Computing and Informatics, Vol. 35, 2016, 111–127

TIME SERIES TREND ANALYSIS BASED ON K-MEANS AND SUPPORT VECTOR MACHINE

Van Vo

Faculty of Information Technology
Industrial University of Ho Chi Minh City, Vietnam
e-mail: vttvan@iuh.edu.vn

Jiawei Luo

School of Information Science and Engineering
Hunan University, China
e-mail: luojiawei@hnu.edu.cn

Bay Vo*

Division of Data Science
Ton Duc Thang University, Ho Chi Minh City, Vietnam
&
Faculty of Information Technology
Ton Duc Thang University, Ho Chi Minh City, Vietnam
e-mail: vodinhbay@tdt.edu.vn, bayvodinh@gmail.com

Abstract. In this paper, we apply both supervised and unsupervised machine learning techniques to predict the trend of financial time series based on trading rules. These techniques are K-means for clustering the similar group of data and support vector machine for training and testing historical data to perform a one-day-ahead trend prediction. To evaluate the method, we compare the proposed method with traditional back-propagation neural network and a standalone support vector machine. In addition, to implement this combination method, we use the financial time series data obtained from Yahoo Finance website and the experimental results also validate the effectiveness of the method.

* corresponding author

Keywords: Machine learning, time series trend analysis, support vector machines, k-means clustering

Mathematics Subject Classification 2010: 68-T05

1 INTRODUCTION

Time series is a collection of monitoring data items identified clearly through repeating measurements from time to time. The studies about time series analysis [1, 2, 3] are attracting the interest of many researchers in both scientific and practical applications. Recently, the time series predictive analysis is of a great importance in many domains of science and engineering, such as finance, electricity, environment, and ecology.

Time series predictive analysis can be examined as a matter of a procedure to establish a mapping between the input and output data. The time series prediction can be divided into two categories depending on the predicted time period: short term and long term. The short term prediction is related to one-step-ahead prediction; the goal of a long term prediction is to predict values for several steps ahead.

Machine learning is one of the artificial intelligence type which gives computers the ability to learn without explicit programming. Supervised and unsupervised learning are two common techniques of machine learning with building and studying ideas that can learn from data. Supervised learning generates a function that maps inputs to outputs, these functions are also called labels because they are usually supplied by labelling the training examples. In the classification issue, the learner estimates a mapping function into classes by looking at the input and output examples. Unsupervised learning designs a set of inputs, like clustering. In this case, the labels are not known during training.

Support vector machines (SVMs) [4, 5, 6, 7, 8] are gaining the popularity due to their many attractive features and their promising empirical performance. Since the decision surface of the SVM is parameterized by a large set of support vectors and accompanying weights, the machine is considerably slower in the test phase than other learning machines such as neural network and decision trees. Being a universal learning machine, the support vector machine (SVM) suffers from expensive computational cost in the test phase due to a large number of support vectors, and greatly impacts its practical use.

The clustering study plays an important role in data analysis and pattern classification. It has many applications in data compression, data mining and so on. Clustering problem has a purpose to partition a set of data points into non-overlapping subsets. In the past several decades, many efficient clustering algorithms [9, 10] have been developed. Among these developed clustering algorithms, the K-means algorithm is the oldest and the most popular one due to its simplicity and effectiveness.

To implement the combination idea of clustering and training support vector machines for prediction, we are interested in the method with the combination of clustering K-means and SVMs. This method comes from the problem that it merges an unsupervised learner with a supervised learning algorithm, while eliminating the need for labeled training instances for SVMs learning. In this paper, we propose a study with the input value samples to be clustered by K-means; the similar samples are in the same cluster. Then, each cluster will be trained and tested by SVMs to predict output values. This method accelerates significantly the response of SVMs classifiers by reducing the number of support vectors. The experiment has three parts, first we cluster the similarity data to several groups depending on the initial value of K, and then we implement the training and testing historical samples in the K clusters for prediction, after that we make the performance comparison of our method with the traditional SVM and artificial neural networks (ANN). To attest the effectiveness of the proposed framework, we use the stock price data set obtained from Yahoo Finance.

The rest of this paper is organized as follows: The predictive analysis and clustering related works are listed in Section 2. Section 3 presents our combined research on time series clustering by K-means and time series trend analysis by SVM. In Section 4, the empirical results are summarized and discussed. Finally, in Section 5, we conclude our work and propose our future works.

2 RELATED WORK

Time series analysis becomes an interesting and important research area due to its frequent appearance in many distinct applications, especially in finance studies [11]. In recent times, the increasing use of time series data has launched various researches in the field of data and knowledge management. Based on the time series analysis, different mining tasks can be found and classified into four areas: pattern discovery and clustering, classification, rule discovery and summarization [1]. Some research issues concentrate on one of these areas, while the others may focus on more than one of the above mentioned processes.

