Data Mining for Financial Time Series

Odeta Shkreli

Abstract— Financial data analysis is a complicated process and has attracted many researches proposing numerous methods and techniques that can be applied and implemented by the mean of information technology. Data mining techniques can be considered as the most advanced techniques, the application of which in the financial sector has given positive results and has proved their worthiness and the need for further research and development. Data mining is the heart of knowledge discovery process that is used to extract useful knowledge from data. This research studies the data mining process and the methods and techniques that can be applied in financial time series data, with the purpose of performing different kind of financial analysis. By analyzing the data, we can understand the volatility, seasonal effects, trends, liquidity etc., make predictions and take appropriate decisions. These kind of analysis are useful to better understand the financial environment, improve financial management and control and support decision making.

Index Terms— financial analysis, data mining, knowledge discovery, prediction, time series.

1 Introduction

Public Finance institutions, responsible for the management of the financial activity of the General Government, require sophisticated financial information systems to perform their day to day activities as well as financial analysis for decision making. In general, financial operations are performed in transactional information systems, which use normalized relational databases. The relational model is composed of:

- Entities: objects that the institution needs to analyze;
- Relationships: describe how entities interact with each other;
- Attributes: represent entities characteristics; and
- Operations such as insert, delete and update, which are very fast due to the normalized form of the database [1].

These type of systems can produce various kind of reports, mainly to track and monitor the financial activity, financial statements, fiscal reports, statutory reports and so on. An effective and efficient financial management process, requires constant financial analysis performed on the financial information, which include not only actual accounting data, but also financial data from other sources, historic data and planned and forecasted data. In order to perform analysis, these data need to be integrated in a single environment. Transaction information systems can hardly store big volumes of data; they are usually designed to store data related to their activity and not as repositories. The queries performed for a specific analysis, are usually very complex and require retrieving of data from joining many tables. The more tables are included in a query, the more aggregations are made (calculations such as summarizing, aver-age, etc.), the longer will it take for the query to complete. On one hand, analysis generation time will depend on the complexity of the queries it needs to perform, and on the other hand, this will impact significantly the performance of the transactional system itself, which main goal is to perform transactions quickly and effectively. For these reasons, transactional information systems are not suited for complex analysis.

The most appropriate solution for conducting financial analysis and support decision making is the use of business intelligence system. In general, Business Intelligence (BI) is defined as: "an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance" [2].

In [3] and [4] the authors identify that a good BI system should provide these tools: end-user reporting, query and reporting, OLAP, dashboard/screen tools, data mining tools, and planning and modeling tools.

The financial data that needs to be analyzed consist mainly on numerical values of revenues and expenditures (actual, planned and forecasted), recorded at equal time intervals and are categorized as time series data. Time series databases are widely used in many applications such as market stock analysis, economic forecasts, budget analyzes, and so on.

The development of statistical methods has produced a set of data analysis techniques that are useful to confirm predefined hypothesis. Such techniques are, however, inadequate in the process of discovering new correlations and dependences between data, which on the other side, grow in quantity, dimensions and complexity.

The global process of information analysis and processing, with the purpose of extracting knowledge to support decisions, is known as Knowledge Discovery (KD). The KD process is characterized by the following steps:

- Data Selection: During this phase is performed the identification and extraction of the data that are needed for the analysis;
- Preprocessing: During this phase the data selected for analysis undergo the de-noising process;
- Data transformation: During this phase the selected data are transformed in forms for appropriate mining;
- Data Mining: During this phase intelligent methods and algorithms are applied to the data to identify and extract patterns;
- Knowledge presentation: During this phase visualization and knowledge representation techniques are used to present mined knowledge to users;

Data mining is the process of exploring and analyzing large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data, through different methods and techniques.

Data mining techniques of time series data are the most used and efficient techniques for analyzing financial times series, so this study is focused on the data mining techniques for financial time series.

