

Previsão de séries temporais multivariadas através da estrutura codificadora baseada em atenção



Shengdong Du ^a, Tianrui Li ^{a, *}, Yan Yang ^a, Shi-Jinn Horng ^{b, *}

^a Escola de Ciência da Informação e Tecnologia, Universidade de Jiaotong sudoeste, Chengdu 611756, China

^b Departamento de Ciência da Computação e Engenharia da Informação, Universidade Nacional de Ciência e Tecnologia de Taiwan, Taipei 10607, Taiwan

artille info

Histórico do artigo:

Recebido em 22 de março de 2019

Revisado em 9 de dezembro de 2019

Aceito em 31 de dezembro de 2019

Disponível online em 15 de janeiro de 2020

Comunicado pelo Dr. Wojciech Froelich

Keywords:

Séries temporais multivariadas

Atenção temporal

Previsão de várias etapas

Codificador-decodificador

Aprendizado profundo

Redes de memória de longo prazo

abstract

A previsão de séries temporais é uma técnica importante para estudar o comportamento dos dados temporais e prever valores futuros, que é amplamente aplicado em muitos campos, por exemplo, a previsão da qualidade do ar, previsão de carga de energia, monitoramento médico e detecção de intrusões. Neste artigo, propomos, em primeiro lugar, um novo modelo de codificador de atenção temporal para lidar com o problema de previsão de séries temporais multivariadas. É uma estrutura de aprendizagem profunda de ponta a ponta que integra o vetor de contexto tradicional de codificação e vec-tor de atenção temporal para o aprendizado de representação temporal conjuntamente, que é com base em camadas bidirecionais de longo prazo de redes de ory (Bi-LSTM) com mecanismo de atenção temporal como a rede codificante para aprender adaptativamente a dependência a longo prazo e características ocultas de correlação de dados temporais multivariados. Resultados experimentais extensivos em cinco conjuntos de dados típicos de séries temporais multivariadas mostraram que nosso modelo tem o melhor desempenho de previsão em comparação com os métodos de linha de base.

© 2020 Elsevier B.V. Todos os direitos reservados.

1. Introdução

A previsão de séries temporais tem recebido muita atenção nas últimas décadas devido às suas importantes aplicações em muitos campos [1], incluindo a previsão de fluxo de tráfego [2], poluição do ar [3], previsão [3], detecção de anomalias de séries temporais [4], análise de monitoramento médico [5], detecção de intrusão de rede, etc. De um modo geral, os dados da série temporal podem ser descritos como um conjunto de observações em ordem cronológica. Seu tipo pode ser dividido em séries temporais univariadas e séries temporais multivariadas, que têm as características de uma grande escala em tamanho de dados, alta dimensionalidade e mudança constante. Muitos estudiosos têm estudado o problema da previsão de séries temporais, especialmente com base em modelos estatísticos clássicos como ARIMA [6] e máquina típica modelos de aprendizagem como HMM [8], SVR [7] e ANN [9–11]. Mas muitos métodos tradicionais empregam principalmente o modelo estatístico para pesquisar a evolução dos dados temporais. Por exemplo, Zhang propôs um método híbrido que combina tanto ARIMA quanto ANN para modelagem de séries temporais lineares e não lineares [9]. Pai et al. apresentou um método híbrido SSVR para previsão de séries temporais, empregando modelos SARIMA e SVR [7]. Sapankevych e Sankar forneceram um levantamento de aplicações de previsão de séries temporais usando um modelo clássico de aprendizado de máquina: Suportam Máquinas vetoriais (SVM),

* autores correspondentes.

Endereço de e-mail: trli@swjtu.edu.cn (T. Li).

<https://doi.org/10.1016/j.neucom.2019.12.118> 0925-2312/©

2020 Elsevier B.V. Todos os direitos reservados.

which nos fornece uma visão sobre as vantagens e desafios usando o SVM para previsão de séries temporais [12].