Clustering methods can be broadly divided into three main categories [9, 10]: overlapping, partitional, and hierarchical. A hierarchical clustering [12, 13] method works by grouping data objects into a tree of clusters. There are generally two types of hierarchical clustering methods: agglomerative and divisive. The clustering procedure finally terminates when the number of iterations exceeds the maximum allowed number of iterations or convergence. A neural clustering method, the self-organizing map (SOM) [14], is used for pattern discovery. Ghaseminezhad [14] presented a novel SOM-based algorithm that can automatically cluster discrete groups of data using an unsupervised method. Hidden Markov model (HMM) is a common model based algorithm adopted in time series clustering [15].

Currently, most of the researches on time series prediction proposed hybrid approaches based on the classical approach aims to reduce time and increase efficiency.

Wichard [16] suggested a hybrid strategy in order to cope with the different seasonal features of the time series. Behnamian et al. [17] proposed hybrid approach that is simply structured, and comprises two components: a particle swarm optimization (PSO) and a simulated annealing (SA). The performance of the proposed method is evaluated using standard test problems and compared with those of related methods in literature, ARIMA and SARIMA [2] models. Statisticians have studied to obtain better forecasts for long years and by these studies hybrid methods have been improved in the literature. Aladag et al. [18] suggested a new hybrid approach combining Elman's Recurrent Neural Networks (ERNN) and ARIMA models.

A hybrid fuzzy time series approach is proposed by Egrioglu et al. [19] in order to reach more accurate forecasts. In this hybrid approach, fuzzy c-means clustering method and artificial neural networks are employed for fuzzification and defining fuzzy relationships, respectively. Ismail et al. [20] provided the combination of least square support vector machine with the self-organizing maps (also known as SOM-LSSVM) for time series prediction.

Stock markets have been researched with many tasks [1] in order to extract the useful patterns and predict their movements in the short term or long term future data. There are various approaches in predicting the movement of stock market and variety of prediction techniques has been used by stock market analysts. These tasks [11] has been growing interest in financial time series prediction in recent years as accurate prediction of financial prices has become an important problem in investment decision making.

3 TIME SERIES TREND ANALYSIS METHOD

3.1 K-Means

Unlike static data, the feature of time series data is values changed over the time, so we chose K-means [2, 10, 12] to apply our method because this algorithm is the most popular method of partition-based clustering, and the Figure 1 shows the brief flowchart of K-means algorithm with time series data. In each iteration, the winner cluster is found and its center is updated accordingly. The initial cluster centers can be chosen in various ways, e.g. chosen arbitrarily or by some sequences. Also, the number of clusters is a critical parameter to be determined. It can be fixed beforehand or can vary during the clustering process. The clustering procedure finally terminates when the number of iterations exceeds the maximum allowed number of iterations or convergence.

While patterns can be directly discovered from time series, a major problem is that time series data mostly increase linearly with time. This will cause the storage needs to increase rapidly and slow down the pattern discovery process exponentially. Therefore, an effective mechanism for compressing the huge amount of time series data, especially historical data, is needed. This not only reduces the size of storage, but also maintains an acceptable level of information for the discovery process.

First, K-means algorithm selects randomly specific number of K centers. Then, the purpose is to pick up each object to the nearest center. At the time all the data objects were included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the earlier formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum. The variants of K-means algorithm differ in several parameters such as the initialization of clusters, the definition of similarity, or the definition of cluster representativeness.

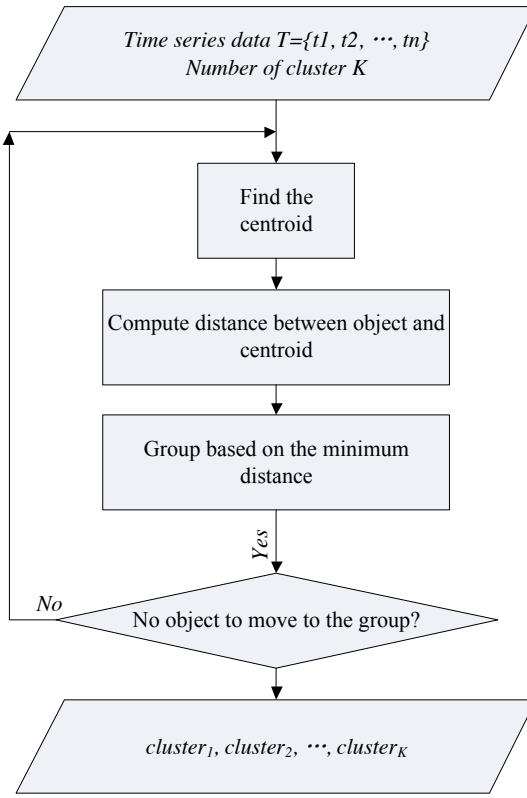

Figure 1. The K-means algorithm with time series analysis

Given a number of clusters K and the data set $T = \{t_1, t_2, \ldots, t_n\}$ with $t_i = (x_i, y_i)$, the flowchart of K-means for time series data is shown in Figure 1. The intuitive idea of K-means is defined as these steps:

Step 1. Choose randomly K instances to be the initial centroids.

