FEATURE CLUSTERING WITH SELF-ORGANIZING MAPS AND AN APPLICATION TO FINANCIAL TIME-SERIES FOR PORTFOLIO SELECTION

Bruno Silva
Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal, Campus do IPS, Setúbal, Portugal
bruno.silva@estsetubal.ips.pt

Nuno Marques
Departamento de Informática, Faculdade de Ciências, Universidade Nova de Lisboa, Monte da Caparica, Portugal
nmm@di.fct.unl.pt

Keywords: Self-organizing maps, Correlation hunting, Feature clustering, Time-series clustering, Portfolio selection.

Abstract: The portfolio selection is an important technique for decreasing the risk in the stock investment. In the portfolio selection, the investor's property is distributed for a set of stocks in order to minimize the financial risk in market downturns. With this in mind, and aiming to develop a tool to assist the investor in finding balanced portoflios, we achieved a generic method for feature clustering with Self-Organizing Maps (SOM). The ability of neural networks to discover nonlinear relationships in input data makes them ideal for modeling dynamic systems as the stock market. The method proposed makes use the remarkable visualization capabilities of the SOM, namely the Component Planes, to detect non-linear correlations between features. An appropriate metric - the improved R_v coefficient - is also proposed to compare Component Planes and generate a distance matrix between features, after which an hierarchical clustering method is used to obtain the clusters of features. Results obtained are empirically sound, although at this moment we do not provide mathematical comparisons with other methods. Results also show that feature clustering with the SOM presents itself as a viable method to cluster time-series.

1 INTRODUCTION

The portfolio selection is an important technique for decreasing the risk in stock investments. In the portfolio selection, the investor's property is distributed for a set of stocks in order to minimize the financial risk in market downturns. If some stocks happen to drop, this loss can be minimized and/or counterbalanced by others that have an opposite historical behavior. Therefore, reliable tools in the selection process can be of great assistance to investors. These tools should give a competitive edge over others as he/she can identify the performing stocks with minimum effort. With this in mind, and aiming to develop a tool to assist the investor in finding balanced portfolios, we achieved a generic method for feature clustering with Self-Organizing Maps (SOM) (Kohonen, 2001). The SOM is a very popular artificial neural network algorithm based on competitive and unsupervised learning, and is primarily used for the visualization of nonlinear relations of multidimensional data and dimensionality reduction. The SOM is able to project high-dimensional data in a lower dimension, typically 2D, while preserving the relationships among the input data, thus electing it as a data-mining tool of choice (Vesanto, 2000; Vesanto, 1999; Flexer, 1999; Himberg et al., 2001). This non-linear projection produces a 2D pattern map that can be useful in analyzing and discovering patterns in the input space. The ability of neural networks to discover nonlinear relationships in input data makes them ideal for modeling dynamic systems as the stock market. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends or correlations that are too complex to be noticed by either humans or other computer techniques.

At this point, we believe it is necessary to clarify the differences between the terms *clustering* and *feature clustering*: *Clustering*, in its traditional sense, is the unsupervised classification of observations (patterns, data items or feature vectors) into groups (clusters). Each of these observations has a set of attributes (features) that characterizes it and define the dimensionality of the data-space. Emerging data mining applications (Clark et al., 2008; Barraquand and Martineau, 1995; Kumar, 2001; Kohonen et al., 2000) place special requirements on clustering techniques, such as the ability to handle high dimensionality, assimilation of cluster descriptions and usability. Regarding the clustering of high dimensional data, an object typically has from dozens to hundreds of attributes in which the domains of the attributes are large. *Feature clustering*, on the other hand, can be defined as a technique to cluster the features that characterize the observations, forming groups of features that are somewhat dependent of each other and/or correlated. It has several applications in data-mining, namely inspecting correlations in data and dimensionality reduction, i.e., with a better understanding of the interaction among the various features, knowledge can be derived from a problem at hands and explanations obtained; in addition, for high dimensional datasets, feature clustering can pose itself as an alternative to feature selection and dimensionality reduction algorithms, treating each cluster as a single feature. We will extend the application of feature clustering to the clustering of time-series, given particular datasets.

