Ngày 29/10/2010

Nhóm Kiên – Long

Mục lục

[Chương 1 BỨC TRANH LỚN VỀ CÁC MÔ HÌNH 1](#_Toc276080256)

[Chương 2 CHI TIẾT MÔ HÌNH TIẾP CẬN 1](#_Toc276080257)

[2.1 DỮ LIỆU 1](#_Toc276080258)

[TIỀN XỬ LÝ DỮ LIỆU 1](#_Toc276080259)

[2.2 MÔ HÌNH SVM 2](#_Toc276080260)

[2.3 CÁC ĐỘ ĐO 4](#_Toc276080261)

[2.4 KIỂM THỬ 5](#_Toc276080262)

[2.5 CÁC TÀI LIỆU TRÍCH DẪN 6](#_Toc276080263)

[Chương 3 ĐẶC TRƯNG CỦA CÁC MÔ HÌNH 7](#_Toc276080264)

[3.1 SVR + GA 7](#_Toc276080265)

[Đặc trưng 7](#_Toc276080266)

[Kết quả 8](#_Toc276080267)

[3.2 FUZZY SVR 8](#_Toc276080268)

[Đặc trưng 8](#_Toc276080269)

[Kết quả 9](#_Toc276080270)

[3.3 SOM + SVR 9](#_Toc276080271)

[Đặc trưng 9](#_Toc276080272)

[Kết quả 10](#_Toc276080273)

[3.4 SVM + Dynamic Kernel 11](#_Toc276080274)

[Đặc trưng 11](#_Toc276080275)

[Kết quả 12](#_Toc276080276)

[3.5 SVM + FUZZY + GA 13](#_Toc276080277)

[Đặc trưng 13](#_Toc276080278)

[Kết quả 14](#_Toc276080279)

[3.6 SVM + Box theory 14](#_Toc276080280)

[Kết quả 15](#_Toc276080281)

[3.7 Recurrent ε – SVR 15](#_Toc276080282)

[Kết quả 16](#_Toc276080283)

[Chương 4 KẾT LUẬN 18](#_Toc276080284)

[4.1 Về mô hình tiếp cận 18](#_Toc276080285)

[4.2 Về các vấn đề cải tiến 18](#_Toc276080286)

# BỨC TRANH LỚN VỀ CÁC MÔ HÌNH

Xem file excel: SVR\_Overall.xlsx

# CHI TIẾT MÔ HÌNH TIẾP CẬN

## DỮ LIỆU

* Khá nhiều mô hình đề xuất một số loại dữ liệu đầu vào khác ngoài giá đóng cửa. Tuy nhiên, chưa có công trình nào thực sự nghiên cứu sâu về vấn đề này để có thể trả lời câu hỏi đâu là những đầu vào tối ưu nhất. Hơn nữa, với những thị trường khác nhau, câu trả lời cho vấn đề trên không bất biến[5]. Vì vậy, tại thời điểm này, **đầu vào dành cho mô hình sẽ chỉ là giá đóng cửa hàng ngày**. Mọi cải tiến về đầu vào sẽ được tìm hiểu sâu hơn trong giai đoạn sau.
* Về dữ liệu:
  + Trước tiên, ta sẽ sử dụng dữ liệu được mô tả trong các bài báo (cụ thể là chọn dữ liệu của thị trường chứng khoán Đài Loan) với mục đích là kiểm tính đúng đắn của các mô hình do mình cài đặt.
  + Sau đó, ta mới áp dụng cho dữ liệu của chứng khoán Việt Nam. Hiện tại, việc chọn lựa các mã chứng khoán nào, trong khoảng thời gian nào tạm thời chưa làm.

### TIỀN XỬ LÝ DỮ LIỆU

* Dự đoán giá
* Scale lại trong khoảng (0, 1)

Trong đó:

: giá đóng cửa tại thời điểm t

: giá đóng cửa cao nhất

: giá đóng cửa thấp nhất

* SMA (Simple Moving Average): k-days SMA đơn giản là giá của ngày x sẽ bằng trung bình của giá của k ngày trước đó (kể cả ngày x.)
* EMA (Expotential Moving Average): k-days EMA được tính như sau:
* EMA1 = x1.
* EMAi = EMAi-1\*(1 – 2/k) + xi\*2/k với i chạy từ 2 đến k.