Step 2. For each instance, assign it to the cluster the centroid of which is the closest to the instance

Step 3. For each cluster, recompute its centroid based on the instances in that cluster.

Step 4. If the convergence criterion is satisfied, then stop; otherwise, go back to Step 2.

The clustering of K-means stops if one of these criteria is satisfied: no reassignment of instances to different clusters; or no change of centroids; or insignificant decrease in the sum of squared error as the Equation (1).

$$E = \sum_{k=1}^{K} \sum_{x \in C_k} d(x, m_k)^2 \tag{1}$$

where C_k is the k^{th} cluster, m_k is the centroid of cluster C_k, and $d(x, m_k)$ is the distance between instance x and centroid m_k. In our method, Euclidean distance is considered to determine the distance between each object and the cluster centers.

$$d(x, m_k) = \|x, m_k\| = \sqrt{(x_1 - m_{k1})^2 + (x_2 - m_{k2})^2 + \ldots + (x_n - m_{kn})^2} \tag{2}$$

3.2 Support Vector Machine

SVMs are developed based on statistical learning theory given by Vapnik [4] to resolve the issues of data classification and data regression problem [5, 6, 7]. The SVMs are based on the principle of structural risk minimization, which has proved to be more efficient than the empirical risk minimization.

SVMs implement a learning machine algorithm that performs learning from examples in order to predict the values depending on previous data. The goal of support vector regression is to generate a model which will give out prediction of unknown output values based on the known input parameters. In the learning phase, the formation of the model is performed based on the known training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $\{x_i\}$ are input vectors, and y_i outputs associated with them. Each input vector consists of numeric features. In the phase of practical application, the trained model on the basis of new inputs $\{x_1, x_2, \ldots, x_n\}$ makes prediction of output values $\{y_1, y_2, \ldots, y_n\}$.

We use this machine learning technique for proposed method because the features of the SVMs which implement a learning machine algorithm that performs learning from examples in order to predict the values depend on previous data in the prediction model. In this model, the historical and current values of time series are used as the inputs: $\{y(t+1), y(t+2), \ldots, y(t+h)\} = F(y(t), y(t-1), \ldots, y(t-m+1))$ where h represents the number of ahead predictions, F is prediction model and m is a size of regressor.

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 $y_i \in \{-1; +1\}$ for, N is the total number of training samples. SVMs are based on the computation of a linear function in a high dimensional feature space where the

input data are mapped via a nonlinear function. The linear function to distinguish as the following:

$$f(x) = w^T \cdot \phi(x) + b \tag{3}$$

$\Phi X \to H$, $w \in H$, b is a threshold value. The $\Phi(x)$ represents the high-dimensional feature spaces which is nonlinearly mapped from the input space. The purpose of SVMs is to find an optimal hyperplane that the margin between the two classes reaches the maximum value. It means finding an optimal weights w and threshold b as well as to define the criteria for finding an optimal set of weights. Besides, in order to ensure generalization, a slack variable is included to make easier conditions subclass. The coefficients w and b are estimated by minimizing:

$$\min_{w,b,\xi} \frac{1}{2} w^T w^T + C \sum_{i=1}^{N} \xi_i \tag{4}$$

subject to

$$\begin{aligned} y_i(w^T \phi(x_i) + b) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, \forall i \in [1, N], \end{aligned}$$

where $C > 0$ is regularization parameter, ξ is slack variable.

With the Sequential Minimal Optimization [21] algorithm, the Equation (4) can be solved by the problem of quadratic programming

$$\max_{\alpha} L(\alpha) \equiv \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \cdot \phi(x_j), \tag{5}$$

$$0 \leq \alpha_i \leq C, \forall i \in [1, N] \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0$$

where α_i, α_j are Lagrange multipliers

After obtaining the value of the problem in Equation (5), we will have the optimal value w^* and b^* of the hyperplane. Only samples $\alpha_i \geq 0$ are involved in the support vector. Finally, the classification decision function is as Equation (6):

$$f(x) = \text{sgn}(\sum_{i=1}^{N} \alpha_i y_i \phi(x_i)^T \cdot \phi(x_i) + b^*). \tag{6}$$

Support that $K(x_i, y_i) = \Phi(x_i) \cdot \Phi(y_i)$. For the financial time series data, a nonlinear transformation, nonlinear Gaussian function (RBF – Radial Basis Function) can be chosen as the Kernel function in Equation (7).

$$K(x_i, y_i) = \exp\left(-\gamma \|x_i - y_i\|^2\right) \tag{7}$$

More about SVMs can be found in [5, 8]. For the experiments we used a publicly available library LibSVM – A Library for Support Vector Machines [22] which we integrated into our program for the time series analysis.

