Clustering of time series using a similarity between segments and bands determined by patterns of technical analysis

R. Basagoiti & E. Juaristi
Faculty of Business Studies of Mondragon University, Spain

Abstract

The representation based on segments, continuous or discontinuous, has been used for dimensionality reduction of temporal data. This reduction is essential for the posterior process of data mining. In this work, patterns of technical analysis, like those defined in Lo *et al.* (2002), are used to extract the extremes to get two different representations. The first one is based on segments drawn between extremes of the patterns, STA. The second one is based on the best harmonic of the Fourier decomposition between the same extremes, FTA. Due to the subjective nature of technical analysis some parameters are considered in the process of extreme extraction and pattern selection. Once a representation is adopted, similarities between segments defined in Keogh *et al.* (2002) and between bands suggested by us are used for clustering and the results compared with those obtained with the Euclidean and the lower bound of the dynamic time warping distance defined in Keogh.
Keywords: data approximation, dimensionality reduction, time series clustering, pattern extraction.

1 Introduction

With the rapid increase of stored data, the interest in the discovery of hidden information has exploded in the last decade. The focus has mainly been on classification, clustering, query by content and relationship finding. Treat data with temporal dependencies is an important problem. A time series data is a sequence of real values, each of which represents a value measured at a point in time. We can find examples of time series data in diverse sources and applications, such as stock prices and currency exchange data. There has been

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060081

much recent work on adapting data mining algorithms to time series data. Algorithms that work with time series need to compute the similarity between them. Looking specifically to stock time series, technical analysis is widely used for prediction of the market. In Lo *et al.* [1] and in Dong and Zhou [4], the presence and the information generated by these patterns was studied. The subsequence-matching problem was considered using patterns of technical analysis in Rafiei and Mendelzon [5] and Agrawal *et al.* [6]. In Gavrilov *et al.* [7] the authors studied many others distances to be used for mining the stock market. Distances between models generated using classical time series analysis are also used. See Kalpakis *et al.* [8] and Bagnall and Janacek [9], for example.

The representations adopted in this work are a piecewise linear approximation in STA, and with the same segmentation, the best harmonic, for each subsequence in FTA, Alvarez [10], Bloomfield [11]. The piecewise linear representation is a good technique for reducing the complexity of the raw data. Multiple segmentations algorithms have been developed. See Keogh *et al.* [12], for example. For the STA and FTA representations, the segmentation is obtained via the extremes of the time series that constitute a pattern previously determined, like *head-shoulder-head*. This way, the segmentation used is dependent of the domain of application. An evolutionary approach has been used in Chung *et al.* [13] for domain dependent pattern extraction.

2 Representations of time series

2.1 Detection of patterns

If we need to retain what is happening in a graphic, a relatively good strategy is to remember the position of the turning points and reconnect them. It is common to think that two graphics are similar if their turning points are similar. If we assume this, we can think that two series of stocks are similar if they contain the same pattern of technical analysis. In our case, patterns of technical analysis are defined as in Lo *et al.* [1]; a kernel smoother is used to smooth the time series and its first derivative to extract the extremes. When the extremes of the series that constitute some pattern have been selected a representation is adopted. Two parameters, *h* (used in the kernel smoother) and *lgp* (percentage of the time series' length used by the pattern) are definitive at the time to decide which is or not an extreme and in the selection of a certain pattern.

2.2 The representations adopted

A piecewise linear representation with k continuous linear segments is used in STA to represent the time series. It is enough to preserve the length and the left value for each segment (the right value will be seen in the next segment). In the case of FTA, the best harmonic is traced between extremes. Both are displayed in Figure 1.

Each pattern is defined using 5 extremes, E_1 ,E_2 ,E_3 ,E_4 ,E_5 and eight different patterns are considered (SHS, ISHS, BTOP, BBOT, TRIT, TRIB, RTOP, RBOT). L_1, L_2 , L_3 , L_4 , L_5 are the time periods between the extremes.

