DỰ BÁO TRÊN CHUỖI THỜI GIAN SỬ DỤNG BÀI TOÁN TÌM KIẾM TƯƠNG TỰ
PREDICTION IN TIME SERIES USING SIMILARITY SEARCH PROBLEM

Nguyễn Thành Sơn
Trường đại học Sư phạm Kỹ thuật TP.HCM

Ngày tòa soạn nhận được bài 17/3/2015, ngày phản biện đánh giá 03/4/2015, ngày chấp nhận đăng 15/4/2015

TÓM TẮT

Bài toán dự báo trên chuỗi thời gian là bài toán quan trọng trong nhiều lĩnh vực và đã nhận được nhiều sự quan tâm từ các nhà nghiên cứu trong những năm gần đây. Trong bài báo này, chúng tôi nghiên cứu cách sử dụng bài toán tìm kiếm tương tự vào bài toán dự báo trên chuỗi thời gian có xu hướng hoặc theo mùa. Phương pháp này được thực hiện như sau: (1) Trích một chuỗi giá trị trên chuỗi thời gian ngay trước khoảng thời gian muốn dự báo, (2) Sử dụng chuỗi này để tìm k lân cận gần nhất (hoặc các lân cận trong phạm vi một ngưỡng tương tự T cho trước) của nó trong dữ liệu quá khứ, (3) Trích các chuỗi (có chiều dài bằng với chiều dài muốn dự báo) ngay liền sau mỗi chuỗi lân cận tìm được, và (4) Chuỗi dự báo được xác định bằng cách tính trung bình các chuỗi tìm được trong bước (3). Kết quả thực nghiệm cho thấy cách tiếp cận này cho kết quả (về độ chính xác và thời gian thực thi) có thể cạnh tranh được khi so sánh với kết quả dự báo trên chuỗi thời gian có xu hướng hoặc theo mùa sử dụng mạng nơ ron nhân tạo (ANN). Trong thực nghiệm, chúng tôi cũng xem xét ảnh hưởng của k và T đến độ chính xác của dự báo.

Từ khóa: *Chuỗi thời gian, dự báo, tìm kiếm tương tự.*

ABSTRACT

Time series forecasting problem is very important problem in several domains and has received a lot of interest from researchers in recent years. In this paper, we investigate the use of pattern matching technique in seasonal or trend time series prediction. This method is performed as follows: (1) This technique retrieves the sequence prior to the interval to be forecasted, (2) This sequence is used as a sample for searching k-nearest neighbors or neighbors within a threshold T in historical data, (3) Sequences next to these found patterns are retrieved (the length of them are equal to the prediction interval), and (4) The forecasted sequence is calculated by averaging the sequences found in the 3rd step. The experimental results showed that this approach produces competitive results on seasonal or trend time series in comparison to artificial neural network (ANN) in terms of prediction accuracy and time efficiency. In our experiment, we also examine the impact of parameter values k and T on the predictive accuracy.

Keywords*: time series, prediction, similarity search.*

I. GIỚI THIỆU

Một chuỗi thời gian là một chuỗi các số thực. Mỗi số biểu diễn một giá trị đo được tại những khoảng thời gian bằng nhau. Dữ liệu chuỗi thời gian tồn tại trong nhiều ứng dụng của các lĩnh vực khác nhau như khoa học, kỹ thuật, kinh tế, tài chính, y học, quản lý hành chính, v.v… .

Dự báo trên chuỗi thời gian là một trong những công việc thách thức và phức tạp nhất trong khai phá dữ liệu chuỗi thời gian. Hệ thống dự báo chuỗi thời gian dự báo các giá trị tương lai của chuỗi thời gian bằng cách xem xét dữ liệu thu thập được trong quá khứ. Độ chính xác của dự báo trên chuỗi thời gian sẽ là cơ sở cho nhiều tiến trình ra quyết định và vì vậy việc nghiên cứu cải tiến độ hiệu quả của các phương pháp dự báo sẽ không bao giờ kết thúc. Các phương pháp dự báo thường được

chia thành ba loại: dự báo ngắn hạn, trung hạn và dài hạn.

- Dự báo ngắn hạn là dự báo những gì sẽ xảy ra trong khoảng thời gian ngắn ở tương lai như ngày, tuần, tháng.
- Dự báo trung hạn là dự báo những gì sẽ xảy ra trong khoảng thời gian dài hơn ở tương lai như một năm, hai năm.
- Dự báo dài hạn là dự báo những gì sẽ xảy ra trong nhiều năm ở tương lai.

Công việc của chúng tôi là nghiên cứu cách sử dụng bài toán tìm kiếm tương tự trong dự báo trên chuỗi thời gian dạng mùa hoặc có xu hướng. Đầu tiên, phương pháp này trích một chuỗi ngay trước khoảng thời gian cần dự báo. Sau đó, chuỗi này được dùng như một mẫu để tìm kiếm *k* lân cận gần nhất hay các lân cận trong phạm vi một ngưỡng *T* cho trước. Tiếp theo, trích các chuỗi (có độ dài bằng với độ dài của khoảng thời gian cần dự báo) liền ngay sau mỗi lân cận tìm được. Cuối cùng, chuỗi dự báo được xác định bằng cách tính trung bình các chuỗi vừa tìm được ở bước trước.