Nhận thấy có 2 điểm khác nhau giữa SMA và EMA:

* K-day SMA làm giảm k - 1 số điểm dữ liệu, trong khi EMA thì giữ nguyên.
* EMA đặt trọng lên dữ liệu gần hơn là SMA.
* Dự đoán xu hướng

Đối với mô hình dự đoán xu hướng, giá đóng cửa phải được tinh chỉnh về dạng return. Đây sẽ là dữ liệu đầu vào thực tế:

 với I là giá [4]

## MÔ HÌNH SVM

* Sử dụng hàm Gaussian kernel và hàm tổn thất là ε-insensitive loss function.
* Về việc chọn các tham số: rất quan trọng, ta cần phải chọn lựa kỹ càng.
* Có 3 tham số quan trọng cần được xác định là C (regularization parameter), ε, và σ2.
* Để xác định các tham số này, phương pháp cơ bản nhất là dùng cross validation. Ta có thể chọn từ từ C và σ2 bằng cách cho tăng dần theo số mũ [2]. Cũng có tác giả chọn ε= 0, 001 với lý lẽ là người ta đã chỉ ra rằng SVR không bị ảnh hưởng bởi ε miễn là ε là một giá trị hợp lý (như thế nào là hợp lý?), với C và σ2 thì tác giả chỉ chọn trong một tập do tác giả tự đưa ra: và [3]. Rõ ràng, ta cảm thấy cách làm như vậy không được logic cho lắm.
* Một phương pháp rất hợp lý là dùng GA để chọn ra bộ 3 tham số này [3].

**Do vậy, ta sẽ chọn GA để giải quyết vấn đề chọn bộ 3 tham số trên.**

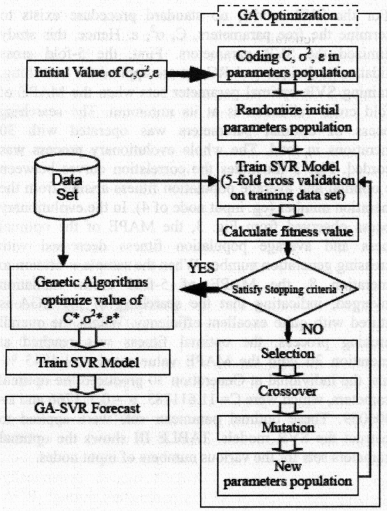


Figure : Mô hình GA-SVR

* Ngoài ra, còn một tham số nữa cũng rất quan trọng là số node đầu vào của SVR [1]. Ta có thể phát biểu bài toán hồi qui với SVR như sau:

Tìm f sao cho: pricek = f(pricek-1, …, pricek-n)

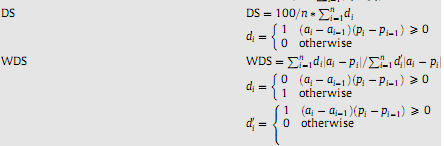
Ở đây, n chính là số node đầu vào của SVR.

Phương pháp hiện giờ là cho n chạy từ 1 – 5 và chọn ra trường hợp tốt nhất [1].

## CÁC ĐỘ ĐO

Để có thể so sánh được giữa các mô hình thì bắt buộc trong mỗi nhóm, các mô hình phải có cùng độ đo.

* Nhóm dự đoán giá
* Bốn độ đo lỗi:
  + Normalized Mean Square Error (NMSE)
  + Mean Square Error (MSE)
  + Mean Absolute Percentage Error (MAPE)
  + Mean Absolute Error (MAE)
* Ngoài ra, còn có 2 độ đo để đánh giá thêm xu hướng:
  + Directional Symmetry (DS): càng lớn càng tốt.
  + Weighted Directional Symmetry (WDS): phạt những lỗi mà xu hướng sai và thưởng những lỗi mà xu hướng đúng. Cảng nhỏ càng tốt.



* Nhóm dự đoán xu hướng

Có 3 độ đo thường được sử: Mean squared error(MSE), Normalized mean squared error(NSME), Squared correlation coefficient(SCC). Trong đó MSE thể hiện sự sai khác cục bộ còn SCC thể hiện sự sai khác toàn cục.