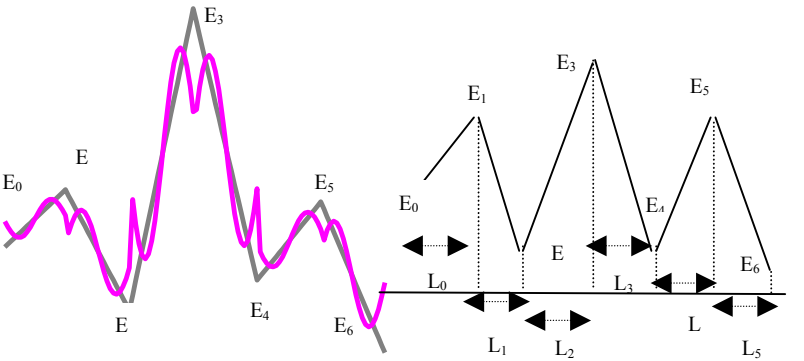

Figure 1: FTA and STA representations adopted using the extremes extracted from a series.

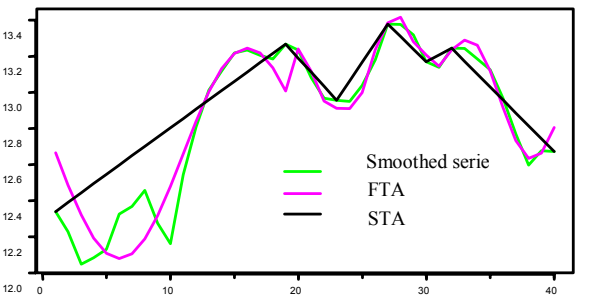

Figure 2: An example of a smoothed time series with STA and FTA representations.

The investigation is reported using Telefonica's daily closing prices from January 1990th to July 2004th. Twenty seven different databases are generated with different window lengths l = 40 , 90 and 180 (with $l/4$ overlap), three different values of h =0.6, 0.9, 1.4, and three values for *lgp*, 25%, 20% and 15% of the length of l.

Having a time series X with length N and a reconstruction $\hat{X}$, two measures of reconstruction error are calculated: the relative mean error and the square of the mean quadratic error, eqn (1):

$$RME = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - \hat{x}_i|}{|x_i|} \quad , \qquad SQME = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \qquad (1)$$

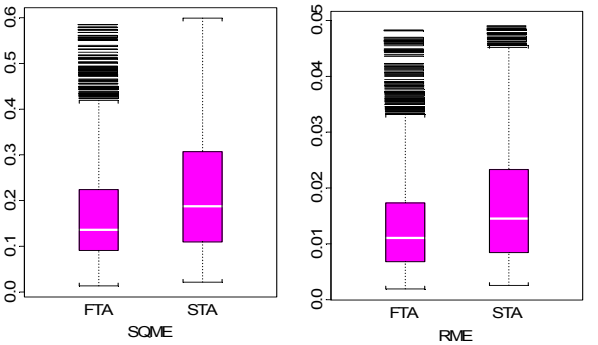

Figure 3: A boxplot of SQME and RME calculated in the representations STA and FTA for all the databases.

2.3 Results of the pattern extraction process

The process of extreme extraction depends on the *l*, *lgp* and *h* considered. The *h* parameter used in the kernel smoother is important because it determines the neighbourhood, too large and the weighted average will be too smooth and hide the nonlinearities of interest, too small and the averages will be too variable and include lots of noise. The length of the patterns found, *lgp,* is also important, the more time a patterns lasts, the more reliable it is for technical analysis. For this, the average of the number of extremes and the standard deviation are evaluated for different *l* and *h* values. As can be seen in Figure 4, the average number of extremes is dependent on the *h* smoothing parameter: the smaller *h* is, the bigger the number of extremes found, but the standard deviation is bigger.

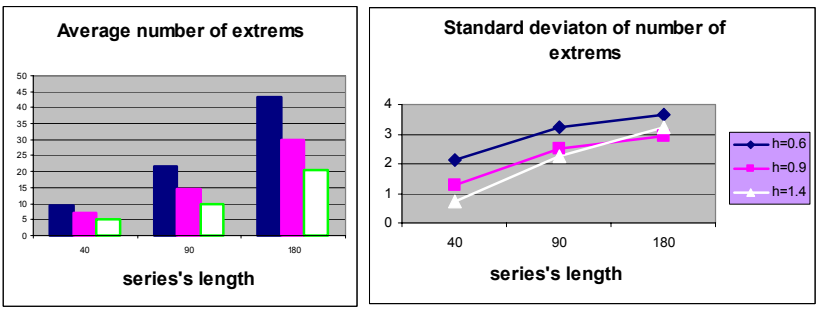

Figure 4: Average number of extremes and standard deviation.

