A news-based financial time series discretization

Danilo Di Stefano[1] and Valentino Pedirođa[2]
[1]Es.Tec.O Research Labs, via Giambellino 7, 35121 Padova, Italy
[2]Dept. of Mechanical Engineering, University of Trieste, via Valerio 10, 34100 Trieste
email: danilo.distefano@esteco.com, pediroda@units.it

Keywords: recursive Self-Organizing Maps, time series analysis, statistical document analysis

***Abstract*—** In this paper a new method for financial time series discretization that allows to take into account qualitative features about financial indicators is proposed. Qualitative features are extracted from financial news web sites and they are inserted into the learning phase of a recursive Self Organizing Map by means of a suitable parameter derived from statistical analysis of document collections. A postprocessing phase based on unsupervised clustering by U-Matrix method leads to the actual discretization of the time series. A real case application to a stock closing price series reveals that the inclusion of qualitative features leads to a more compact discretization of the series. This could be useful if a compact coding of the series is sought, for example in the preprocessing phase of a forecasting methodology.

1 Introduction

Today a growing amount of financial news is largely available on dedicated web sites. Financial analysts attempt to extract from them any useful information for more efficient understanding of the market behavior and also to develop more reliable forecastings. The scientific and technological support to this activity mainly consists in automating extraction of relevant information from document collections (*Information Retrieval* methodologies) and in representing it in a useful way (*Knowledge Representation* techniques).

Recently, methods were proposed that directly integrate the contribution of financial news in a forecasting methodology [1] [2] [3]. They all assume that news contents and financial market tend to move accordingly, in the sense that news could influence market as well as market could influence news, so they're somewhat coupled. They forecast future behavior stating that if a similar news comes up, a corresponding market response should occur. The main problem of this kind of approaches is that they need the correct time window between news release and market reaction. If we translate that in a machine-learning-like terminology, we could say that they use a supervised learning: news is the input and market indicators are the corresponding output shifted by a suitable time window. This kind of approach was applied with the aim to "beat the market", in the sense that if a financial analyst could know in advance the effect of a given news on the behavior of a group of stocks he could act on the market with a deeper knowledge.

The present study develops a novel news integration methodology based on unsupervised learning. News is considered as a background for financial indicators in the sense that its informational content drives the dynamics of memory structures in a recursive self-organized mapping of the financial indicators. Indeed, one of the possible applications of unsupervised methodologies as Self-Organizing Maps (SOM) to a time series is to discretize it (some examples of the application of SOM to time series clustering and prediction are reported in [4] [5] [6]). In a pictorial sense, the discretization of a time series consists of its subdivision into a few horizontal blocks and it has to be distinguished from time series segmentation, aiming to subdivide the time series into vertical blocks (see figure 1).

Segmentation and discretization are used as preprocessing techniques for the statistical analysis of a time series. Segmentation allows the analysis of different blocks of the time series focusing on particular events such as crashes or on possible trends. Discretization acts as a kind of noise reduction method allowing to encode the time series on a restricted base. For example, it is possible to characterize the different blocks by their mean value and re-build a new time series formed by the sequence of the mean values of the succeeding blocks. Another possibility is to assign a letter to each blocks. In this way the time series becomes a sequence of symbols on a short alphabet allowing the application of statistical methodologies for sequence analysis, such as Markov Models.

The main contribution of this work is not intended to be a forecasting technique by itself, but rather a methodology to determine a more reliable discretization of a time series by integrating the informational content of the news. First, the method performs a statistical analysis of a collection of financial news gathered from the web. It extracts a couple of parameters ready to be inserted in the learning phase of the recursive SOM. Then, a postprocessing phase based on clustering of the resulting map leads to determine a characteristic discretization of the financial time series under study.