3.3 Our Method

Our proposed method for time series trend analysis uses and defines some concepts in Section 3.3.1. The details of a combination of machine learning techniques are shown in Section 3.3.2.

3.3.1 Definitions

Definition 1. The time series is series of numerical measurements related through time, $T = \{(t_1, y(t_1)), (t_2, y(t_2)), \ldots, (t_n, y(t_n))\}$ where the variable t_i marks the value taken from the series at the specific time point $y(t_i)$ [2].

If the total number of data points in this series is known in advance, the time series is called static and that time series has a length n. In the case, the data points are arriving 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 2. A time series database **D** is a set of unordered time series possibly of different lengths [2]. The time series database is especially meaningful and useful when dealing with historical data.

Finance time series database has several general characteristics such as: multiple stocks identified by the ticker symbol, having multiple attributes (ticker name, timestamp, open, high, low, close, volume and adjust close).

Definition 3. Trend of a one-day-ahead is the trending of the next day $(t+1)$ of the current date t, we suppose that the trend has three values as upward, downward and no-trend. The trend analysis is the component of a time series that represents variations of low frequency in a time series, the high and medium frequency fluctuations being out. Determination of the trend [23] is done in the following way:

- Closing value must lead (lag) the 25-day moving average.
- 25-day moving average must lead (lag) the 65-day moving average.
- 25-day moving average must have been rising for at least 5 days.
- 65-day moving average must have been rising for at least 1 day.

If the movement cannot be classified as either upward or downward, it means the value is no-trend.

Definition 4. Exponential Moving Average (EMA) [24] is one of the most used indicators in technical analysis. Time series indicator function EMA is a series derived from others; the EMA is calculated with the Equation (8).

$$EMA_N(t) = (\alpha * P(t)) + ((1 - \alpha) * EMA_N(t - 1)) \tag{8}$$

where $P(t)$ is a current price at the time t; α is the smoothing factor $\alpha = 2/(1+N)$; N is a number of time periods.

Definition 5. The trading rules for our method are simply defined as:

- If the predicted trend of the next day = *upward* then *should buy*, else if *already bought* then *should hold*.
- If the predicted trend of the next day = *downward* then *should sell*, else if have not got them then *should not buy*.
- If the predicted trend of the next day = *no-trend* then *should hold*, else if have not got them then *should not buy*.

Definition 6. In the case, the trend is increasing but the predicted value is decreasing, we call it a wrong prediction and vice versa. If this prediction ratio is lower, it means the reliability model is higher. The accuracy of the model is defined below.

$$Accuracy\ (\%) = number\ of\ samples\ correctly\ classified \Big/ \sum samples$$

3.3.2 Proposed Method

The proposed method for time series trend analysis is divided into four main subprocesses as shown in Figure 2. There are several tasks such as data collection, data transformation, training samples and trading rules. The first component is collecting data from finance time series website into time series database D. We focus on possible interactions with financial data and the storage system. The given architecture of data collecting process from website has the essential objective, that is to save and display the historical data with the quality of the data in a short query time. This time series database has many stock symbols, each stock symbol includes several attributes. We use daily closing data for the predicted trend model.

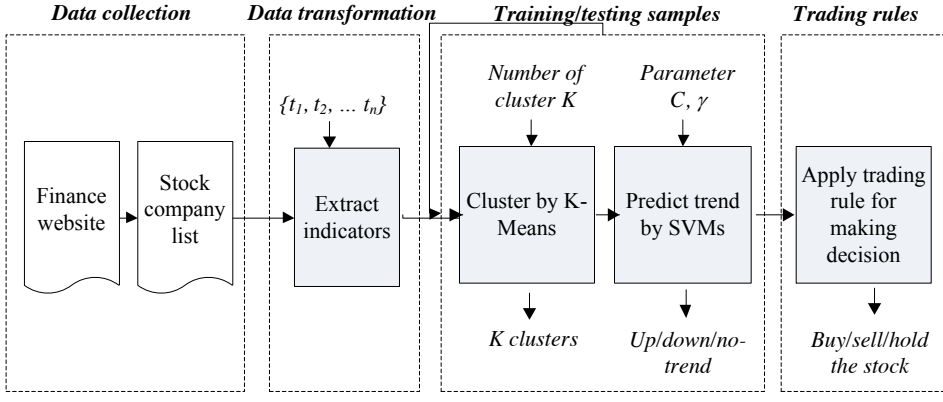

Figure 2. The proposed method of time series trend analysis

The second component of the method is data transformation. The task of data transformation changes the set of values from the data format of a source data