Com a chegada da era dos big data, os dados de séries temporais multivariadas e multicanais estão aumentando explosivamente. Em muitos casos, os dados de séries temporais multivariadas têm características de dependência dimensional e espacial-temporal elevadas, ou contêm dados barulhentos, o que dificulta ser modelado efetivamente por métodos estatísticos clássicos [10]. Além disso, esses métodos tradicionais enfrentam dificuldades no processamento de big data, especialmente dados temporais multivariados massivos e complexos. Portanto, os métodos de pré-dição de séries temporais baseadas em dados são cada vez mais favorecidos pelos pesquisadores, e alguns pesquisadores fizeram progressos em muitas fields, por exemplo, a previsão da poluição do ar [3, 13], previsão de fluxo de tráfego [14], detecção de anomalias [4], e previsão de fluxo de multidão urbana [15]. Recentemente, o deep learning [16, 17] tem sido usado em muitas áreas e obteve os melhores resultados em muitos conjuntos de dados de benchmark, por exemplo. imagem [18] e processo de vídeo [19], compreensão da linguagem natural [20]. Embora os métodos tradicionais ainda possam ser usados para modelagem de

séries temporais, os métodos de previsão baseados em aprendizagem profunda estão se tornando mais populares [21–23].

Além disso, os métodos tradicionais de previsão de séries temporais também enfrentam os dois seguintes desafios: em primeiro lugar, muitos métodos clássicos resolvem principalmente o problema de previsão de passo único, que é limitado em aplicativos práticos de monitoramento e aviso prévio com base na modelagem de séries temporais. A previsão da série temporal em passo único

geralmente não ajuda muito em um aplicativo real de aviso antecipado, pois é difícil prever o que vai acontecer depois de uma condição de várias etapas. Além disso, a previsão em várias etapas é mais complicada do que a previsão em etapa única, que deve considerar mais condições, por exemplo, acúmulo de erros e degradação do desempenho de previsão [24]. Em segundo lugar, em alguns casos, há correlações entre variáveis em dados de séries temporais multivariadas, e uma melhor compreensão é frequentemente obtida pela modelagem de todas as variáveis relacionadas juntos do que apenas modelando uma variável. Por isso, é significativo estudar o modelo de previsão com base em dados multivariáveis da série temporal.

Em resposta às duas questões acima, um codificador baseado em Bi-LSTM– decodificador com mecanismo de atenção é proposto pela primeira vez, que visa alcançar o propósito de aprender adaptativamente o características de dependência temporal implícita de dados de séries temporais multivariadas. E as contribuições deste artigo incluem os seguintes aspectos:

- 1) Em primeiro lugar, propomos um novo modelo de codificador baseado em atenção temporal, decodificador e aplicamos-no às tarefas de previsão multivariada de séries temporais. O Bi-LSTM com mecanismo de atenção é usado para codificar as representações ocultas de dados de séries temporais multivariadas como o vetor do contexto temporal, e o outro LSTM é usado para decodificar a representação oculta para previsão. Através deste processo de ponta a ponta, características dependentes ocultas de longo prazo e recursos de correlação não lineares podem ser aprendidos com os dados da série temporal multivariada bruta.
- 2) Em segundo lugar, introduzimos o mecanismo de atenção temporal ser-interpolador a rede de encoder network e decodificador, que pode selecionar estados ocultos codificadores relevantes em todos os passos de tempo para várias etapas para várias etapas previsão com mais precisão, de modo a melhorar a capacidade de representação do nosso modelo de dados dinâmicos de séries temporais multivariadas.
- 3) Finalmente, demonstramos a eficácia do nosso modelo testando-o em cinco conjuntos de dados reais de séries temporais multivariadas, e os resultados do experimento mostraram que nosso modelo tem o melhor desempenho de fundição em comparação com os métodos de linha de base.

O resto do papel é organizado da seguinte forma: a Seção 2 dá uma visão geral dos trabalhos relacionados sobre modelagem e previsão de séries temporais. A Seção 3 expõe a motivação da pesquisa do modelo proposto, analisa a arquitetura geral da estrutura de codificação de atenção e descreve a teoria e o processo relevantes. Detalhes. A Seção 4 é o conteúdo de análise experimental. Um experimento comparativo baseado em cinco conjuntos de dados de séries temporais multivariadas é realizado, e os resultados experimentais são analisados detalhadamente. Finalmente, damos as conclusões do estudo e discutimos futuras prioridades de pesquisa.

