RESEARCH ARTICLE

ỨNG DỤNG KHAI THÁC MẪU TUẦN TỰ VÀO BÀI TOÁN DỰ ĐOÁN XU HƯỚNG GIÁ CỔ PHIẾU

Nguyễn Tuấn Dũng, Trần Minh Thái

Khoa Công nghệ Thông tin, Đại học Ngoại ngữ - Tin học Tp.HCM

dungnt1@huflit.edu.vn, thaitm@huflit.edu.vn

TÓM TẮT—Dự đoán xu hướng cổ phiếu là sự hỗ trợ cần thiết cho các nhà đầu tư. Việc dự đoán chính xác và nhanh chóng đang được các nhà nghiên cứu áp dụng bằng nhiều mô hình khác nhau. Phương pháp dự đoán bằng cách khai thác trên dữ liệu lịch sử, biểu đồ nến là một trong những công cụ phân tích kỹ thuật được các nhà đầu tư sử dụng để lập chiến lược giao dịch cổ phiếu. Trong đó, ứng dụng khai thác dữ liệu vào dự đoán xu hướng cổ phiếu là cách tiếp cận mới. Trong bài báo này, chúng tôi đề xuất mô hình sử dụng kỹ thuật khai thác dữ liệu áp dụng vào việc dự đoán xu hướng cổ phiếu. Mô hình dự đoán dựa vào thuật toán khai thác mẫu con tuần tự trên tập dữ liệu lịch sử cổ phiếu. Bên cạnh đó, kỹ thuật xác định mẫu con thông qua độ tương tự cũng được trình bày trong bài báo. Dữ liệu thực nghiệm được thu thập trên trang https://finance.yahoo.com. Kết quả thực nghiệm của mô hình được đề xuất có độ chính xác trung bình tốt hơn so với mô hình truyền thống như SVM và LSTM.

Từ khoá—Khai thác dữ liệu, mẫu tuần tự, dự đoán xu hướng cổ phiếu, biểu đồ nến phiếu.

I. GIỚI THIỆU

Hiện nay, sự bùng nổ thông tin trong nhiều lĩnh vực như thị trường chứng khoán tạo ra lượng thông tin giao dịch mỗi giây được lưu lại là rất lớn. Thị trường chứng khoán là nơi các nhà đầu tư giao dịch chứng khoán làm tăng hay giảm khoản đầu tư ban đầu của mình. Nhiều phương pháp và kỹ thuật đã nghiên cứu dự đoán xu hướng cổ phiếu nhằm hạn chế rủi ro cho các nhà đầu tư. Thông thường, các nhà đầu tư sử dụng phân tích cơ bản và phân tích kỹ thuật để phân tích dự đoán nhằm lập chiến lược giao dịch cổ phiếu cho riêng mình. Một cách tiếp cận mới về dự đoán xu hướng là phương pháp khai thác dữ liệu. Khai thác dữ liệu là kỹ thuật tìm ra các thông tin ẩn, các mối quan hệ trong dữ liệu và khám phá các mẫu phổ biến được ứng dụng rộng rãi. Như việc tìm các mẫu hành vi khách hàng để cải thiện dịch vụ cung cấp mạng, giải quyết vấn đề gian lận trong các ngân hàng và bảo hiểm, nhận biết nhu cầu của học sinh để tăng sự hỗ trợ, v.v... Khai thác dữ liệu giống như một phương pháp máy học có thể dự đoán mẫu tương lai dựa vào khai thác dữ liệu.

Trong bài báo này, chúng tôi đề xuất mô hình sử dụng kỹ thuật khai thác mẫu con tuần tự trên dữ liệu lịch sử giao dịch cổ phiếu (biểu đồ nến Nhật). Mô hình có đề xuất về việc sàng lọc mẫu cho phù hợp với bài toán đặt ra khác với việc sàng lọc bằng ngưỡng hỗ trợ tối thiểu (minSup) do người dùng định nghĩa trong thuật toán khai thác mẫu phổ biến truyền thống. Mô hình đề xuất cho phép dự đoán xu hướng cổ phiếu ngày thứ 6 sau đó với mẫu dự đoán là 5 ngày. Bài báo được thể hiện các nội dung với bố cục như sau: Mục 2 trình bày các định nghĩa. Mục 3 tóm tắt các công trình nghiên cứu liên quan. Mục 4 và 5 trình bày thuật toán đề xuất và các kết quả thực nghiệm. Cuối cùng, kết luận và các hướng nghiên cứu tiếp theo được thể hiện trong Mục 6.

II. ĐỊNH NGHĨA BÀI TOÁN

Cho tập I = {i_1, i_2, …, i_n} gồm n phần tử phân biệt còn gọi là các sự kiện (item). Một tập sự kiện itemset là tập không có thứ tự khác rỗng, gồm các sự kiện. Mỗi itemset được biểu diễn trong cặp dấu ngoặc tròn. Cặp dấu ngoặc tròn được loại bỏ để đơn giản hóa ký hiệu cho các tập sự kiện với chỉ một sự kiện đơn. Ví dụ, (A, B, C) biểu diễn 1 tập sự kiện gồm 3 sự kiện là A, B và C.

Một chuỗi tuần tự sequence, ký hiệu $S = \langle e_1, e_2, \dots, e_m \rangle$, là một tập có thứ tự các tập sự kiện, với mỗi e_i ($1 \le i \le m$) là một tập sự kiện. Các sự kiện trong tập sự kiện được sắp xếp theo thứ tự từ điển, ký hiệu $\succ_{lex}$.