2 Data Mining Techniques for Financial Time Series Data

The purpose of data mining for time series data is the extraction of all meaningful knowledge from the data. A time series contains some or all of the following components:

1. Trend - the overall direction of the series, ascending or descending over time;
2. Seasonality - regular variations in the time series that is caused by re-occurring events [5];
3. Random component - additional fluctuations in the series that may be attributed to noise or other random events.

Therefore, time series data can be classified in five main categories: seasonal, trended, seasonal and trended, stationary and chaotic [6].

In order to gain knowledge from financial time series data, they have to undergo a series of processes, that we will describe in the following sections

2.2 Identifying the Data

This is the first phase of knowledge discovery and has a particular importance for gaining the needed knowledge. In this phase it is performed the so called business understanding, where managerial need for new knowledge and business objectives are analyzed and clear goals are defined. Following the business understanding, the next activity to be performed is to identify the relevant data and the data sources. Careful identification and selection of data and data sources should be performed, as well as identification of the most relevant variables, in order to make it easier for data mining algorithms to discover useful and accurate knowledge. After the selection of the data we can move to the next phase of KD.

2.3 Preprocessing

Usually, financial time series data contain noises, that needs to be removed in order to perform data mining. Noises can be defined as random errors or variances of a measured variable [7]. There are some techniques that can be used to remove noises, that we will be analyzed below.

2.3.1 Smoothing

This is one of the most popular methods for de-noising in time series data. By averaging historical values, it can identify random noise and separate it to make the data pattern consistent [6]. This is a simple technique and very accurate for short-term predictions.

In general, smoothing techniques use as a bases the concept of moving average. In this section we will introduce some of the most used methods for the predictions of financial time series data.

Simple Moving Average (SMA): SMA is calculated from the value itself and its neighbors, which can be ahead or behind in the series. In our study we will consider the values before the current value. The number of previous values included is often referred to as the “window” or “lag”, so if we want to consider the current value and four previous ones this would be considered a simple moving average of lag 5 (SMA5).

Weighted Moving Average (WMA): A simple moving average assigns equal importance to all data points being averaged, however if this is considered unsuitable a higher weighting can be applied to certain data points elevating their importance in the average and thus generating a weighted moving average [8].

Exponential Moving Average (EMA): EMA is an extension of the weighted moving average [9]. This method gives greater importance to the recent data, making the averaged values closer to reality. Weighting factors descend exponentially, resulting in the emphasis falling on the recent values though not discarding the older ones totally. This is the main method used for smoothing financial time series data.

Holt-Winters Smoothing Models: This method uses three parameters α, β, γ, one for the level, one for the trend and one for the seasonality, that define the smoothing degree. Initially, an α value is used to dictate the amount of smoothing, putting more emphasis to the recent data. Then, if the trend is present in the data set, a second smoothing is applied using the value β and it is called double smoothing. Finally, if the seasonal component is present in the data set, a third smoothing is performed using the γ parameter.

2.3.2 Auto-Regression

Regression refers to the study of the impact of known variables (independent) on an unknown variable (dependent). In time series, previous data have impact on the current data and this is of great importance especially in financial time series data. Auto-Regression refers to the prediction of the current values, using the previous values. In this model, the value of a variable of interests is predicted using a linear combination of the previous values of the same variable. The term Auto-Regression indicates that it is a regression of the variable of itself. Auto-regression includes several models, such as:

Auto-Regressive Moving Average (ARMA) – This model combines the moving average with auto-regression. In general terms, an ARMA(p,q) model uses the previous p values with auto-regression and the moving average derived from the last q values.

Auto-Regressive Integrated Moving Average (ARIMA): This is the most commonly used method in financial time series data. In financial time series data, the trend component is usually present. In order to account for trend, they are initially transformed in stationary data sets, and modeling is then performed on the transformed data after which they are returned to the initial state. One method for removing trend is differencing [10]. Differencing method replaces the actual values with the values of the differences between them.

2.3.3 Finding Similarities

This technique is used to remove noises from financial time series data, through finding similar patterns. Various techniques can be used such as: the longest common subsequence, Euclidean and time warping distance functions, and indexing techniques such as nearest neighbor classification. These techniques find similar patterns and can be used to facilitate pruning of irrelevant data from a given dataset [11].