The work is organized as follows: section 2 shortly describes the statistical analysis of document collections, section 3 describes the method in detail and section 4 presents an application with real world data.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

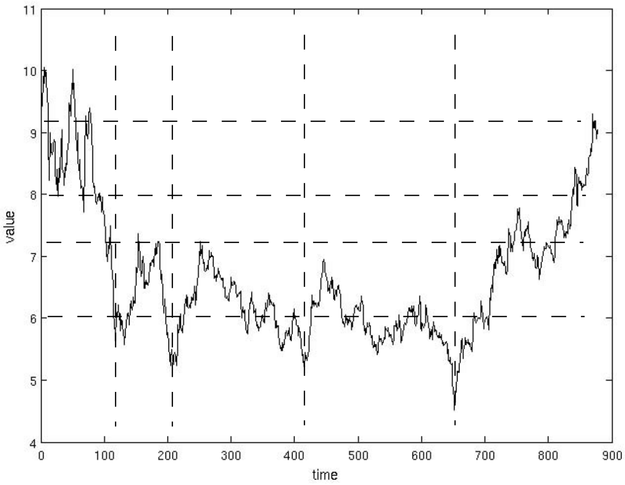

Figure 1: Example of time series segmentation (vertical dotted lines) and discretization (horizontal dotted lines).

2 Statistical analysis of document collections

The first step in a statistical analysis of a collection of textual documents is the choice of the variables describing the collection itself. There are various ways to do it [7], a possible choice being the Salton's vector space model [8]: a single document is treated as a point in an m-dimensional space, where m is a subset of words chosen by a smart word frequency scheme named Term-Frequency Inverse-Document-Frequency (*tf-idf*). *tf* is the term frequency of a word, determining the relevance of the word in the document, and *idf* is proportional to the number of documents containing this word, determining its relevance in the collection. Therefore *tf-idf* scheme selects a word as relevant for the collection if this word is more relevant for a single or few documents and less for all the others.

This way, a vector with following components is assigned to each document:

$$(tf_1 \log(n/df_1)), ..., ..., (tf_m \log(n/df_m)) \qquad (1)$$

where tf_m is the frequency of the m-th word in the document and df_m is the log-transformed normalized frequency of the m-th word in the collection of n documents. Note that now it is possible to evaluate the similarity of two documents by the calculation of their distance in this vector space. A suitable distance criterion is the scalar product between normalized vectors: parallel vectors are the most similar, orthogonal vectors are the less.

Therefore, the entire collection of n documents can be described by the Term Document Matrix (TDM), an $n \times m$ matrix having each row describing a document in the m-dimensional vector space. This is a compact representation of the documents and it allows the numerical treatment of the collection by suitable statistical tools. For example, it is possible to clusterize the TDM grouping similar documents to each other. The most representative words in each of the clusters could be interpreted as a short abstract of the documents that it contains.

Dimensionality reduction techniques like Principal Components Analysis (PCA) could also be applied to the TDM to reduce the complexity of the m-dimensional space to smaller dimension. For example, by the use of PCA it is possible to visualize the entire document collection in the space of the principal eigenvectors (eigenvectors corresponding to the highest eigenvalues of the covariance matrix of the TDM). Besides visualization, PCA offers the ratio between the first and the second principal eigenvalues: this entity can be intended as an indicator of the possibility to establish preferential directions in the vector space representing the variability of the words describing the documents. In this way, if TDM has a dominant principal component we could say that there exists an efficient way to describe the entire document collection by using a single linear combination of words. This in turn could be interpreted as a dominant argument in the news that could somehow drive the informational content of the collection. This is a purely qualitative assumption. This methodology, used within the learning cycle of a recursive SOM, contributes effectively to the discretization of the series.

The text preprocessing phase can finally be resumed as follows. At first, daily-based financial news collections are gathered from the web; then, Term Document Matrix is build for each daily collection and calculation of the ratio between the first two principal eigenvalues is performed for each of them.

3 Methodology

3.1 Recursive SOM

Self Organizing Maps [9] are a well known tool for unsupervised cluster analysis by the use of the U-Matrix codebook representation [10]. The straightforward application of classical SOM to structured data, i.e. data presenting a sequential order (DNA strings, for example) or a temporal dependence (this is the case of stock prices), gives unsatisfactory results because of the lack of suitable structures reflecting the order/temporal dependence of the training data.

