	0.0111	0.315	no
6/10/1987	56.9241	26.8292	116.3636	0.0111	0.32	yes
7/10/1987	61.5812	44.1176	125.9259	0.0125	0.34	yes
8/10/1987	59.8394	63.3333	121.8182	0.0131	0.335	no
9/10/1987	67.7082	85.7143	127.1187	0.0164	0.375	yes
12/10/1987	79.5656	91.8033	161.6667	0.027	0.485	yes
13/10/1987	85.1652	96.2617	187.3016	0.043	0.59	yes
14/10/1987	81.72	93.7931	171.6418	0.0539	0.575	no
15/10/1987	72.2742	85.4545	149.2958	0.0585	0.53	no
16/10/1987	61.9863	68.8312	149.2063	0.057	0.47	no
19/10/1987	57.5734	44.8529	146.6667	0.0529	0.44	no
26/10/1987	43.2159	17.7305	96.875	0.0393	0.31	no
27/10/1987	41.5012	3.9474	92.0635	0.0267	0.29	no
29/10/1987	41.0626	1.8182	89.0625	0.0161	0.285	no
30/10/1987	44.8304	4.1958	92.6471	0.0099	0.315	yes
2/11/1987	53.388	23.9316	117.9104	0.0109	0.395	yes
3/11/1987	53.388	55.5556	105.3333	0.0115	0.395	no
4/11/1987	47.5118	73.3333	70.1031	0.0077	0.34	no
5/11/1987	50.1952	74.2424	61.8644	0.0066	0.365	yes
6/11/1987	47.0844	51.6667	58.2609	0.0033	0.335	no

Prediction – Επεξεργασία Δεδομένων (3)

- Το πεδίο «Invest?» παράγεται με τον αλγόριθμο:

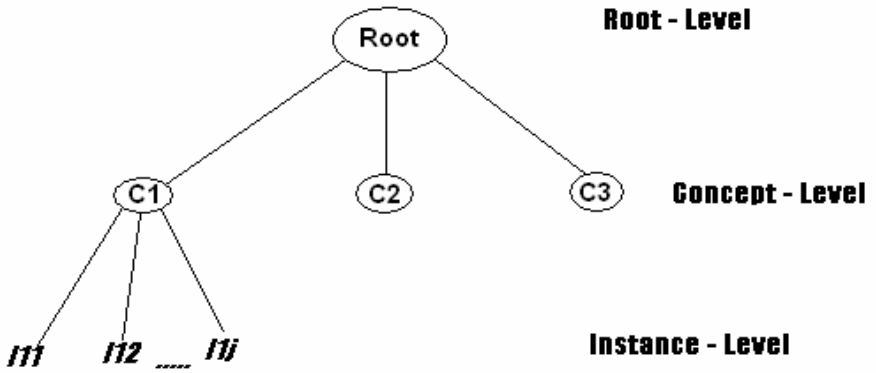
```
for(i=1; i<N; i++){
        if(x[i]>x[i-1])
                y[i].assign("yes");
        else if(x[i]<x[i-1])
                y[i].assign("no");
        else if(x[i]=x[i-1]){
                if(x[i]<x[i+1])
                        y[i].assign("yes");
                else
                        y[i].assign("no");
}
```

- Σύμφωνα με αυτόν, ελέγχεται η τιμή κλεισίματος της i-οστής μέρας με την (i-1) μέρα για να διαπιστωθεί εάν θα είναι μεγαλύτερη ώστε να επενδύσει κανείς (yes), διαφορετικά να μην επενδύσει (no).

Classification/Prediction – Αλγόριθμος

- Για την κατασκευή του μοντέλου πρόβλεψης χρησιμοποιήσαμε το εμπορικό πακέτο iDA (iData Analyzer), το οποίο ενσωματώνει ως συνιστώσα (component) τον αλγόριθμο ESX.
- Ο ESX επενεργεί σε ένα ιεραρχικό μοντέλο τριών επιπέδων (three – level concept hiercarchy).