Cơ sở dữ liệu (CSDL) tuần tự (Sequence Database): CSDL tuần tự, ký hiệu *SDB*, là danh sách các chuỗi tuần tự, được biểu diễn dưới dạng $SDB = \{S_1, S_2, \dots, S_{|SDB|}\}$, trong đó |SDB| là số lượng chuỗi tuần tự trong SDB, và S_i ($1 \le i \le |SDB|$) là chuỗi tuần tự thứ i trong SDB. Ví dụ: Cho một CSDL có 5 khách hàng mua thực phẩm trong 3 tháng của một siêu thị (Bảng 1)

Bảng 1. *CSDL giao dịch mua thực phẩm*

Mã khách hàng	Thời gian	Mặt hàng mua
001	05/01/2020	Sữa
002	06/01/2020	Đường, Mì
005	10/02/2020	Gạo
004	11/02/2020	Gạo

002	13/02/2020	Cháo, Trà, Cà phê
003	20/02/2020	Sữa, Thịt, Cà phê
004	31/02/2020	Sữa
004	09/03/2020	Cháo, Cà phê
002	15/03/2020	Sữa
001	28/03/2020	Gạo

Với dữ liệu trong Bảng 1 có thể biểu diễn thành CSDL tuần tự gồm các chuỗi tuần tự mua sắm của từng khách hàng như trong Bảng 2. Trong đó, ký hiệu các sự kiện a, b, c, d, e, f tương ứng với các mặt hàng đường, sữa, …

Bảng 2. *CSDL tuần tự*

SID	Sequence
001	⟨a, b⟩
002	⟨(c, d), (e, f, h), a⟩
003	⟨a, i, h⟩
004	⟨b, a, (e, h)⟩
005	⟨b⟩

Trong Bảng 2 thể hiện CSDL chuỗi tuần tự SDB gồm có 5 chuỗi tuần tự, |SDB| = 5, và 3 sự kiện phân biệt I = {A, B, C}. Các chuỗi tuần tự có định danh lần lượt là 001, 002, 003, 004, 005 trong cột SID, thông tin chuỗi tuần tự được thể hiện trong cột Sequence. Chuỗi tuần tự S = ⟨(c, d), a, (e, f, h) ⟩ gồm có 3 tập sự kiện. Tập sự kiện thứ nhất là (c, d) có 2 sự kiện c và d. Tập sự kiện thứ hai là a có 1 sự kiện a. Cuối cùng là tập sự kiện (e, f, h) có 3 sự kiện là e, f và h.

Kích thước chuỗi (size of sequence): số tập sự kiện (itemset) có trong chuỗi S, ký hiệu |S|.

Độ hỗ trợ (support): độ hỗ trợ của chuỗi tuần tự S trong CSDL tuần tự được định nghĩa là tổng số chuỗi tuần tự trong CSDL có chứa S, ký hiệu .

$$\text{support}_{SDB}(S) = |\{\langle SID, S\rangle | (\langle SID, S\rangle \in SDB) \wedge (S \sqsubseteq SBD)\}|$$

Chiều dài chuỗi (length of sequence): được tính dựa vào số lượng sự kiện có trong chuỗi S. Chuỗi có k sự kiện được ký hiệu là k-sequence.

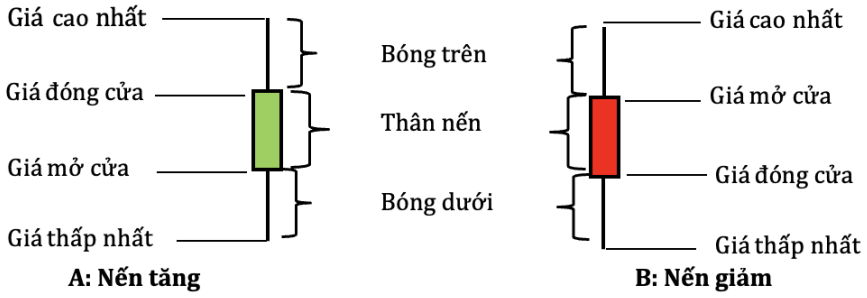
Chuỗi cha và chuỗi con tuần tự (supersequence và subsequence): một chuỗi $S_A=(a_1,a_2,\ldots,a_n)$ được gọi là chuỗi con của chuỗi $S_B=(b_1,b_2,\ldots,b_m)$ nếu và chỉ nếu tồn tại dãy số $1 \le i_1 < i_2 < \ldots < i_n <= m$ sao cho $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$,…, $a_n \subseteq b_{in}$. Khi đó ta nói chuỗi S_A là chuỗi con của S_B hay chuỗi S_B là chuỗi cha của S_A.

Tiền tố (prefix): một chuỗi $S_A=(a_1,a_2,\ldots,a_n)$ được gọi là tiền tố của chuỗi $S_B=(b_1,b_2,\ldots,b_m)$ nếu và chỉ nếu với mọi n < m và $a_1=b_1, a_2=b_2,\ldots,a_n=b_n$.

Mẫu tuần tự (sequential pattern): Cho trước một ngưỡng hỗ trợ tối thiểu, ký hiệu *minSup*, được xác định trước bởi người dùng. Trong đó, 0 < *minSup* ≤ 1. Một chuỗi S được xem là chuỗi phổ biến khi và chỉ khi độ hỗ trợ của S lớn hơn bằng ngưỡng hỗ trợ tối thiểu, sup(S) ≥ *minSup*. Khi đó, S được gọi là mẫu tuần tự [1].

Khái niệm cây nến Nhật: Cây nến Nhật do người Nhật phát minh vào năm 1600. Mỗi cây nến biểu diễn dữ liệu giá mở cửa, đóng cửa, cao nhất, thấp nhất. Cây nến gồm 2 phần real body (thân nến) và shadow (bóng nến) (Hình 1). Hình chữ nhật gọi là thân nến thể hiện phạm vi thời gian mở và đóng phiên giao dịch. Thân nến màu đỏ (nến giảm như Hình 1-B) tức giá đóng cửa của phiên giao dịch thấp hơn giá mở cửa. Ngược lại, thân nến có màu xanh (nến tăng như Hình 1-A) tức giá đóng cửa của phiên giao dịch lớn hơn giá mở cửa.