To overcome this problem, various modifications to the SOM algorithm have been proposed (for a detailed review see [11]). The most important modifications are: Temporal Kohonen Map [12], which is based on leaky integrators to mine the temporal dependence using a time-window based input; in this way, the map learns to distinguish short sequences of input data; RecSOM [13], based on recurrent connections acting on time-delayed copies of the activity of the entire map; SOMSD [14], which also is based on recurrent connections, but acting on a compressed representation of the activity of the map (only the location of the last winner is stored); MergeSOM [15], another recurrent approach representing the activity of the map not by the location of the last winner neuron but rather by its content.

For our purposes, the RecSOM method was chosen, because it offers the possibility to clearly separate the contribution of the current input from that of past inputs in the learning phase. A brief explanation of RecSOM functioning follows.

In standard SOM each map unit matches his codebook vector $\mathbf{w}_i$ against the current input vector $\mathbf{s}(t)$ in the learning cycle. The updating rule for the codebooks is:

$$\Delta \mathbf{w}_i = \gamma h_{ik}(\mathbf{s}(t) - \mathbf{w}_i) \tag{2}$$

where k is the best-matching unit, h_{ik} is the neighborhood function and γ is the learning rate.

RecSOM defines the state of the map at time t as its activity at time t, represented by the vector $\mathbf{y}(t)$ defined by an exponential transfer function applied to the quantization error of each of the N neurons of the map (see equation 4 for the definition of the quantization error)

$$y_i = \exp(-E_i) \tag{3}$$

with $i = 1, ..., N$. Each i-th unit of the map is then equipped by two vectors, codebook vectors $\mathbf{w}_i$ and $\mathbf{c}_i$ that are respectively matched against the current input $\mathbf{s}(t)$ and the activity of the map at the previous time step $\mathbf{y}(t-1)$. The selection of the best-matching unit is performed by a linear combination of the quantization errors corresponding to $\mathbf{s}(t)$ and $\mathbf{y}(t-1)$:

$$E_i = \alpha \|\mathbf{s}(t) - \mathbf{w}_i\|^2 + \beta \|\mathbf{y}(t-1) - \mathbf{c}_i\|^2 \tag{4}$$

with $\alpha > 0$ and $\beta > 0$. The updating rules for $\mathbf{w}_i$ and $\mathbf{c}_i$ are an extension of the original SOM updating rule of equation 2:

$$\Delta \mathbf{w}_i = \gamma h_{ik}(\mathbf{s}(t) - \mathbf{w}_i) \tag{5}$$

$$\Delta \mathbf{c}_i = \gamma h_{ik}(\mathbf{y}(t-1) - \mathbf{c}_i) \tag{6}$$

where the symbols have the same meaning as in equation 2. A scheme of the RecSOM updating rule is displayed in figure 2.

3.2 News integration

The main characteristic of RecSOM is the ability to train its neurons to recognize temporal patterns. As a result, it gives a topologically ordered two-dimensional map of these temporal patterns. The visualization by means of the U-Matrix method and the application of a suitable partitioning algorithm allow to determine the clustering structure of the patterns. In the case of a time series, each cluster represents a group of patterns having a similar evolution in time. Once we have the cluster structure of the time series, it is possible to state the discretization of the series based on the cluster each data belongs to.

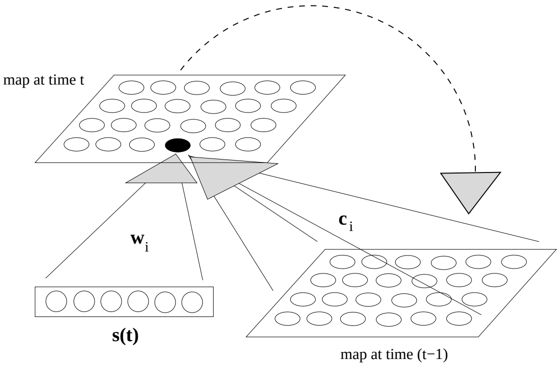

Figure 2: Recursive architecture of the RecSOM. The current input $\mathbf{s}(t)$ is processed concurrently with the $\mathbf{y}(t-1)$ vector that denotes the previous activity of the map.