The average number of patterns found for different *l, h* and *lgp* values are also considered. Figure 5 shows that the smoothing parameter *h*, together with *l* and *lgp* are crucial in the pattern extraction process.

The standard deviation and the average length of segments for time series with recognized pattern are shown in Figure 6, plotted for different lengths, smoothing parameters and minimum pattern lengths.

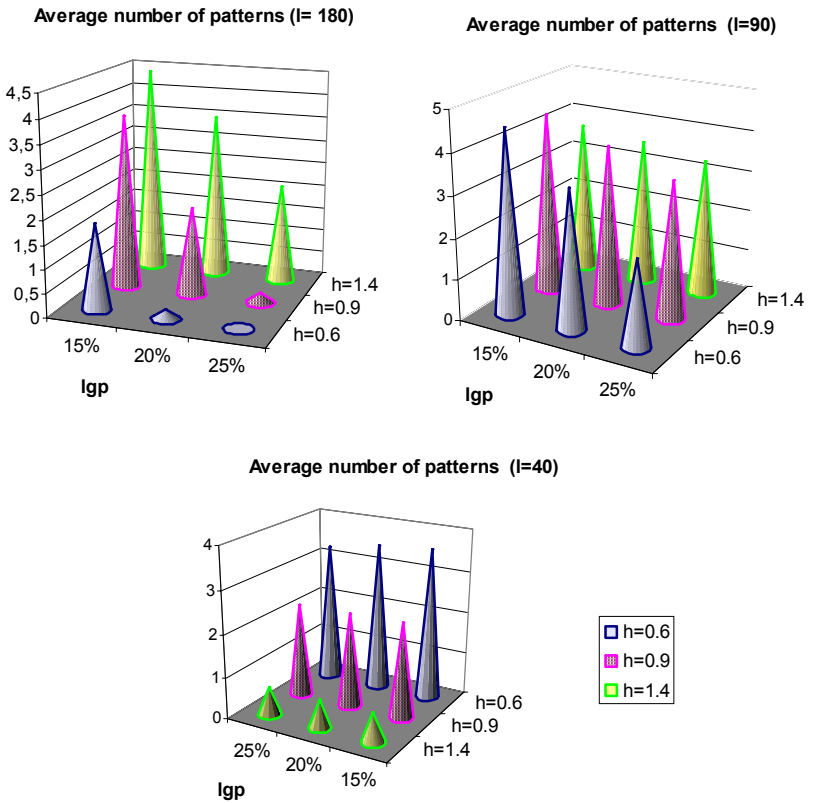

Figure 5: Number of patterns found for different time series length *l*, smoothing parameter *h* and minimum pattern length *lgp*.

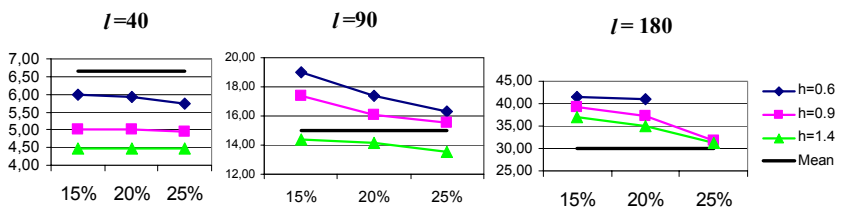

Figure 6: Standard deviation of segment length for different *l, lgp* and *h*

The slope of the movements that occur before and after the patterns are observed trying to evaluate the predictive power of the patterns found. The patterns are classified as Continuity predictors (TRIT, TRIB, RTOP, RBOT), Change Up predictors (ISHS, BBOT) and Change Down predictors (SHS, BTOP). In Figure 7 the percentage of success of each class is shown.