Phần ở trên và dưới thân nến gọi là bóng nến, phần ở trên gọi bóng trên có đỉnh là mức giá cao nhất và phần ở dưới gọi là bóng dưới có đáy là mức giá thấp nhất của phiên giao dịch. Cây nến có thể vẽ trong các khung thời gian quan sát khác nhau (như phút, giờ, ngày, …). Trong phạm vi của bài báo này, chúng tôi sử dụng cây nến được quan sát theo khung thời gian là 1 ngày. Ví dụ xét Hình 1, khung thời gian quan sát cây nến theo ngày, cây nến tăng (Hình 1-A) đáy thân nến là giá giao dịch đầu tiên hay gọi là giá mở cửa và đỉnh thân nến là giá kết thúc phiên giao dịch hay gọi giá đóng cửa. Đỉnh bóng trên tương ứng giá cao nhất và đáy bóng dưới tương ứng giá thấp nhất. Ngược lại, cây nến giảm (Hình -B) giá mở cửa nằm trên thân nến, giá đóng cửa nằm dưới thân nến, và giá cao nhất, thấp nhất giống nến tăng.



Hình 1. *Cấu trúc cây nến Nhật [2] với A: Nến tăng và B: Nến giảm.*

III. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Những năm gần đây, dữ liệu giao dịch cổ phiếu, thông tin mạng xã hội, tin tức tài chính, các blog của mạng máy tính, …. được xem như "vàng". Khai thác dữ liệu nhằm tìm kiếm thông tin ẩn và khám phá tri thức từ dữ liệu. Việc khai thác thông tin để tìm thông tin hữu ích cho các mục đích khác nhau như phát hiện gian lận, dự đoán tương lai, phân tích hành vi, ... ngày càng phổ biến và cần thiết cho các công ty, doanh nghiệp trong nhiều lĩnh vực khác nhau [1].

Thị trường chứng khoán sinh ra lượng thông tin giao dịch rất lớn mỗi giây, việc dự đoán xu hướng cổ phiếu rất cần thiết cho các nhà đầu tư có chiến lược giao dịch, giảm rủi ro và có thể tăng lợi nhuận [3]. Biểu đồ nến nhật là một trong những công cụ của phương pháp phân tích kỹ thuật thường dùng phổ biến để phân tích xu hướng của cổ phiếu. Biểu đồ là tập hợp nhiều cây nến được biểu diễn theo khung thời gian nhất định (như năm, tháng, ngày).

Các nhà phân tích cơ bản tin rằng giá cổ phiếu chịu ảnh hưởng của dữ liệu kinh tế vĩ mô, dữ liệu giao dịch chứng khoán cũ [4]. Các thông tin như lãi suất, tỷ giá hối đoái, giá tiêu dùng, báo cáo tài chính, tỷ suất cổ tức, ... được xem như là biến số kinh tế. Mô hình dự đoán chứng khoán thường sử dụng mô hình hồi quy tuyến tính hay phi tuyến đòi hỏi các tham số xác định trước để cố gắng tìm ra mối quan hệ giữa lợi nhuận chứng khoán với các biến số kinh tế tài chính. Trong đó, họ giả định rằng giá cổ phiếu có thể dự đoán dựa vào việc phân tích dữ liệu giá cổ phiếu lịch sử [5]. Họ phân tích dữ liệu về sự thay đổi giá cổ phiếu, khối lượng giao dịch, … Phân tích kỹ thuật phổ biến thường phân tích mẫu hình biểu đồ và chỉ báo kỹ thuật [6]. Chỉ báo kỹ thuật là các tín hiệu chuyên biệt cho biết giá cổ phiếu trong một khoảng thời gian nhất định trong lịch sử.

Đầu tiên, cách tiếp cận bằng phương pháp máy học có giám sát SVM được đề xuất trong dự đoán xu hướng cổ phiếu [7]. Trong bài báo này, nhóm tác giả của bài báo đề xuất 2 phần gồm lựa chọn đặc trưng và xây dựng mô hình dự đoán. Các đặc trưng được chọn lựa bằng SVM và được đánh giá xếp hạng. Mô hình dự đoán là SVM có hàm chuyển đổi bán tuyến tính (quasi-linear kernel) gần giống bộ phân loại phi tuyến, nó tạo ra đường biên phân loại tuyến tính đa nội bộ và phép nội suy. Mô hình SVM dự đoán xu hướng thị trường chứng khoán trên dữ liệu lịch sử của thị trường chứng khoán Đài Loan. Kết quả thực nghiệm mô hình dự đoán tốt xu hướng thị trường chứng khoán.

Ngoài ra, phương pháp học sâu cũng ứng dụng trong các dự đoán cổ phiếu như dự đoán lợi nhuận cổ phiếu [8]. Dự đoán giá cổ phiếu dựa vào dữ liệu lịch sử và các chỉ báo trên dữ liệu 10 cổ phiếu ở Mỹ và Đài Loan [9]. Dự đoán xu hướng dựa vào đánh trọng số thời gian cho các dữ liệu dự đoán, xu hướng chứng khoán được xác định dựa vào lý thuyết và thực tiễn. Cuối cùng, mạng LSTM dùng để dự đoán xu hướng tương lai cho dữ liệu có tính chất thời gian. Thực nghiệm trên chỉ số CSI 300 đạt độ chính xác 83,91% [10]. Với lợi thế xử lý dữ liệu chuỗi thời gian (time series) mô hình LSTM được áp dụng nhiều trong các bài toán dự đoán rất tốt.

Gần đây, nhiều mô hình dự đoán cổ phiếu đã được nghiên cứu. Trong đó, mô hình sử dụng thuật toán khai thác mẫu tuần tự là một tiếp cận mới trong dự đoán xu hướng cổ phiếu. Kỹ thuật khai thác mẫu tuần tự được Agrawal và Srikant giới thiệu lần đầu tiên vào năm 1995 [11]. Các tác giả định nghĩa trình tự mua hàng của khách hàng trong CSDL giao dịch và được biểu diễn thành CSDL tuần tự với thuật toán khai thác mẫu tuần tự phổ biến là AprioriAll, AprioriSome [12]. Bài báo dự báo lợi nhuận [13] trình bày kỹ thuật khai thác dữ liệu phân tích mức độ liên quan của biến số và mạng noron dùng phân loại và ước tính giá trị lợi nhuận cổ phiếu trên chỉ số S&P 500. Nhiều phương pháp biểu diễn chuỗi thời gian khác nhau như dựa trên mô hình tổng quát, dựa trên phép chuyển đổi hay dựa trên miền thời gian. Biểu diễn chuỗi thời gian giúp quá trình tính toán đơn giản hơn và giảm được số chiều thuận lợi quá trình dự đoán [14]. Việc biểu diễn chuỗi thời gian dựa trên miền thời gian được sử dụng rộng rãi trong các ứng dụng tài chính.