This paper introduces a method for feature clustering using Self-Organizing Maps (SOM) (Kohonen, 2001) and an application to stock market time-series data to aid in a stock portfolio selection. Therefore, the main contributions of this paper are a:

- **Feature Clustering Method based on Self-organizing Maps.** We explore the visual properties of the SOM (introduced in Section 3), namely a particular visualization technique called *Component Planes* (Section 3.1), which can be seen as a "sliced" version of the SOM, to obtain correlations between features. Each feature in the SOM has its own component plane. If two component planes are similar to each other, this means that the two features associated with them are correlated; we also **propose the use of an appropriate metric - the *modified R_v coefficient*** (Smilde et al., 2009) - to compare component planes that enables us to transform the feature-space into a distance matrix that can be used to obtain the feature clusters by a hierarchical clustering method. This method is generic and can be applied to any problem. However, we illustrate the process in detail (Section 3.2) using a simple well-known dataset - the *Iris* dataset.
- **Application of Financial Time-series Clustering for Portfolio Selection.** We illustrate the use of proposed method with time-series (Section 4). The presented method is applied to aggregate similar stocks based on their historical behavior. Given historical prices of a set of stocks - having an observation per each day of the stock market, in which the features are the stock values - the method is able to cluster the stocks into disjoint clusters, from which a stock portfolio can be selected. This differs from applications where the time-series are the observations themselves and where other algorithms can be applied, namely *k-means* and other techniques presented in Section 2.

Additionally, in Section 2 we make an overview of traditional methods used for feature clustering and time-series clustering, as well as common applications, enlightening the particular strengths of the SOM is such situations. To the best of our knowledge, the SOM has not been used before to form stock portfolios based on full stock time-series. However, we present other works that used the SOM for similar tasks and other problems related to the stock market. Finally, in Section 5 we present our final conclusions and future work.

2 RELATED WORK

As said before, one of the applications of feature clustering can be *dimensionality reduction*. A variety of approaches for reducing dimensionality of a data space have been developed. Classical statistics techniques include principal component analysis (PCA) and factor analysis, both of which reduce dimensionality by forming linear combinations of features (Pearson, 1901). For the principal component analysis technique, a lower-dimensional representation is found that accounts for the variance of the features, whereas the factor analysis technique finds a representation that accounts for the correlations among the features. Unfortunately, dimensionality reductions obtained using these conventional approaches conflict with the requirements placed on the assimilation aspects of data mining. For example, PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called *principal components*. These principal components represent "new" features that are often hard to understand, since they do not represent original features directly. This is sufficient in classification tasks, but it is a poor method if we are also interested in extracting knowledge from the dimensionality-reduction procedure. Data mining applications often require cluster descriptions that can be assimilated and used by users because insight and explanations are the primary purpose for data mining. The advantages of using SOMs for correlation detection are their robustness against noise in data, i.e., outliers that harm the correct results for common statistical methods, since they affect the variance of the data; the use of large SOMs, called *emergent* SOMs, allowing the detection hidden correlations within the data, that can be missed by the conventional methods (Silva and Marques, 2007; Ultsch and Herrmann, 2005; Ultsch and Morchen, 2005); and the visualization capabilities of the SOM that makes the interpretation of the results richer. Feature clustering can be an alternative in cases where one needs dimensionality reduction for clustering purposes without losing any underlying information for result interpretation. The most common application of feature clustering found in literature is in text-mining. In (Dhillon et al., 2003; Baker and McCallum, 1998) feature clustering is considered a powerful alternative to feature selection (Wang et al., 2009) for reducing the dimensionality of text data. Feature selection involves selecting the features that have more ”weight” in the description of each observation, i.e., the ones that are sufficient to distinguish observations. The consequence of this is the loss of information (the remaining features) that can be useful in extracting further knowledge. Using feature clustering in detriment of feature selection can be obtained by treating each cluster as a single feature, thus reducing the dimensionality of the data. As an example of this in the domain of text-mining, in (Yang and Pedersen, 1997) feature clustering is found to be more effective than feature selection.

In terms of time-series clustering, it has attracted increasing interest in the last decade, particularly for long time series such as those arising in the bioinformatics (Potamias, 2002) and financial domains (Dose and Cincotti, 2005) and are mainly addressed by using techniques such as *singular value decomposition* (SVD), *discrete Fourier transform* (DFT), and *discrete wavelet transform* (DWT) (Zhang et al., 2006). Regarding the SOM, two relevant works in literature are found in (Fu et al., 2001), applied also to the stock market, and (Hammer et al., 2005). In these two works the time-series are the observations themselves and use the traditional clustering capabilities of the SOM, either by extracting representative features from the original time-series, using other distance metrics to compare time-series or using recursive-SOMs (Voegtlin, 2002). Our work, differs in the way that time-series are represented. A time-series is represented by a feature, i.e., an observation is a moment in time, e.g., a day, and each feature contains the values for each variable of interest, e.g., a stock. In consequence, to cluster time-series we need to cluster the features themselves. This way, we preserve all of the aspects of the original SOM which has been validated in hundreds of applications in the last years. Other applications of the traditional SOM clustering capabilities to stock-picking can be also found in literature, as well as financial problems in general (Deboeck, 1998). In (Khan et al., 2009) the SOM is used as a method of classification of selected stocks into a fixed number of classes, given properties for each stock as a result of technical analysis. Another work (Stankevicius, 2001) uses a similar method to form a portfolio of stocks, based on the clustering of the data provided by the SOM. Both works use the traditional SOM clustering technique (Flexer, 1999) to group the stocks, since they are described by technical and fundamental indicators.