2. Obras relacionadas

A chave para a modelagem de séries temporais é projetar métodos eficazes de representação de recursos para dados dinâmicos de séries temporais, que é um tópico desafiador e enfrenta problemas como alta dimensão, dinâmico e incerteza, e envolve uma ampla gama de aspectos e sempre foi o foco dos pesquisadores. Métodos tradicionais de previsão de séries temporais geralmente usam modelos estatísticos ou métodos típicos de learning de máquinas [1], por exemplo, ARIMA [6], SVR [7], HMM [8], e ANN [9]. E Ahmed et al. fez uma análise comparativa e pesquisa sobre métodos de previsão de séries temporais com base em modelos tradicionais de aprendizagem de máquinas [25]. Park et al. propôs um modelo de HMM contínuo para lidar com a tarefa de previsão da série de tempo financeiro [8].

De Gooi-je e Hyndman revisaram os últimos 25 anos de pesquisa sobre previsão de séries temporais e descobriram que ainda há muitos problemas em necessidade de mais pesquisas [1]. Hassan et al. apresentou um método de fusão com um mecanismo lógico difuso para processar dados temporais não lineares [26]. A Fu forneceu uma revisão abrangente dos métodos de mineração de dados de séries temporais [10]. Em termos de problemas de previsão de séries temporais multipassos, Taieb et al. revisou as estratégias existentes para previsões de vários passos e as comparou em termos teóricos e práticos [27].

Nos últimos anos, com o crescimento explosivo de big data spatio-temporal, muitos dados temporais coletados geralmente têm volumes massivos, dependência espaço-temporal, características dinâmicas e não lineares. Devido à complexidade e incerteza da representação dinâmica de séries temporais multivariadas, o aprendizado de máquina clássico e métodos estatísticos são difíceis de lidar com a evolutiva características dos dados temporais grandes [28]. Muitos pesquisadores tentam propor métodos de previsão orientados por dados, como o aprendizado profundo para lidar com o problema [17, 29]. O deep learning é considerado o método de aprendizagem de representação mais bem sucedido e fez progressos significativos em muitas áreas, por exemplo, imagens, vídeos, áudios e tarefas de compreensão da linguagem natural [18–20]. Além disso, o modelo de previsão de séries temporais baseadas em deep learning tornou-se um ponto de acesso à pesquisa [21, 30–33]. Por exemplo, Yao et al. desenvolveu um modelo de previsão de séries temporais que integra CNN e RNN [35]. Zheng et al. propôs um novo modelo de aprendizagem profunda baseado na CNN para a classificação de séries temporais multivariadas [34].

O codificador-decodificador é um estrutura sequenciado popular, que é um método simplificado e automático para modelagem de dados sequenciais [20, 36, 37]. Ele geralmente usa uma rede neural profunda (por exemplo, RNN ou LSTM) para codificar a entrada do modelo em um vetor fixo, e então emprega outra rede neural profunda para decodificar a sequência de saída de destino do vetor fixo para previsão [38]. Chorowski et al. melhorou o desempenho do modelo de reconhecimento da fala, ampliando o mecanismo de atenção [37]. Sutskever et al. propôs uma abordagem de ponta a ponta para o aprendizado de sequências e demonstrou que a estrutura simples pode alcançar um bom desempenho [38]. O modelo codificador-decodificador também pode ser usado para processamento de dados temporais. Por exemplo, Malhotra et al. desenvolveu um novo modelo de detecção de anomalias multi-sensores, que usou o LSTM como componente codificador e decodificador [48]. Park et al. empregava o modelo codificador-decodificador para analisar o padrão implícito dos dados trajetória [49]. Kuznetsov e Mariet primeiro realizaram uma análise profunda do modelo de aprendizagem profunda seq2seq e usado para previsão de séries temporais [39].