system into the data format of the destination system. In the data transformation process, data extraction is the process of retrieving data out of data sources for further data processing or data storage. Therefore the import into the extracting system is commonly followed by data transformation and possibly the addition of metadata prior to exporting them to another representation in the data processing. In this component, we extract the indicator of time series stock data which was received from financial website. In this study, Exponential Moving Average (EMA) in Definition 4 is used as indicator function. EMA is increasingly preferred by technical analysts over other moving average methods, and the EMA represents an excellent compromise between the excessively sensitive weighted moving average and the overly slow simple moving average.

The main purpose of our method for time series trend analysis is executed in the third component. In this component, we implement the combination method of clustering by K-means algorithm and training samples by support vector machine in order to predict the trend of one-day-ahead data. The result of the combination method would be used for making decisions by the predefined trading rules. K-means algorithm collected samples of the training data to each cluster which has similar characteristics. We choose K-means because this algorithm is a well-known non-hierarchical clustering method and requires the user to assign the number of clusters present in the dataset. The idea of this combination approach takes the advantages of K-SVMeans [28], this is a clustering algorithm for multitype interrelated datasets that integrates the K-means clustering with the SVM. The K-SVMeans is a clustering algorithm K-means for different datasets where clustering together with one data type learns a classifier in another, and the classifiers effect the clustering decisions made by the clusterer.

For each cluster, we train sub-sets with regularization parameters C, γ (cross validation) according to BRF kernel function in Equation (7).

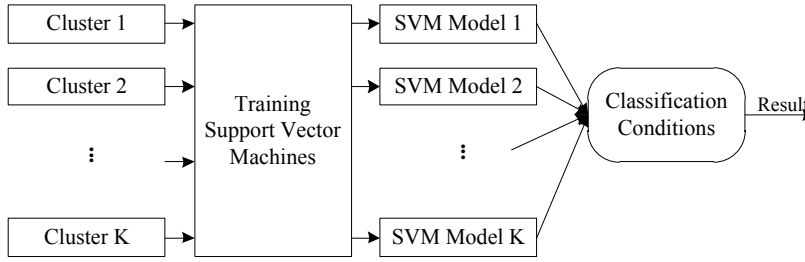

Figure 3. The details of training and testing samples by multi-SVMs

In order to make quickly the training process, we choose "one against one" strategy [25, 26] for multiclass classification SVMs. By the way indicated for a problem with N classes, $N(N-1)/2$ SVMs are trained to distinguish the samples of one class from the total samples of another class. In this method, classification of an unknown pattern is done according to the maximum voting, where each SVM votes for

one class. To implement this strategy, we use LibSVM [22] for multiclass learning problems.

This component applies multi-SVMs which are usually implemented by the combining several SVMs constructs K binary classifiers. The i^{th} classifier output function ω_i is trained taking the examples from ω_i as positive and the examples from all other classes as negative. For a new example x, the "one against one" strategy assigns it to the class with the largest value of ρ_i. The stop condition of the training and testing sample by multi-SVMs will be accepted if the training accuracy of classifier in training process is the best if considering both accuracy and implementation time according to "one against one" strategy. In this case, the output of the process will be the *"SVM model i^{th}"*, then we can choose this SVM model to implement the testing process for making the prediction of trend. Our combination method predicts the trend and outputs three values corresponding the class labels *"increase"*, *"decrease"* and *"no trend"*. To determine the trend of the present day, we used the Definition 3.

Step 1: The input parameters include: the kernel parameter γ, the regularization parameter C, and the number of clusters K.

Step 2: The K-means clustering algorithm is run on the original data and all cluster centres are regarded for building classifiers.

Step 3: SVM classifiers are built on the cluster's data.

Step 4: The input parameters are adjusted by the heuristic searching strategy.

Step 5: Back to the Step 1 to test the new combination of input parameters and stop if the combination is acceptable according to testing accuracy and response time.

4 EXPERIMENTAL EVALUATION

4.1 Experimental Environment and Dataset

The experimental dataset used the financial stock time series data. In the data collection process we obtained daily stock prices from Yahoo Finance [29]. The experiments were implemented on Windows 8 operating system with a 2.4 GHz Intel Core Duo processor and 4 GB of main memory. We tested our method with three different stock companies' data (AAPL – Apple Inc., IBM – International Business Machines Corporation and HPQ – Hewlett-Packard Company). Three different data sets of these time series stock data are shown in the Table 1. This data was used as a training set and testing set for experimental evaluation.