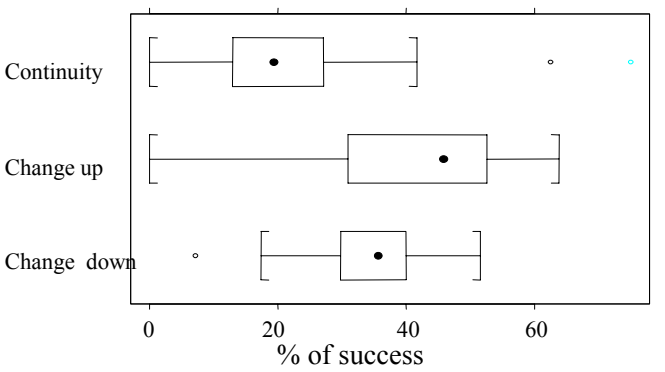

Figure 7: Predictive power of the technical analysis patterns.

3 Similarity measures

Similarity between time series is an area where much research has been reported. The distance used in STA is one defined in Keogh [3]. Given two segments (A, B) of the same length,

$$A = \left(\left(x'_1, y'_1 \right), \left(x'_2, y'_2 \right) \right), \quad B = \left(\left(x_1, y_1 \right), \left(x_2, y_2 \right) \right)$$

the distance between them is defined as, eqn (2):

$$D_S(A,B) = AW * BW * \left| \left(y'_1 - y_1 \right) - \left(y'_2 - y_2 \right) \right|, \quad AW, BW \text{ are the weights.} \quad (2)$$

This distance is defined between segments of the same length so, we interpolate when they are of different lengths. The weights used are according to the quartiles of the segments' original lengths, *ls* , for each time series. (*ls* ≤ Q1 →W=0.05; Q1≤ *ls* <Q2→ W=0.1 ; Q2≤ *ls* <Q3→ W=0.2; Q3≤ *ls* → W=0.65).

With the FTA representation the distance used is defined between bands. A band, for each subsequence, is parallel to the corresponding segment of STA and it is traced over the first maximum and the first minimum value of the reconstruction with the best harmonic. Two bands are similar if they have the same slope and the same width. Db is defined in eqn (3):

$$A = \left(\left(x'_1, y'_1 \right), \left(x'_2, y'_2 \right), \text{width1} \right); B = \left(\left(x_1, y_1 \right), \left(x_2, y_2 \right), \text{width2} \right),$$
$$D_B(A,B) = AW * BW * \left(\left| \left(y'_1 - y_1 \right) - \left(y'_2 - y_2 \right) \right| + \left| \text{width1} - \text{width2} \right| \right) \quad (3)$$

In Figures 8 and 9 two similar time series using Ds and Db distances respectively, are displayed.

3.1 Clustering

Algorithms for clustering similar time series are important in areas as diverse as computational biology, physics, speech recognition and econometrics.

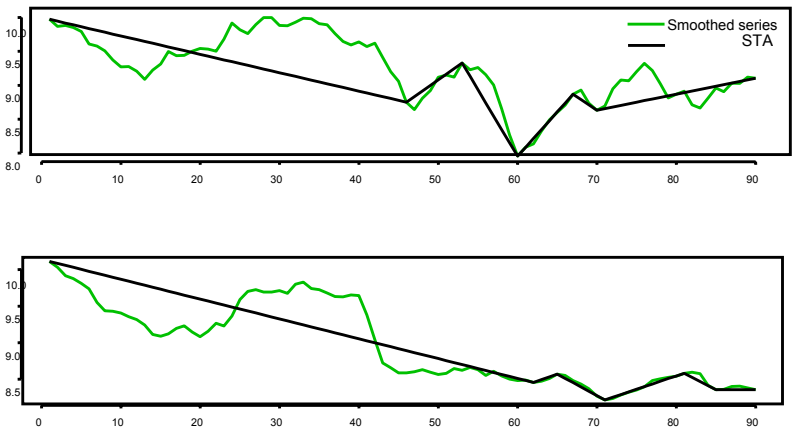

Figure 8: Similar time series using Ds.