No entanto, a estrutura codificadora-decodificadora ainda não foi bem pesquisada para modelagem de dados temporais. Os métodos de pesquisa existentes raramente consideram a modelagem eficaz de sequências temporais de longo prazo a partir da perspectiva do mecanismo de atenção, que é a razão para nos motivar a investigar neste jornal. Em addition, a previsão de vários passos para a frente é mais difícil do que a previsão de etapa única de dados de séries temporais multivariadas e a arquitetura de aprendizado sequencial pode lidar com este problema bem. Neste artigo, diferente dos métodos estatísticos clássicos e de aprendizagem de máquina, propomos um modelo de fundição multivariada de séries temporais de várias etapas através de uma sequência baseada em atenção para sequência estrutura de aprender. Experimentos em cinco conjuntos de dados de séries temporais multivariadas mostraram que o modelo proposto pode efetivamente prever o valor da série temporal em diferentes condições.

3. Metodologia

3.1. Problema e definições

Os dados da série temporal são geralmente sequências de valores (em forma discreta ou contínua) medidos ao longo do tempo. A atualização dinâmica, a certeza e a alta dimensionalidade das séries temporais tornam-no diferente de outros dados, como imagem, texto e assim por diante. O fundição de séries temporais sempre foi uma área de pesquisa muito importante das tarefas de mineração de dados, cujo objetivo é prever a mudança dos valores futuros tem-poral. E o intervalo de tempo de observação de diferentes tipos de dados da série temporal é muitas vezes diferente e decidido pelas especificações do sensor, por exemplo, o intervalo de tempo de observação de dados da série de tempo de fluxo de tráfego talvez 5 min, 15 min, 60 min, etc. e o intervalo de tempo de observação de uma série de tempo PM2.5 geralmente é de uma hora.

Uma série temporal multivariada normalmente contém vários valores variáveis em cada etapa de tempo de observação. Cada variável de séries temporais multivariadas depende não apenas do seu valor passado, mas também de outros valores variáveis em alguns casos. Essas correlações potenciais são cruciais para modelar dados de séries temporais multivariadas. Como aprender as características de correção das múltiplas variáveis monitoradas de conjuntos de dados multivariados de séries temporais é importante para modelar os dados temporais, por isso focamos na previsão de séries temporais multivariadas neste artigo.

Então damos uma definição geral da tarefa de previsão de séries temporais multivariadas [41, 42], que tem objetivo é antecipar o valor futuro tem-poral $ft+1$ no momento $t+1$ ou $ft+p$ no tempo $t+p$ com base no conjunto de dados histórico da série temporal. A entrada do modelo inclui a própria Ft e as outras variáveis de séries temporais multivariadas. Tomando como exemplo a previsão de fluxo de tráfego, ele não só inclui a variável de fluxo, mas também inclui outras variáveis, por exemplo, velocidade de tráfego, densidade de tráfego e comprimento da estrada.

$$L \left(\left[\begin{array}{c} MD = m d_{1, \dots, t-l} \dots m d_{1, t} \end{array} \right]_{\text{modelo}} \rightarrow \left[\begin{array}{c} ft \\ ft+p \end{array} \right] \right) \quad (1)$$

Como mostrado na fórmula acima, a tarefa de previsão multivariada da série tempo multi-etapa neste artigo é representada como prever os próximos valores p de dados de séries temporais $[ft, ft+1, \dots, ft+p]$ dado um conjunto de dados de séries temporais multivariadas de seu tory $MD = \{ m d_{ij} \mid i = 1, 2, \dots, n; j = t-l, \dots, t-1, t \text{ no passado} \}$, onde l denota o tamanho do pesquisado dos dados históricos, n representa um número variável de dados de entrada, e p denota o tamanho de previsão de vários passos para a frente.

3.2. Visão geral do quadro proposto

Uma estrutura de previsão multivariada de séries temporais de várias etapas através da estrutura codificadora-decodificadora baseada em atenção é proposta neste pa-per (como mostrado na Fig. 1), que tem três componentes: Bi-LSTM como componente codificador, um LSTM como componente decodificador e uma camada de contexto de atenção temporal como componente de atenção. O Bi-LSTM é usado para aprender a representação oculta de dados de entrada com comprimentos arbitrários, que extrai os recursos de depen-dency temporal profundo do tempo

multivariado série e, em seguida, usa a camada de atenção temporal para construir variáveis de espaço latentes (vetores de contexto de atenção poral). O decodificador LSTM é responsável por prever o valor da série temporal de futura, que se baseia nas variáveis de espaço latentes geradas. Fig. 1 mostra a ilustração gráfica do quadro proposto, que pode modelar os dados da série temporal multivariada em um processo de ponta a ponta.