  - ✓ Root – level
  - ✓ Concept – level
  - ✓ Instance - level



Prediction – Αποτελέσματα ΕΤΕ (1)

- Για τη μετοχή ΕΤΕ: Σύνολο στιγμιοτύπων 5319
  - #Training set = 3221 (60% των συνολικών στιγμιοτύπων)
  - #Test set1=2088 (40% των συνολικών στιγμιοτύπων)
  - #Test set2=10 στιγμιότυπα (Validation test set)

- Οπότε προκύπτει το μοντέλο με τον «confusion matrix»
- Το μοντέλο παρουσιάζει
  accuracy = 61%
- 36.9%<Error rate<41.1%
  για διάστημα εμπιστοσύνης 95%
  του test set.

Confusion Matrix		
	Computed Class	
	yes	no
yes	560	454
no	355	713
Percent Correct:		61.0%
Error: Upper Bound		41.1%
Error: Lower Bound		36.9%

Prediction – Αποτελέσματα ETE (2)

- Επιπλέον με τον RuleMaker (component του iDA) μπορούμε να παράγουμε κανόνες αφού πρώτα ορίσουμε τις παρακάτω παραμέτρους:
  - Minimum correctness = 75% (παράγονται οι κανόνες με error rate=<25%)
  - Minimum rule coverage = 50% (καλύπτει το 60% και πάνω των στιγμιοτύπων)

Rules for Class yes
288 instances
****

47.49 <= Stochastic Oscillator <= 85.66
:rule accuracy 75.09%
:rule coverage 75.35%

48.76 <= Relative Strength Index <= 63.31
:rule accuracy 75.00%
:rule coverage 86.46%

47.49 <= Stochastic Oscillator <= 85.66
and 48.76 <= Relative Strength Index <= 63.31
:rule accuracy 90.91%
:rule coverage 65.97%

**Total Percent Coverage = 95.83%

Rules for Class no
339 instances
****

15.03 <= Stochastic Oscillator <= 57.47
:rule accuracy 75.06%
:rule coverage 94.99%

36.57 <= Relative Strength Index <= 51.45
:rule accuracy 75.18%
:rule coverage 93.81%

36.57 <= Relative Strength Index <= 51.45
and 15.03 <= Stochastic Oscillator <= 57.47
:rule accuracy 88.01%
:rule coverage 88.79%

**Total Percent Coverage = 100.00%

Prediction – Αποτελέσματα ΕΤΕ (3)

- Εφόσον κατασκευάσαμε το μοντέλο μας μπορούμε να το ελέγξουμε με τα στιγμιότυπα που χαρακτηρίσαμε ως Test set #2 (Validation test set).

- Τα αποτελέσματα έχουν ως εξής:

1	Date	MACD	Stochastic Oscillator	RSI	Momentum	Close	Invest?	Computed Invest
2	20/12/2006	-0.2107	48.5596	47.1469	101.3994	34.78	yes	no
3	21/12/2006	-0.2138	19.3548	45.5476	102.6097	34.6	no	no
4	22/12/2006	-0.2077	14.9757	46.4176	101.7009	34.68	yes	no
5	27/12/2006	-0.1841	20.3791	48.8383	102.0468	34.9	yes	no
6	28/12/2006	-0.1500	39.0375	50.7969	104.3427	35.08	yes	no
7	29/12/2006	-0.1350	53.5033	48.7856	102.346	34.9	no	no
8	2/1/2007	-0.0452	80.3572	58.7523	105.1522	35.92	yes	yes
9	3/1/2007	0.0545	85.446	61.7396	104.9133	36.3	yes	yes
10	4/1/2007	0.1637	94.2966	64.7761	105.0343	36.72	yes	yes
11	5/1/2007	0.2642	95.8042	66.2854	104.5866	36.94	yes	yes

Prediction – Αποτελέσματα TITK (1)

- Για τη μετοχή TITK: Σύνολο στιγμιοτύπων 4852
  - #Training set = 2851 (60% των συνολικών στιγμιοτύπων)
  - #Test set1=1901 (40% των συνολικών στιγμιοτύπων)
  - #Test set2=100 στιγμιότυπα (Validation test set)

- Οπότε προκύπτει το μοντέλο με τον «confusion matrix»:
- Το μοντέλο παρουσιάζει
  accuracy = 58%
- 39.7%<Error rate<44.3%
  για διάστημα εμπιστοσύνης 95%
  του test set.