The presented feature clustering method lays its grounds on feature correlation detection by the SOM. The method of detecting feature correlations in the SOM is commonly referred in literature as *correlation hunting*. In (Vesanto and Ahola, 1999; Pal, 2001), a first method for this task was presented, which shares some similarities with the presented work, i.e., the use of the component planes to detect correlations. It was used to re-organize the presentation of the component planes to the user, given their similarities to interpret more easily the correlations between features. In this work we extend that idea to obtain a feature clustering method and propose a very recent metric - the improved R_v coefficient - that can be applied directly to component planes, instead of the Pearson's correlation coefficient used in the cited works. We also show that this method can be used with time-series to cluster similar time-series in their overall behavior.

3 THE SELF-ORGANIZING MAP AND FEATURE CLUSTERING

The Self-Organizing Map (SOM) was introduced by Kohonen in the early 1980s. It can be visualized as a sheet-like neural-network array, whose neurons become specifically tuned to various input vectors (observations) in an orderly fashion. They provide a way of representing multidimensional data on much lower dimensional spaces - usually two dimensions. The training of these maps results in networks that store ”compressed” information in their neurons, through a competitive learning algorithm. Also, topological relationships in the input data are maintained on trained maps, allowing the application of visualization techniques.

Since this is a well-known algorithm and due to space constraints we refer the reader to (Kohonen, 2001) for a detailed description of the SOM training algorithm.

3.1 Visualizations of the SOM

Although the SOM is a powerful data-mining tool, humans still play an important role in terms of interpreting the information that can be extracted from trained maps. The map by itself is not of much use without visualization techniques that enhance particular properties of the underlying data. The SOM can be used efficiently in data visualization due to its ability to represent the input data in two dimensions, and different SOM visualizations offer information of correlations between data components and of the cluster structure of the data. Two pertinent visualizations of the SOM for this paper are introduced using the well-known Iris dataset. Each sample of the dataset has four features, describing some measurements, in *cm*, of the flowers: Sepal Length, Sepal Width, Petal Length, and Petal Width. A particular characteristic of this dataset is that the class of Iris-Setosa flowers is linearly separable from the other two. Also, the attributes Petal Length and Petal Width are highly correlated. Figure 1 plots the four Iris features by sample. Samples are ordered by class. Sample ordering was considered as time, so each sample is presented on the same time instant. This clearly illustrates that Petal Length and Petal Width are highly correlated. Also the relation with feature clustering on a time-series should become clear. Let us consider a SOM feature map for representing this dataset. The depicted figures are from a 15 × 20 SOM.

Unified Distance Matrix or simply U-Matrix, visualizes the cluster structure of the SOM. A matrix of distances between the weight vectors of adjacent neurons on the map is formed, after which some representation for the matrix is chosen, for example a color scale. The U-Matrix of the Iris dataset is shown in Figure 2. The lighter the color between two map units is, the smaller is the relative distance between their weight vectors. Given this, dark areas on the maps usually identify boundaries between clusters in the underlying data.

Component Planes is a representation that visualizes relative component values in the weight vectors of the SOM. The illustration can be considered as a sliced version of the SOM, where each plane shows the distribution of one weight vector component (one feature). Using the distributions, correlations between different components can be studied. The four component planes for the Iris dataset are illustrated in Figure 3.

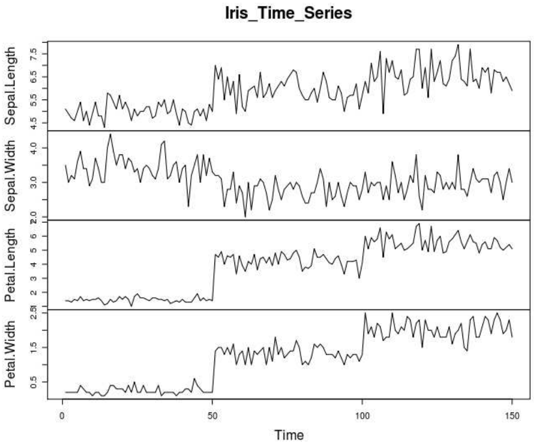

Figure 1: Iris dataset features. Each sample is considered as a time instant. Dataset is plotted ordered by class.

