Expert Systems with Applications 37 (2010) 5993–5998

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Analyzing foreign exchange rates by rough set theory and directed acyclic graph support vector machines

Ping-Feng Pai [a,b,*], Shi-Yu Chen [a], Chao-Wei Huang [a], Ya-Hsin Chang [a]

[a] Department of Information Management, National Chi Nan University, Taiwan
[b] Department of International Business Studies, National Chi Nan University, Taiwan

ARTICLE INFO

Keywords:
Foreign exchange
Rough set theory
Directed acyclic graph support vector machines
Immune algorithms
Tabu search algorithms

ABSTRACT

Rough set theory (RST) and directed acyclic graph support vector machines (DAGSVM) are two emerging techniques in dealing with classification problems. The RST approach is able to select important features and generate rules from data. The SVM technique is powerful in solving classification problems with high generalization ability by applying the structure risk minimization principle. However, one particular model cannot capture all data patterns easily. This investigation presents a hybrid RST and DAGSVM model (RSTDAGSVM) to exploit the unique strengths of both RST and SVM in analyzing the movements of exchange rates. In the proposed hybrid model, the RST approach is used to extract the rules of exchange rate changes; and the DAGSVM technique is employed to deal with situations that cannot be included in the RST model. In addition, an immune algorithm and tabu search (IA/TS) method is applied to select parameters of SVM models. Experimental results reveal that the developed model achieves more accurate prediction results than either the RST model or the DAGSVM model on its own. Thus, the presented RSTDAGSVM model is a promising alternative for analyzing exchange rates.

1. Introduction

The forecasting of exchange rates has long been regarded as one of the most challenging applications of modern financial prediction (Beran & Ocker, 1999; Fernando, Simon, & Julian, 1999), and numerous models have been depicted to provide investors with more precise predictions. Due to their general nonlinear function mapping capabilities and data driven characteristics, machine learning models have received increasing attention in exchange rate forecasting (Chun & Kim, 2003; Hann & Steurer, 1996; Kiani & Kastens, 2008; Qi & Wu, 2003; Tang, Zeng, Nourbakhsh, & Shen, 2009; Wu, 1995; Zhang & Hu, 1998; Zhang & Wan, 2007).

RST (Pawlak, 1982) is a useful mathematical tool in identifying hidden knowledge, characterized by vague and uncertain information, and in generating decision rules. RST has been successfully applied to problems with vagueness and uncertainty of information, and it has provided many exciting results in a considerably wide range of fields, such as medicine (Tsumoto, 2000), travel modeling (Witlox & Tindemans, 2004), location services (Sikder & Gangopadhyay, 2007), analysis of customer complaints (Yang, Liu, & Lin, 2007), financing (Sanchis, Segovia, Gil, Heras, & Vilar, 2007), learning models of low-level medical data (Brtka, Stotić, & Srdić, 2008), and image segmentation (Mushrif & Ray, 2008). SVM (Vapnik & Cortes, 1995) is a classification technique and has better generalization ability than that of traditional approaches. SVM was originally designed for two-class classifications. However, in many situations, the ability to categorize multiple-class is required for SVM model. Therefore, some multi-class classification techniques have been proposed for SVM models. The most typical approach for multi-class problems is the so called one-versus-rest (1-v-r) method classifying one class from the other class (Vapnik, 1995). The disadvantage of the 1-v-r method is the learning time scales linearly with the number of classes. One-versus-one (1-v-1) combining all possible two-class classifier is another methodology for dealing with multi-class problems (Friedman, 1996; Knerr, Personnaz, & Dreyfus, 1990). Knerr et al. (1990) combined the "AND" gate with the 1-v-1 approach. Friedman (Nguyen & Skowron, 1995a, 1995b) and KreBel (KreBel, 1999) used the Max Wins algorithm to enhance the 1-v-1 method. However, the drawback of the 1-v-1 approach is the size of classifier grows super-linearly with the number of classes. Therefore, the execution time of 1-v-1 approach is much worse than the 1-v-r method. Proposed by Platt, John, and Nello (2000), DAGSVM is one of the most popular approaches for multi-class classification methods. The training stage of the DAGSVM is the same as 1-v-1 model. However, the DAGSVM method uses a rooted binary directed acyclic graph to test the model. Therefore, the testing time of using DAGSVM model is less than that of the 1-v-1 approach.

* Corresponding author. Address. Department of Information Management, National Chi Nan University, Puli, Nantou 545, Taiwan. Fax: +886 49 2915205.
E-mail address: paipf@ncnu.edu.tw (P.-F. Pai).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.006

In this study, a hybrid RSTDAGSVM model was developed to analyze the movements of exchange rates. The remainder of this study is organized as follows. Section 2 introduces the methodologies used in this study. Section 3 presents the developed RSTDAGSVM model and uses a numerical example to demonstrate the performance of the proposed model. Conclusions and recommendations for future research are finally drawn in Section 4.

2. Methodologies

2.1. Rough set theory

RST uses information systems to represent knowledge and deal with vague data. An information system containing condition attributes and decision attributes is expressed as follows:

$$\mathrm{IS} = (U, \Omega, V, f) \tag{1}$$

where U denotes a nonempty finite set (namely the universe) with n objects $\{p_1, p_2, \ldots, p_n\}$, Ω is a nonempty finite set with m attributes $\{q_1, q_2, \ldots, q_m\}$. V is called the range of $U \times \Omega$, and $f: U \times \Omega \to V$ is an information function where $f(p, q) \in V$ for every $p \in U$, $q \in \Omega$. Moreover, let $Q \subseteq \Omega$ and $(x, y) \in U \times U$. Here, x and y are two objects.