Trong thực nghiệm, chúng tôi so sánh phương pháp dự báo đề xuất với phương pháp ANN. Phương pháp ANN được chọn để so sánh vì nó là phương pháp thường được dùng để dự báo trên chuỗi thời gian trong những năm gần đây và có khả năng dự báo tốt hơn trên dữ liệu chuỗi thời dạng gian phi tuyến, phức tạp khi so sánh với các phương pháp truyền thống ([22]). Chúng tôi cũng xem xét ảnh hưởng của tham số k và T tới độ chính xác của dự báo.

Kết quả thực nghiệm cho thấy cách tiếp cận này cho kết quả (về độ chính xác và thời gian thực thi) có thể cạnh tranh được khi so sánh với kết quả dự báo trên chuỗi thời gian có xu hướng hoặc theo mùa sử dụng mạng nơ ron nhân tạo (ANN).

Phần còn lại của bài báo được tổ chức như sau. Trong phần 2, chúng tôi trình bày tóm tắt các kiến thức nền tảng và các kết quả nghiên cứu liên quan của các tác giả khác. Phần 3 mô tả phương pháp dự báo trên chuỗi thời gian do chúng tôi đề xuất. Phần 4 là phần đánh giá bằng thực nghiệm phương pháp đề xuất trên các tập dữ liệu thực. Phần 5 là kết luận và hướng phát triển.

II. KIẾN THỨC LIÊN QUAN VÀ CÁC NGHIÊN CỨU TRƯỚC ĐÂy

1. Kiến thức liên quan

- **Độ đo Euclid.**

Độ đo Euclid là phương pháp đơn giản để đo độ tương tự của các chuỗi thời gian. Cho hai chuỗi thời gian Q = {q1, …, qn} và C = {c1, …, cn}, độ đo Euclid giữa Q và C được định

$$D(Q,C) \equiv \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2}$$

- **Biểu diễn MP_C (Middle Point and Clipping)**

Đây là phương pháp thu giảm số chiều chuỗi thời gian do chúng tôi đề xuất trong nghiên cứu trước đây [17]. Phương pháp này có thể đươc tóm tắt như sau:

Cho một chuỗi thời gian *C* có chiều dài *n*. *C* được chia thành *m* đoạn bằng nhau (*m* do người dùng chọn). Các điểm giữa của mỗi đoạn đươc trích ra và được chuyển đổi thành chuỗi nhị phân, trong đó điểm giữa được chuyển thành 1 nếu nó nằm trên đường trung bình, ngược lại thì nó bằng 0. Giá trị trung bình và chuỗi nhị phân tương ứng được lưu lại như đặc trưng của chuỗi.

- **Cấu trúc chỉ mục đa chiều dùng cho chuỗi thời gian**

Cấu trúc chỉ mục đa chiều thông dụng là R-tree và các biến thể của nó ([6], [1]). Một R-tree là một cây cân bằng tương tự như B-tree. Trong một cấu trúc chỉ mục đa chiều như R-tree hay R*-tree, mỗi nút được kết hợp với một vùng bao hình chữ nhật nhỏ nhất (MBR-Minimum Bounding Recrangle). Một MBR tại một nút là vùng bao nhỏ nhất bao các nút con của nó. Mỗi phần tử trong nút lá chứa một MBR của chuỗi thời gian và một con trỏ đến đối tượng dữ liệu nguyên thủy được bao bởi MBR. Điểm yếu của R-tree là các MBR trong các nút trên cùng một mức có thể phủ

lấp nhau. Sự phủ lấp (overlap) này có thể làm giảm hiệu quả thực thi của việc tìm kiếm dựa vào chỉ mục.

Chỉ mục Skyline được đề xuất bởi Li et al., 2004 [16]. nhằm khắc phục tình trạng phủ lấp (overlap) giữa các hình chữ nhật chặn bên trong các MBR của các chuỗi bằng cách định nghĩa một vùng bao mới gọi là *vùng bao đường chân trời* (Skyline Bounding Region - SBR) thay cho MBR. Vùng bao SBR dùng để xấp xỉ và biểu diễn một nhóm các chuỗi thời gian theo hình dạng chung của chúng. Một SBR được định nghĩa trong cùng không gian *thời gian-giá trị* như chuỗi thời gian. SBR cho phép chúng ta định nghĩa một hàm khoảng cách là chặn dưới của khoảng cách giữa một câu truy vấn và một nhóm các chuỗi thời gian. vùng bao SBR chỉ bao gồm một vùng duy nhất và không xảy ra tình trạng phủ lấp. Bằng thực nghiệm, các tác giả cho thấy chỉ mục đường chân trời có thể cải thiện hiệu quả của bài toán tìm kiếm tương tự lên gấp 3 lần [16]