3.3. A estrutura de aprendizagem codificador-decodificador

LSTM é um esquema popular para aprender recursos de de-pendency temporal de longo prazo de dados de séries temporais brutas [43]. A unidade típica LSTM contém cinco componentes: i_t, f_t, s_t, o_t, h_t como o portão de entrada, para- obter portão, célula de memória, portão de saída e estado oculto da unidade LSTM, respectivamente. Como o canto inferior direito da Fig. 1 mostra, um bloco de células LSTM típical inclui três portões e uma unidade de célula de memória, que tem a capacidade de esquecer ou informações de memória (de-termine como muitas informações devem ser transferidas para a próxima célula) com base nesses componentes.

$$i_t = \sigma(U^{(i)}x_t + W^{(i)}h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(U^{(f)}x_t + W^{(f)}h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(U^{(o)}x_t + W^{(o)}h_{t-1} + b_o) \quad (4) \quad s_t$$

$$= \tanh(U^{(c)}x_t + W^{(c)}h_{t-1} + b_c) \quad (5)$$

$$st = ft \cdot st-1 + it \cdot s_t \quad (6)$$

$$ht = \tanh ot(st) \quad (7)$$

Como mostrado nas fórmulas acima, para um valor de série de tempo de entrada em t time-step, a unidade LSTM calcula um estado oculto h_t e um estado de memória s_t que é um codificação da série de tempo de entrada val-ues até t time-step, σ é a função de ativação, e é a multiplicação em termos de elementos.

No entanto, o LSTM tradicional tem um ponto fraco. Ele só pode aprender o contexto anterior de dados de séries temporais, e não pode aprender o contexto de ward de dados de mesma sequência. Então usamos o Bi-LSTM como o codificador em, que pode simultaneamente processar dados sequenciais em direções diferentes- ent através de duas camadas ocultas interconectadas: uma recção di-rection processo usa uma camada oculta para a frente de $t = 1$ a T ; o processo de outra direção usa uma camada oculta para trás de $t = T$ a 1 . O processo clássico de computação Bi-LSTM é o seguinte:

$$i_t = \sigma(\bar{U}^{(i)}\bar{x}_t + \bar{W}^{(i)}\bar{h}_{t-1} + \bar{b}_i) \quad (8) \quad ft = \sigma(\bar{U}^{(f)}\bar{x}_t + \bar{W}^{(f)}\bar{h}_{t-1} + \bar{b}_f) \quad (9)$$

$$\bar{o}_t = \sigma(\bar{U}^{(o)}\bar{x}_t + \bar{W}^{(o)}\bar{h}_{t-1} + \bar{b}_o) \quad (10)$$

$$\bar{s}_t = \tanh(\bar{U}^{(c)}\bar{x}_t + \bar{W}^{(c)}\bar{h}_{t-1} + \bar{b}_c) \quad (11)$$

$$st = ft \cdot it \quad \bar{s}_{t-1} + \bar{s}_t \rightarrow \bar{s}_t \quad (12)$$

$$h_t = \bar{o}_t \odot \tanh(\bar{s}_t) \quad (13)$$

$$i_t = \sigma(U^{(i)}x_t + W^{(i)}h_{t-1} + b_i) \quad (14)$$

$$f_t = \sigma(U^{(f)}x_t + W^{(f)}h_{t-1} + b_f) \quad (15) \quad ft$$

$$o_t = \sigma \left(U \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} + W \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} + b_o \right) \quad (16)$$

$$\tilde{s}_t = \tanh \left(\tilde{U} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} + \tilde{W} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} + \tilde{b}_c \right) \quad (17)$$

