Ngày 14/10/2010

Nhóm Kiên - Long



Rào trước: với hai hướng là dự đoán giá và dự đoán xu hướng sẽ có những phần chung và những phần riêng. Dưới đây, những phần nào mà không chỉ rõ là dự đoán giá hay dự đoán xu hướng thì có nghĩa là chung; ngược lại, là riêng.

# 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 σ2bằng 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.**

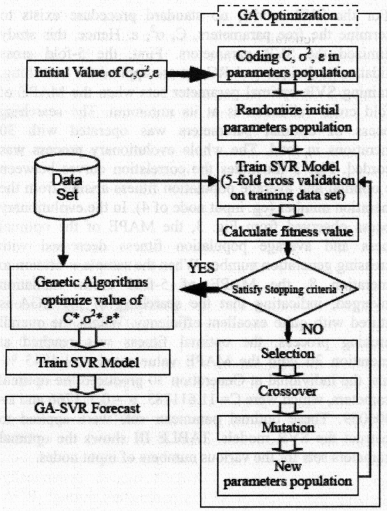


Figure 1: Mô hình GA-SVR

* 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.

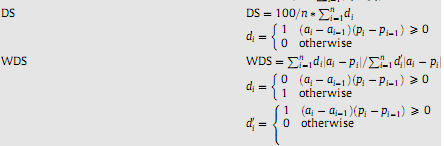
Phương pháp hiện giờ là cho n chạy từ 1 – 5 và chọn ra trường hợp tốt nhất [1].

# 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