2 Các nghiên cứu trước đây

Nhiều phương pháp dự báo chuỗi thời gian đã được giới thiệu và đưa vào ứng dụng trong thực tế. Một số phương pháp thường được sử dụng cho bài toán dự báo dữ liệu chuỗi thời gian như phương pháp làm trơn theo hàm mũ (exponential smoothing) ([7]), mô hình ARIMA (autoregressive integrated moving average) ([3],[13],[14]), mạng nơ ron nhân tạo (artificial neural network – ANN) ([2], [4], [8], [9], [21], [22]) và máy véc tơ hỗ trợ ([15], [19]). Trong đó, phương pháp làm trơn theo hàm mũ và mô hình ARIMA là các mô hình tuyến tính vì chúng chỉ có thể nắm bắt được các đặc trưng tuyến tính của chuỗi thời gian, còn ANN là một mô hình phi tuyến đã được sử dụng cho bài toán dự báo dữ liệu chuỗi thời gian. Tuy nhiên, vấn đề mô hình ANN có thể xử lý một cách hiệu quả dữ liệu có tính xu hướng và tính mùa hay không đang là một vấn đề gây bàn cãi vì có những nhận định trái ngược nhau trong cộng đồng nghiên cứu về dự báo dữ liệu chuỗi thời gian [22].

Năm 2007, Nayak và te Braak đã đề xuất phương pháp dự báo cho dữ liệu thị trường chứng khoán sử dụng thuật toán gom cụm [18]. Phương pháp này dựa trên ý tưởng là một cụm được hình thành quanh một biến cố có thể được dùng để ước lượng cho biến cố ở tương lai. Cụm đó cần được xác định với bán kính nhỏ nhất có thể.

Cũng trong năm 2007, Troncoso và các cộng sự đã đề xuất một phương pháp dự báo được gọi là phương pháp dự báo dựa vào chuỗi mẫu (pattern sequence-based forecasting – PSF) [20]. Phương pháp này sử dụng thuật toán k-Means để gom cụm dữ liệu và phát sinh ra một chuỗi các nhãn phân cụm. Cuối cùng phương pháp thực hiện dự báo dựa trên các nhãn này. Cách tiếp cận này đã giới thiệu một phương pháp luận mới có thể cung cấp các qui luật dự báo dựa trên các nhãn dữ liệu thu được một cách tự động từ thuật toán gom cụm. Năm 2011, phương pháp này đã được ứng dụng dự báo giá thị trường điện và nhu cầu sử dụng điện [5]. Tuy nhiên, qua thực nghiệm chúng tôi thấy rằng kết quả dự báo phụ thuộc vào số cụm và việc xác định số cụm tốt nhất bằng cách gom cụm nhiều lần để chọn ra số cụm tốt nhất sẽ tốn nhiều thời gian. Ngoài ra, trong một số trường hợp bất thường, nếu các mẫu tìm kiếm không có trong tập huấn luyện, phương pháp này không thể dự báo các biến cố ở tương lai ngay cả khi chiều dài của mẫu là 1.

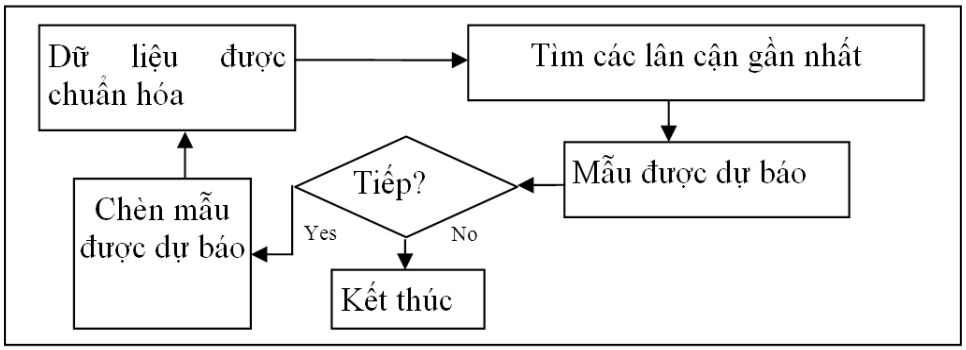
Năm 2009, Jang và các cộng sự đề nghị một phương pháp dự báo chuỗi thời gian chứng khoán dựa vào thông tin motif [12]. Sau khi phát hiện ra motif quan trọng nhất trong một chuỗi thời gian, motif đó được chia làm hai phần: tiền tố (prefix) và hậu tố (postfix). Nếu mẫu hiện hành của dữ liệu chuỗi thời gian khớp với tiền tố của motif, thì ta có thể dự đoán trị của bước thời gian kế tiếp dựa vào hậu tố của motif. Do giải thuật phát hiện motif được dùng trong công trình này không được hữu hiệu, nên độ chính xác dự báo và độ hữu hiệu về thời gian tính toán của phương pháp dự báo dựa vào motif chưa cao.

