Stream Time Series Approach for Supporting Business Intelligence

Van Vo[1,2], Luo Jiawei[1,*] and Bay Vo[3]

[1]*School of Information Science and Engineering, Hunan University, China.*
[2] *Faculty of Information Technology, Ho Chi Minh University of Industry, Vietnam.*
[3]*Information Technology College, Ho Chi Minh, Vietnam.*
luojiawei@hnu.edu.cn, vothithanhvan@hui.edu.vn, vdbay@itc.edu.vn

Abstract

Business intelligence has an important role in effective decision making to improve the business performance and opportunities by understanding the organization's environments through the systematic process of information. This paper proposes a novel framework based on data mining technologies for making a prediction of business environment. We present a business intelligence model to predict the business performance by using dimensionality reduction as preprocessing data then applying Sequential Minimal Optimization based on the Support Vector Machine algorithm to generate future data. To examine the approach, we apply them on stock price data set obtained from Yahoo Finance.

Keywords: *dimensionality reduction, stream time series, business intelligence predictive analytics, knowledge management*

1. Introduction

Business intelligence contains a set of ideas about methods and procedures in order to improve business performances and decisions, using information from multiple sources and applying past experience to develop an exact understanding of business dynamics. There are two important issues of business intelligence, the first issue is gathering, analysis and distribution of information. The second objective is supporting the strategic decision-making process. In the technical view of business intelligence, it usually includes the processes or applications and technologies for collecting, storing and analyzing data, and for providing access to data to help management better business decisions.

Predictive analytics help organizations look forward then makes some decisions that suitable for the future needs. The combined knowledge is used to take many activities that can improve business management. Predictive analytics has a lot of techniques of data mining and statistics that examine current and historical values to make predictions about future data. The information is important to have decision-making. Time series predictive analytics become an interesting and important research area due to its frequent appearance in many distinct applications, especially in business intelligence.

In recent times, the increasing use of time series [4, 22] data has activated various researches in the field of data and knowledge management. Time series data are described as large, with high dimensionality and that needs continuous update. Moreover, the time series data are usually considered as a whole instead of individual numerical fields. Time series research includes these tasks such as indexing, classification, clustering and representation of time series.

* Corresponding Author. Tel: +86 731 888 21971.

Business intelligence (BI) is the process of transforming raw data into useful information for more effective strategic, operational insights, and decision-making purposes so that it yields real business benefits. We are interested in creating a business intelligence framework with dimensionality reduction by matching predefined sample in order to reduce the number of data points before applying the prediction techniques. It is a suitable approach for the stream data environment, which supports of low prediction memory usage and high accuracy of the future values. So in this paper, we intended a short introduction to BI with the emphasis on getting historical data, then implement the reducing unimportant point technique of time series stocks, apply prediction technique Sequential Minimal Optimization (SMO) [15, 30] based on Support Vector Machine (SVM) [2, 27, 29] in order to get future data. The experiment extracts and evaluates the accuracy of future values depends on the number of history using for prediction. To attest the effectiveness of our approach we use the stock closing price series sets getting from Yahoo Finance.

The paper is organized as follows. In the next section we present some related researches to business intelligence with prediction analytics and the associated problems with time series. Section 3 describes the proposed algorithm and pseudo-code design, clarifying the process of testing and training prediction model. Section 4 then presents details of our approach including these problems: getting and filtering data, preprocessing data by dimensionality reduction and using historical data to predict future data. Section 5 presents the evaluations for our approach. Finally we summarize and conclude this work in Session 6.

2. Related Work

Business intelligence refers to a management philosophy and tool that help organizations manage and refine business information to make effective decisions. The first meaning of business intelligence is related to information and knowledge of the organization [23], which describe the business environment. In additional, business intelligence is a systematic process by which organizations achieve, analyze and distribute the info [13, 16].

The technology categories of business intelligence mainly include data warehouse or data mart, On-Line Analytical Processing (OLAP), and data mining [17]. More specifically, data warehouse or data mart [26] is the fundament infrastructure to save historical data, and data mining is its main component to discover trends, identify patterns and analyze data, while OLAP is the set of front-end analyzing tools. With regard to data warehouse and OLAP, we can consider them as components for next-generation database systems. Comparing to developing software packages, researchers in industrial informatics and enterprise systems tend to be more interested in applying business intelligence in the industrial environment [23].

Based on the time series analysis, different mining tasks can be found in the literature and they can be roughly classified into four areas: pattern discovery and clustering, classification, rule discovery and summarization [4, 22]. Some research issues concentrate on one of these areas, while the others may focus on more than one of the above processes. The fundamental problem is how to represent the time series data [17, 25, 28]. Mostly, there are many kinds of time series data related research, such as finding similar time series , subsequence searching in time series [7, 21], dimensionality reduction and segmentation [28] .One of the common approaches is transforming the time series to another domain for dimensionality reduction [19, 25] followed by an indexing mechanism. These researches have been studied in considerable detail by both database and pattern recognition communities for different domains of time series data.

Dimensionality reduction is the process of reducing the number of variables or points under specific consideration. Dimensionality reduction is one of the most important

preprocessing procedures for analyzing a stream time series environment. There are some typical methods for time series dimensionality reduction in order to represent time series in lower dimensional spaces including the Discrete Fourier Transform, Discrete Wavelet Transform, Piecewise Linear Approximation, Piecewise Aggregate Approximation, Singular Value Decomposition and Adaptive Piecewise Constant Approximation (APCA) [4, 12]. Time series are highly correlated data, so that, the representation techniques use a scheme that aims at reducing the dimensionality of time series by projecting the original data onto lower dimensional spaces and processing the query in those reduced spaces. This scheme is widely used in time-series data mining literature. Some data points may contribute to the overall shape of the time series while others may only have little influence on the time series or they may even be discarded. These points are therefore more important than other data points in the time series. Several approaches are based on important points such as Landmark points, Extreme points and Perceptually Important Points (PIP) [5, 28].