The integration of the news is based on the gathering and on the preprocessing of the document collections, which is performed in the following way. For each trading day, news are collected from financial web sites and then the calculation of TDM is performed by the *tf-idf* scheme. The ratio between the two principal eigenvalues is also calculated. The amount of this ratio is associated with the ”homogeneity” of the daily news: for example, if the ratio is very close to 1, one can argue that there is not a news argument overriding all the others, reflecting the absence of a preferential direction in the TDM; conversely, if this ratio is far from 1, it is possible to deduce that there is a ”breaking news” able to induce a particular directionality in the TDM.

As seen in the previous section, RecSOM has a structure able to computationally translate this idea. Each neuron is equipped with two kinds of reference vectors $\mathbf{w}_i$ and $\mathbf{c}_i$, $\mathbf{w}_i$ driving the map to learn the current input and $\mathbf{c}_i$ conditioning the learning to retain some aspects of the state of the map at the preceding cycle. In this way, each input data is processed by the map taking into account its effective context. The relative value of α and β in equation 4 states the influence of the context in the learning phase with respect to the value of the current input.

The following hypothesis is made: a direct relation is established between the amount of the ratio $eig2/eig1$ and the relative amount of α and β, $eig2$ and $eig1$ being respectively second and first principal eigenvalues of TDM. When $eig2/eig1$ is close to 1, the influence of the context is comparable to the influence of the current value: the informational content of the news for that trading day does not present aspects able to influence markedly the financial indicators, so α and β in equation 4 are fixed to almost equal values; converse, when $eig2/eig1$ is far from 1, it is possible to assume that there is an argument more important than all the others, which is able to influence in a more effective way the evolution of the financial indicators. In this case, a correspondingly greater value than that of α is

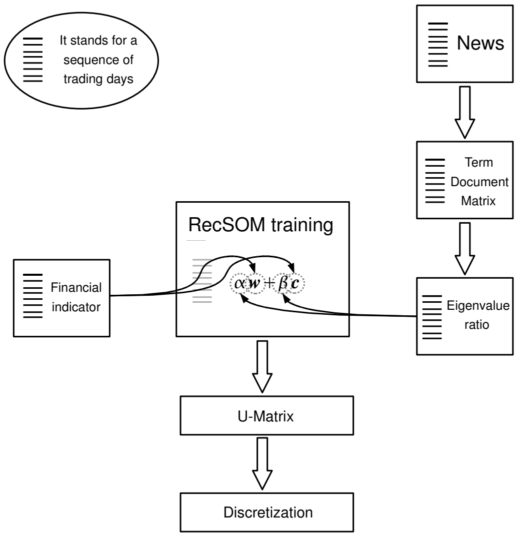

Figure 3: The global scheme of the proposed methodology

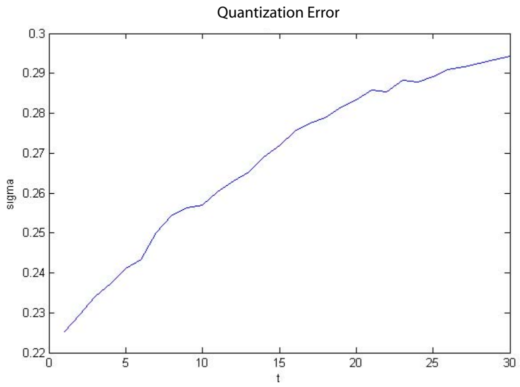

Figure 4: Temporal quantization error for the RecSOM. t is the index for the last thirty trading days starting from the current input.

assigned to β.

Hence, the news contribution consists mainly in driving the learning of the RecSOM by the modulation of the effect of the context. Figure 3 shows a scheme of the process described so far.

4 Application to the FIAT stock

The proposed methodology was applied to the time series of the FIAT stock (F.MI Yahoo ticker) daily closing price starting from 13th October 2003 to 21st April 2005. The choice of this period was driven by the free availability of financial news from the web sites of interest. The total number of gathered news amounts to 40,000 documents, i.e. approximately 80 news for each trading day. News and financial data were gathered from Italian *Yahoo Finance* web site.