Năm 2010 và 2012, Huang và các cộng sự đề xuất một chiến lược kết hợp k-lân cận gần nhất với mô hình máy véc tơ hỗ trợ bình phương tối thiểu (least square support vector machine – LS-SVM) để dự báo dài hạn trên dữ liệu chuỗi thời gian [10][11].

III. PHƯƠNG PHÁP ĐỀ XUẤT

Chúng tôi sử dụng thuật toán tìm k lân cận gần nhất hoặc tìm lân cận trong phạm vi một ngưỡng cho trước dựa trên một cấu trúc chỉ mục đa chiều như chỉ mục đường chân trời.

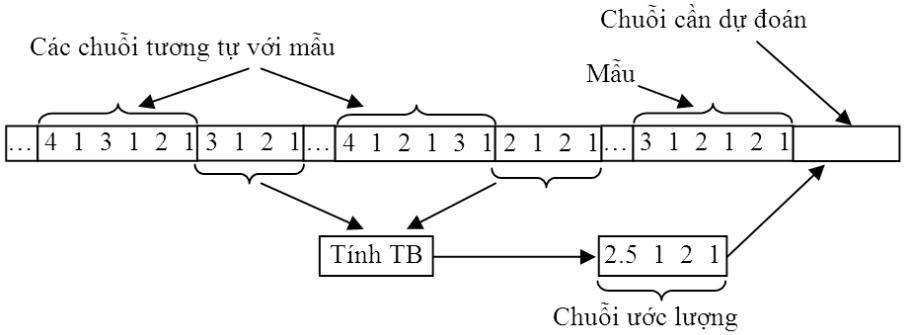
Cách tiếp cận k-lân cận gần nhất là một trong những kỹ thuật dự báo phi tham số (non-parametric), hiểu theo nghĩa người dùng không phải biết trước mối quan hệ lý thuyết nào giữa các trị xuất và các trị nhập trong bài toán dự báo, do đó nó rất tự nhiên và trực giác. Ý tưởng chính của cách tiếp cận này là nhận dạng các mẫu trong quá khứ khớp với mẫu hiện hành và dùng tri thức về cách mà chuỗi thời gian biến đổi trong quá khứ trong những tình huống tương tự để dự báo về biến đổi trong tương lai. Ngoài ra, với cách tiếp cận k-lân cận gần nhất này, các mẫu dự báo có thể được hồi tiếp trở lại vào tập dữ liệu để sử dụng cho các lần dự báo sau, nhờ vậy tầm (horizon) của dự báo có thể được kéo dài theo yêu cầu (kỹ thuật này được gọi là dự báo lặp – iterated prediction). Hình 1 trình bày ý tưởng cơ bản của cách tiếp cận này.



Hình 1. Ý tưởng cơ bản của cách tiếp cận dựa trên phương pháp so trùng mẫu.

Thuật toán dự báo chuỗi thời gian dựa vào kỹ thuật k-lân cận gần nhất được thực hiện như sau: Cho một trạng thái (mẫu) hiện hành có chiều dài w trong chuỗi thời gian có chiều dài n (w << n) và chúng ta phải dự đoán chuỗi có chiều dài m (m ≤ w) sẽ xảy ra ở bước kế tiếp theo thời gian (tức là dự báo m bước về phía tương lai). Đầu tiên, thuật toán sẽ tìm kiếm k lân cận gần nhất hay các lân cận trong một ngưỡng T cho trước đối với mẫu đó. Sau đó, thuật toán lấy các chuỗi có chiều dài m nằm kế cận bên phải của các lân cận gần nhất tìm được ở bước trên. Cuối cùng, chuỗi dự báo được ước lượng bằng cách tính trung bình cộng các chuỗi vừa thu được. Trong trường hợp cần dự báo cho các chuỗi khác nữa, chuỗi ước lượng có thể được chèn vào cuối tập dữ liệu để dự báo cho các mẫu tiếp theo.

Hình 2 minh họa bằng thí dụ thuật toán được đề xuất và hình 3 trình bày các bước chính của thuật toán này.



Hình 2. Minh họa thuật toán được đề xuất.

Chú ý là trong trường hợp m < w, chúng ta có thể dùng một biến để lưu tích lũy các chuỗi ước lượng cho tới khi m bằng với w. Khi đó, chúng ta có thể chèn chuỗi tích lũy được vào trong cấu trúc chỉ mục mà không cần phải xây dựng lại cấu trúc chỉ mục khi quay lại thực hiện bước 1.

Chúng tôi ứng dụng phương pháp MP_C [17] kết hợp với chỉ mục đường chân trời [16] vào bài toán dự báo dựa trên việc so trùng mẫu để dự báo trên dữ liệu chuỗi thời gian có xu hướng hoặc biến đổi theo mùa. Chỉ mục Skyline đươc chọn sử dụng vì nó nhiều ưu điểm hơn so với R*-tree.