$$s_t = f_t \odot s_{t-1} + i_t \odot \tilde{s}_t \quad (18)$$

$$h_{-t} = o_t \odot \tanh(s_t) \quad (19)$$

$$h_t = h_{-t} \parallel h_t \quad (20)$$

Como mostrado nas fórmulas acima, a seta representa a direção pró-processing e h_t denota a saída oculta final de Bi-LSTM, que é concatenada pela saída para a frente h_t e para trás – saída da ala h_t . Usamos o Bi-LSTM como componente codificador e um LSTM como componente decodificador, que prevê o valor da série temporal de várias etapas, gerando valor-alvo condicionado ao contexto temporal anterior e o valor gerado anteriormente. Em outras palavras, o codificador Bi-LSTM é responsável por codificar os dados de entrada para uma geração de contexto C_T . É um vetor fixo como a representação temporal dos dados de entrada. O decodificador LSTM, então

previsão. decodifica C_T e gera uma sequência de alvos (f_1, f_2, \dots, f_P) como o valor de previsão de vários passos para a frente. A probabilidade da sequência de destino pode ser calculada da seguinte forma:

$$p(f_1, f_2, \dots, f_P | X_1, X_2, \dots, X_T) = \prod_{t=1}^P p(f_t | C_T, f_1, f_2, \dots, f_{t-1}) \quad (21)$$

onde (X_1, X_2, \dots, X_T) é a entrada modelo da série temporal multivariada da história, X_i denota as variáveis de observação de passo-tempo i de dados temporais, e (f_1, f_2, \dots, f_P) é o p -tar de previsão de várias etapas cujo comprimento P pode diferir do comprimento de entrada T (chamado tamanho de pesquisa ou tamanho da janela). Como mencionado acima, apresentamos o processo de cálculo de um codificador-decodificador clássico [20], que pode ser usado para modelar dados de séries temporais multivariadas e completar um tarefa de previsão de várias etapas.

3.4. Mecanismo de atenção temporal

No entanto, a estrutura clássica de aprendizagem profunda codificador-decodificador tem defeitos que o codificador deve comprimir toda a representação oculta das informações temporais passadas em um vetor de contexto de comprimento fixo C_T . Portanto, a capacidade de previsão se