The use of these visualizations give us some insight of the structure and correlations of the underlying data. As it can be seen in the U-Matrix, there is a clear identification of two clusters in the Iris dataset, that conforms to the characteristics of the dataset. By inspecting the component planes, it can be immediately seen that there is a high correlation of the Petal Length and Petal Width features, i.e. the respective component planes are similar. Sometimes these component planes can be useful in interpreting the type of samples that belong to a cluster, comparing them to the U-Matrix. In conjoining these two visualizations techniques, it can be seen that the top cluster is formed by flowers that have small values for Petal Length and Petal Width.

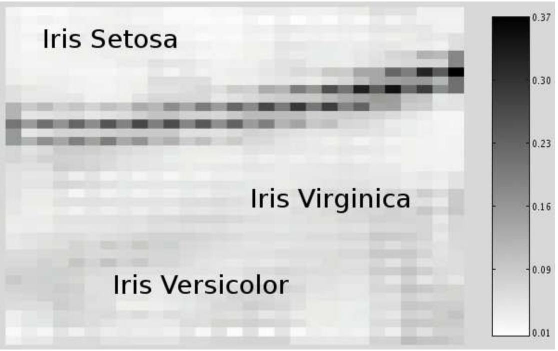

Figure 2: U-Matrix for the Iris dataset. It is visible that two clusters are defined in the SOM map. Labeling the map, one can see that one contains the Iris-Setosa flowers and the other contains the Iris-Versicolor and Iris-Virginica flowers.

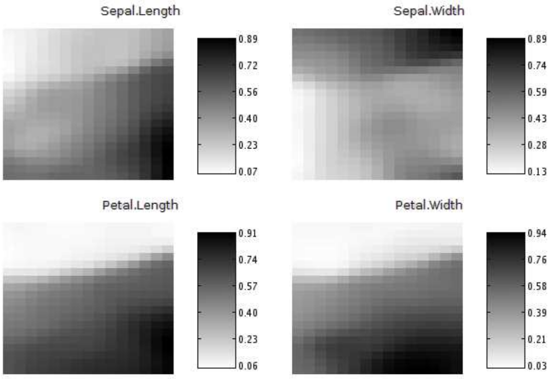

Figure 3: Component Planes for a trained SOM with the Iris dataset.

3.2 Feature Clustering with the SOM

In cases where the number of features is relatively low, it suffices the use of the component planes visualization to detect correlated features. But a problem arises when their number is very high - it is difficult for the observer to detect similarities and group them when there are dozens of features to analyze.

Based on the component planes, which mathematically are matrices of real values, it is necessary to use a metric that can give a measure of how two matrices resemble one another, i.e., $\sigma(C_i, C_j)$, where σ can be an appropriate resemblance measure. The R_v coefficient (Escoufier, 1973) was initially chosen and deemed appropriate for this task. It is defined as *a similarity coefficient between positive semi-definite matrices*. However, in early experiments we noted that, especially for increasing larger maps (by visual comparison of the component planes), that the results of the original R_v coefficient deteriorated, i.e., for larger maps this similarity measure was giving high scores for pairs of component planes that were not very similar. This was confirmed by the work in (Smilde et al., 2009), where it was proved that the original R_v coefficient depended on the size of the matrices. To solve this, **we applied the modified R_v coefficient** described in (Smilde et al., 2009). As a consequence of the definition, positive semi-definite matrices are square, symmetric and their diagonal elements are always equal or greater than zero. For rectangular matrices (the SOM map is usually rectangular) and to apply this measure between component planes (C_i and C_j), the method involves transforming the matrices (component planes) into positive semi-definite matrices by multiplying each matrix by its transpose. In addition, the modified R_v coefficient also requires the diagonals to be subtracted from the resulting matrices. This is formalized in Equation 1.

Table 1: Distance matrix for features of the Iris dataset using the modified R_v coefficient. S.L - Sepal Length; S.W. - Sepal Width; P.L. - Petal Length; P.W. - Petal Width.

	S.L.	S.W	P.L.	P.W.
S.L.	0.00	1000.00	359.10	396.31
S.W.	1000.00	0.00	1000.00	1000.00
P.L.	359.10	1000.00	0.00	126.93
P.W.	396.31	1000.00	126.93	0.00

$$\begin{aligned} S &= C_i C_i^T - diag(C_i C_i^T) \\ T &= C_j C_j^T - diag(C_j C_j^T) \end{aligned} \qquad (1)$$

The similarity measure between two component planes is given by Equation 2.

$$\sigma(C_i, C_j) = \frac{trace\{S^T T\}}{\sqrt{(trace\{S^T S\}) \times (trace\{T^T T\})}}, \qquad (2)$$

where the *trace* operation is the sum of the diagonal elements.