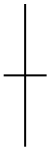
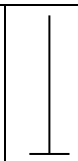
IV. ĐỀ XUẤT MÔ HÌNH

Trong phần này, chúng tôi trình bày một mô hình đề xuất, với tên gọi DMSP-TS, sử dụng phương pháp khai thác mẫu con tuần tự được định nghĩa lại với mô hình tìm mẫu con có độ phù hợp cao nhất, dữ liệu lịch sử được mã hóa theo các quy tắc nhất định và xu hướng cổ phiếu được xác định bằng sự thay đổi của đường giá đóng cửa. Độ tương tự của mẫu có đặt trọng số được đề xuất để cải thiện kết quả của dự đoán.

A. MÃ HÓA CÂY NẾN

Mỗi cây nến được vẽ với các quan hệ giá khác nhau. Cấu trúc cây nến thường có 4 giá: giá cao nhất, giá thấp nhất, giá mở cửa và giá đóng cửa. Để có thể áp dụng thuật toán khai thác mẫu tuần tự, cây nến được mã hóa thành các sự kiện theo các ký hiệu a, b, c, … Tuy nhiên, với những quan hệ giá tạo nên rất nhiều kiểu cây nến, việc mã hóa cây nến sẽ theo quy định được trình bày trong Bảng 4.

Bảng 3. *Mã hóa các kiểu cây nến theo Bảng 4*

Kiểu nến												
Kí hiệu	a	b	c	d	e	f	g	h	i	j	k	l

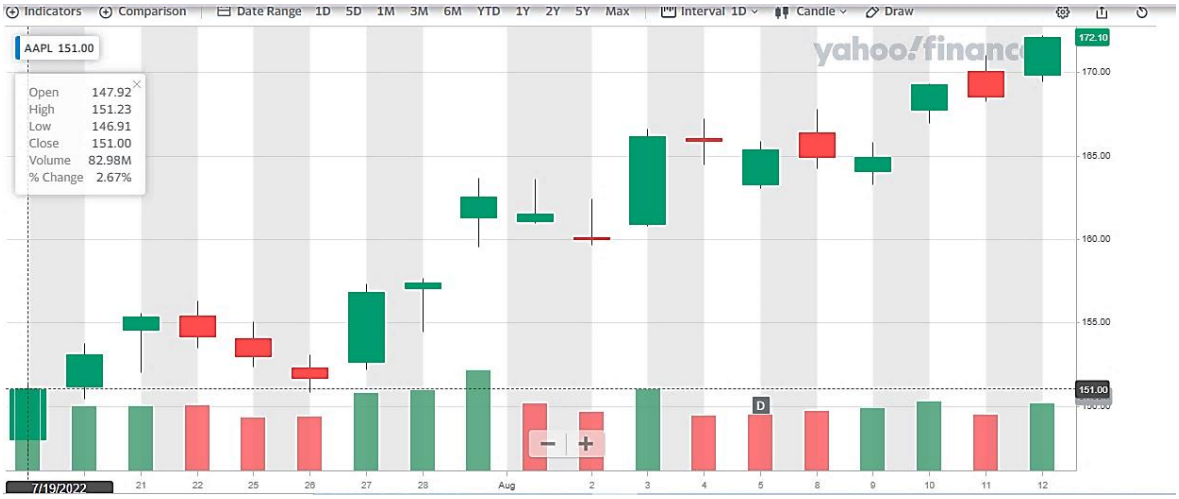
Bảng 4. *Các quy định mã hóa cây nến*

Quy định	NT có : H > C C > O O > L	NT có : H > C C > O O = L	NT có : H = C C > O O > L	NT có : H = C C > O O = L	NG có : H > O C < O C > L	NG có : H > O C < O C = L
Kí hiệu	a	b	c	d	e	f
Quy định	NG có : H = O C < O C > L	NG có : H = O C < O C = L	Nến có : H > O C = O C > L	Nến có : H = O O = C C > L	Nến có : H > C C = O O = L	Nến có : H = O O = C C = L
Kí hiệu	g	h	i	j	k	l

Trong Bảng 4 mô tả quy định mã hóa cây nến có H là giá cao nhất, C là giá đóng cửa, O là giá mở cửa, L là giá thấp nhất, NT là nến tăng, NG là nến giảm.

B. XÁC ĐỊNH XU HƯỚNG CỔ PHIẾU

Xu hướng được xác định theo các điểm thay đổi giá trên đường giá đóng cửa, có nghĩa giá đóng cửa đang tăng sau đó giảm và ngược lại. Xu hướng tăng khi giá đóng cửa sau cao hơn giá đóng cửa trước ngược lại là xu hướng giảm và xu hướng được gán sau điểm thay đổi. Xu hướng tăng được gán nhãn là 1 còn xu hướng giảm gán nhãn là -1 và xu hướng không tăng cũng không giảm thì được gán nhãn 0. Xu hướng được minh họa Hình 2.



Hình 2. *Đường giá đóng cửa và xu hướng được gán tại các điểm thay đổi giá.*

Thể hiện mẫu khai thác trong thuật toán: Nhằm phù hợp với bài toán dự đoán xu hướng nên mẫu khai thác trong thuật toán có thêm xu hướng gọi là mẫu nến gồm 2 phần: chuỗi nến và xu hướng của chuỗi nến kí hiệu $pc = \{\langle sc\rangle, xh\}$ trong đó:

$\langle sc\rangle$: là chuỗi nến được mã hóa các cây nến thành kí hiệu a, b, c, …

xh : là xu hướng của chuỗi nến.