Confusion Matrix		
	Computed Class	
	no	yes
no	622	329
yes	466	483
Percent Correct:		58.0%
Error: Upper Bound		44.3%
Error: Lower Bound		39.7%

Prediction – Αποτελέσματα TITK (2)

- Κατασκευή κανόνων με τις προκαθορισμένες ρυθμίσεις (Minimum correctness = 75%, Minimum rule coverage = 50%):

```
Rules for Class yes
273 instances
*******************************

49.94 <= Relative Strength Index <= 69.01
 and 49.19 <= Stochastic Oscillator <= 93.18
  :rule accuracy 83.63%
  :rule coverage 69.23%


  **Total Percent Coverage = 69.23%
```

```
Rules for Class no
296 instances
*******************************

38.79 <= Relative Strength Index <= 54.41
 and 9.76 <= Stochastic Oscillator <= 61.61
  :rule accuracy 82.19%
  :rule coverage 81.08%

**Total Percent Coverage = 81.08%
```

Prediction – Αποτελέσματα TITK (3)

- Ελέγχουμε το μοντέλο με το Test Set #2 (Validation test set) και λαμβάνουμε τα ακόλουθα αποτελέσματα:

Από τα 100 instances
τα 58 έχουν προβλεφθεί
σωστά.

Date	Relative St	Stochastic	Momentun	MACD	Close	Invest?	computed Invest?
14/8/2006	51	22.8349	99.3158	0.1256	37.74	no	no
16/8/2006	50	15.6524	99.3133	0.1029	37.6	no	no
17/8/2006	54	25.4548	99.6335	0.1183	38.06	yes	no
18/8/2006	59	63.3589	101.5175	0.1846	38.8	yes	yes
21/8/2006	62.2665	91.2088	104.0827	0.2692	39.26	yes	yes
22/8/2006	61.6159	91.3934	104.1999	0.3281	39.2	no	yes
23/8/2006	61.3857	85.9259	103.1053	0.3693	39.18	no	yes
24/8/2006	61.694	79.6813	102.4021	0.4005	39.22	yes	yes
25/8/2006	50.5996	54.2987	100.7916	0.344	38.2	no	no
28/8/2006	47.1946	26.5218	99.2651	0.2672	37.82	no	no
29/8/2006	49.6892	13.0953	100.9544	0.2234	38.08	yes	no
30/8/2006	47.9986	17.4075	100	0.1729	37.9	no	no
31/8/2006	56.0503	37.5001	102.8087	0.1987	38.8	yes	no
1/9/2006	59.6397	60.9375	104.4681	0.2528	39.28	yes	yes
4/9/2006	55.7574	74.0602	102.207	0.264	38.9	no	yes
5/9/2006	51.2216	59.854	99.0206	0.2338	38.42	no	no
6/9/2006	51.0354	33.9844	97.8095	0.206	38.4	no	no
7/9/2006	50.0551	15.6779	97.7041	0.1742	38.3	no	no
8/9/2006	49.0407	9.0909	97.4987	0.1398	38.2	no	no
11/9/2006	44.0424	7.109	96.0734	0.0722	37.68	no	no
12/9/2006	46.0907	10.2151	99.1099	0.0317	37.86	yes	no
13/9/2006	51.6021	29.5302	101.4807	0.0386	38.38	yes	no
14/9/2006	51.3845	53.4247	100.7353	0.042	38.36	no	no
15/9/2006	54.0945	81.119	101.8997	0.0637	38.62	yes	yes
18/9/2006	52.6361	79.5775	99.2268	0.0711	38.5	no	yes
19/9/2006	52.3826	77.8571	97.9633	0.0746	38.48	no	yes
20/9/2006	51.3181	63.3094	98.7147	0.0705	38.4	no	no

Prediction – Συμπεράσματα

Η ποιότητα – ακρίβεια των αποτελεσμάτων εξαρτάται από τους εξής παράγοντες:

- Ποιότητα – Θόρυβο των δεδομένων.
- Από το είδος των δεδομένων (ημερήσια – εβδομαδιαία κλπ) που χρησιμοποιούνται ως είσοδος.
- Από το πόσο εμπλουτισμένα με πληροφορία είναι τα δεδομένα που χρησιμοποιούμε.
- Από το είδος της εξόδου – πρόβλεψης (Invest yes/no), στην περίπτωση που εξετάσαμε (Categorical output) μειώνει την ακρίβεια του μοντέλου και αυξάνει το σφάλμα. Επιθυμητό είναι η έξοδος να προσδιορίζεται από περισσότερες των δύο κλάσεων.

Ερωτήσεις!