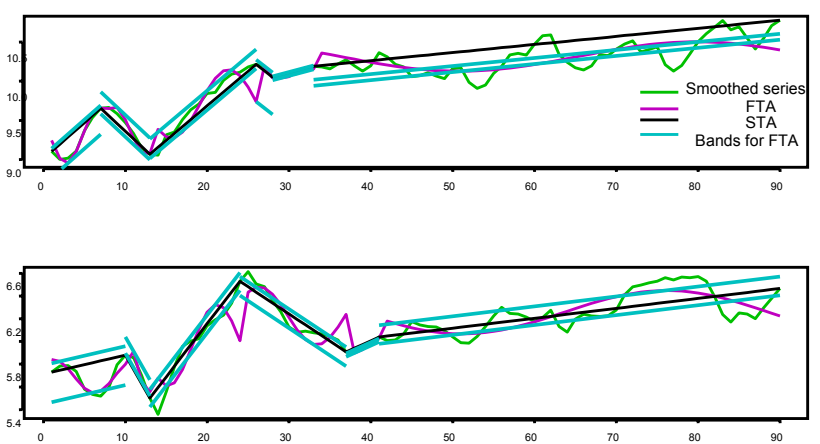

Figure 9: Similar time series using Db.

Clustering of time series can be used as a mechanism with three objectives:

- Find stable relationships between clusters and not between individual time series.
- Find stable relationships only within clusters.
- Identifying structural similarities in the processes that generate time series.

In this section the results are reported using 23 databases, excluding the databases without patterns. For each database the distance matrix for Ds, Db , Euclidean and LB_Keogh is calculated and the *hclust* and *cutree* commands of

S-plus used to get five different clusters (K = 11, 9, 7, 5 ,3). The clusters obtained are compared using the similarity between clusters presented in Gavrilov et al. [7].

Given two clusters, $C=C_1,...C_k$ and $C'=C_1',..C_k'$, the similarity between them is calculated with eqn (4) :

$$Sim(C,C') = \frac{\left(\sum_i \max_j Sim(C_i, C_j')\right)}{k}, \qquad Sim(C_i, C_j') = 2\frac{\left|C_i \cap C_j'\right|}{\left|C_i\right| + \left|C_j'\right|} \quad (4)$$

The results obtained are better for Db than Ds, mainly when compared to LB_Keogh with optimal values for k=9, as can be seen in Figure 10 and Figure 11.

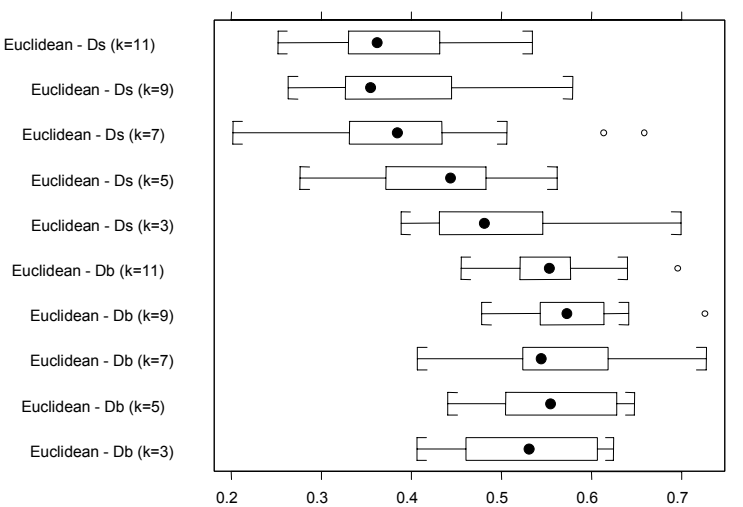

Figure 10: Similarities between clusters in STA(Ds distance), FTA (Db distance) and Euclidean using the hclust command of S-plus and cutting the dendogram with 11, 9, 7, 5 and 3 clusters.

4 Conclusions

A distance between segments doesn't pick up the fluctuation of a time series between extremes. Db is defined to improve Ds: besides the parallelism between segments the width of a great smoothing, using the highest harmonic, is also considered. This way, the results obtained in the reconstruction errors of FTA are slightly smaller and in the similarities between clusters, Db is better than Ds.