In the following, two groups of results are reported with the aim to better investigate the effect of the inclusion of the news informational content on the discretization of this time series: one group related to the inclusion of news contribution and the other group related only to the financial numerical indicator. In the latter case, α and β in equation 4 are fixed to suitable values during the entire learning phase.

4.1 FIAT stock: RecSOM of the financial indicator

The RecSOM was trained on the values of the closing price with the following parameters: a squared map with 100 neurons, 200 training epochs (one epoch corresponds to a number of cycles equal to the number of input data) for a total number of 68800 learning cycles, with $\alpha = 0.9$ and $\beta = 0.1$. The calculation required a time execution of approximately 30 min on a AMD Opteron64 processor at 2.3 GHz with 2Gb RAM. The main computational limitation is represented by the huge amount of memory required by RecSOM, because for each neuron it is necessary to store an $N \times N$ matrix (the map activity vector $\mathbf{y}(t)$), N being the number of neurons.

The ordering of the input data in the resulting map is mainly driven by their numerical value as in a standard SOM, but their relative position in the map is calculated based on the sequential order they were presented to the map, thanks to the contribution of the $\mathbf{c}$ vectors to the calculations of the best-matching unit distances.

Figure 4 represents the temporal quantization error, a quantity which has the same meaning of the standard quantization error of SOM but build up in such a way to reflect the learning of a 30 element sequence of previous input data. As can be seen, the obtained RecSOM has learned the time dependence of the input data with a sufficient accuracy also on a time window of 30 trading days. The U-Matrix is displayed in figure 5. The detection of a cluster structure from the U-Matrix is performed by a simple area filling algorithm [16] with a threshold of 0.02, showing the presence of six clusters. Displaying this cluster structure onto the sequential representation of the series gives a discretization of the series itself (see figure 6). Then, the original time series is potentially compressed to a sequence of symbols (one for each of the clusters). In table 1, the mean value and standard deviation for the input data contained in each cluster are reported.

As can be seen from figure 6, each cluster represents a particular temporal pattern of the time series. For example, cluster 1 groups the input patterns displaying absolute peaks for the closing price, cluster 2 displays patterns that show a rapid succession of positive and negative trends, and so on for the other clusters.

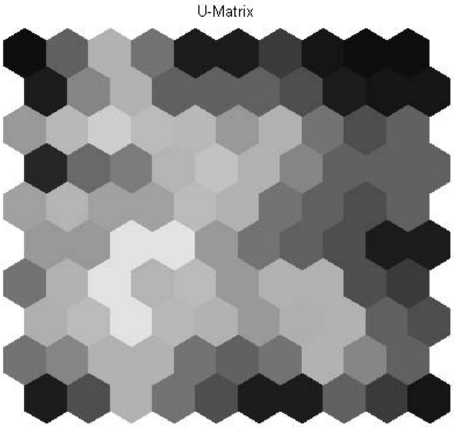

Figure 5: U-Matrix for the $\mathbf{w}$ codebooks of the RecSOM. Large distance is shown as light color.

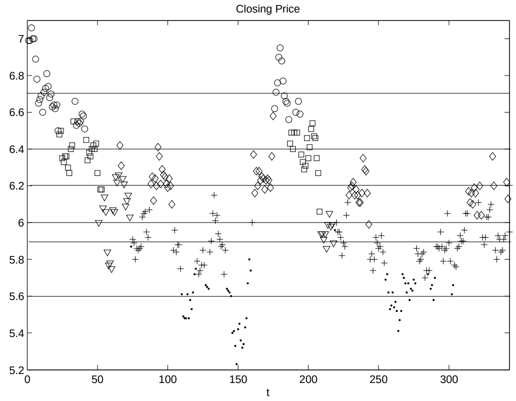

Figure 6: Discretization of the time series. Each cluster is represented by a different symbol (see table 1). Horizontal lines represent the mean of the values of each cluster.