One of the common statistical time series predictive analytics approaches is the Autoregressive Integrated Moving Average (ARIMA) [9, 14, 18]. Box and Jenkins proposed a general ARIMA model to cope with the modeling of non-stationary time series. In the past, various approaches have been presented for time series prediction, including fuzzy-based paradigm, neural-net computing model, neural fuzzy hybrid system, and others. Recently, related works on predicting values involve hybrid techniques for example neuro-fuzzy system with autoregressive integrated moving average. Mehdi et al proposed NFS-ARIMA [14] with the new computational intelligence approach to have better performance for time series prediction. Wong, *et al.*, [20] suggested an adaptive time variant prediction model based on window size of fuzzy time series. Joshi and Kumar [11] presented a fuzzy time series model based on non-determinacy index by incorporating intuitionistic fuzzy sets.

Recent predictive analysis is developing in the direction of data mining and intelligence computing. For data mining prediction techniques, Men et al. apply the Least Squares Support Vector Machines [27] method based on time series to actual load forecasting. This methodology based on time sequence can assure higher accuracy and faster convergence speed as compared to the other traditional time series method and it also can discover the global optimal solution. Yang, *et al.*, [30] provide the approach of support vector regression (SVR) based on the Sequential Minimal Optimization [15] algorithm to build the model and predict single time series by mining computation.

3. Problem Statement and Definitions

The technical approach considers business intelligence as a set of algorithms and techniques that supports the process of saving, recovery, manipulation and analysis of data and information. However, in the overall view of our approach, there are two important issues in this research. The first is the gathering, analysis and distribution of information. The second is predictive analysis to support the strategic decision-making process.

Definition 3.1: A time series as a sequence of variables, $x_1, x_2, x_3 \ldots x_n$, where the variable x_i marks the value taken by the series at the specific time point (y_i). In the case, the total number of data points in the time series is known in advance, this time series is called static,

and that time series has a length *n*. If the data points are arrived at continuously, the value of *n* represents the number of data points seen in the time series so far, which is, the so-called time series streaming.

Definition 3.2: Stream time series is defined as a set of time series data, T_1, $T_2 \ldots$ and T_n, each time series T_i including *m* ordered points at the current timestamp *(m-1)*, that is, $T_i = \{t_{i0}, t_{i1}, \ldots, t_{i(m-1)}\}$ where t_{ij} is the point at timestamp *j* in T_i.

Let's assume that *n* stream time series only receive data after the timestamps *m*. Otherwise, for each time series T_i, the future values t_{im}, $t_{i(m+1)}$, …, and $t_{i(m+F-1)}$ fitting to timestamps *m*, *(m+1)*, …, and *(m+F-1)*, respectively, arrive in a batch manner at the same timestamp *(m+F)*. At the period from timestamp *m* to *(m+F-1)*, the system doesn't know about *F* future values in each time series.

Definition 3.3: Pattern matching in this research is implemented by sequential scanning in order to find in the original time series stream which matched the pre-defined samples. Given a query sequence $Q(q_1, \ldots q_i)$ and a set of data sequences $P(p_1, \ldots, p_j)$ extracted from a time series dataset, we desire to recognize a subset of *P* that is similar to *Q* with conditions.

Definition 3.4: The predict error is defined as the difference between actual value and predict value for the corresponding period ($predict\ error_t = actual\ value_t - predict\ value_i$).

The accuracy of an *n* time series stream T_1, $T_2 \ldots T_n$ will use the average of each time series. The average is computed with the Equation (1), where E_i and *n* are the predict error and number of time series stream.

$$\overline{E} = \frac{1}{n}\sum_{i=1}^{n} E_i \tag{1}$$

Definition 3.5: We consider that business intelligence model is a framework for gathering the historical data, filtering the necessary data and using them to predict future value. This model helps to improve the performance of the organization.

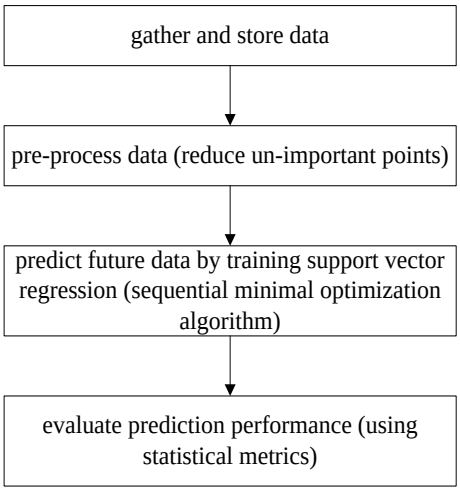


Figure 1. Overview of proposed approach for business intelligence

```
Gathering and storing data function
  Repeat
    Choose stock stream time series
    Choose the time start/end
        Save Ti= {ti0, ti1, …, ti(m-1)}
  Until i = N
Preprocessing data function
  define patterns for eliminating un-important points
  for each time series stream Ti
    eliminating un-important points
    begin
      for each time series stream Ti
        if (the condition Ti[a:b] of matching is sastify)
            choose points and eliminate others
        end if
      end for
    end eliminating function
    save preprocessing data T'i = {t'i0, ti1, …, ti(m-1)}
  end of each time series stream
Prediction process (provide future data depend on history
data)
  input: C, K(xi, yi) (Gaussian Kernel), kernel parameters,
epsilon
  for each stock stream time series
    (1) Initialize b and all α's to 0
    (2) Repeat until KKT [10] satisfied:
      (2.1) Find an example e1 that violates KKT
      (2.2) Choose a second example e2.
        if that fails to result in change,
          randomly choose unbound example.
        if that fails,
          randomly choose example.
        if that fails,
          re-choose e1.
      (2.3) Update α1 and α2 in one step
    (3) Compute new threshold b
  end for
Evaluation prediction performance for decision making
  choose the accuracy approach
  compare the result with actual data
```

Figure 2. The pseudo-code of proposed algorithm

Business intelligence is the ability to perform all its capabilities and then convert them into knowledge of an organization. Business intelligence technologies provide historical, current and predictive prospects of business working performances. With our approach, the workflow is shown in Figure 1, proposed pseudo code is in Figure 2 and the notations are in Table 1. Our approach has four main procedures: gathering then preprocessing data, future prediction and evaluation. The first procedure is data gathering form Yahoo Finance website, with this step we choose the name of finance stock companies to get the historical data for a range of time. The second procedure is data preprocessing and aims to reduce the un-important points of stock stream. In this procedure, we apply the pattern matching with a defined pattern for

each frame to reduce the dimensions. The following procedure is the training and testing of the support vector regression to predict future values. In this process, we apply SMO based on SVM in order to reduce the memory storage of dynamic programming. In the last one, we evaluate the accuracy performance based on indicators of statistical methods.