The process of patterns extraction is evaluated through three different parameters and the results show that *lgp*, *l* and *h* must be selected together and probably estimated again for a new data mining case.

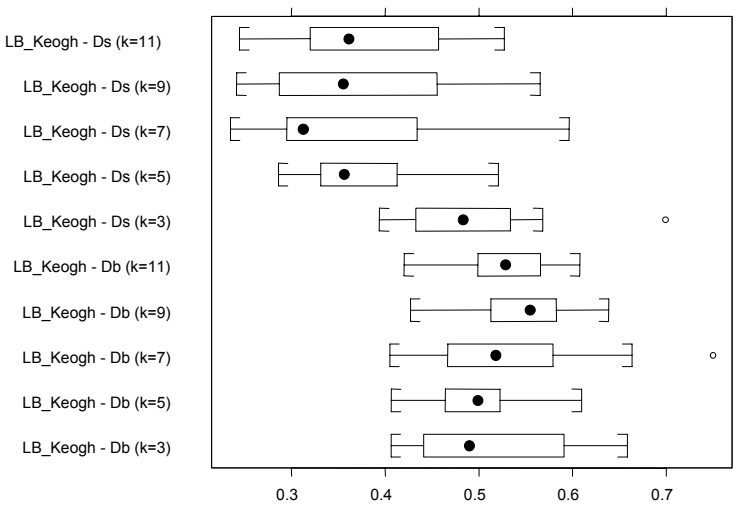

Figure 11: Similarities between clusters in STA(Ds distance), FTA (Db distance) and LB_Keogh using the hclust command of S-plus and cutting the dendogram with 11, 9, 7, 5 and 3 clusters.

Although the work done in this paper focuses in stock data market the proposed pattern extraction mechanism can be applied to any set of patterns given. The two distances are also independent of the domain of application. We have not focused on distances defined only between segments that constitute the patterns or on the predictions power of a pattern not fully materialized. Regarding this second issue, the results obtained in the evaluation of the predictive power shown in Figure 7 can be an advance of the possibilities of the representations adopted.

References

[1] Lo, A.W., Mamaysky, H., Wang, J., Foundations of technical Analysis: Computational Algorithms, Statistical inference, and empirical implementation. The Journal of finance, vol. LV, nº 4, 2000.

[2] Keogh, E., Pazzani, M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Fourth International Conference on Knowledge Discovery and Data Mining KDD, 1998.

[3] Keogh, E. Exact indexing of dynamic time warping. Proceedings of the 28th VLDB Conference. Hon-Kong China, 2002.

[4] Ming Dong, Xu-Shen Zhou: Exploring the Fuzzy Nature of Technical Patterns of U.S. Market. FSKD, pp. 324-328, 2002.

[5] Rafiei, D. Mendelzon, A. Similarity based queries for time-series data 1997, Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, 1997.
[6] Agrawal R., Psaila G., Wimmers, E. Querying shapes of histories. Proceedings de la 21 conferencia BLVD, 1995.
[7] Gavrilov, M., Anguelov, D., Indyk, P., Motwani, R. Mining the stock market: Which measure is Best?. Proceedings of the KDD, USA, Boston, pp. 487-496, 2000.
[8] Kalpakis, K., Gada, D., Vasundhara, P., Distance measures for effective clustering of ARIMA time series, 2001.
[9] Bagnall, A. J., Janacek, G. J. Clustering time series from ARMA Models with clipped data, Technical report CMP-C04-01. School of computing Sciences of East Anglia, 2004.
[10] Alvarez Vazquez.N. Introducción a la econometría. Universidad Nacional de Educación a Distancia, 1999.
[11] Bloomfield, P., Fourier analysis of time series: An introduction. New York: John Wiley, 1976.
[12] Keogh, E., Chu, S., Hart, D., Pazzani, M. An Online Algorithm for Segmenting Time Series**.** Proceedings of the 2001 IEEE International Conference on Data Mining, 2001.
[13] Chung, F., Fu, T., NG. V., Luk, R.W.P., An evolutionary approach to pattern-based time series segmentation, IEEE Transactions on Evolutionary Computation, Volume 8, Issue 5, pp. 471 – 489, 2004.