4.2 Experimental Results and Analysis

The most important problem of the SVM classification or regression is finding the appropriate parameters. Two main parameters needed to decide when implement

	Training Set	Testing Set	Training Samples	Total Samples
Set 1	02 Jan 2007 to 11 Jun 2013	12 Jun 2013 to 13 Dec 2013	1 621	1 751
Set 2	02 Jan 2003 to 11 Jun 2013	12 Jun 2013 to 13 Dec 2013	2 628	2 758
Set 3	02 Jan 1993 to 11 Jun 2013	12 Jun 2013 to 13 Dec 2013	5 148	5 278

Table 1. The description of data set for training and testing samples

the machine learning are C and γ. In this experiment, we apply the Grid search methods and the cross through the assessment (5-fold cross validation, 10-fold cross validation, 20-fold cross validation) to find the optimal values for these parameters. Their value is limited to about: $C \in [2^{-5}, 2^{15}]$ and $\gamma \in [2^{-15}, 2^{3}]$ in the case of 5-fold cross validation. In the case of 10-fold cross validation, the C and γ will be replaced in the range $[2^{-4}, 2^{12}]$ and $[2^{-12}, 2^{4}]$, respectively. In the case of 20-fold cross validation, the range of C and γ will change to $[2^{-2}, 2^{8}]$ and $[2^{-8}, 2^{5}]$, respectively.

After data collection and data transformation process, we implement the clustering algorithm with K-means algorithm. In this experiment, we investigate to make the decision for choosing the number of clusters K. In the Table 2, we show the evaluation result of the cases $K = \{2, 3, 4, 5\}$. If the dataset is more divided, we would not have enough information for training the SVM. The results show that choosing $K = 2$ gives the best accuracy of the predictions in the case of small data set (Set 1) and choosing $K = 3$ gives the best accuracy in the case of bigger data set (Set 2 and Set 3).

	2-KMs.SVMs			3-KMs.SVMs		
Data Set #	1	2	3	1	2	3
AAPL	82.5	78.1	79.4	80.6	79.3	79.8
HPQ	78.7	80.2	79.7	77.9	80.8	80.2
IBM	77.2	73.6	73.3	76.5	75.4	74.8
	4-KMs.SVMs			5-KMs.SVMs		
Data Set #	1	2	3	1	2	3
AAPL	80.2	77.9	78.9	79.7	77.2	77.4
HPQ	76.5	79.1	77.8	77.6	79.2	79.1
IBM	75.2	73.1	72.5	72.9	73.4	71.2

Table 2. The accuracy of the combination method according to the number of clusters K

We test the performance on AAPL, IBM and HPQ with the data in Table 1. The ranges include 1 751 samples, 2 758 samples, 5 278 samples in the first, second and the last data set, respectively. We use 1 621 samples, 2 628 samples and 5 148 samples of Set 1, Set 2 and Set 3, respectively, for the testing process. The same 130 samples are applied for each testing of the model. The results of testing set are described in Table 3. The testing model determines the trend of current day and outputs the results such as upward, downward or no-trend.

	Downward			No-Trend			Upward		
Data Set #	1	2	3	1	2	3	1	2	3
AAPL	28	27	27	67	68	68	35	35	35
IBM	68	68	66	50	49	51	12	13	13
HPQ	44	44	42	53	55	53	33	31	35

Table 3. The results of testing samples of data in the Sets 1, 2 and 3

In order to prove the effectiveness and efficiency of the method, we make the comparison with other methods such as the traditional Back Propagation Neural Networks (BPNN) [6, 27] and SVM. BPNN is a three-layer model of an artificial neural network, the learning process is done through the back propagation. We choose Tanh as the activation function in this experiment because this function gives the better accurate results than other activation function for non-linear data transformation. Several parameters of BPNN are used in the testing process shown in Table 4.

Parameters	**Values**
Number of hidden layers for BPNN	3
Input node	10
Hide node	4
Output node	3
Activation function	Tanh
Learning coefficient	0.5
Momentum coefficient	0.1