Input: Chuỗi thời gian D có chiều dài n_1, tập kiểm tra TS có chiều dài n_2, chiều dài cửa sổ w, số lân cận gần nhất k (hoặc ngưỡng T) và chiều dài chuỗi cần dự báo m ($m \le w < n_2$ and $w << n_1$).

Output: Chuỗi ước lượng S có chiều dài m.

1. Thu giảm số chiều các chuỗi con có chiều dài w trong D và chèn chúng vào trong một cấu trúc chỉ mục đa chiều (nếu cần).
2. Lấy chuỗi S (mẫu) có chiều dài w nằm trước vị trí chuỗi ta phải dự báo trong TS.
3. Tìm k lân cận gần nhất (hay các lân cận nằm trong phạm vi ngưỡng T) của S.
4. Với mỗi lân cận gần nhất tìm được ở bước 3, khôi phục chuỗi có chiều dài m nằm kế cận nó trong D.
5. Tính trung bình cộng các chuỗi tìm được ở bước 4.
6. Trả lại kết quả ước lượng ở bước 5.
7. Chèn chuỗi ước lượng ở bước 5 vào D để dự báo các mẫu tiếp sau và quay lại bước 1 (nếu cần).

Hình 3. Các bước chính của thuật toán dự báo theo phương pháp đề xuất.

Chú ý là trong trường hợp m < w, chúng ta có thể dùng một biến để lưu tích lũy các chuỗi ước lượng cho tới khi m bằng với w. Khi đó, chúng ta có thể chèn chuỗi tích lũy được vào trong cấu trúc chỉ mục mà không cần phải xây dựng lại cấu trúc chỉ mục khi quay lại thực hiện bước 1.

IV. ĐÁNH GIÁ BẰNG THỰC NGHIỆM

1. Môi trường và dữ liệu thực nghiệm

Chúng tôi so sánh sự thực thi của phương pháp dự báo đề xuất với sự thực thi của phương pháp ANN. Thực nghiệm được thực hiện trên bốn tập dữ liệu thực: Temperatures at Savannah International Airport, Fraser River (FR), Milk production (MP) and Carbon Dioxide (CD). Phương pháp đề xuất được cài đặt bằng Microsoft Visual C# trên laptop Core i3, Ram 2GB. ANN (sử dụng Spice-Neuro) với cấu trúc sau: 12 nút input, 3 nút output cho hai tập dữ liệu MP và CD, 12 nút output cho các tập dữ liệu khác. Hai phương pháp dự báo được so sánh sự thực thi trên tất cả các đoạn của tập dữ liệu kiểm tra và sau đó tính lỗi trung bình trong khoảng dự báo.

Các tập dữ liệu được chia thành hai tập con theo tỉ lệ xấp xỉ là 9:1. Trong đó lấy khoảng 90% làm tập huấn luyện và khoảng 10% làm tập kiểm tra. Các tập dữ liệu dùng trong thực nghiệm như mô tả sau:

- Tập dữ liệu Temperatures at Savannah International Airport, từ 1/1910 đến 12/2010. Tập huấn luyện được chọn từ 1/1910 đến 12/2000 và tập kiểm tra từ 1/2001 đến 12/2010.
- Tập dữ liệu Fraser River dataset, từ 1/1913 đến 12/1990. Tập huấn luyện được chọn từ 1/1913 đến 12/1982 và tập kiểm tra từ 1/1983 đến 12/1990.
- Tập dữ liệu Milk Production, từ 1/1962 đến 12/1975. Tập huấn luyện được chọn từ 1/1962 đến 12/1971 và tập kiểm tra từ 1/1972 đến 12/1975.
- Tập dữ liệu Carbon Dioxide dataset, từ 1/1959 đến 12/2008. Tập huấn luyện được chọn từ 1/1959 đến 12/1998 và tập kiểm tra từ 1/1999 đến 12/2008.

Tất cả các tập dữ liệu trên được lấy từ web site: http://www.datamarket.com**. Hình**

4 minh họa hình dạng của các tập dữ liệu thực nghiệm dưới dạng đồ họa.