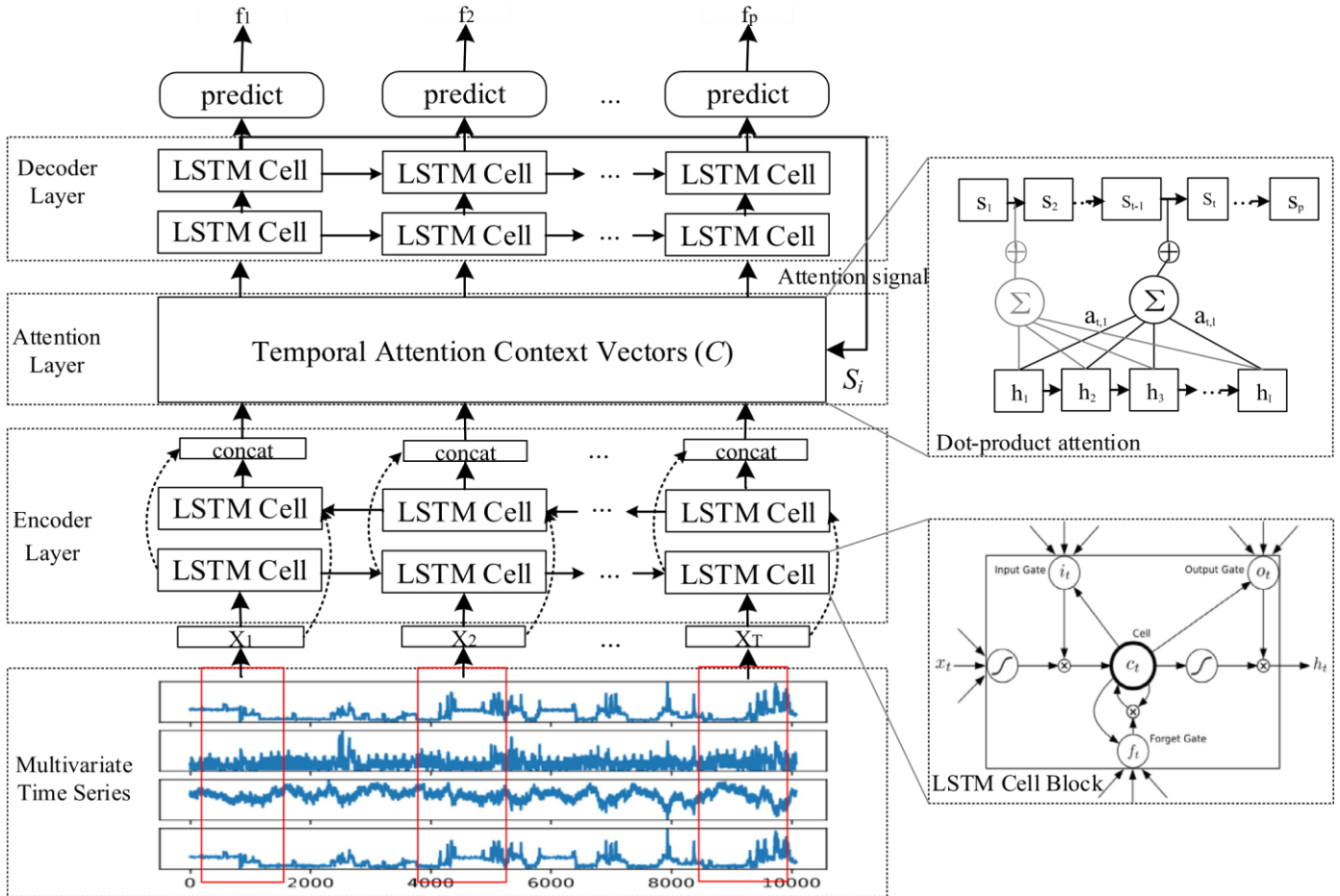


Fig. 1. Ilustração gráfica da estrutura de previsão de séries temporais multivariadas através de um modelo decodificador baseado em atenção temporal (MTSMFF para abreviar), onde a camada de contexto de atenção temporal é projetada para selecionar os melhores quadros da representação oculta do codificador e fazer com que o decodificador atenda-lhe para obter mais precisão degradará gradualmente à medida que o tempo de entrada aumenta [44]. Diante dessa dificuldade, este artigo propõe primeiro um mecanismo de atenção