Table 4. The parameters for BPNN for experimental evaluation

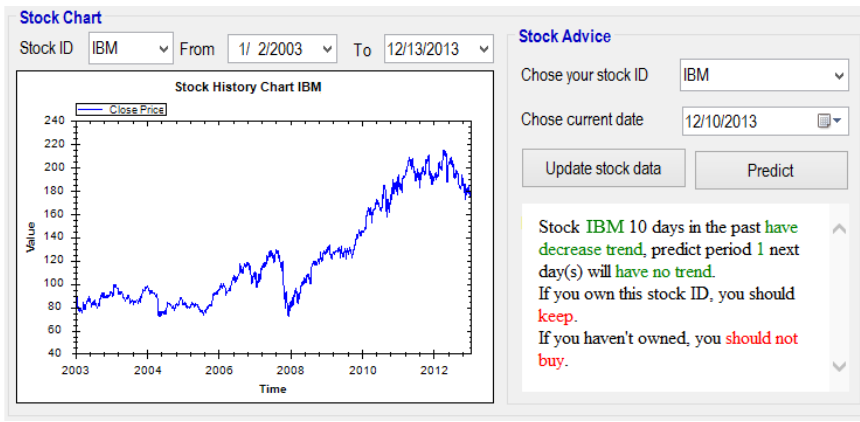

Figure 4. The advice according to the trading rules

With the approach of artificial neural network models, the BPNN requires the large training data set. And the ability of BPNN to specialize is low because of the reason that they happen too often matched by the local optimal value that it achieved. Meanwhile, the SVM algorithm is best evaluated by overcoming the above disadvantages and effectiveness of multi-dimensional data and nonlinear fluctuations of the stock. This measure improvement has brought many good results in different stock symbols.

Depending on the result of trend prediction, we can make a plan with the trading rules in Definition 5. Considering the case when one-day-ahead value of stock symbol is upward, downward or no-trend, the action making will be buying, holding or selling. Figure 4 shows the result of investigating stock symbol GOOG with training and testing data in Set 2 (Table 1). Stock plan making can be defined as the process of getting choices among the possible options, the result of trend analysis can help to make a good choice.

Stock Symbol	**SVMs** Accuracy (%)	**BPNN** Accuracy (%)	**KMs.SVMs** Accuracy (%)
AAPL	81.2	81.9	82.5
IBM	78.1	77.5	78.7
HPQ	76.8	76.1	77.2

Table 5. The result of the combination method according to the number of clusters K

In the Table 5, we listed the results about the accuracy of the stock companies with three approaches: SVM, BPNN, the combination method of K-means and SVM. These results are one-day-ahead prediction values. From the results of Table 5, we can prove that the proposed method has a higher precision value.

5 CONCLUSIONS

We have studied the problem of time series trend analysis with both supervised and unsupervised learning machines. Applying these techniques, we have proposed a method with the combination of clustering K-means algorithm and training SVM algorithm for the problem of trend prediction. This method used K-means for clustering the input data, then from each cluster trained SVM classification to predict the output result of a time series trend such as upward, downward or no-trend. Our experiment results showed that the proposed combination method has a higher accuracy than BPNN or the traditional SVM. In addition, to implement this combination method, we use the financial time series data obtained from Yahoo Finance website and the experimental results also validate the effectiveness of the method.

In the future, we will improve our method by using K-means and SVM with probability estimates in order to have a better accuracy for the trend predictive analysis. In addition, we will build a complete system for the trend prediction with predefinded trading rules.

REFERENCES

[1] Fu, T.: A Review on Time Series Data Mining. Engineering Applications of Artificial Intelligence, Vol. 24, 2011, No. 1, pp. 164–181.

[2] Han, J. W.—Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufman Publishers, 2nd Edition, 2006.

[3] Chounta, I. A.—Avouris, N.: Towards a Time Series Approach for Classification and Evaluation of Collaborative Activites. Computing and Informatics, Vol. 34, 2015, No. 3, pp. 588–614.

[4] Vapnik, V.: Statistical Learning Theory. Wiley, 1998.

[5] Basak, D.—Pal, S.—Patranabis, D. C.: Support Vector Regression. Neural Information Processing, Vol. 11, 2010, No. 10, pp. 203–224.

[6] Turker, N.—Gunes, F.: A Competitive Approach to Neural Device Modeling: Support Vector Machines. International Conference on Artificial Neural Networks, 2006, pp. 974–981.

[7] Cherkassky, V.—Ma, Y.: Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. Neural Networks, Vol. 17, 2004, No. 1, pp. 113–126.

[8] Smola, A. J.—Scholkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing, Vol. 14, 2004, pp. 199–222.

[9] Jain, A.—Dubes, R.: Algorithms for Clustering Data. Prentice Hall, 1988.

[10] Guan, H. S.—Jiang, Q. S.: Cluster Financial Time Series for Portfolio. Proceedings of International Conference on Wavelet Analysis and Pattern Recognition, 2007, pp. 851–856.

[11] Atsalakis, G. S.—Valavanis, K. P.: Surveying Stock Market Forecasting Techniques – Part II: Soft Computing Methods. Expert Systems with Applications, Vol. 36, 2009, No. 3, pp. 5932–5941.

[12] Liao, T. W.: Clustering of Time Series Data – A Survey. Pattern Recognition, Vol. 38, 2005, No. 11, pp. 1857–1874.

[13] Karypis, G.—Han, E. H.—Turker, V.: Chameleon: Hierarchical Clustering Using Dynamic Modeling. Computer, Vol. 32, 1999, No. 8, pp. 68–75.