2. Tiêu chuẩn đánh giá

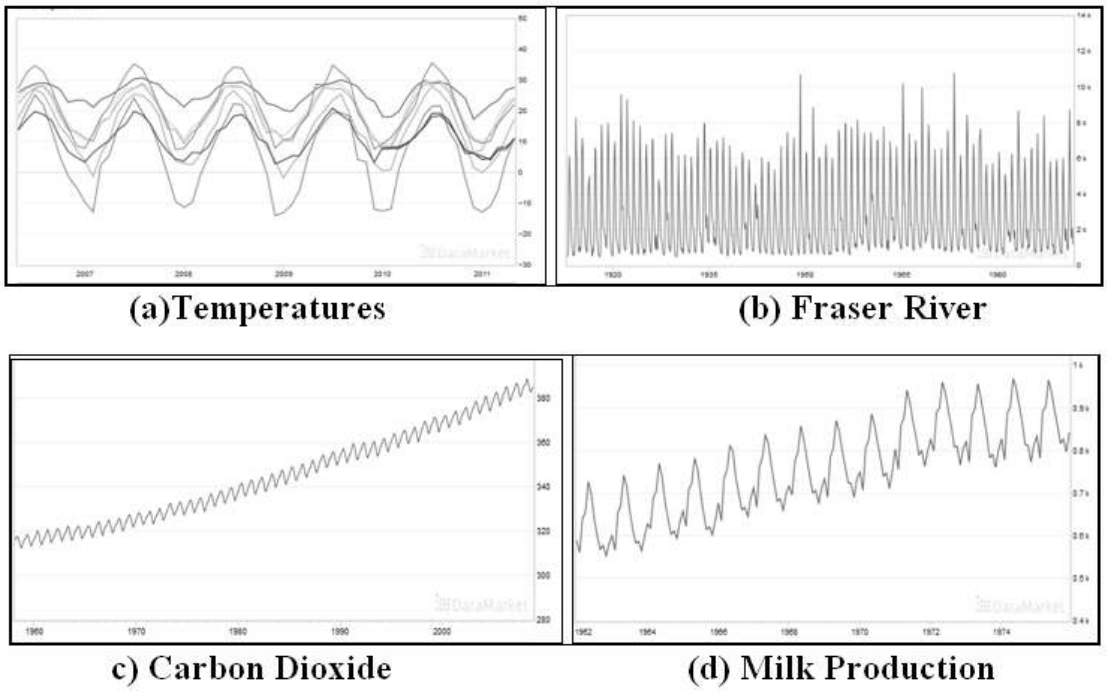
Trong bài báo này, chúng tôi sử dụng hai tiêu chuẩn đánh giá thường dùng là Lỗi trung bình tương đối so với x_{mean} (MER - Mean Error Relative) và Lỗi trung bình tuyệt đối (MAE - Mean Absolute Error) được định nghĩa như sau [5]:

- $$MER = 100 \times \frac{1}{N}\sum_{i=1}^{N}\frac{\left|x_{model,i} - x_{obs,i}\right|}{x_{mean}}$$ với x_{mean}

Trong đó, x_{obs} là giá trị quan sát được, x_{model} là giá trị tính được bởi mô hình tại thời điểm i, x_{mean} là giá trị trung bình trong khoảng thời gian xem xét và N là chiều dài của chuỗi dự báo.

- Lỗi trung bình tuyệt đối (MAE).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|x_{model,i} - x_{obs,i}\right|$$



Hình 4. Minh họa hình dạng bốn tập dữ liệu thực nghiệm.

3. Kết quả thực nghiệm

Để xem xét ảnh hưởng của k và ngưỡng T tới độ chính xác của dự báo, chúng tôi tiến hành thực nghiệm với các giá trị k và T khác nhau sau đó tính trung bình lỗi dự báo. Bảng 1 là các lỗi dự báo của thực nghiệm trên tập dữ liệu Frazer River với k thay đổi từ 1 đến 10.

Bảng 1. Lỗi dự báo của thực nghiệm trên tập Frazer River với k khác nhau

K	MER (%)	MAE	k	MER (%)	MAE
1	26.62	0.055	6	24.31	0.050
2	29.20	0.060	7	23.29	0.048
3	23.74	0.049	8	22.70	0.047
4	22.46	0.046	9	23.00	0.047
5	24.39	0.050	10	22.66	0.047

Kết quả thực nghiệm cho thấy lỗi dự báo sẽ khác nhau khi thực nghiệm với các giá trị k khác nhau. Trong thực nghiệm này, ta có thể thấy lỗi dự báo là nhỏ nhất với k bằng 4.

Bảng 2 là kết quả lỗi dự báo khi thực nghiệm trên tập dữ liệu Frazer River với các giá trị T khác nhau. Kết quả thực nghiệm cho thấy lỗi dự báo sẽ khác nhau khi thực nghiệm với các giá trị T khác nhau. Trong thực nghiệm này, ta có thể thấy lỗi dự báo là nhỏ nhất với T bằng 0.21.

Bảng 2. Lỗi dự báo của thực nghiệm trên tập Frazer River với *T* khác nhau.

T	0.15	0.17	0.19	0.21	0.23	0.25
MER (%)	27.94	27.05	25.64	23.11	25.29	25.91
MAE	0.056	0.055	0.052	0.047	0.051	0.052

Bảng 3 là lỗi dự báo của thực nghiệm trên tập dữ liệu Frazer River với giá trị k tốt nhất khi thực nghiệm dự báo sử dụng bài toán k lân cận gần nhất (k-NN) và giá trị T tốt nhất khi dự báo sử dụng bài toán tìm kiếm lân cận theo ngưỡng T (Range search). Lỗi dự báo được tính cho từng năm. Dòng cuối của bảng là lỗi dự báo trung bình trong tám năm. Kết quả thực nghiệm cho thấy lỗi dự báo trong cả hai trường hợp là xấp xỉ nhau.