This modified coefficient takes values between in the range of $[-1, 1]$. The value of $R_v = 1$ says that the two component planes are identical, while $R_v = -1$ says that they are very different. In this work, the resulting coefficient is adjusted to the range $[0, 1]$, by $(R_v + 1)/2$. Using this coefficient to assess the similarity of two component planes, a distance matrix is then generated using Equation 3 to assign a distance between component planes C_i and C_j. It is common for a distance matrix to satisfy the triangle inequality, so Equation 3 takes this into account.

$$dist(C_i, C_j) = \sqrt{1 - \sigma(C_i, C_j)} \times S_c, \qquad (3)$$

where S_c is a scale factor to ease the results visualization and prevent floating-point arithmetic precision errors in operations that use the values from this distance matrix. For the component planes presented in Figure 3 the resulting distance matrix using the modified R_v coefficient is presented in Table 1, with $S_c = 1000$.

It can be seen that Sepal Length is very distant from Petal Length and Petal Width. Due to the scaling applied to the modified R_v coefficient, the value of 1000 says that the features are not in any way correlated. These values are in conformance with the component planes depicted in Figure 3. After this matrix is generated, we proceed to the clustering of these component planes (features) by their relative distance using some appropriate method. *Hierarchical clustering* was chosen as it presents some properties that are interesting for this task, in particular the fact that hierarchical clustering can be represented by a two dimensional diagram, known as *dendrogram*, which illustrates the merges made at each successive stage of analysis. From this dendrogram the features in each cluster can be extracted automatically, besides giving additional information about the similarity of features. The dendrogram for this example can be seen in Figure 4. Boxes are shown grouping features for $k = 2$ clusters, using a cut-off algorithm which establishes a horizontal line that crosses k links in the dendrogram to obtain the members of the clusters. Based on this clustering we are able to present the clustering of the Iris dataset features, for $k = 2$. Feature clustering provided two clusters: **Cluster 1** containing Petal Lenght, Petal Width and Sepal Length and **Cluster 2** containing Sepal Width. If we look at Figures 1 and 3, we notice that this should be expected when we impose a $k = 2$ (*Sepal Width* feature is too different from the other three features). This result also illustrates the problems of imposing a priori a value for k, and relates to the automatic determination of the number of clusters discussed in future work.

4 METHOD APPLICATION TO STOCK PORTFOLIO SELECTION

In this section, the feature clustering method detailed earlier is applied to historical stock prices data. Given historical prices of a set of stocks (having an observation per each day of the stock market, in which the features are the stock values) the goal is to cluster similar stocks in their historical behavior as an assistance to stock portfolio selection. All results were obtained with a custom SOM toolbox - *netSOM* (Silva and Marques, 2007), that is being continuously upgraded with new features, such as feature clustering. It has distributed processing capabilities to train large maps and/or datasets. Those functionalities were not necessary for this relatively small task. The hierarchical clustering is provided by the *R language*, through the call of an external process to the interpreter, and results are handled by the toolbox to provide the overall results.

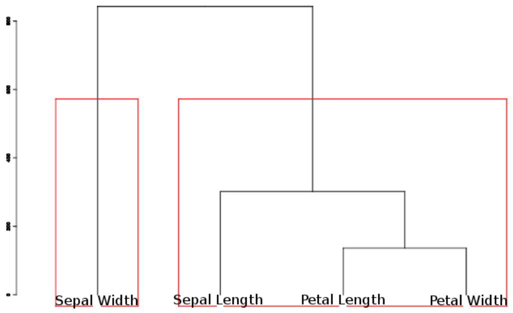

Figure 4: Resulting dendrogram from hierarchical clustering for the features in the Iris dataset.

4.1 Description of the Dataset and Data Pre-processing

The original dataset describes historical data for 49 stock prices and the *gold* commodity price[1]. It comprises information gathered from 1998 to 2009. The complete list of features is presented in Table 2. The dataset contained missing values. Given that we were dealing with time-series, the method of *last observation carried forward* (LOCF) was applied (standard procedure in financial data analysis). Also, where missing values were present due to the fact that a particular stock only appeared in the stock market after 1998, the opening price was used as the base value, i.e., applying LOCF backwards. The resulting dataset is composed by 2928 sample objects with 50 features. Before feeding this dataset to the SOM, all features values were normalized in the range $[0, 1]$.

4.2 Results and Discussion

A 20×30 SOM map was trained by *netSOM*, with the described pre-processed dataset. The presented feature clustering method yielded the $k = 10$ clusters presented in Table 3. The choice for this value was purely empirical to allow for a large variance between clusters. Validation of the clusters formed by the method was only done by visual inspection of the original time-series. Due to space constraints only the smaller clusters are presented in Figure 5.