Searching for similarities results in finding data sequences that differ very little from the sequence of given queries. Many query of similarities in time series find a set of sequences containing subsequences that are similar to the sequence of a given query. Searching for similarities requires data or dimension reductions and transformation of time series data. Through this technique we are able to identify similar objects even though they are not mathematically identical, emphasizing the most visible features of these objects. Similarity measures techniques are consistent with human intuition and are universal in the sense that they allow to identify or distinguish arbitrary objects. They abstract from distortions and are invariant to a set of transformations.

Similarity measures techniques can be classified in four main categories: shape-based distances, which compares the general shape of the times series; edit-base distances which compares two time series based on the minimal number of operations required to transform one series into the other; feature-based distance, which compares the features of the times series using any kind of distance function; and structure-based distance, which compares the higher-level structures found in the times series.

2.4. Data Transformation

In order to provide a strong foundation for subsequent data mining, time series data needs to be transformed into other formats [12]. Using data transformation, a continuous time series can be discretized into meaningful subsequences and represented using real-valued functions [13]. The three most commonly used techniques for the transformation of the financial time series are:

Fourier Transformation: This method uses sinus and cosinus functions to decompose the input values into harmonic wave forms.

Wavelet Transformation: Wavelet is a powerful data reduction technique that allows prompt similarity search over high dimensional time series data [14].

Piecewise Linear Representation: It refers to the approximation of time series 'T', of length 'n' with 'K' straight lines. Such a representation can make the storage, transmission and computation of data more efficient [15].

2.5. Data Mining

In this section we will describe the data mining techniques that can be used for financial time series. The following tasks may be used individually or simultaneously depending on the knowledge discovery we want to achieve.

2.5.1. Query by Content

Query by content includes data mining techniques that retrieve a set of solutions that are similar to the query provided by the user. The user should as well specify the amount of similarity, which is called similarity measure. Similarity measures can be calculated using the Euclidian distance function or other alternative distance functions. The query can be performed on the whole database (in the general case), or the user may want to specify a threshold to retrieve time series whose similarity with the defined query does not exceed the threshold. This technique is highly dependent on the data, and in many cases users need only a set of solutions by defining the number of series it should contain, without knowing how far they will be from the query. This need can be fulfilled by applying the K-Nearest Neighbor algorithm where the user specifies the similarity measure and the K number of series that are most similar to the given query.

Query by content can be used to reduce time series dimensionality, handle scaling and gaps [16], noise, query constraints and time warping.

The trend of query by content is being focused on streams lately. In his research Kontaki [17] treated both cases of time series: static and streaming where he proposes the use of an incremental computation of Discrete Fourier Transform to adapt to the stream update frequency. Whereas Lian and Chen in [18] proposed the polynomial, DFT, and probabilistic approaches, to predict future unknown values and answer queries based on the predicated data.

2.5.2. Clustering

Clustering is an important data mining technique that classifies common groups called clusters. Clustering can be defined as the process of partitioning a set of data in subsets, where data in a subset is a cluster, in a way that they are similar to each other and different from the data of the other clusters. The series in a cluster are as similar to each other as possible within each cluster. Time series clustering methods are partitioning, hierarchical and model based.

A partitioning method constructs k partitions of the data, where each partition represents a cluster containing at least one object. Clustering of time series is useful for data streams, which can be implemented by using auto-regressive models (which was described previously), k-mean algorithm, where each cluster is represented by the mean value of the objects in the cluster and the k-medoids algorithm, where each cluster is represented by the most centrally located object in a cluster [19].

A hierarchical clustering method creates a hierarchical decomposition of objects. The hierarchical method may be agglomerative or divisive, depending on how hierarchical decomposition will occur. The agglomerative approach, otherwise referred to as the bottom-up approach, starts by placing each object in a single cluster. It then merges clusters that are similar to each other in larger cluster, until all objects are in a single cluster or until a certain condition is met (highest level of hierarchy). The divisive approach, otherwise referred to as the top-down approach, starts by having all objects in a cluster. At each subsequent iteration, the cluster is divided into smaller cluster, until each object is in a single cluster.