Ví dụ : $pc = \{\langle a, b, d, d, c, a, e\rangle, 1\}$ chuỗi nến $sc = \langle a, b, d, d, c, a, e\rangle$ có xu hướng tăng.

C. THUẬT TOÁN DMSP-TS

Thuật toán DMSP-TS gồm ba phần chính: (1) kiểm tra mẫu nến có phải chuỗi con hay không, (2) thuật toán khai thác mẫu nến, (3) độ tương tự và độ phù hợp.

Bảng 5. *Các mô đun trong thuật toán DMSP-TS: (a) CheckSubSeq, (b) SeqPatternMin*

Method: CheckSubSeq(X,Y)
Input: X,Y là chuỗi tuần tự
Output: 1 hay 0

```
1.  m, n //chiều dài của chuỗi tuần tự  X, Y
2.  if (m>n):
3.    return 0
4.  i, k=0
5.  while i<m :
6.    if k==n : # Y rỗng
7.      return 0
8.    for j in range(k,n):
9.      if X[i]==Y[j]:
10.   i+=1
11.   k=j+1
12.   break
13. elif j==(n-1) :
14.   return 0
15.  return 1
```

(a) CheckSubSeq

Method: SeqPatternMin(setP)
Input: setP //tập mẫu nến
Output: Psc //tập chi tiết mẫu nến

```
16.  n=len(setP)//chiều dài mẫu nến
17.  for i in range(0,n):
18.    sp = 0 // số lần xuất hiện chuỗi con giống xu hướng
19.    sup = 0 //số lần xuất hiện chuỗi con
20.    sc1 = setP[i][0:-1]
21.    xh1 = setP[i][-1]
22.    for j in range(0,n):
23.      sc2 = setP[j][0:-1]
24.      kt = CheckSubSeq(sc1,sc2)
25.      if kt == 1:
26.        sup += 1
27.        xh2 = setP[j][-1]
28.        if xh1 == xh2:
29.          sp += 1
30.    if ExistPset(setP[i],Psc) == False:
31.      acc = sp/sup
32.      Psc = addP(sc1,xh1,sup,sp,acc)
33.  return Psc
```

(b) SeqPatternMin

Bảng 5 trình bày thuật toán được đề xuất. Thuật toán (a) kiểm tra xem mẫu nến có phải là chuỗi con của mẫu nến khác trong tập mẫu nến, lấy chiều từ trái sang phải làm điểm bắt đầu so sánh các phần tử, nếu 2 phần tử của 2 mẫu nến giống nhau thì cả 2 mẫu nến phần tử sẽ di chuyển tới phần tử kế tiếp (dòng 9, 10, 11), ngược lại phần tử của mẫu nến trong tập mẫu nến giữ nguyên vị trí. Kết quả (a) là 1 (dòng 15) mẫu nến là chuỗi con của mẫu nến khác trong tập nến, còn kết quả là 0 (dòng 3, 7, 14) mẫu nến không phải là chuỗi con.

Tiếp theo, thuật toán (b) khai thác tập mẫu nến, dựa theo thuật toán khai thác mẫu con tuần tự nhưng không dùng ngưỡng hỗ trợ tối thiểu (minSup) mà tần suất của mẫu nến được tính bằng độ chính xác mẫu (dòng 31) theo công thức 1, các mẫu nến trong tập mẫu nến lần lượt kiểm tra chuỗi con của mẫu nến khác không (dòng 24). Từ dòng 25 đến dòng 29 mô tả cách tính tần suất mẫu nến giống xu hướng và tần suất xuất hiện mẫu nến từ đó tính độ chính xác của mẫu, cuối cùng kết quả được lưu lại chi tiết các thông tin của mẫu nến (dòng 32). Quá trình khai thác mẫu con tuần tự được thực hiện từ đầu cho đến hết mẫu trong tập mẫu nến.

Công thức tính độ chính xác của mẫu nến như sau:

$$acc = \frac{số\ lượng\ pc\ có\ xh\ giống\ xh\ chuỗi\ nến}{tổng\ số\ chuỗi\ nến\ giống\ chuỗi\ nến\ pc} \qquad (1)$$

Ví dụ : Cho mẫu nến $p = \{\langle a, a, e\rangle, 1\}$ có 2 mẫu nến $p2 = \{\langle a, a, e\rangle, 1\}$ và $p1 = \{\langle a, a, e\rangle, -1\}$ có chuỗi nến giống p nhưng khác xu hướng độ chính xác acc(p)= 1/2 = 0.5.

Trong một số trường hợp dự đoán mẫu nến có cùng độ chính xác, mô hình so khớp truyền thống dự đoán xu hướng cho kết quả không được tốt. Do vậy, chúng tôi đề xuất độ tương tự của hai mẫu giúp giải quyết vấn đề mẫu nến có cùng độ chính xác bằng cách đặt trọng số cho các phần tử. Trọng số đặt là 5 cho các phần tử đầu tiên và giảm dần cho các phần tử phía sau, độ tương tự được tính theo cả hai chiều (từ trái sang phải và ngược lại) của 2 chuỗi nến.

$$w_i = \begin{cases} 5, & i = 1 \\ m - i + 1, & 1 < i \leq m \end{cases} \qquad (2)$$

w_i : trọng số mẫu nến phần tử thứ i, và m chiều dài mẫu nến.

Độ tương tự của 2 chuỗi tuần tự được tính theo công thức:

$$sim(X, Y) = \frac{Max\ (sim_1, sim_2, \dots sim_k)}{|m - n| + 1} \qquad (3)$$

Trong đó, $sim_k = \sum_{i,j=0}^{m,n} w_i \times w_j$ với phần tử thứ i giống phần tử thứ j là độ tương tự trường hợp thứ k, có k trường hợp tương tự, m,n là chiều dài của 2 mẫu nến X và Y. Hàm Max lấy số lớn nhất của độ tương tự.