Table 1. The symbols and their descriptions

Symbol	***Description***
T_i	the i^{th} time series stream ($T_{i,} \in \{T_1, \ldots T_n\}$ $i=1..n$)
T_i'	the i^{th} time series stream after applying preprocessing data
n, p	the number of time series before and after preprocessing data
m	the number of data points in the stream ($t_{i0,} t_{i1,} \ldots t_{i(m-1)}$)
H, F	the number of historical data used for prediction and predicted value
C	regularization constant of support vector machine regression
$K(x_i, y_i)$	Kernel function (Gauss Kernel)
α, ξ	Lagrange multiplier
$T_i[a:b]$	The subsection of T_i from a to b, $t_a, t_{a+1}, \ldots, t_b$

4. Our Approach

4.1. Business Intelligence

Currently, most of the businesses have implemented business intelligence to improve their decision making. Many business intelligence models are available and each model has different type of requirement. In our approach, the term business intelligence can be explained as a collection of approaches for gathering, storing, analyzing data that help users to gain insights and make better fact-based business decisions.

Storing data is concerned with making sure the data are filed and stored in appropriate ways to ensure it can be found and used for analysis. The advantages of modern databases are that they allow multi-dimensional formats so we can store the same data under different categories – also called data marts or data warehouse access layers.

We apply a single data repository that allows analysis over data. Gathering data can come in many formats and basically refers to the automated measurement and collection of performance data. Gathering data is concerned with collecting or accessing data which can then be used to inform decision making. With this research in Figure 3, we gather stock price data from Yahoo Finance website then filter the necessary data for the prediction period,. A main problem of gathering data is making sure that the related data is gathered in the right way at the right time.

The next component of business intelligence is analyzing the data. Here we take the data that has been gathered and inspected, transformed and modeled it in order to gain new insights. Data analysis comes in many streams, we consider both quantitative and qualitative.

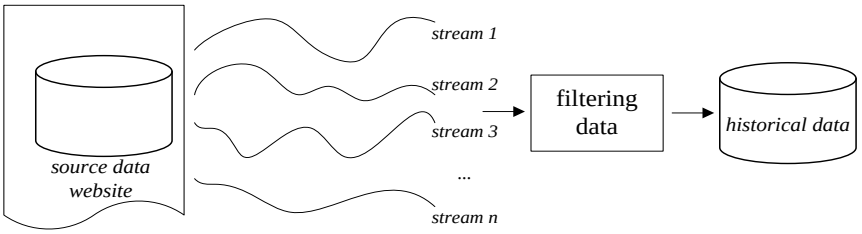


Figure 3. Gathering, filtering and storing data

4.2. Reducing points by matching sample

With the aim of making less time and memory for implementing the framework, we apply the pattern matching technique [6] for reducing some points of time series data, these points are called unimportant points. Suppose that the original time series T_i changed to T_i' after matching and eliminating some points. Our method for choosing and eliminating points is not only based on the fluctuation parameter λ_v but also the time duration parameter λ_t. The λ_v is defined as the average of these value points in time during. The time duration parameter is defined as a sliding window with equals the number of successive points, λ_t = w = 5 as in the Figure 4.

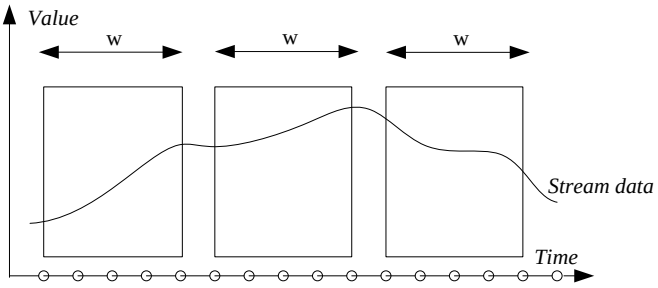


Figure 4. Example of pattern matching width w = 5 successive points

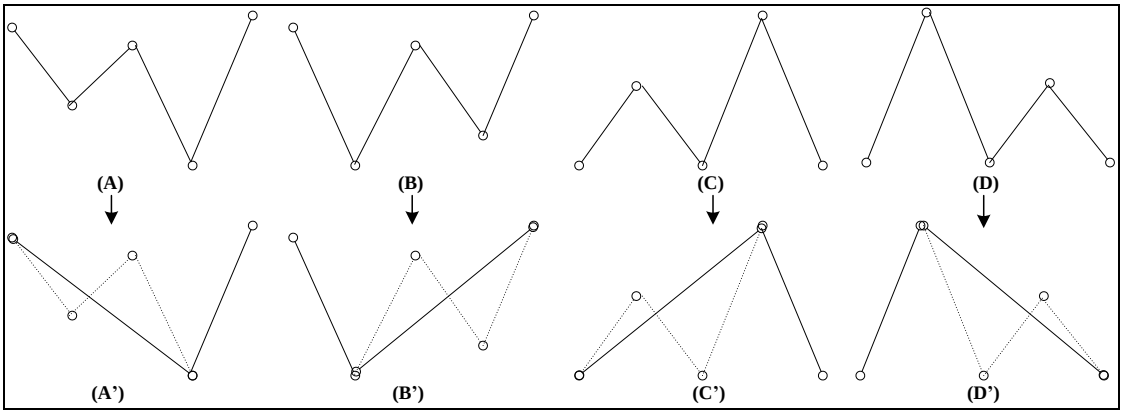


Figure 5. Samples for matching and reducing points

In a stream environment with the solving many time series streams, for a given time series $T_i = \{t_{i1}, t_{i2}, , t_{im}\}$, a window in the examination process of time series T_i is defined at time period j of the i^{th} stream with the width w. The changes (decrease or increase) are shown in

the Figure 5. There are four cases which satisfy both λ_t and λ_v parameters, it means we consider both specific of the fluctuation and time duration. Our strategies for eliminating the points that are not important are shown in detail on the Figure 5. To explain more about our solution, we supposed that the sliding window had 5 successive points p_F, p_2, p_3, p_4, p_L and we examine check if p_3 < *average of* (p_2, p_3, p_4) and (($p_2 < p_4$) *or* ($p_2 > p_4$)) then we keep p_F, p_4, p_L and eliminate p_2, p_3. And another checking is satisfied if (p_3 < *average of* (p_2, p_3, p_4) and (($p_2 > p_4$) *or* ($p_2 < p_4$)) then we keep p_F, p_3, p_L and eliminate p_2, p_4.