Cluster	Mean	Standard Deviation
1 (○)	6.7	0.1
2 (□)	6.4	0.1
3 (▽)	6.0	0.1
4 (◇)	6.2	0.1
5 (+)	5.9	0.1
6 (•)	5.6	0.1

Table 1: Mean and standard deviation for the clusters of the FIAT stock closing price. Symbols refer to figure 6.

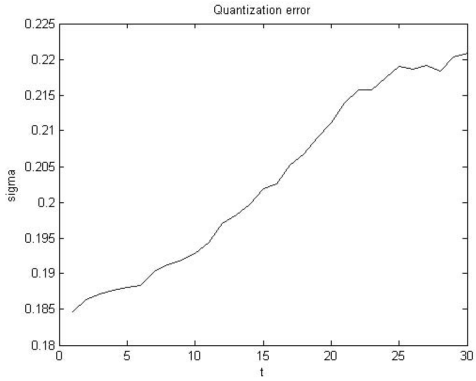

Figure 7: Temporal quantization error for the RecSOM build up with the contribution of the news. t is the index for the last thirty trading days starting from the current input.

4.2 FIAT stock: news integration in RecSOM

In this case, the RecSOM was trained with these parameters: a map of 324 units (18x18), 200 epochs, α and β determined by the criterion described in section 3. A greater number of map units was chosen with the aim to attenuate the unavoidable local instability induced in the learning phase by the updating of the α and β parameters.

The temporal quantization error is reported in figure 7, showing that also in this case the map has learned the time-dependent structure of the input data. The U-Matrix is displayed in figure 8: it shows a very different cluster structure if compared to which reported in figure 5. Cluster borders are more sharply defined and the application of the area-filling algorithm shows the presence of five well-defined clusters based on a threshold of 0.2, an order of magnitude greater than that of the previous case. This stands for a sharper discretization of the time series (see figure 9). Also note that clusters 3 and 5 are rather completely overlapped (table 2).

Looking at figure 9, one could say that the insertion of the news contribution drives the discretization of the series to concentrate on more ”interesting” zones, i.e. those where there are positive peaks (cluster 1,2 and 4) and negative peaks (clusters 5). The central zone, where there is a sequence of positive and negative local trends, is discretized into a unique broad cluster. This could be interpreted as if the insertion of the news was able to concentrate the discretization into zones where there are more interesting changes in the behavior of the stock price, reflecting the fact that the news could act mainly on ”drastic” changes of the stock price, rather than on local adjustments (central zone of the discretization).

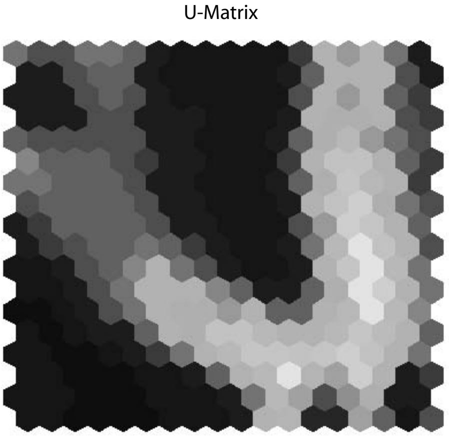

Figure 8: U-Matrix for the **w** codebooks of the RecSOM build up with the contribution of the news. Large distance is shown as light color.

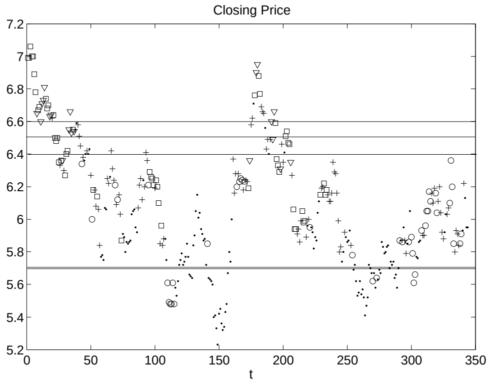

Figure 9: Discretization of the time series resulting from the inclusion of the news in the training phase of the RecSOM. Each cluster is represented by a different symbol (see table 2). Horizontal lines represent the mean of the values of each cluster.