Công thức tính độ phù hợp của mẫu nến như sau:

$$mat(pcur, pc_i) = sim(pcur, pc_i) \times acc_i \qquad (4)$$

$pcur$: mẫu nến dự đoán.

pc_i : mẫu nến thứ i trong tập mẫu nến

acc_i : độ chính xác của mẫu nến theo công thức 1

V. KẾT QUẢ THỰC NGHIỆM

Thực nghiệm trên dữ liệu lịch sử thu thập từ trang https://finance.yahoo.com trong khoảng thời từ 04/01/2021 cho đến 12/05/2021. Các mã cổ phiếu là các công ty trong chỉ số NASDAQ-100. Mỗi cổ phiếu có thuộc tính ngày ghi cổ phiếu giá cao nhất, giá thấp nhất, giá mở cửa, giá đóng cửa, giá đóng cửa điều chỉnh, khối lượng giao dịch. Thời gian biểu diễn cây nến đơn vị 1 ngày. Để đánh giá mô hình mỗi mã cổ phiếu được chia 2 phần theo thứ tự: 80% dữ liệu đầu làm tập huấn luyện và 20% dữ liệu sau làm tập kiểm tra.

Mô hình SVM thường được dùng để dự đoán xu hướng cổ phiếu theo phương pháp phân loại được sử dụng trong nhiều bài báo. Trong bài báo này, mô hình SVM sử dụng phương pháp Support Vector Classification (SVC) để dự đoán xu hướng cho ngày thứ 6 với mẫu kiểm thử. Các ngày trong mẫu kiểm thử được nối lại với nhau thành một vector có 20 phần tử theo thứ tự thời gian và giá gồm "giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa" [14]. Xu hướng là kết quả so sánh "giá đóng cửa" [14] của ngày kế tiếp mẫu kiểm thử (tức ngày thứ 6) với ngày cuối của mẫu kiểm thử (tức ngày thứ 5). Nếu giá đóng cửa ngày thứ 6 lớn hơn giá đóng cửa ngày thứ 5 thì xu hướng cổ phiếu tăng được gán nhãn bằng 1, còn giá đóng cửa ngày thứ 6 nhỏ hơn giá đóng cửa ngày [14] thứ 5 là xu hướng giảm được gán nhãn bằng -1, cuối cùng giá đóng cửa ngày thứ 6 bằng giá đóng cửa ngày thứ 5 thì xu hướng đi ngang được gán nhãn bằng 0. Mô hình SVM sử dụng các tham số phạt C=1.0 để biên lề không hẹp quá, kernel='linear' để chia các vector theo dạng tuyến tính.

Mô hình LSTM thường sử dụng dự đoán trên dữ liệu chuỗi thời gian, hay chuỗi tuần tự. Mô hình dự đoán các phụ thuộc xa rất hiệu quả. Trong các nghiên cứu dự đoán cổ phiếu thường sử dụng mô hình LSMT để đoán xu hướng

tương lai. Trong bài báo này, mô hình LSTM sử dụng phương pháp hồi quy để dự đoán xu hướng ngày thứ 6 có đầu vào là mẫu kiểm thử. Đầu vào của mô hình LSTM tương tự như đầu vào của mô hình SVM là các ngày trong mẫu kiểm thử được nối lại với nhau theo thứ tự thời gian và giá. Xu hướng được xác định cũng giống như việc xác định xu hướng trong mô hình SVM. Mô hình LSTM trong các nghiên cứu dự đoán và thực tiễn thường được xây dựng các tầng LSTM chồng lên nhau và cuối cùng thường là tầng Dense. Mô hình LSTM trong bài báo được xây dựng gồm 2 tầng LSTM với units=50 để lưu thông tin quan trọng của dữ liệu đầu vào mô hình, tầng Dropout có tham số rate=0.1 để ẩn bớt các nút mạng tránh vấn đề overfitting và cuối cùng là tầng Dense với units=1 để đưa ra kết quả dự đoán xu hướng cổ phiếu ngày thứ 6. Ngoài ra, các tham số cho quá trình huấn luyện mô hình có loss=mean_squared_error, optimizer=adam, epochs=100, batch_size=32.

Thực nghiệm chia thành 2 nhóm: Nhóm 1 thực nghiệm chủ yếu so sánh 2 mô hình khai thác mẫu con tuần tự, mô hình đề xuất với mô hình truyền thống, nhóm 2 thực nghiệm so sánh mô hình đề xuất với mô hình dự đoán phổ biến SVM , LSTM trong dự đoán xu hướng.

Lấy lần lượt 5 ngày liên tiếp theo thứ tự thời gian và xu hướng ngày thứ 6 được xác định bằng giá đóng cửa làm mẫu dự đoán xu hướng, các mẫu nến dự đoán được lệch nhau 1 ngày, có nghĩa cửa sổ trượt là 1 ngày.

Độ đo chính xác (ACC) sử dụng làm thước đo đánh giá hiệu suất mô hình. Kết quả thực nghiệm cho thấy mô hình đề xuất có độ chính xác trung bình (Bảng 7) cao hơn so với mô hình truyền thống, vì việc so khớp mẫu có độ chính xác cao với mẫu muốn dự đoán có nhiều mẫu có cùng độ chính xác nên thuật toán sẽ lấy mẫu đầu tiên có cùng độ chính xác cao nhất mà không xét đến vị trí thứ tự phần tử khi so khớp dẫn đến kết quả dự đoán không tốt. Thứ tự của các phần tử trong mẫu là một đặc trưng của dữ liệu, mô hình đề xuất độ tương tự có trọng số để xác định mẫu phù hợp nhất để dự đoán xu hướng và giải quyết được vấn đề có cùng độ chính xác. Bảng 6 minh họa 5 cổ phiếu có độ chính xác cao của mô hình truyền thống (ACC MHTH), và mô hình đề xuất (ACC MHDX), tuy nhiên có một số mã cổ phiếu thì ngược lại có thể do ảnh hưởng của nền kinh tế thế giới, dịch bệnh toàn cầu, … Tuy nhiên, Bảng 7 chứng tỏ hiệu quả mô hình đề xuất so với mô hình truyền thống trên tập dữ liệu thực nghiệm.