Trong đó là giá trị đầu ra thực sự, là giá trị đầu ra dự đoán, , là các giá trị trung bình của chúng.

Ngoài ra, để tính hiệu quả dự đoán xu hướng của mô hình, người ta thường sử dụng công thức tính độ chính xác sau [4]:

## KIỂM THỬ

* Nhóm dự đoán xu hướng

Sử dụng 2 kiểu kiểm thử [6]:

* + - Two-period: dữ liệu chia 2 phần, 1 phần huấn luyện, 1 phần kiểm (giống các nghiên cứu khác).
    - Multi-period: dữ liệu huấn luyện là phần dữ liệu liền trước của phần muốn kiểm thử.

Chẳng hạn có 100 ngày, để dự đoán ngày thứ 15 ta dùng dữ liệu từ ngày 1-14. Để dự đoán ngày 16 ta dùng dữ liệu từ ngày 2-15, vân vân…

Trong mỗi kiểu kiểm thử sẽ chia ra theo nhiều chu kỳ khác nhau: 5, 10, 30, 60 ngày.

## CÁC TÀI LIỆU TRÍCH DẪN

1. Kuan-Yu Chen, Chia-Hui Ho, *An Improved Support Vector Regression Modeling for Taiwan Stock Exchange Market Weighted Index Forecasting*, IEEE 2005
2. Yu-kun Bao, Zhi-tao Liu, Lei Guo, Wen Wang , *Forecasting stock composite index by fuzzy support vector machines regression*, IEEE, 2005
3. Sheng-Hsun Hsu, JJ Po-An Hsieh, Ting-Chih Chih, Kuei-Chu Hsu, *A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression*, Expert Systems with Applications 36 (2009) 7947–7951
4. Shiyi Chen, Kiho Jeong, Wolfgang K.Hardle, *Recurrent Support Vector Regression for a Nonlinear ARMA Model with Applications to Forecasting Financial Returns*, SFB 649 Discussion Paper, 2008
5. QinghuaWen \*,Zehong Yang, Yixu Song, Peifa Jia, *Automatic stock decision support system based on box theory and SVM algorithm*, Expert Systems with Applications, 2010
6. Deng-Yiv Chiu \*, Ping-Jie Chen, *Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm*, Expert Systems with Applications, 2009

# ĐẶC TRƯNG CỦA CÁC MÔ HÌNH

## SVR + GA

### Đặc trưng

Ở đây, sử dụng hàm Kernel là Gaussian với hàm tổn thất là . Tác giả có 2 cải tiến như sau:

* *Dùng GA để chọn ra bộ tham số tốt nhất (C, , ) cho mô hình SVR*. Thông thường, để chọn ra bộ tham số cho SVR người ta cho các tham số tăng dần theo số mũ và chọn ra bộ tốt nhất (dựa trên cross-validation). Nhận thấy ngay với cách làm này miền các giá trị được chọn là miền rời rạc. Với GA, miền được chọn liên tục và do đó cho phép cải tiến độ tốt của bộ tham số.
* *Nhận ra được tầm quan trọng của số node đầu vào của mô hình SVR.* Tuy nhiên, ở đây vẫn dùng phương pháp thử và sai (cho số node chạy từ 1-5.) Việc tìm cách chọn ra số node đầu vào một cách bài bản được tác giả định hướng là công việc tương lai.

Công thức tiền xử lý của tác giả để scale về (0, 1) của tác giả cũng khá lạ:

Trong đó:

: giá đóng cửa tại thời điểm t

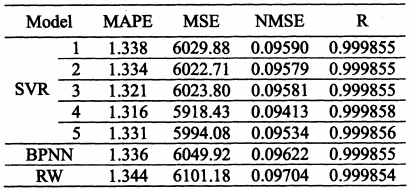
: giá đóng cửa cao nhất

: giá đóng cửa thấp nhất

(Lý lẽ của tác giả là “This scaling for original data points helps to improve the forecasting accuracy [3]” với [3] C. C. Chang and C. J. Lin, LIBSVM:A library for support vector machines, Retrieved May 20, 2004, from National Taiwan University,Department of Computer Science and Information Engineering Website: http://www.csie.edu.tw/~cjlin/papers/libsvm.pdf, 2001.)