Bảng 3. Lỗi dự báo của thực nghiệm trên tập Frazer River với giá trị *k* và *T* tốt nhất.

Year	MER (%)		MAE	
	k-NN	Range search	*k*-NN	Range search
1	24.27	21.87	0.06	0.06
2	18.94	16.75	0.04	0.03
3	28.48	22.39	0.06	0.05
4	15.15	26.86	0.03	0.05
5	25.77	22.66	0.05	0.05
6	32.20	28.52	0.06	0.05
7	18.57	20.86	0.04	0.04
8	21.12	25.02	0.04	0.05
Mean	23.06	24.16	0.05	0.05

Bảng 4 là lỗi dự báo của thực nghiệm trên tập dữ liệu Temperatures at Savannah International Airport. Lỗi dự báo được tính cho từng năm. Dòng cuối của bảng là lỗi dự báo trung bình trong mười năm.

Do giới hạn số trang của bài báo, trong bảng 5 chúng tôi chỉ trình bày kết quả tổng hợp từ thực nghiệm trên các tập dữ liệu khác nhau. Các giá trị trong bảng là lỗi dự báo trung bình trong các năm thực hiện dự báo.

Kết quả thực nghiệm cho thấy mặc dù lỗi dự báo trong một vài năm của phương pháp do chúng tôi đề xuất lớn hơn lỗi dự báo của phương pháp ANN, nhưng lỗi dự báo trung bình trong các năm dự báo của phương pháp do chúng tôi đề xuất luôn nhỏ hơn lỗi dự báo của phương pháp ANN. Chỉ có trường hợp thực nghiệm trên tập Carbon Dioxide, lỗi trung bình MAE khi sử dụng k lân cận gần nhất là lớn hơn một ít so với lỗi dự báo trung bình MAE của ANN. Tuy nhiện lỗi trung bình MER khi sử dụng k lân cận gần nhất thì nhỏ hơn lỗi dự báo trung bình MER của ANN.

Bảng 4. Lỗi dự báo của thực nghiệm trên tập Temperatures at Savannah International Airport.

Year	MER(%)		MAE	
	k-NN	ANN	*k*-NN	ANN
1	7.555	17.814	0.043	0.065
2	6.779	11.666	0.039	0.059
3	8.316	11.523	0.047	0.039
4	6.288	10.239	0.035	0.036
5	7.652	8.921	0.042	0.039
6	8.329	10.053	0.047	0.040
7	7.570	9.590	0.044	0.044
8	7.767	11.335	0.045	0.053
9	5.004	8.298	0.029	0.035
10	14.542	14.394	0.081	0.049
Mean	7.980	11.383	0.045	0.046

Bảng 5. Lỗi dự báo trung bình khi thực nghiệm trên các tập dữ liệu khác nhau.

Dataset	MER (%)		MAE	
	k-NN	ANN	*k*-NN	ANN
Frazer River	23.06	24.16	0.05	0.06
Milk Production	8.06	14.73	0.09	0.10
Carbon Dioxide	3.38	3.61	0.037	0.032

Bên cạnh việc đánh giá về độ chính xác, chúng tôi còn so sánh hai phương pháp dự báo về thời gian thực thi. Bảng 6 là thời gian thực thi (tính theo giây) của hai phương pháp dự báo thực nghiệm trên bốn tập dữ liệu. Kết quả thực nghiệm cho thấy phương pháp dự báo sử dụng *k* lân cận gần nhất luôn thực thi nhanh hơn khi so sánh với phương pháp ANN.

Bảng 4. Thực nghiệm về thời gian thực thi của hai phương pháp dự báo trên bốn tập dữ liệu.

Dataset	ANN	*k*-NN
Temperatures	50	0.262
Milk Production	4	0.464
Carbon Dioxide	37	1.261
Frazer River	58	0.199

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài báo này, chúng tôi đã đề xuất phương pháp dự báo trên chuỗi thời gian dạng mùa hoặc có xu hướng sử dụng bài toán tìm kiếm tương tự. Trong cách tiếp cận này, chúng tôi sử dụng phương pháp thu giảm số chiều MP_C kết hợp với chỉ mục Skyline cho bài toán tìm kiếm tương tự nhằm tăng nhanh tốc độ tìm kiếm. Chúng tôi cũng xem xét ảnh hưởng của k và T đến độ chính xác của dự báo. Thực nghiệm cho thấy với các giá trị k và T thích hợp, phương pháp dự báo sử dụng bài toán tìm kiếm tương tự sẽ cho kết quả tốt hơn so với ANN về độ chính xác và thời gian thực thi khi dự báo trên chuỗi thời gian dạng mùa hoặc có xu hướng.

Trong tương lai, chúng tôi dự định sẽ nghiên cứu cách xác định giá trị tốt nhất cho k và T một cách tự động cho bài toán dự báo sử dụng bài toán tìm kiếm k lân cận gần nhất hoặc sử dụng bài toán tìm lận cận trong một ngưỡng T.