The interpretation of the results must take into account the normalization that was made to the values of the different stocks (time-series). A good example is immediately visible in **Cluster 1**, where the method clustered GOLDS and VOW.GY. By inspecting the original time-series in Figure 5 they do not seem similar at first sight. However, given that the values of both time-series were scaled individually to the range of $[0, 1]$, VOW.GY is the financial product that best approaches the behavior of gold. The distance matrix had the value of 400 between these two products, showing that they are not too closely correlated. This is where the generated dendrogram can be helpful, which presents these distances also. Nevertheless, the component planes also show this particular fact. Another important detail that this first

[1] *gold* price can be helpful to detect other financial products with similar behavior, which is particularly relevant in states of market crisis, since gold is considered a safe commodity for investors.

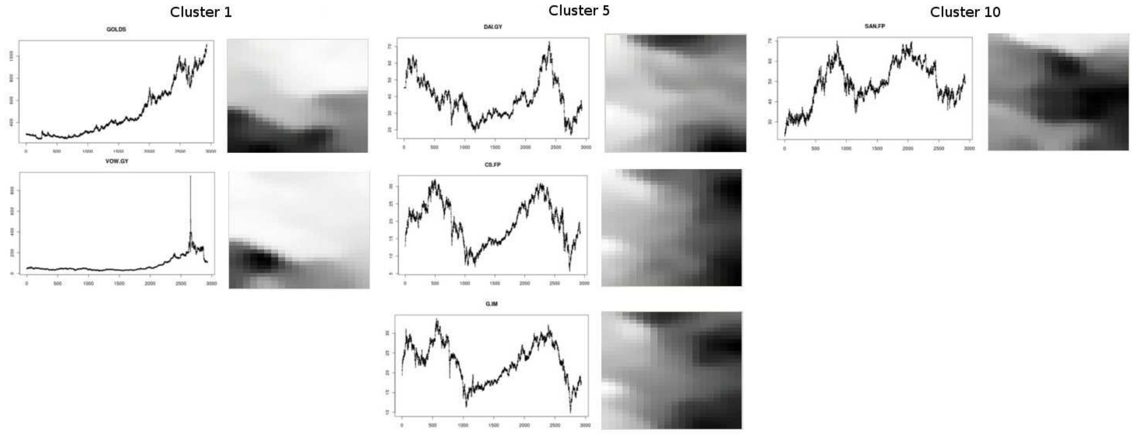

Figure 5: Cluster members showing the component planes and original time-series. Cluster members (from top to bottom): Cluster 1 - GOLDS, VOW.GY; Cluster 5 - DAI.GY, CS.FP, G.IM; Cluster 10 - SAN.FP.

cluster illustrates is an outlier in the VOW.GY time series, either from an error in the dataset creation, or an abnormal spike in the price of this stock. The SOM is robust to noise and filters these kind of outliers. In **Cluster 5** it is interesting that both insurance companies, i.e., CS.FP (AXA SA) and G.IM (ASSICURAZIONI GENERALI) were clustered together, being closely related by component plane and time-series comparison. DAY.GY (DAIMLER AG-REGISTERED SHARES) was also added to this cluster, apparently because it shares common behavior from half the available period of time forward. This particular fact is another major advantage of using SOM for feature clustering, i.e.,besides being able to detect non-linear correlations, it is also able to detect partial correlations. **Cluster 10** only has one member, i.e., SAN.FP (SANOFI-SAVENTIS) because its historical behavior distances from all other financial products. This discovery is relevant for stock portfolio selection. Other clusters seem to group financial institutions together, e.g., **Cluster 7**. Given that consecutive clusters are somewhat correlated (a consequence for being obtained through the dendrogram), this seems very coherent. Other stocks that were clustered together with financial institutions probably depend closely from the performance of these ones. The visual comparison of the members of the other clusters, proved that this method is reliable for clustering similar time-series in their behavior. This kind of analysis can be extremely useful in portfolio selection, paying attention to which stocks behave similarly and selecting an equilibrated set of stocks. In the simplest of ways, one can chose a stock from each cluster. Overall, this application also shows that emergent SOMs are able to detect similar time-series in their historical behaviors and should be regarded as an interesting approach to such problems.