### Kết quả

* Ở đây, ta so sánh giữa SVR, BPNN (back propagation NN), RW trên 3 độ đo lỗi: NMSE (Normalized Mean Square Error), MSE (Mean Square Error), MAPE (Mean Absolute Percentage Error.)
* Nhìn chung, cả 3 đều cho kết quả tốt, trong đó SVR tốt hơn 2 cái còn lại. Hệ số tỉ lệ giữa giá dự đoán và giá thực của cả 3 đều gần = 1, trong đó SVR là gần 1 nhất.
* Về khảo sát số node đầu vào: khi ta tăng số node đầu vào thì độ đo lỗi MAPE giảm nhưng khi số node này quá 4 thì MAPE lại tăng. Do đó, số node đầu vào = 4 được đánh giá là tối ưu.



## FUZZY SVR

### Đặc trưng

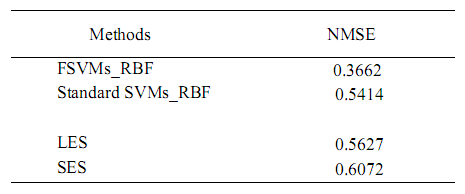
Bằng cách tiền xử lý bằng 10 days-EMS và sử dụng hàm fuzzy để đặt giá trị liên thuộc cho mỗi điểm dữ liệu trong chuỗi time-series, dụng ý của tác giả là *đặt trọng lên dữ liệu gần*, giảm ảnh hưởng của nhiễu.

Hàm liên thuộc: được xác định đơn giản như sau:

* Chọn L > 0 là cận dưới của hàm liên thuộc.
* Ta có hàm liên thuộc theo thời gian: si = f(ti)
  + Ta có: sn = f(tn) = 1 và s1 = f(t1) = L với n là tổng số điểm dữ liệu training.
  + Tùy vào ta muốn hàm liên thuộc là tuyến tính (si = ati + b) hay bậc 2 (si = a(ti - b)2 + c), ta sẽ dễ dàng tìm được công thức cho hàm liên thuộc.

### Kết quả

* Ở đây, ta so sánh giữa FSVM\_RBF, SVM\_RBF, LES (Linear expotential smoothing), SES (Simple expotential smoothing) với độ đo lỗi NSME.
* Trong đó, FSVM\_RBF cho kết quả tốt nhất.



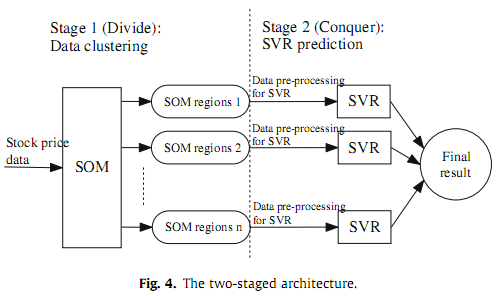
## SOM + SVR

### Đặc trưng

Hướng tiếp cận của bài báo xuất phát từ nhận định: bản chất của chuỗi giá chứng khoán là không ổn định. Do đó, mối quan hệ giữa một chuỗi giá chứng khoán và sự dự đoán cho ngày kế (tức hàm f mà ta cần tìm) là không cố định.

🡪 Ta kết hợp giữa SOM và SVR theo tư tưởng **chia để trị** như sau:

* Trước hết, với SOM ta sẽ chia toàn bộ không gian đầu vào thành các vùng mà ở đó các điểm dữ liệu có sự phân bố thống kê tương tự nhau.
* Kế đến, ta áp dụng SVR trên từng vùng.



Một điểm khác cũng đáng để ý ở đây là cách tiền xử lý dữ liệu của họ:

* Input: từ giá đóng cửa ban đầu, họ chuyển về 5 biến đầu vào: RDP-5, RDP-10, RDP-15, RDP-20 và EMA15 (với lý lẽ là làm cho sự phân bố dữ liệu tăng thêm tính đối xứng.)
* Output: từ giá đóng cửa ban đầu 🡪 3-day EMA 🡪 RDP+5.
* Những giá trị RDP nằm ngoài độ lệch chuẩn +-2 được coi là nhiễu và được thay thế bằng giá trị ở biên.
* Dữ liệu được scale lại trong khoảng [-0.9; 0.9]