A model-based approach assumes a model for each of the clusters and attempt to best fit the data to the assumed model.

Neural networks are widely used in the model-based approach using Adaptive Resonance Theory (ART) and Self-Organizing Maps (SOM).

In order to apply a clustering technique to a time series database, we should first calculate the distance or similarity between the time series that will be compared. Several distance/similarity measures that can be used in financial time series data were explained in section 2.2.3.

2.5.3. Classification and Regression

Classification is a data mining tool that distinguishes series based on a known model and assigns the series to the correspondent model. It learns patterns from past data (a set of information-traits, variables, features-on characteristics of the previously labeled items, objects, or events) in order to place new instances (with unknown labels) into their respective groups or classes [20]. Time series classification task consists in training a classifier and label new time series for e given labeled set of time series. There are several techniques that can be used where the most popular in financial time series are Piecewise Representation (explained in the previous section), that it is robust to noise; nearest neighbor algorithm with Dynamic Time Warping are widely used as classifiers and ARMA models. Whereas classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions. That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels. The term prediction refers to both numeric prediction and class label prediction. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data [21].

Regression analysis is a predictive modeling technique that investigates the relationships between dependent and independent variables and is widely used for time series modeling. Classification and regression may need to be proceeded by other relevant analysis in order to identify valuable attributes for classification and regression [21].

2.5.4. Segmentation

The segmentation task is to create approximate time series, by reducing its dimensions, while the series retains its key features. There are many algorithms that can be used to accomplish temporal segmentation, but we will describe the three major algorithms used for segmenting financial time series:

Sliding window: In sliding windows, an initial segment of the time series is first developed and then is grown until it exceeds some error threshold [22]. A full segment is created until the whole series is analyzed. The Sliding Window algorithm works by anchoring the left point of a potential segment at the first data point of a time series, then attempting to approximate the data to the right with increasing longer segments. At some point i, the error for the potential segment is greater than the user- specified threshold, so the subsequence from the anchor to i-1 is transformed into a segment. The anchor is moved to location i, and the process repeats until the entire time series has been transformed into a piecewise linear approximation. This is a simple technique, but has shown poor performance with many real-life datasets [15].

Top-Down: This algorithm takes into account every possible partition of the time series. The time series is recursively partitioned until some stopping criteria is met. The Top-Down algorithm works by considering every possible partitioning of the times series and splitting it at the best location. Both subsections are then tested to see if their approximation error is below some user-specified threshold. If not, the algorithm recursively continues to split the subsequences until all the segments have approximation errors below the threshold [15].

Bottom-Up: This algorithm complements the Top-Down algorithm. It begins by creating finest possible approximation of the time series, so that n/2 segments are used to approximate the n length time series. It calculates the cost of merging neighboring segments. It merges the lowest cost pair and continues to do so until a stopping criterion is met [15].

2.5.5. Prediction

Time series are usually very long and smooth, meaning that subsequent values are within predictable ranges of one another [23]. Prediction is the process of modeling variable dependencies to forecast the next values in the series. It uses time series representation and finding similarities as well as statistical components such as model selection and statistical learning.

For performing this task auto-regressive models are commonly used. Other more complex approaches include neural networks and cluster function approximatio

3. Conclusion

In this research we analyzed shortly the need for advanced tools and techniques to perform financial analysis of the data and concluded that the most appropriate solution would be the use of data mining. Financial data are an obvi-ous example of time series, therefore, our study was based on the data mining techniques and algorithms that can be used for mining financial time series data, with the purpose of preparing meaningful analysis and extract knowledge from the available data. We analyzed the knowledge discov-ery processes and presented the most appropriate tech-niques, that can be used separately or simultaneously de-pending on the type of analysis that the end user needs to perform. We presented the techniques that can be used to preprocess the time series in order to make them ready for mining application. We then focused on the data mining tasks, analyzing both data mining styles: descriptive and predictive. Depending on the final purpose, a user can use one or the other, or even a combination of both.