Bảng 6. *So sánh 5 cổ phiếu với độ chính xác của 2 mô hình*

Mã chứng khoán	Tên công ty	ACC MHTH	ACC MHDX
AMAT	Applied Materials, Inc.	50%	78.5%
MRVL	Marvell Technology, Inc.	71.4%	78.5%
AMD	Advanced Micro Devices, Inc.	21.4%	78.5%
INTU	Intuit Inc.	35.7%	64.2%
CMCSA	Comcast Corporation	57.1%	64.2%

Bảng 7. *Bảng so sánh hiệu suất 2 mô hình dự đoán.*

Mô hình dự đoán	Độ chính xác trung bình	Thời gian thực hiện
Truyền thống	47.5%	0.879 giây
Đề xuất	51.3%	0.891 giây

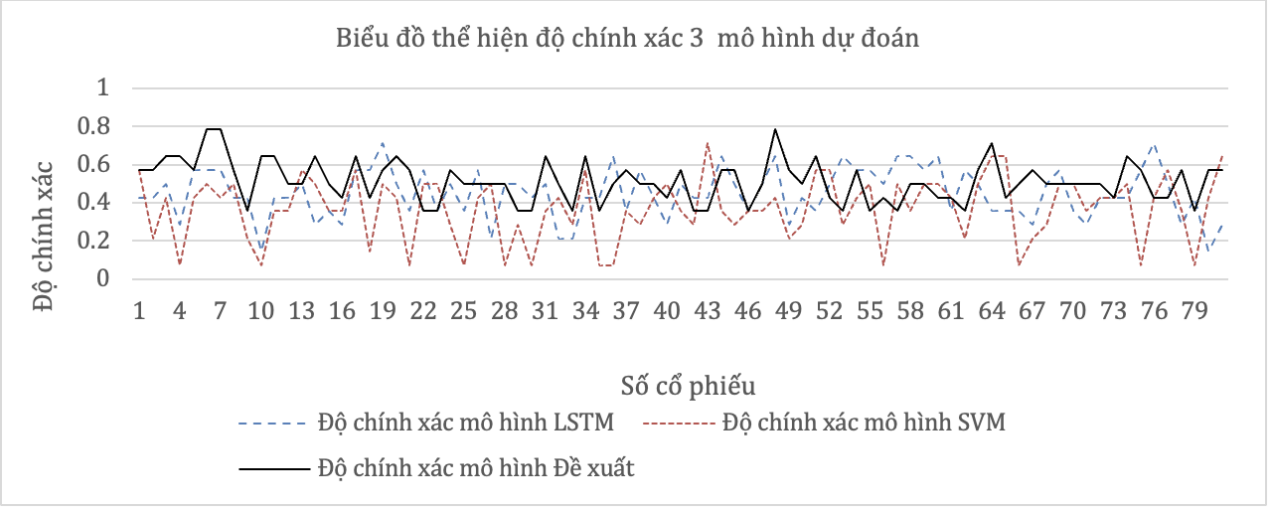
Nhóm 2 thực nghiệm mô hình SVM sử dụng phương pháp phân loại để dự đoán xu hướng, dữ liệu 5 ngày liên tiếp theo thứ tự thời gian được nối với nhau để giữ nguyên thứ tự của chuỗi nến và nó tạo thành 1 vector có 20 phần tử làm đầu vào của mô hình SVM với các tham số **C=1.0, kernel='linear'**. Xu hướng dự đoán ngày thứ 6 tăng được gán nhãn 1, xu hướng giảm là -1 và xu hướng không tăng cũng không giảm là 0.

Mô hình LSTM tỏ ra hiệu quả trong dự đoán phụ thuộc xa dữ liệu chuỗi thời gian. Mô hình sử dụng phương pháp hồi quy để dự đoán xu hướng, đầu vào mô hình giống mô hình SVM với tham số **units=50 dropout=0.1 và các tham số loss=mean_squared_error, optimizer=adam, epochs=100, batch_size=32.** Kết quả dự đoán nếu lớn hơn 0 là xu hướng tăng, nhỏ hơn 0 là xu hướng giảm và bằng 0 là xu hướng không tăng cũng không giảm.

Bảng 8 cho thấy kết quả thực nghiệm mô hình LSTM thấp hơn do mô hình cần nhiều dữ liệu được thu thập trong thời gian dài, mô hình đề xuất có hiệu quả hơn với dữ liệu có thời gian ngắn phù hợp với sự biến động phức tạp của thị trường chứng khoán. So sánh thời gian thực hiện mô hình SVM có thời gian nhỏ nhất nhưng độ chính xác không cao, cũng giống mô hình LSTM cần thêm dữ liệu. Mô hình đề xuất đạt được kết quả tốt trong một vài điều kiện nhất định phù hợp với dự đoán xu hướng ổn định trong thời gian ngắn.

Bảng 8. *Kết quả của các mô hình dự đoán trong thực nghiệm nhóm 2*

Mô hình dự đoán	Độ chính xác trung bình	Thời gian thực hiện
Đề xuất	51.3%	0.891 giây
SVM	36.9%	0.302 giây
LSTM	45.0%	44.755 giây



Hình 3. *So sánh độ chính xác của 3 mô hình.*

Hình 3 thể hiện mô hình đề xuất có ưu điểm hơn, nhiều cổ phiếu có độ chính xác trên 60%, ngược lại mô hình SVM có nhiều cổ phiếu có độ chính xác thấp.

VI. KẾT LUẬN

Bài báo trình bày cách tiếp cận phương pháp khai thác mẫu con tuần tự vào việc dự đoán xu hướng cổ phiếu với mô hình đề xuất DMSP-TS. Mô hình đề xuất dựa trên thuật toán khai thác mẫu con tuần tự có hiệu chỉnh cho phù hợp với bài toán đặt ra. Ngoài ra, chúng tôi đề xuất độ tương tự của mẫu nến kết hợp với độ chính xác của mẫu nến để cải thiện kết quả dự đoán. Độ tương tự được đặt trọng số tỏ ra hiệu quả trong dự đoán và giúp giải quyết vấn đề mẫu nến có cùng độ chính xác khi so khớp. Bài báo còn trình bày cách mã hóa dữ liệu cổ phiếu đa biến và cách xác định xu hướng. Kết quả thực nghiệm đạt hiện quả cao tốt hơn so với các phương pháp khác.