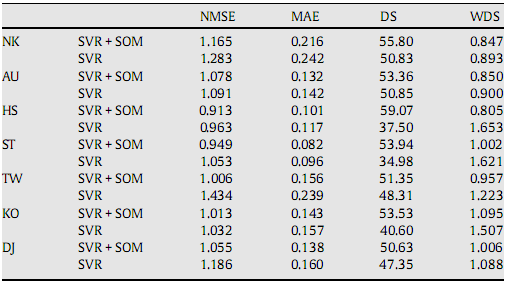
Cách chọn bộ tham số của tác giả:

* Tham số : họ lý luận rằng “Tay and Cao (2001a) showed that SVRs are insensitive to e, as long as it is a reasonable value” (thế nào là hợp lý?). Và họ chọn luôn .
* Tham số C và : họ đưa ra một tập các giá trị của 2 tham số này và sau đó dùng 10-fold cross validation để chọn ra bộ tốt nhất.
  + {2,1,0.5,0.1,0.01,0.001,0.0001}
  + {1000,750,500,100,50,2}

### Kết quả

Để ý:

* NMSE, MAE, WDS: càng nhỏ càng tốt.
* DS: càng lớn càng tốt.



## SVM + Dynamic Kernel

### Đặc trưng

Một cách sơ bộ, ý tưởng là không dùng các hàm Kernel chuẩn như là polynomial hay Gaussian (những kernel này bắt buộc phải thỏa hai điều kiện: symmetric và PSD); thay vào đó, ta dùng các hàm mà không phải chịu những ràng buộc này, cụ thể ở đây là Dynamic Time Warping.

Bài báo tiến hành rất nhiều thí nghiệm để khảo sát tính hiệu quả của các chuỗi input/ouput khác nhau, so sánh giữa các dynamic kernel, so sánh giữa các loại SVM và khảo sát tính hiệu quả của chiều dài chuỗi time-series. Từ đây, họ có được 2 kết luận chính như sau:

* Thứ 1, việc ta lấy một chuỗi time-series dài về quá khứ không cho kết quả tốt hơn so với việc lấy một chuỗi time-series ngắn.
* Thứ 2, ta nên tập trung kỹ càng vào việc tiền xử lý để có thể bắt được các mẫu. Một cách mà người ta cho là hiệu quả là sử dụng vector đầu vào nhiều chiều (i.e. chỉ sử dụng giá đóng cửa là vector đầu vào 1 chiều) như sau:

Sử dụng bộ 5 loại ROC (vector đầu vào 5 chiều): ROC1, ROC2,ROC3,ROC5,ROC8.

Trong đó, ROC (Rate Of Change) được tính như sau:

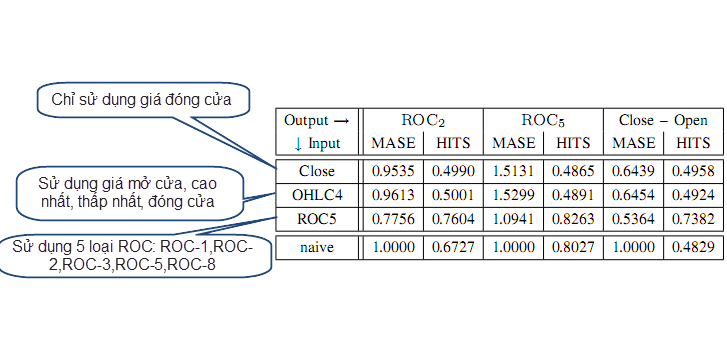


### Kết quả

#### Khảo sát tính hiệu quả của các chuỗi input/output khác nhau

Để ý:

* MASE càng nhỏ càng tốt.
* HITS càng lớn càng tốt.



Như ta có thể thấy, Input là ROC5 cho kết quả tốt nhất.

#### So sánh giữa các Dynamic Kernel

Ở đây, ta so sánh giữa Dynamic Time Warping kernel (DTW) và longest common subsequence kernel (LCSS). Kết quả cho thấy DTW không những chính xác hơn hẳn mà còn nhanh hơn LCSS.