References

[1] M.K. Sandhu, A. Kaur, R. Kaur, Data Warehouse Schemas, International Journal of Innovative Research in Advanced Engineering (IJIRAE), 4(2) 2015.

[2] GARTNER. Business Intelligence (BI) [online] http://www.gartner.com/it-

glossary/business-intelligence-bi

[3] S. Rouhani, S. Asgari, and S. V. Mirhosseini, Review Study: Business Intelligence Concepts and Approaches, American Journal of Scientific Research. 50. pp. 62-75.

[4] W. W. Eckerson, What Are Performance Dashboards, Available at http://bpmpartners.com/documents/Chapter1Excerpt.pdf.

[5] So, M. K. P. and R. S. W. Chung. Dynamic seasonality in time series. Computational Statistics and Data Analysis, 70(0), 2014. pp. 212 – 226, 2014.

[6] P. Cortez, M. Rocha, & J. Neves, Evolving time series forecasting neural network models. In Proceedings of the Third International Symposium on Adaptive Systems, Evolutionary Computation, and Probabilistic Graphical Models, 2001, pp. 84-91.

[7] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques Third Edition.

[8] L. Stevens, Essential Technical Analysis: Tools and Techniques to Spot Market Trends. Wiley Trading. Wiley, 2002. ISBN 9780471273813

[9] K. Ord, Charles holt's report on exponentially weighted moving averages: an introduction and appreciation. International Journal of Forecasting, 20(1), 2004, pp. 1 - 3.

[10] T. C. Mills, The Foundations of Modern Time Series Analysis. Palgrave Macmillan, 2011. ISBN 9780230290181.

[11] M. Vlachos, D. Gunopulos, & G. Das, (2004). Indexing time series under conditions of noise. In M. Last, A. Kandel, & H. Bunke (Eds.), Data Mining in Time Series Database (Series in Machine Perception and Artificial Intelligence Volume 57) (pp. 67-100). New Jersey, USA: World Scientific.

[12] A. Lerner, D. Shasha, Z. Wang, X. Zhao & Y. Zhu, Fast algorithms for time series with applications to finance, physics, music, biology and other suspects. In Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 2004, pp. 965-968. New York, USA: ACM Press.

[13] Chung, F.L., Fu, T.C., Ng, V. & Luk, R.W.P. An evolutionary approach to pattern-based time series segmentation, IEEE Transactions on Evolutionary Computation, 8(5), 2004, pp. 471-489.

[14] I. Popivanov, RJ. Miller, Similarity search over time series data using wavelets. In Proceedings of the Eighteenth International Conference on Data Engineering, 2002, pp. 212-221

[15] E. Keogh, S. Chu, D. Hart & M. Pazzani, Segmenting time series: a survey and novel approach. In M. Last, A. Kandel, & H. Bunke (Eds.), Data Mining in Time Series Database, Series in Machine Perception and Artificial Intelligence Volume 57, 2004, pp. 1-21.

[16] M. Vlachos, D. Gunopulos & G. Kollios, Discovering similar multidimensional trajectories. In Proceedings of the 18th International Conference on Data Engineering. IEEE Computer Society, 2002, pp. 673–684.

[17] M. Kontaki, A. Papadopoulos & Y. Manolopoulos, Adaptive similarity search in streaming time series with sliding windows. Data & Knowledge Engineering, 63, 2007, pp. 478–502.

[18] X. Lian and L. Chen, Efficient similarity search over future stream time series. IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 1, 2008, pp. 40–54.

[19] T. W. Liao, Clustering of time series data—a survey, Pattern Recognition 38 (2005) 1857 – 1874.

[20] E. Turban, R. Shadra, D. Delen, D. King, A managerial approach to understanding business intelligence systems.

[21] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques Third Edition.

[22] H. Shatkay, S. Zdonik, Approximate queries and representations for large data sequences. In Proceedings of the 12th International Conference on Data Engineering, 1996, pp. 536–545.

[23] D. Shasha, Y. Zhu, High Performance Discovery in Time Series: Techniques and Case Studies. Springer, 2004