TÀI LIỆU THAM KHẢO

[1] N. Beckman, H.P. Kriegel, R. Schneider and B. Seeger, "*The R*-tree: An efficient and robust access method for points and rectangles*", Proc. of 1990 ACM-SIGMOD Conf., Atlantic City, NJ, May 1990, pp. 322-331.

[2] S. D. Balkin and J. K. Ord, "*Automatic neural network modeling for univariate time series*", International Journal of Forecasting, vol.16, 2000, pp. 509–515.

[3] C. Chatfield, *Time-series forecasting*, New York, NY, Chapman and Hall, Inc., 2000

[4] E. Cadenas and W. Rivera, **"***Short term wind speed forecasting in La Venta, Oaxaca, México, using artificial neural networks***"**, Renewable Energy, vol. 34, no. 1, 2009, pp. 274–278.

[5] F. M. Álvarez, A. Troncoso, J. C. Riquelme and J. S. A. Ruiz, **"***Energy Time Series Forecasting Based on Pattern Sequence Similarity***",** IEEE Trans. on Knowledge and Data Engineering, vol. 23, No. 8, Aug. 2011, pp. 1230 – 1243.

[6] A. Guttman, **"***R-trees: a Dynamic Index Structure for Spatial Searching***",** Proc. of the

ACM SIGMOD Int. Conf. on Management of Data, June 18-21, 1984, pp. 47-57.

[7] S. Gelper, R. Fried, and C. Croux, "*Robust forecasting with exponential and Holt-Winters smoothing***",** Journal of Forecasting, vol. 29, 2010, pp. 285-300.

[8] M. Ghiassi, H. Saidane, and D. K. Zimbra, **"***A dynamic artificial neural network model for forecasting series events***",** International Journal of Forecasting, vol.21, 2005, pp. 341–362.

[9] S. Heravi, D. R. Osborn and C. R. Birchenhall, "*Linear versus neural network forecasting for European industrial production series*", International Journal of Forecasting, vol.20, 2004, pp. 435–446.

[10] Z. Huang and M. L. Shyu, **"***k-NN Based LS-SVM Framework for Long-Term Time Series Prediction***,"** in The 11th IEEE International Conference on Information Reuse and Integration (IRI 2010), Tuscany Suites & Casino, Las Vegas, Nevada, USA, 2010, pp. 69-74

[11] Z. Huang and M.-L. Shyu, **"***Long-Term Time Series Prediction using k-NN Based LS-SVM Framework with Multi-Value Integration***,"** in Recent Trends in Information Reuse and Integration, K. K. a. M. T. Tansel Ozyer, Ed. Springer Vienna, 2012, ch. 9, pp. 191-209.

[12] Y. Jiang, C. Li, J. Han, **"***Stock temporal prediction based on time series motifs***,"** in Proc. of 8th Int. Conf. on Machine Learning and Cybernetics, 2009.

[13] I.-B. Kang, **"***Multi-period forecasting using different models for different horizons:An application to U.S. economic time series data***",** International Journal of Forecasting, vol.19, 2003, pp. 387–400.

[14] J. H. Kim**, "***Forecasting autoregressive time series with biascorrected parameter estimators***",** International Journal of Forecasting, vol.19, 2003, pp. 493–502.

[15] K. J. Kim, **"***Financial time series forecasting using support vector machines***",** Neurocomputing, vol. 55, 2003, pp. 307-319.

[16] Q. Li, I. F. V. Lopez, and B. Moon, **"***Skyline Index for Time Series Data***",** IEEE Trans. on Knowledge and Data Engineering, vol.16, No. 6, 2004

[17] N. T. Son, D. T. Anh, **"***Time Series Similarity Search based on Middle Points and Clipping***",** Proceedings of the 3rd Conference on Data Mining and Optimization (DMO 2011), Putrajaya, Malaysia, June 28-29, 2011, pp.13-19.

[18] R. Nayak, and te Braak, **"***Temporal Pattern Matching for the Prediction of Stock Prices***",** In (Ong, K.-L. and Li, W. and Gao, J., Eds.) Proceedings 2nd International Workshop on Integrating Artificial Intelligence and Data Mining (AIDM 2007), pp. 99-107.

[19] Y.Radhika and M.Shashi, **"***Atmospheric Temperature Prediction using Support Vector Machines***,"** International Journal of Computer Theory and Engineering, vol. 1, no. 1, 2009, pp. 55-58.

[20] A. Troncoso, J. M. Riquelme, J. C. Riquelme, A. Gómez, and J. L. *Time series prediction: Application to the short term electric energy demand* Martínez, «", LNAI 3040, Springer, 2004, pp. 577- 586.

[21] G. Tkacz, **"***Neural network forecasting of Canadian GDP growth***",** International Journal of Forecasting, vol.17, 2001, pp. 57–69.

[22] G. P. Zhang, M. Qi, **"***Neural Network Forecasting for Seasonal and Trend Time Series***",** European Journal of Operational Research, vol. 160, 2005, pp. 501-514.