Tuy nhiên, mô hình đề xuất có kết quả dự đoán còn thấp vì vậy hướng phát triển tiếp theo là xác định xu hướng theo 1 khoảng giá nào đó. Ngoài ra, thông tin trên mạng xã hội và tin tức về tài chính cũng là yếu tố ảnh hưởng xu hướng cổ phiếu. Do đó, việc kết hợp với mô hình xử lý ngôn ngữ để trích xuất đặc trưng có liên quan và phân loại cổ phiếu trước khai thác để tăng hiệu quả dự đoán xu hướng. Cuối cùng mô hình cần thực nghiệm và đánh giá trên nhiều tập dữ liệu của nhiều thị trường chứng khoán ở nhiều quốc gia khác nhau.

VII. TÀI LIỆU THAM KHẢO

[1] Kamber, J. Han and Micheline, Data Mining: Concepts and Techniques, San Francisco: Diane Cerra, 2006.

[2] S. Nison, Japanese Candlestick Charting Techniques, Prentice Hall Press, 2001.

[3] A. Kumar and M. Chaudhry, "Review and Analysis of Stock Market Data Prediction Using Data mining Techniques," in *International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, 2022.

[4] A. J. N. K. M. C. A. A. I. Milon Biswas, "Stock Market Prediction: A Survey and Evaluation," *International Conference on Science & Contemporary Technologies (ICSCT),* pp. 1-6, 2021.

[5] L. Zhe, "Research on China's stock exchange markets: Problems and improvement," in *International Conference on Education and Management Technology*, Cairo, Egypt, 2010.

[6] J. Chen, "technical analysis of stocks and trends," 6 12 2021. [Online]. Available: https://www.investopedia.com/terms/t/technical-analysis-of-stocks-and-trends.asp.

[7] Y. Lin, H. Guo and J. Hu, "An SVM-based approach for stock market trend prediction," in *International Joint Conference on Neural Networks (IJCNN)*, 2014.

[8] S. D. &. M. Esterabi, "Predicting stock returns of Tehran exchange using LSTM neural network and feature engineering technique," *Multimedia Tools and Applications,* vol. 80, no. 13, p. 19947–19970, 2021.

[9] Z. L. N. H. B. V. &. J. C.-W. L. Jimmy Ming-Tai Wu, "A graph-based CNN-LSTM stock price prediction algorithm with leading indicators," *Multimedia Systems,* vol. 27, no. 1, 2021.

[10] Z. Zhao, R. Rao, S. Tu and J. Shi, "Time-Weighted LSTM Model with Redefined Labeling for Stock Trend Prediction," in *International Conference on Tools with Artificial Intelligence (ICTAI)*, 2018.

[11] R. Agrawal and R. Srikant, "Mining sequential patterns," *Proc. of IEEE International Conference on Data Engineering,* pp. 3-14, 1995.

[12] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering* , 2002.

[13] S. T. David Enke, "The use of data mining and neural networks for forecasting stock market returns," *Expert Systems with Applications,* vol. 29, no. 4, pp. 927-940, 2005.

[14] Q. P. Z. Z. Y. S. H. H. N. N. A. Tian Han, "A pattern representation of stock time series based on DTW," *Statistical Mechanics and its Applications Volume 550,* vol. 550, 2020.

[15] Forex, "Forex Pro Center," 2018. [Online]. Available: https://forexprocenter.com/hoc-trade-all-mo-hinh-gia/cac-dang-bieu-do-co-ban.html.

APPLICATION OF SEQUENTIAL PATTERN MINING FOR STOCK PRICE TREND PREDICTION

Nguyen Tuan Dung, Tran Minh Thai

ABSTRACT— Stock trend prediction is an essential support for investors. Accurate and fast prediction is being applied by researchers using various models. The method of prediction by mining historical data, the candlestick chart is one of the technical analysis tools used by investors to create a stock trading strategy. In particular, the application of data mining to predict stock trends is a new approach. In this paper, we propose a model using data mining techniques to predict stock trends. The predictive model is based on a sequential pattern mining algorithm on a historical data set of stocks. Identifying patterns through similarity is also presented in the paper. Experimental data were collected on https://finance.yahoo.com. The experimental results of the proposed model have better average accuracy than traditional models such as SVM and LSTM.

Keywords —Data mining, sequential pattern, stock trend prediction, candlestick chart.

TS. Trần Minh Thái tốt nghiệp cử nhân CNTT năm 2001 và thạc sỹ Tin học năm 2006 ĐH Khoa học Tự nhiên – ĐH Quốc gia TP.HCM, nhận bằng tiến sỹ CNTT năm 2017 do ĐH Quốc gia TP.HCM cấp; từng là giảng viên và quản lý khoa CNTT trường CĐ CNTT TP.HCM từ 2002 đến 2015. Từ 2015 đến hiện tại, anh là giảng viên và là trưởng bộ môn HTTT thuộc khoa CNTT trường ĐH Ngoại ngữ - Tin học TP.HCM. Lĩnh vực nghiên cứu chính của anh liên quan đến vấn đề khai thác dữ liệu, ẩn dữ liệu, xử lý dữ liệu lớn và nhận dạng.

Cử nhân Nguyễn Tuấn Dũng tốt nghiệp đại học ngành CNTT và là nhân viên kỹ thuật phòng máy tại trường ĐH Ngoại ngữ - Tin học TP.HCM. Hiện tại, anh đang là học viên cao học ngành CNTT tại ĐH Công nghệ Thông tin – ĐH Quốc gia Tp.HCM. Lĩnh vực nghiên cứu chính là Khai thác dữ liệu