[14] Ghaseminezhad, M.. H.—Karami, A.: A Novel Self-Organizing Map (SOM) Neural Network for Discrete Groups of Data Clustering. Applied Soft Computing, Vol. 11, 2011, No. 4, pp. 3771–3778.

[15] Bicego, M.—Cristani, M.—Murino, V.: Unsupervised Scene Analysis: A Hidden Markov Model Approach. Computer Vision and Image Understanding, Vol. 102, 2006, No. 1, pp. 22–41.

[16] Wichard, J. D.: Forecasting the NN5 Time Series with Hybrid Models. International Journal of Forecasting, Vol. 27, 2011, No. 3, pp. 700–707.

[17] Behnamian, J.—Fatemi Ghomi, S. M. T.: Development of a PSO-SA Hybrid Metaheuristic for a New Comprehensive Regression Model to Time-Series Forecasting. Expert Systems with Applications, Vol. 37, 2010, No. 2, pp. 974–984.

[18] Aladag, C. H.—Egrioglu, E.—Kadilar, C.: Forecasting Nonlinear Time Series with a Hybrid Methodology. Applied Mathematics Letters, Vol. 22, 2009, pp. 1467–1470.

[19] Egrioglu, E.—Aladag, C. H.—Yolcu, U.: Fuzzy Time Series Forecasting with a Novel Hybrid Approach Combining Fuzzy C-Means and Neural Networks. Expert Systems with Applications, Vol. 40, 2013, No. 3, pp. 854–857.

[20] Ismail, S.—Shabri, A.—Samsudin, R.: A Hybrid Model of Self-Organizing Maps (SOM) and Least Square Support Vector Machine (LSSVM) for Time-Series Forecasting. Expert Systems with Applications, Vol. 38, 2011, No. 8, pp. 10574–10578.

[21] John, C. P.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research, 1998.

[22] Chang, C. C.—Lin, C. J., LIBSVM: A Library for Support Vector Machines, 2001. Available on: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[23] Nair, B. B.—Mohandas, V. P.—Sakthivel, N. R.: A Genetic Algorithm Optimized Decision Tree-SVM based Stock Market Trend Prediction System. International Journal on Computer Science and Engineering, Vol. 2, 2010, No. 9, pp. 2981–2988.

[24] Colby, R. W.: The Encyclopedia of Technical Market Indicators. 2nd Edition. McGraw-Hill, 2003.

[25] Duan, K. B.—Keerthi, S. S.: Which Is the Best Multiclass SVM Method? An Empirical Study. Proceedings of the 6th International Workshop on Multiple Classifier Systems, 2005, pp. 278–285.

[26] Hamamura, T.—Mizutani, H.—Irie, B.: A Multiclass Classification Method Based on Multiple Pairwise Classifiers. International Conference on Document Analysis and Recognition, 2003, pp. 809–813.

[27] Dayhoff, J. E.: Neural Network Architectures. An introduction. Van Nostrand Reinhold Co., 1990.

[28] Bolelli, L.—Seyda, E.—Zhou, D.—Giles. C. L.: K-SVMeans: A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets. Proceedings of International Conference on Web Intelligence, 2007, pp. 198–204.

[29] Yahoo Finance. Available on: `http://finance.yahoo.com`.

Van Vo received her M.Sc. degree in computer science from the Faculty of Information Technology, University of Science, Vietnam National University of Ho Chi Minh, Viet Nam. In 2013, she received her Ph.D. degree in computer science and engineering from the School of Information Science and Engineering, Hunan University, Republic of China. Her researches embrace machine learning, knowledge management and related data mining problems.

Jiawei Luo is 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 bioinformatics. She has a vast experience in implementing national projects on bioinformatics. She has authored many research articles in leading international journals.

Bay Vo is Associate Professor from 2015. He received his Ph.D. degree in computer science from the University of Science, Vietnam National University of Ho Chi Minh, in 2011. He is Dean of Faculty of Information Technology, Ho Chi Minh City University of Technology. His research interests include association rule mining, classification, incremental mining, distributed databases, and privacy preserving in data mining. He serves as an associate editor of the ICIC Express Letters, Part B: Applications (indexed by Scopus and EI), a member of the review board of the International Journal of Applied Intelligence (Springer, indexed by SCI, Scopus and EI), and an editor of the International Journal of Engineering and Technology Innovation. He also served as co-chair of several special sessions such as ICCCI 2012; ACIIDS 2013, 2014, 2015, 2016; KSE 2013, 2014; SMC 2015; as reviewer of many international journals and conferences such as IEEE-TKDE, IEEE-SMC: Systems, Information Sciences, Knowledge-Based Systems, Soft Computing, PLOS ONE, etc. He has published around 70 journal/conference publications including 29 SCI(E) articles.