Future Work should address the validation of the method and comparison with other statistical methods. In terms of finding the optimal number of feature clusters, we enumerate three possible ways of achieving this: through dendrogram link inconsistency analysis (Stanberry et al., 2003), model based analysis (Fraley and Raftery, 1998) or allowing another SOM to cluster the previous results and obtaining the correct number of clusters with Ward's method (Ward, 1963). The first method has the drawback of needing to specify the link inconsistency coefficient and does not present itself has a reliable approach. The second method is deemed to be very good, since it analyses several models of possible distribution of samples, but does not work directly with the obtained distance matrices - one would need to apply a multidimensional scaling (MDS) to obtain a 2D projection of the features and then apply this method. Finally, the SOM approach could make use of the automatic clustering abilities of the SOM, by using, for example, the Ward method (Ward, 1963) to identify the correct number of clusters. It can be used in two ways, either by applying the MDS and clustering the points, or by clustering directly the component planes, but loosing a precise manner of comparing component planes (the R_v coefficient) in detriment of a simple euclidean distance, as in (Vesanto and Ahola, 1999).

5 CONCLUSIONS

A feature clustering method based on Self-Organizing Maps was presented in this paper. Feature clustering has many applications, namely inspecting correlations in data, dimensionality reduction and time-series clustering. Some considerations on how to apply feature clustering to these applications were made

Table 2: Original features for the historical stock prices dataset. Only features regarding stock prices and the gold commodity were used.

Date	Date of observation
GOLDS	GOLD SPOT $/OZ (Comdty) PX-LAST
AGN-NA	AEGONNVPX-LAST
AI-FP	AIRLIQUIDESA
ALV-GY	ALLIANZSE-REG
ALO-FP	ALSTOM
ABI-BB	ANHEUSER-BUSCHINBEVNV
MT-NA	ARCELORMITTAL
CS-FP	AXASA
SAN-SQ	BANCOSANTANDERSA
BAS-GY	BASFSE
BAYN-GY	BAYERAG
BBVA-SQ	BANCOBILBAOVIZCAYAARGENTA
BNP-FP	BNPPARIBAS
CA-FP	CARREFOURSA
ACA-FP	CREDITAGRICOLESA
CRH-ID	CRHPLC
DAI-GY	DAIMLERAG-REGISTEREDSHARES
BN-FP	DANONE
DBK-GY	DEUTSCHEBANKAG-REGISTERED
DB-GY	DEUTSCHEBOERSEAG
DTE-GY	DEUTSCHETELEKOMAG-REG
EOAN-GY	E.ONAG
ENEL-IM	ENELSPA
ENI-IM	ENISPA
FTE-FP	FRANCETELECOMSA
GSZ-FP	GDFSUEZ
G-IM	ASSICURAZIONIGENERALI
IBE-SQ	IBERDROLASA
INGA-NA	INGGROEPNV-CVA
ISP-IM	INTESASANPAOLO
OR-FP	L'OREAL
MC-FP	LVMHMOETHENNESSYLOUISVUI
MUV-GY	MUENCHENERRUECKVERAG-REG
NOKV-FH	NOKIAOYJ
PHIA-NA	PHILIPSELECTRONICSNV
REP-SQ	REPSOLYPFSA
RWE-GY	RWEAG
SGO-FP	COMPAGNIEDESAINT-GOBAIN
SAN-FP	SANOFI-AVENTIS
SAP-GY	SAPAG
SU-FP	SCHNEIDERELECTRICSA
SIE-GY	SIEMENSAG-REG
GLE-FP	SOCIETEGENERALE
TIT-IM	TELECOMITALIASPA
TEF-SQ	TELEFONICASA
FP-FP	TOTALSA
UCG-IM	UNICREDITSPA
UNA-NA	UNILEVERNV-CVA
DG-FP	VINCISA
VIV-FP	VIVENDI
VOW-GY	VOLKSWAGENAG

Table 3: Ten clusters generated by the presented method for 50 financial products.

Cluster 1
GOLDS VOW.GY
Cluster 2
AI.FP BAS.GY BN.FP DB1.GY DG.FP ENI.IM EOAN.GY FP.FP IBE.SQ RWE.GY UNA.NA
Cluster 3
ALO.FP ALV.GY CA.FP DTE.GY FTE.FP TIT.IM VIV.FP
Cluster 4
ABI.BB BAYN.GY GSZ.FP MT.NA TEF.SQ
Cluster 5
CS.FP DAI.GY G.IM
Cluster 6
ACA.FP BNP.FP CRH.ID ENEL.IM GLE.FP ISP.IM REP.SQ SAN.SQ SGO.FP SU.FP
Cluster 7
BBVA.SQ DBK.GY INGA.NA OR.FP UCG.IM
Cluster 8
MC.FP SAP.GY SIE.GY
Cluster 9
MUV2.GY NOK1V.FH PHIA.NA
Cluster 10
SAN.FP

during the related work survey. For time-series clustering, particular relevance was given and the presented method was applied to a dataset describing historical data for 49 stock prices and the *gold* commodity, comprising information gathered from 1998 to 2009. The method was able to cluster stocks with similar historical behavior, showing that the SOM can be used to cluster time-series by their overall behavior. The method revealed good empirical results, based on the visualization of the original time-series and common sense, namely the method was able to cluster together insurance companies in one cluster, and financial institutions in other two close clusters, as well as revealing a particular stock that had an historical behavior different from all others. This kind of analysis is relevant in portfolio selection, since composing a portfolio with too much similar stocks in their historical behavior is to be avoided as a means to lower exposure to risk. Instead, diversification is the main goal.