4.3. Stream time series prediction

First, SVM estimates the regression of a set of linear function that are defined in a high dimensional feature space. Second, SVM carries out the regression estimation by risk minimization, where the risk is measures using Vapnik's ε-insensitive loss function. And SVM implements the structural risk minimization principle which minimizes the risk function consisting of the empirical error and regularized term. Given the training sample set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ $(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$, where R^n is the space of input sample, R is the space of output sample, N is the total number of training samples.

The fundamental idea of support vector regression [27, 29, 19] is to map the vector x (input vector) into high dimensional feature space by nonlinear mapping function Φ and then to perform linear regression in the feature space. This transformation is realized by Kernel function $K(x_i, y_i) = \Phi(x_i) \cdot \Phi(y_i)$. Kernel function may be Gaussian, polynomial or neural network non-linearity. If Kernel is linear, it can be written as follows:

$$f(x) = w.\phi(x) + b \tag{2}$$

$\Phi : \chi \rightarrow H$, $w \in H$, b is a threshold value. The $\Phi(x)$ represents the high-dimensional feature space which is nonlinearly mapped from the input space. It depends on the way we choose Kernel parameters, the predicted result also evaluated by statistical metrics.

The goal is to find an optimal weights w and threshold b as well as to define the criteria for finding an optimal set of weights. The coefficients w and b are estimated by minimizing

Minimize: $$E(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{3}$$

SVMs have been applied successfully to both *classification* and regression tasks. Support vector machine has many ways to optimize the quadratic programming, SMO is the best way to resolve that problem. The reason we applied the prediction with SMO is the performance reducing satisfaction, SMO algorithm just calls the kernel matrix iteration, therefore the performance is improved significantly. This implementation will reduce the main memory at run time.

The prediction using the SVM algorithm makes the operation speed a bit slower in a big data set environment. Particularly in our approach, the stream time series environment with many time series streams $T_1, T_2, \ldots, T_n$ with $T_i = \{t_{i0}, t_{i1}, t_{i2}, \ldots t_{i(m-1)}\}$, if each one of the time series puts its own kernel matrix into the main memory, the primary memory will be overflow easier.

SMO is an iterative algorithm for solving the optimization problem of the SVM. In general SMO algorithm implements the dividing problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers, the smallest possible problem involves two such multipliers. Then, for any two multipliers ξ and ξ^*, the constraints are reduced to:

$$\begin{aligned} 0 \le \xi, \xi^* \le C \\ y_1\xi + y_1\xi^* = k \end{aligned} \tag{4}$$

where C is an SVM hyper parameter and this reduced problem can be solved analytically. Our algorithm proceeds several steps as follow:

1. *Kernel function will be chosen as a Gauss Kernel function.*
2. *Choose the C and γ*
3. *For each time series:*

   Step 1. Find out the first Lagrange multiplier ξ that violates the Karush-Kuhn-Tucker (KKT) [10] condition for optimization the problem.

   Step 2. Select the second multiplier ξ^*, then optimize the pair of multipliers (ξ, ξ^*).

   Step 3. Repeat the steps 1 and 2 until convergence.

At the time two Lagrange multipliers, ξ, ξ^* satisfy the KKT conditions, it has meaning that the problem has been solved. Although the SMO algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers to accelerate the rate of convergence.

We select the first Lagrange multiplier by using the external loop of the SMO algorithm to enable the Lagrange multiplier to optimize. Choosing the second Lagrange multiplier is according to maximizing the step length of the learning during joint optimization. |E1 - E2| is used to approximate the step size in SMO [15, 30].

Support that $K(x_i, y_i) = \Phi(x_i) \cdot \Phi(y_i)$. Kernel function is known as a function of the input space. Accordingly, the dot product in the feature space is equivalent to the Kernel function of the input space. Thus, instead of directly to the value of the scalar product, we made indirectly through Kernel function. For the stock data a non-linear transformation, nonlinear Gaussian function (RBF-Radial Basis Function) can be chosen as the Kernel function.

$$K(x_i, y_i) = \exp\left(-\gamma \left\| x_i - y_i \right\|^2\right) \tag{5}$$

4.4. Predictive analysis evaluation

Prediction accuracy is the technique to measure the exactness of predictive analysis. An obvious way to assess the quality of the learned model is to see on how long term the predictions given by the model are accurate. The objective of prediction system is to efficiently predict the *n.F* values for the *n* time series streams with the *predict error* as low as possible and the accuracy as high as well.

We use statistics to evaluate the simulation effect and predictive validity of the prediction model. The prediction performance is evaluated using the following statistical metrics,

namely, the normalized mean squared error (NMSE), mean absolute error (MAE). NMSE and MAE are the measures of the deviation between the actual and predicted values. The smaller the values of NMSE and MAE, the closer are the predicted time series values to the actual values (a smaller value suggests a better predictor) [1, 3, 24].

In order to ensure the accuracy of prediction results, the model must be evaluated accurately and its performance generalized before it can be used to predict. The prediction error is defined as the difference between the actual value and the predicted value for the corresponding period. In this paper, the accuracy of *n* streams time series $T_1, T_2 \ldots$ *and* T_n will use as the average.

5. Experimental Evaluation

5.1. Experimental Environment and Dataset

The experimental dataset used the financial stock time series data. In the gathering data process we filtered then stored daily closing stock prices from Yahoo Finance (http://finance.yahoo.com). The experiments were implemented on Windows XP operating system with a 2G AMD PC and 2 GBs of main memory.

We tested our approach with six different stock companies (HBC-HSBC Holding plc, UN-Unilever NV, MSFT-Microsoft Corporation, AIG - American International Group, Inc., HMC-Honda Motor Co. Ltd., SYMC-Symantec Corporation, PEP-Pepsico, Inc.) and 189 stock ticker's daily closing price during the time from 01 January 2012 to 01 October 2012. After the preprocessing process we implement the training and testing sample in order to predict future values.