Indiscernibility arises from an inability to differentiate among objects in a distinct set, and results in identical information derived from different observations. The indiscernibility relation of x and y in terms of Q is defined as follows:

$$IND(Q) = \{(x, y) \in U \times U : f(x, q) = f(y, q) \forall q \in Q\} \tag{2}$$

This indiscernibility relation partitions the universe U into a family of equivalence classes. The equivalence classes of the relation, $IND(Q)$, are called the Q-elementary sets in IS, and $[x]_{IND(Q)}$ represents the Q-elementary set containing the objective $x \in U$. In RST, knowledge of objects is presented in a decision table.

Lower and upper approximation is the second fundamental concept of RST. Let $Q \subseteq \Omega$, and $X \subseteq U$. The Q-lower approximation of $X(Q_L)$ and the Q-upper approximation of $X(Q_U)$ are then defined respectively as follows:

$$X(Q_L) = \{x \in U : [x]_{IND(Q)} \subseteq X\} \tag{3}$$

$$X(Q_U) = \{x \in U : [x]_{IND(Q)} \cap X \neq \phi\} \tag{4}$$

Every object in the lower approximation set of X must be in X. If an objective is in the upper approximation set of X, then it might or might not be in X.

The reduct, represented by $RED(Q)$, and core, expressed by $CORE(Q)$, are two fundamental RST concepts in knowledge reduction. The reduction of attributes eliminates some irrelevant or redundant attributes without reducing the quality of the approximation of the information system based on the original set of attributes. The indiscernibility relation of an attribute set remains unchanged when redundant attributes are removed. A reduct is a basic component of an information table, of which the core is the intersection of all reducts. The relation between reducts and the core can be represented as follows:

$$CORE(Q) = \cap RED(Q) \tag{5}$$

Reducts can be derived through the discernibility matrix and Boolean operations (Nguyen & Skowron, 1995a, 1995b). The discernibility matrix is a set that can differentiate between two objects or sets.

The rule generation from decision table to classify new objects is one of the most significant functions of RST. Rules are produced by the condition attributes based on the decision table. Thus, a decision rule can be written as "IF condition(s) THEN decision(s)". The prediction of a new objective is performed by matching its description to one of the rules. A support of a rule is represented as Eq. (6)

$$Supp = CARD(\|\Omega \cap V\|) \tag{6}$$

where $CARD$ is the cardinality of the set.

Rules with higher support values are more general, and express more information in the data set which can be classified by decision rules. Additionally, the coverage factor expressed as Eq. (7) is another essential criterion for measuring the performance of the RST. The coverage factor shows the degree to which the reasons for a decision can be trusted in terms of the data (Pawlak, 2002)

$$Coverage = \frac{Supp}{CARD(\|V\|)}. \tag{7}$$

2.2. Directed acyclic graph support vector machines

Originally, SVM was designed for two-class classification. By determining the separate boundary with maximum distance to the closest points of the training data set, SVM obtains a category decision. SVM is able to prevent a possible misclassification efficiently by minimizing structural risk. Consequently, SVM classifier owns better generalization ability than that of traditional classifying approaches. Supposed a training data set $Tr = \{X_i, Y_i\}_{i=1}^{n}$, where $X_i \in \Re^n$ is the ith input vector with binary output $Y_i \in \{-1, +1\}$, the classification function is expressed as Eq. (8)

$$Y_i = W^T \Gamma(X_i) + b \tag{8}$$

where Γ: $\Re^n \to \Re^m$ is the feature which maps the input space to a high dimensional feature space nonlinearly. The data points are linearly partitioned by a hyperplane defined by the pair $(W \in \Re^m, b \in \Re)$ (Vapnik, 1995). The optimal hyperplane that separates the data is represented by the following equation

$$\begin{aligned} &\text{Minimize} \quad \vartheta(w) = \|W\|^2/2 \\ &\text{Subject to} \quad Y_i[W^T \Gamma(X_i) + b] \geq 1 \quad i = 1, \ldots, n \end{aligned} \tag{9}$$

where $\|W\|$ is the norm of a normal weights vector of the hyperplane. This constrained optimization problem is obtained by a primal Lagrangian form formulated as Eq. (10):

$$L(W, b, \alpha) = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{n} \alpha_i [Y_i(W^T \Gamma(X_i) + b) - 1] \tag{10}$$

where α_i represent Lagrange multipliers. Using Karush–Kuhn–Tucker conditions, the solutions of the dual Lagrangian problem, α_i^*, determine the parameters w^* and of the optimal hyperplane. Finally, the decision function is depicted by Eq. (11):

$$\begin{aligned} D(X_i) &= \mathrm{sgn}(W_0^T \Gamma(x_i) + b_0) \\ &= \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* Y_i K(X, X_i) + b_0\right), \quad i = 1, \ldots, N \end{aligned} \tag{11}$$

Any function satisfies Mercer's conditions (Friedman, 1996) is a good candidate for the kernel function, $K(X, X_i)$. Besides, the value of $K(X, X_i)$ is expressed as the inner product of two vectors X and X_i in the feature space. In this work, the Gaussian radical basis function is used for the SVM classifier model.

With a single node without arcs pointing into it, a directed acyclic graph (DAG) is a graph whose edges have an orientation and no cycles. To conduct a function of classification, nodes of a DAG have either zero or two arcs leaving them (Platt, Cristianini, & Shawe-Taylor, 2000). Furthermore, for a problem of K classes, a DAG has K leaves labeled by classes and has $K(K-1)/2$ internal nodes organized in a triangular shape. The classification operation is performed by starting at the root node in a DAG. Then, the