REFERENCES

Baker, L. D. and McCallum, A. K. (1998). Distributional clustering of words for text classification. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103, New York, NY, USA. ACM.

Barraquand, J. and Martineau, D. (1995). Numerical valuation of high dimensional multivariate american securities. *Journal of Financial and Quantitative Analysis*, 30(03):383–405.

Clark, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M., Gehan, E., and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8:37–49.

Deboeck, G. J. (1998). Financial applications of self-organizing maps. In *NEURAL NETWORK WORLD*, volume 8, pages 213–241.

Dhillon, I. S., Mallela, S., and Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of Machine Learning Research*, 3:1265–1287.

Dose, C. and Cincotti, S. (2005). Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*, 355(1):145 – 151.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*.

Flexer, A. (1999). On the use of self-organizing maps for clustering and visualization. In *Principles of Data Mining and Knowledge Discovery*, pages 80–88.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41:578–588.

Fu, T., Chung, F., Ng, V., and Luk, R. (2001). Pattern discovery from stock time series using self-organizing maps. In *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*, pages 26–29.

Hammer, B., Micheli, A., Neubauer, N., Sperduti, A., and Strickert, M. (2005). Self organizing maps for time series. In *Proceedings of WSOM*, pages 115–122.

Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J., and Simula, O. (2001). The self-organizing map as a tool in knowledge engineering.

Khan, A. U., Bandopadhyaya, T. K., and Sharma, S. (2009). Classification of stocks using self organizing map. *International Journal of Soft Computing Applications*, pages 19–24.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585.

Kumar, V. (2001). *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, Norwell, MA, USA.

Pal, N. R., editor (2001). *Pattern Recognition in Soft Computing Paradigm*, chapter The Self-Organizing Map as a Tool in Knowledge Engineering, pages 38–65. Soft Computing. World Scientific Publishing.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

Potamias, G. (2002). Distance and feature-based clustering of time series: An application on neurophysiology. In *SETN '02: Proceedings of the Second Hellenic Conference on AI*, pages 237–248, London, UK. Springer-Verlag.

Silva, B. and Marques, N. (2007). A hybrid parallel som algorithm for large maps in data-mining. In Neves, J., Santos, M. F., and Machado, J., editors, *New Trends in Artificial Intelligence*, Guimarães. Portugal. Associação Portuguesa para a Inteligência Artificial (APPIA).

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., and van Erk, M. J. (2009). Matrix correlations for high-dimensional data: the modified rv-coefficient. *Bioinformatics*.

Stanberry, L., Nandy, R., and Cordes, D. (2003). Cluster analysis of fmri data using dendrogram sharpening. *Human brain mapping*.

Stankevicius, G. (2001). Forming of the investment portfolio using the self-organizing maps (som). *INFORMATICA*.

Ultsch, A. and Herrmann, L. (2005). The architecture of emergent self-organizing maps to reduce projection errors. In *In Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2005)*, pages 1–6. Verleysen M. (Eds).

Ultsch, A. and Morchen, F. (2005). Esom-maps: tools for clustering, visualization, and classification with emergent som. Technical Report 46, Dept. of Mathematics and Computer Science, University of Marburg, Germany.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent-Data-Analysis*, 3:111–26.

Vesanto, J. (2000). Using som in data mining. Licentiate's thesis in the Helsinki University of Technology.

Vesanto, J. and Ahola, J. (1999). Hunting for Correlations in Data Using the Self-Organizing Map. In Bothe, H., Oja, E., Massad, E., and Haefke, C., editors, *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99)*, pages 279–285. ICSC Academic Press.

Voegtlin, T. (2002). Recursive self-organizing maps. *Neural Networks*, 15(8-9):979–991.

Wang, H., Han, L., Zeng, X., and Zhen, Z. (2009). Feature selection with maximum information metric in text categorization. *Information Science and Engineering, International Conference on*, 0:857–860.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc*.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, H., Ho, T. B., Zhang, Y., and Lin, M.-S. (2006). Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica*.