5.2. Experimental Results and Analysis

This section talks about the experiments and results in this research to evaluate the proposed framework. Our discussion is managed in the following three directions: the gathering and storing historical data, the reducing points by matching predefined samples and the prediction future values. Our works implement and evaluate SMO regression on financial stock data, we use the short-term prediction with daily data.

In the Figure 6, we present the historical data which took from Yahoo Finance. With financial stock data, standard deviation is a representation of the risk associated with price-fluctuations. Hence with low standard deviation indicates that the data points tend to be very close to the mean value, and with high standard deviation indicates that the data points are spread out over a large range of values. For these reasons, the Table 3 provides common statistical of daily stock prices of 6 different companies, the standard deviation of our testing daily stock streams are low, so the predicted values more accurately than other cases. The Figure 7 shows the data after eliminating some points by matching predefined samples. It seems after the preprocessing process, the time series stock companies still keeps the shape of the original trends.

The second experiment, we employ daily observations of difference closing stock price streams covering the period 01 January to 01 October in order to predict future values for next

7 days. To evaluate the prediction process, we must use some statistical method. Note that, the characteristics of financial data are data over time except weekends and some special days.

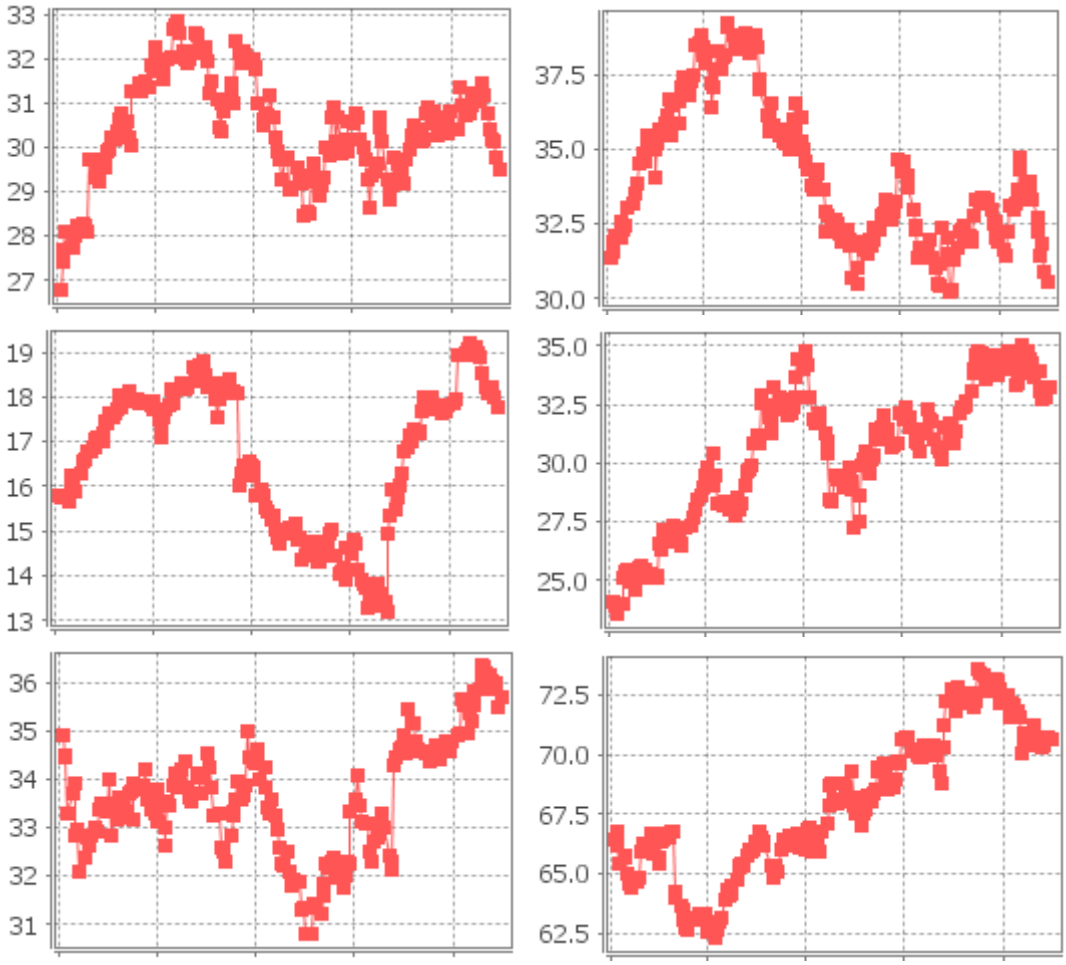


Figure 6. The historical data obtained and filtered from finance website

The mean absolute percentage error (MAPE) measures the accuracy of fitted time series values. It expresses accuracy as a percentage. The mean absolute error (MAE) is the average of the absolute value of the error. The effective improvement in prediction output growth by basing on stock price changes is indicated by lower MAE and MAPE. Table 4 shows both results of MAE and MAPE values in each step-ahead. Each step, MAE is calculated by sum of ($observation_i$ - $prediction_i$)/the number of training data sample. Support vector regression is based on the structural risk minimization (SRM) principle which looks for the minimize an upper bound of the generalization error rather than minimize the empirical error implemented in other methods. Based on SRM principle, support vector regression will have an optimum network structure by striking the right balance between the empirical error and the VC-confidence interval.