Cluster	Mean	Standard Deviation
1 (□)	6.6	0.2
2 (▽)	6.5	0.1
3 (○)	5.7	0.2
4 (+)	6.4	0.2
5 (●)	5.7	0.2

Table 2: Mean and standard deviation for the clusters of the FIAT stock closing price contributed by the news informational content. Symbols refer to figure 9.

5 Conclusions and future work

The presented methodology and its results represent a starting point for the integration of the informational content of the news in financial time series analysis based on unsupervised learning methods. Based on a recursive Self-Organizing Map, this kind of integration does not require quantitative a priori assumptions on time-window length, this being the main crucial point in the application of supervised methodologies to time series analysis.

Future work will concentrate on a more complete test of the method by applying it to an extended period of time and on a more exhaustive news database. The testing will include also its usage as a preprocessing tool for forecasting methods: the encode of the series in a sequence of symbols will allow the application of a Markov Model to the sequence, giving the possibility to estimate the most probable ”next” symbol, i.e. the forecast.

Acknowledgements

This work has been carried out in the research project ”Complex Systems in Economics” funded by the Italian Ministry for Education, University and Research (MIUR). The cooperation of AREA Science Park (Trieste, Italy) is gratefully acknowledged.

References

[1] P.C. Fung, G.X. Yu, H. Lu, ”The Predicting Power of Textual Information on Financial Markets”, *IEEE Intelligent Informatics Bulletin*, Vol. 5, pp 1-10, 2005.

[2] P.C. Fung, G.X. Yu, J.W. Lam,”Stock prediction: Integrating Text Mining Approach Using Real-Time News”, *Proceedings of the 7th IEEE International Conference on Computational Intelligence for Financial Engeneering*, pp 395-402, 2003.

[3] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan ”Mining of Concurrent Text and Time Series”, *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pp. 37-44, 2000

[4] T.W. Liao,”Clustering of Time Series Data - A Survey”, *Pattern Recognition*, Vol. 38, pp. 1857-1874, 2005.

[5] G.A. Barreto,”Time Series Prediction with the Self-Organizing Map: A Review” in *P. Hitzler and B. Hammer, eds., Perspectives on Neural-Symbolic Integration*,Springer-Verlag, 2007.

[6] T.C. Fu, F.L. Chung, V. Ng and R. Luk,”Pattern Discovery from Stock Time Series Using Self-Organizing Maps” in *KDD 2001 Workshop on Temporal Data Mining*, 2001.

[7] G. Salton, ”Introduction to Modern Information Retrivial”, McGraw-Hill, 1983

[8] G.Salton, ”Automatic Text Processing”, McGraw-Hill, 1989

[9] T. Kohonen,”Self-Organized formation of topologically correct feature maps”, *Biological Cybernetics*, Vol. 43, pp. 59-69, 1982.

[10] A. Ultsch, H.P. Siemon, ”Kohonen’s self organizing feature maps for exploratory data analysis”, *Proc. INNC of Int. Neural Network Conf*, 1990.

[11] B. Hammer, A. Micheli, A. Sperduti, M. Strickert, ”Recursive self-organizing network models”, *Neural Networks*, Vol 17(8-9), pp. 1061-1086, 2004

[12] G. Chappel, J. Taylor ”The temporal Kohonen map”, *Neural Networks*, Vol.6, pp. 441-445, 1993

[13] T. Voegtlin, ”Recursive self-organizing maps”, *Neural Networks*, Vol. 15, pp. 979-991, 2002.

[14] M. Hagenbuchner, A. Sperduti, A.C. Tsoi, ”A Self-Organizing Map for Adaptive Processing of Structured Data”, *IEEE Transactions on Neural Networks*, Vol. 14, pp. 491-505, 2003

[15] M. Strickert, B. Hammer, ”Merge SOM for temporal data”, *Neurocomputing*, Vol. 64, pp. 39-72, 2005

[16] D. Opolon, F. Moutarde, ”Fast semi-automatic segmentation algorithm for Self-Organizing Maps”, *Proc. of ESANN 2004*, pp.507-512, 2004

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de