#### So sánh giữa các loại SVM

Ở đây, ta so sánh giữa 4 phiên bản SVM (2 phiên bản đầu là regression, 2 phiên bản sau là classification): . Việc so sánh này được thực hiện trên 2 độ đo là MASE và HITS. Kết quả cho thấy vượt trội so với các phiên bản khác (đứng thứ 2 là .)

#### Khảo sát ảnh hưởng của chiều dài chuỗi time-series

Kết quả cho thấy nếu ta lấy lùi về quá khứ chuỗi time-series dài hơn thì cũng không cho kết quả tốt hơn.

## SVM + FUZZY + GA

### Đặc trưng

Với nhận định các công trình trước sử dụng quá ít hoặc quá nhiều các biến đầu vào (các loại chỉ số) làm cho việc dự đoán không hiệu quả hoặc gây tình trạng nhiễu. Hơn nữa, không có mô hình nào thực sự thể hiện được ảnh hưởng của từng biến đối với chỉ số chứng khoán theo thời gian. Điều này cần một mô hình động để thể hiện sự ảnh hưởng của từng biến qua từng giai đoạn khác nhau. (Ở đây nói về những thời điểm trong năm chứ không phải về chu kỳ dự đoán.)

Đầu tiên, mô hình fuzzy được áp vào từng biến (số lượng biến lên đến hàng chục với các chỉ số kỹ thuật và chỉ số của nền kinh tế vĩ mô) để xác định độ tác động khác nhau của từng biến đó qua các giai đoạn. Để việc này đảm bảo hiệu quả, các tham số cần được xác định là a, a1, b, b1 và khoảng thời gian tác động. Do đó, tác giả áp dụng GA như một cách tối ưu các tham số trên. Cụ thể như sau:

* + Dữ liệu được đưa về [-1;1].
  + Xác định tham số cho từng mô hình fuzzy ở từng chỉ số khác nhau( a1, a, b1,b và khoảng thời gian tác động).
  + Quyết định độ ảnh hưởng của mỗi nhân tố theo từng đoạn trong khoảng thời gian tác động.
  + Sử dụng SVM học để đưa ra độ chính xác. Sau đó GA xem xét đã thỏa mãn chưa, nếu chưa sẽ bắt đầu tiến hóa để tinh chỉnh tham số cho mô hình fuzzy.
  + Cuối cùng khi được mô hình tiến hóa hoàn chỉnh. SVM sẽ được dùng để dự đoán.
  + Ngoài ra, mô hình còn áp dụng 2 kiểu kiểm thử:
    - Two-period: dữ liệu chia 2 phần, 1 phần huấn luyện, 1 phần kiểm (giống các nghiên cứu khác).
    - Multi-period: dữ liệu huấn luyện là phần dữ liệu liền trước của phần muốn kiểm thử.

### Kết quả

Có thể tóm tắt như sau:

* + Với GA, kết quả tốt hơn. Fuzzy cải thiện được độ chính xác.
  + Với SVM, sử dụng multi-period tốt hơn two-period.
  + SVM với 3 loại chỉ số đầu vào (chỉ số kỹ thuật, kinh tế vĩ mô, chỉ số kỹ thuật tương lai) tốt hơn so với chỉ áp dụng 1 hay 2 loại chỉ số.

## SVM + Box theory

Bài toán này theo hướng hỗ trợ ra quyết định mua bán, tuy nhiên có thể xem gần với việc dự đoán xu hướng. Bên cạnh đó, việc áp dụng lý thuyết hộp cũng có thể được xem xét để cải tiến cho dự đoán giá.

Lý thuyết hộp có thể được hiểu rằng giá của chứng khoán tại một thời kỳ nào đó sẽ dao động trong một khoảng giới hạn. Vì thế, sau khi dự đoán được cận trên và dưới của chứng khoán trong một thời kỳ, SVM sẽ được áp dụng để xây dựng nên mô hình ra quyết định.