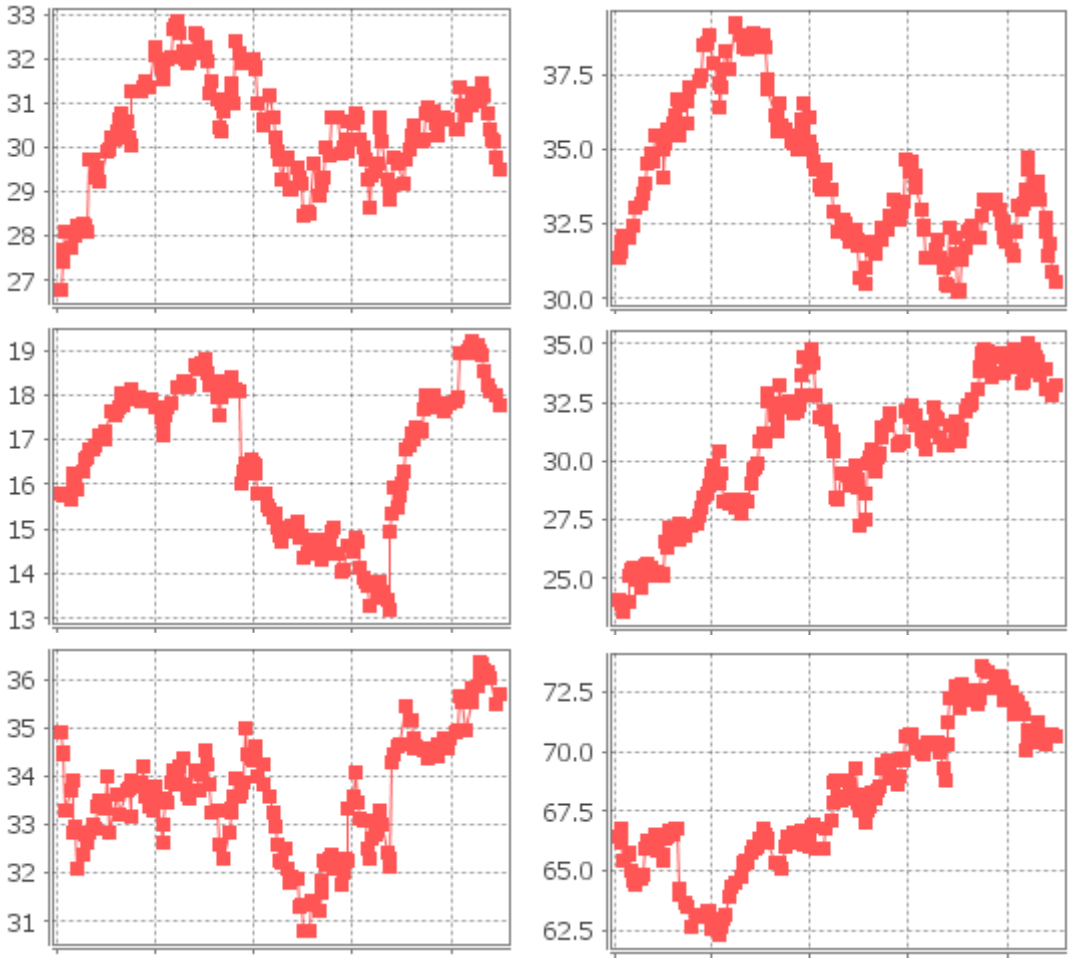


Figure 7. The data after eliminating points by matching predefined pattern

Table 3. Statistics of daily stock streams for prediction

Stock Stream	*Min*	*Max*	*Mean*		*Standard Deviation*	
			Original	*After preprocessing*	*Original*	*After preprocessing*
UN	30.79	36.35	33.608	*33.569*	1.207	*1.214*
MSFT	26.77	32.85	30.428	*30.45*	1.204	*1.183*
AIG	23.54	35.02	30.431	*30.489*	2.925	*2.912*
HMC	30.21	39.2	34.125	*34.015*	2.411	*2.333*
SYMC	13.18	19.2	16.661	*16.638*	1.634	*1.625*
PEP	62.28	73.58	67.922	*68.007*	3.051	*3.011*

The size of the test set is typically about 20% of the total sample, although this value depends on how long the sample is and how far ahead we want to forecast. With time series predictive analysis, one step predicts may not be as relevant as multi step predicts. First we select observation i^{th} for the test set, treat the remaining observations as the training set, compute the error on the test observation. We repeat the first step N times with $i = 1, 2, ..., N$ where N is the total number of observations. Then we calculate the accuracy measures of prediction based on the errors obtained. The SMO algorithm has several input parameters such as complexity C, kernel function, kernel parameters and epsilon. Kernel function used in the experiment is Gaussian RBF. With this experiment we used difference complexity parameter C, the result of six steps ahead is shown in Table 4. In the Table 5, we present the accurate evaluation of first step-ahead of six stock companies and the average of accuracy error as Equation (1).

Table 4. The measurements of six steps ahead with different of support vector regression parameters of MSFT stock closing stream

Step ahead	***1st***	***2nd***	***3rd***	***4th***	***5th***	***6th***
N	***151***	***150***	***149***	***148***	***147***	***146***
C=1.0 and γ = 0.01						
MAE	0.3913	0.4601	0.5135	0.5542	0.5813	0.6092
MAPE	1.284	1.5092	1.6839	1.8166	1.9047	1.9955
C=10.0 and γ = 0.001						
MAE	0.3918	0.4626	0.5177	0.5606	0.5886	0.6167
MAPE	1.2861	1.5181	1.6981	1.8384	1.9296	2.0213
C=10.0and γ = 0.01						
MAE	0.2947	0.3984	0.4489	0.4752	0.4974	0.5122
MAPE	0.9672	1.306	1.4707	1.5551	1.6266	1.6745

Table 5. The accuracy evaluation of first step-ahead (C=1.0 and γ = 0.01)

	MSFT	***HMC***	***SYMC***	***AIG***	***UN***	***PEP***	$\overline{E}$
MAE	0.3913	0.5971	0.3122	0.6495	0.3955	0.5957	0.4902
MAPE	1.284	1.7472	1.9363	2.103	1.1849	0.8811	1.5227

In the Figure 9, we show the predictive value of our approach and the actual value retrieved from Yahoo Finance. This figure includes one line of actual value, two lines of predicted values which different parameters of the SMO algorithm based on SVM [2]. If the larger parameter C of SMO, predicted values and the actual values are nearly identical in a short term prediction.

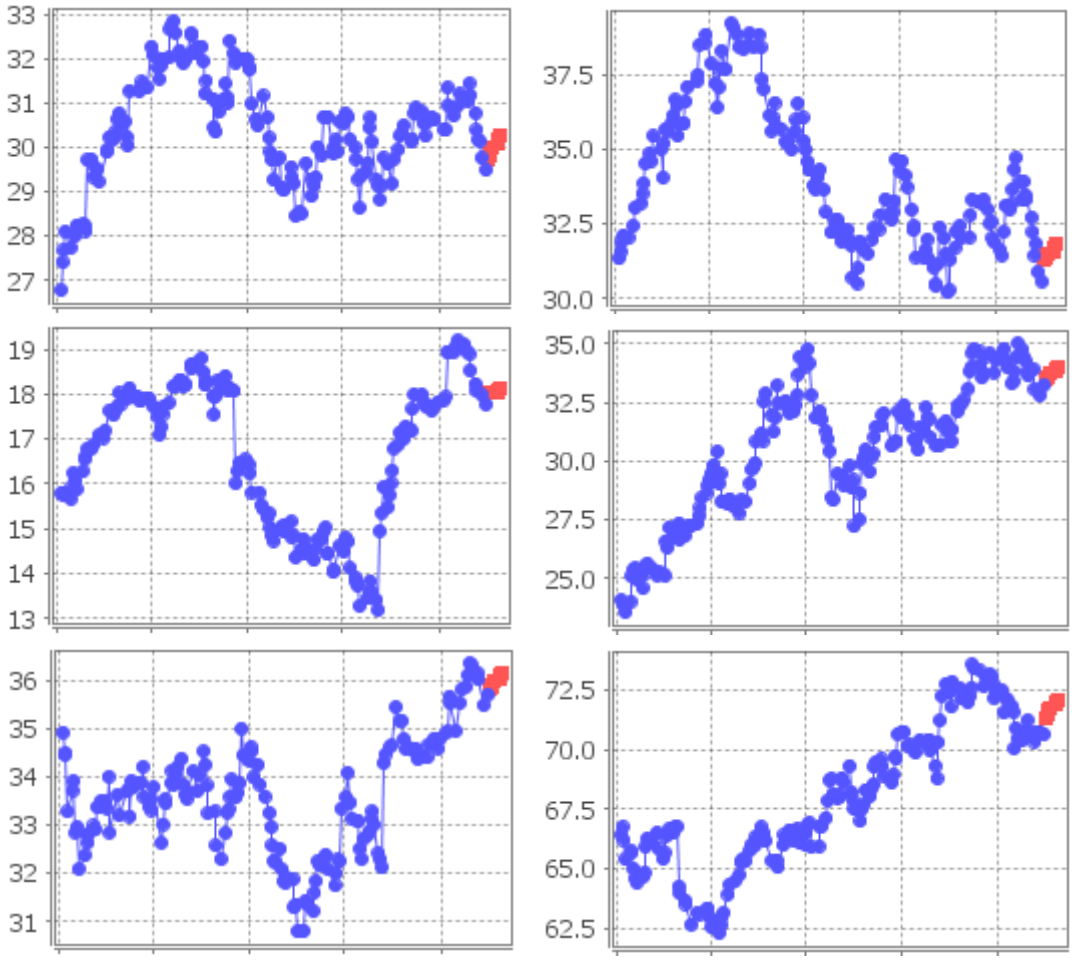


Figure 8. The historical values and future values which computed by SMO with parameters C= 1.0 and γ =0.01

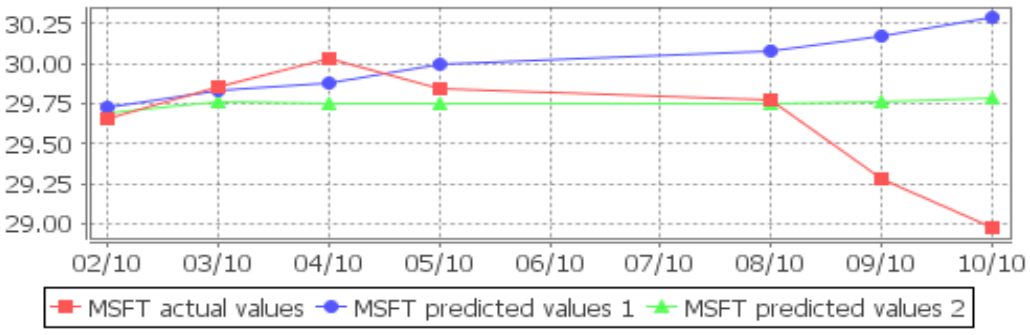


Figure 9. The comparison of actual values and predicted values

6. Conclusion and Future Work

We have studied the problem of time series environment for supporting business intelligence with these tasks: gathering, filtering and storing then preprocessing before use them for prediction purpose.

This paper proposes a framework about stream time series for supporting business intelligence. The prediction future values used the SMO technique based on the optimization of support vector regression and provided the evaluation indicators of accuracy and generalization. We also reduce the large historical data to a smaller data set of matching predefined samples so our performance improves noticeably. After the reducing points method based on the pattern matching predefined samples, the time series generated by our approach still keeps the shape of the original trends. The approach has a meaningful in the environment where the data set is large. Therefore, we applied this work on the stock price data set obtained from Yahoo Finance.

In the future, we propose to provide the complete system for continuous time series. In addition, we plan to extend the current work towards a system pattern discovery, such as in similarity search and finding motifs techniques in the stock time series data.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant no. 61240046) and the Science and Technology Planning Project of Hunan Province (Grant no. 2011FJ3048). We would like to thank the University of Waikato for providing machine learning open source [8].