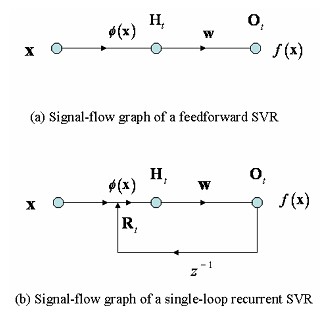
### Kết quả

* Trong việc đầu tư toàn thị trường:
  + Thị trường tăng :
    - Lợi nhuận: 39.11%
    - MSE: 0.0066 – 0.0084
    - SCC: 0.66 – 0.90364
  + Thị trường giảm :
    - Lợi nhuận: 13% hoặc lỗ 21%. (tốt vì thấp hơn tổng dao động của thị trường)
    - MSE: 0.002 – 0.0088
    - SCC: 0.6677 – 0.992
* Trong việc đầu tư riêng, kết quả kiểm với cổ phiếu Microsoft và IBM cho kết quả lợi nhuận từ 5%-74% trong khi việc thua lỗ rất ít và với tỷ lệ -2% đến -13%.

## Recurrent ε – SVR

Thừa nhận rằng với mô hình truyền thống của SVR dữ liệu input được đưa vào chỉ theo một chiều duy nhất mà không có chiều ngược lại để cho biết ảnh hưởng của nó lên mô hình sẽ làm ta không phân biệt được nhiễu.

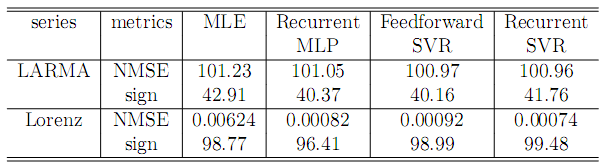
Sử dụng ε -SVR tái hiện (recurrent ε-SVR). Trong đó,việc tái hiện thể hiện ở chỗ luôn có 1 vòng phản hồi toàn cục từ không gian đầu ra tới không gian đầu vào.



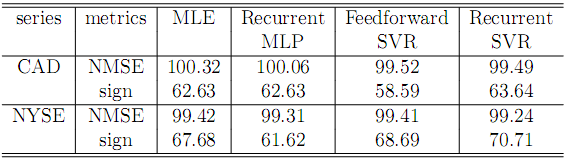
Xem sự biến động của giá cả thực không phải là sự biến đổi tuyến tính. Vì thế, tác giả đã dùng các tiếp cận này dự đoán mô hình phi tuyến ARMA trên dữ liệu mô phỏng.

### Kết quả

* + Ở mô hình mô phỏng *tuyến tính* LARMA, kết quả không tốt cho tất cả phương pháp. Ngay cả recurrent SVR cũng không bằng recurrent ANN (MLE).
  + Ở mô hình mô phỏng phi tuyến ARMA(còn gọi là Lorenz), phương pháp được đề nghị cho kết quả cao nhất. 99.48%.



* + Với dữ liệu thật, phương pháp này cho kết quả cao hơn hết, cao nhất 70.71%



# KẾT LUẬN

## Về mô hình tiếp cận

Tạm thời, để có được một cái nhìn cơ bản nhất trong việc chọn SVR làm hướng tiếp cận, ta sẽ cài đặt theo những phần đã trình bày trong chương 2. Khi đã có được một mô hình cốt lõi và dựa trên việc quan sát hiệu quả mà nó giải quyết bài toán, nhóm sẽ có những phương án để cải tiến thích hợp.

Để đảm bảo thuật toán không kém về hiệu quả dự đoán với những nghiên cứu trước đó. Trước khi áp vào thị trường Việt Nam, cần phải thực thi nó trên 1-2 bộ dữ liệu đã công bố.

## Về các vấn đề cải tiến

Các cải tiến trước đây hầu như tập trung vào tất cả những phần của một mô hình, từ dữ liệu, cách tiền xử lý đến việc chọn tham số và kiến trúc của mô hình.

Một số cải tiến về phần dữ liệu đầu vào tuy có cho một vài kết quả khả quan nhưng các thực nghiệm vẫn thể hiện phần nào đó “tính gượng ép”, không thực sự bộ lộ cơ chế tự nhiên của thị trường. Việc chọn nhiều chỉ số mang theo cách kinh nghiệm tạo cho mô hình một tính chất cục bộ, nghĩa là chỉ áp dụng với riêng thị trường đó, hoặc làm cho nghiên cứu trở nên chủ quan.

Gần đây, với quan sát rằng nhiều biến đầu vào không thực sự cải thiện mô hình, việc cải tiến về cách chọn tham số và kiến trúc mô hình được ưa chuộng hơn, và đem nhiều kết quả khả quan hơn.