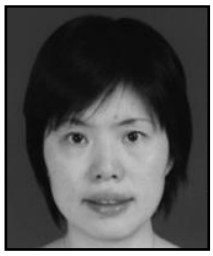
References

[1] K. P. Burnham and D. R. Anderson, "Multimodel Inference: Understanding AIC and BIC in model selection", Sociological Methods and Research, vol. 30, no. 2, **(2004)**, pp. 261-304.
[2] L. J. Cao and F. E. H. Tay, "Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting", IEEE Transactions on neural networks, vol. 14, no. 6, **(2003)**, pp. 1506-1519.
[3] Z. Chen and Y. Yang, "Assessing forecast accuracy measures", http://www.stat.iastate.edu/preprint/articles/2004-10.pdf, **(2004)**.
[4] T. Fu, "A review on time series data mining", Engineering Applications of Artificial Intelligence, vol. 24, no. 1, **(2011)**, pp. 164-181.
[5] T. C. Fu, F. L. Chung, R. Luk and C. -M. Ng, "Representing financial time series based on data point importance", Engineering Applications of Artificial Intelligence, vol. 21, no. 2, **(2008)**, pp. 277-300.
[6] T. C. Fu, F. L. Chung, R. Luk and C. -M. Ng, "Stock time series pattern matching: template-based vs. rule-based approaches", Engineering Applications of Artificial Intelligence, vol. 20, no. 3, **(2007)**, pp. 347-364.
[7] L. Hailin, G. Chonghui and Q. Wangren, "Similarity measure based on piecewise linear approximation and derivative dynamic time warping for time series mining", Expert Systems with Applications, vol. 38, no. 12, **(2011)**, pp. 14732-14743.
[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and H. Ian, "The WEKA data mining software: An update", SIGKDD Explorations, vol. 1, no. 1, **(2009)**, pp. 10-18.
[9] R. J. Hyndman and Y. Khandak, "Automatic Time Series Forecasting: The forecast Package for R", Journal of Statistical Software, vol. 27, no. 3, **(2008)**, pp. 1-22.
[10] S. -T. John and S. Shiliang, "A review of optimization methodologies in support vector machines", Neurocomputing, vol. 74, no. 17, **(2011)**, pp. 3609–3618.
[11] B. P. Joshi and S. Kumar, "Intuitionistic fuzzy sets based method for fuzzy time series forecasting", Cybernetics and Systems: An International Journal, vol. 43, no. 1, **(2012)**, pp. 34-47.
[12] E. Keogh, K. Chakrabarti, M. Pazzani and S. Mehrotra, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases", Knowledge and Information Systems, vol. 3, no. 3, **(2001)** , pp. 263-286.

[13] A. Lönnqvist and V. Pirttimäki, "The measurement of business intelligence", Information Systems Management, vol. 23, no. 1, **(2006)**, pp. 32–40.
[14] K. Mehdi and B. Mehdi, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting", Applied Soft Computing, vol. 11, no. 2, **(2011)**, pp. 2664-2675.
[15] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines", Advances in Kernel Methods, MIT Press Cambridge, **(1999)**, pp. 185-208.
[16] J. Ranjan, "Business justification with business intelligence", The Journal of Information and Knowledge Management Systems, vol. 38, no. 4, **(2008)**, pp. 461–475.
[17] S. Rouhani, M. Ghazanfari and M. Jafari, "Evaluation model of business intelligence for enterprise systems using fuzzy TOPSIS", Expert Systems with Applications, vol. 39, no. 3, **(2012)**, pp. 3764-3771.
[18] C. -C. Wang, "A comparison study between fuzzy time series model and ARIMA model for forecasting Taiwan export", Expert Systems with Applications, vol. 38, no. 8, **(2011)**, pp. 9296-9304.
[19] Q. Wang and V. Megalooikonomou, "A dimensionality reduction technique for efficient time series similarity analysis", Information Systems, vol. 33, no. 1, **(2008)**, pp. 115-132.
[20] W. K. Wong, E. Bai and A. W. C. Chu, "Adaptive time variant models for fuzzy time series forecasting", IEEE Transaction on Systems, Man and Cybernetics-Part B: Cybernetics, vol. 40, no. 6, **(2010)**, pp. 1531-1542.
[21] L. Xian, C. Lei, X. Y. Jeffrey, H. Jimsong and M. Jian, "Multiscale representations for fast pattern matching in stream time series", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 4, **(2009)**, pp. 568-581.
[22] G. E. P. Box and G. Jenkins, "Time series analysis: Forecasting and control", Holden-Day 3rd Edition, Prentice-Hall: New York, **(1994)**.
[23] V. Carlo, "Business Intelligence: Data Mining and Optimization for Decision Making", John Wiley & Sons, **(2009)**.
[24] V. N. Vapnik, "The nature of statistical learning theory", New York: Springer **(1995)**.
[25] A. Camerra, T. Palpanas, J. Shieh and E. Keogh, "iSAX 2.0: Indexing and mining one billion time series", In Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), **(2010)**, pp. 58-67.
[26] A. Martin, L. T. Miranda and V. V. Prasanna, "An Analysis on Business Intelligence Models to Improve Business Performance", In Proceedings International Conference on Advances in Engineering, Science and Management, **(2012)**, pp. 503-508.
[27] D. -Y. Men and W. -Y. Liu, "Application of least squares support vector machine (LS-SVM) based on time series in power system monthly load forecasting", In Proceedings of Power and Energy Engineering Conference (APPEEC), **(2011)**, pp. 1-4.
[28] S. -H. Park, J. -H. Lee, S. -J. Chun and J. -W. Song, "Representation and clustering of time series by means of segmentation based on PIPs detection", In Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 3, **(2010)**, pp. 17 – 21.
[29] J. R. Wang and X. L. Deng, "Selecting training points of the sequential minimal optimization algorithm for support vector machine", In Proceedings of the 2nd International Conference on Control, Instrumentation and Automation (ICCIA), **(2011)**, pp. 456-458**.**
[30] J. F. Yang, Y. J. Zhai, D. P. Xu and P. Han, "SMO algorithm applied in time series model building and forecast", In Proceedings of the 6th International Conference on Machine Learning and Cybernetics, vol. 4, **(2007)**, pp. 2395-2400.

Authors

Van Vo has completed M.Sc. in Computer Science from Faculty of Information Technology under University of Science, Vietnam National University of Ho Chi Minh, Viet Nam. At the moment, she is a PhD candidate in School of Information Science and Engineering, Hunan University, Republic of China. Her researches are machine learning, knowledge management and related data mining problems.



Luo Jiawei is a full professor and vice dean at the School of Information Science and Engineering, Hunan University, Changsha, Republic of China. She holds Ph.D, M.Sc. and B.Sc. degrees in Computer Science. Her research interests include data mining, network security and bio informatics. She has a vast experience in implementing national projects on Bioinformatics. She has authored many research articles in leading international journals.

Bay Vo received his Ph.D degrees in Computer Science from the University of Science, Vietnam National University of Ho Chi Minh, Vietnam in 2011. His research interests include association rules, classification, mining in incremental database, distributed databases and privacy preserving in data mining.