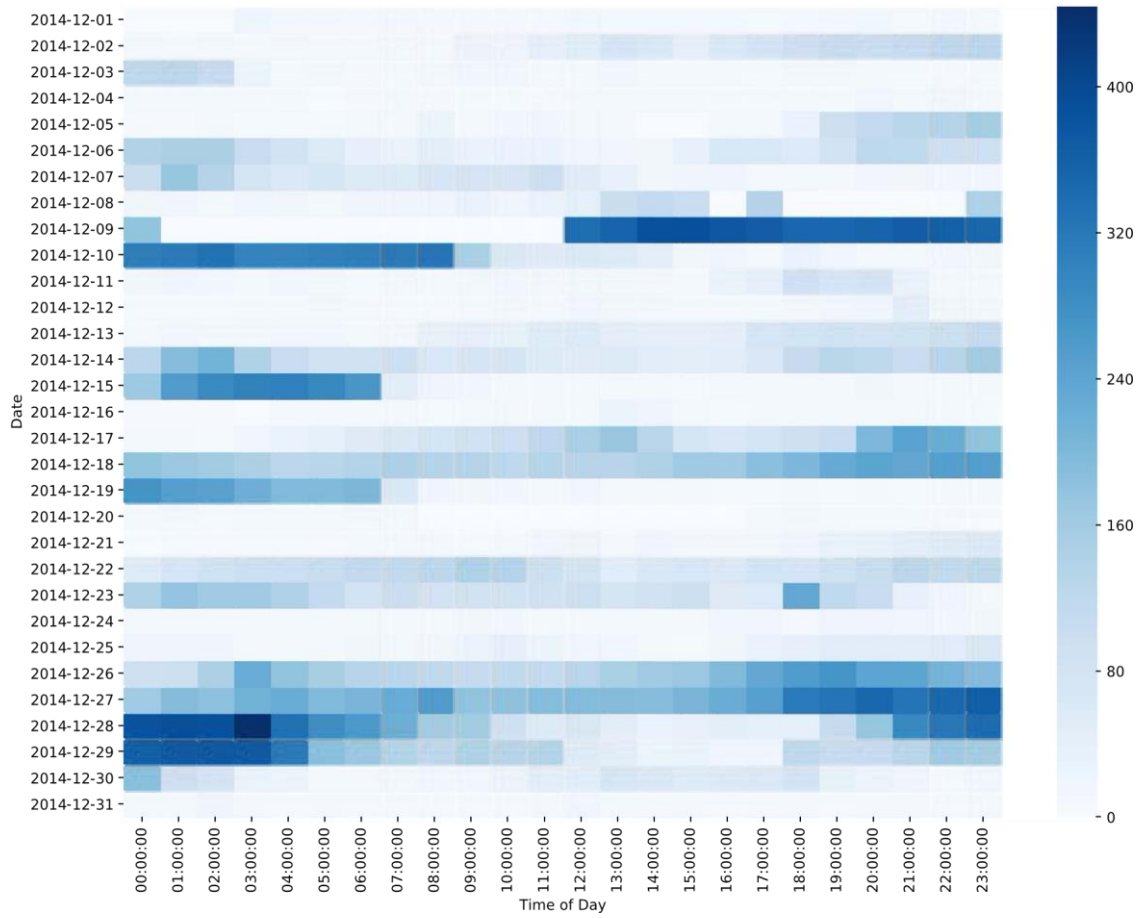


Fig. 2. Visualização do mapa de calor do valor pm2.5 durante um mês (01/12/2014-31/12/2014), do conjunto de dados de qualidade do ar de Pequim [40].

temporal para o processamento de séries temporais multivariadas que prevêem tarefas. Uma camada de atenção temporal é projetada e marcada como a interface entre o codificador e o decodificador. Usamos Bi-LSTM como o codificador em, que pode ter uma sequência de séries de tempo de comprimento variável $x =$

(x_1, x_2, \dots, x_T) como entrada, processando e mantendo seu estado oculto interno h . A cada passo t , o LSTM lê x_t e atualiza seu estado oculto h_t da seguinte forma:

$$h_t = LSTM \left(\mathbf{x}_t, h_{t-1} \right); h_t = LSTM \left(\mathbf{x}_t, h_{t-1} \right); h_t = h_t \quad (22)$$

onde a seta representa a direção de processamento e h_t observa a saída de Bi-LSTM, que é concatenada pela saída dianteira h_t e para trás $-h_t$. Então atenção temporal con-

saída vetores de texto são gerados como uma soma ponderada dos estados ocultos da rede codificador com base no Bi-LSTM, que é usado para selecionar os melhores quadros do codificador escondido representação e fazer o decodificador atender a esses quadros. O processo de computação da camada de atenção temporal pode ser descrito da seguinte forma:

$$e_{i,t} = \text{Alinhar} (s_{i-1}, h_t) = s_{i-1}^{T_i-1} h_t \quad (23)$$

$$= \sum_{k=1}^T \exp(e_{i,t}) \alpha_{i,t} \quad (24)$$

i,k

$$= \sum_{t=1}^T \alpha$$

$$h_{ai,t} \quad (25)$$

Como mostrado nas fórmulas acima, a fórmula (23) representa a computação de alinhamento suave entre o estado oculto s_{i-1} da camada de de-coder e o estado oculto h_t de a camada do codificador. Fórmula

(24) indica os pesos de atenção que correspondem à portance im-portance do período de tempo de entrada no passo de tempo t para pré-dict o valor de saída no passo-hora i , que use o function softmax para normalizar o vetor e_i de comprimento T como a máscara de atenção durante a sequência da série de tempo de entrada. A fórmula $h_{a(25)}$ é o estado final da camada de atenção. O problema do treinamento modal é minimizar a probabilidade negativa de log de um conjunto de treinamento histórico $MD = \{X_i, f_i\}^{mi}$ $=_1$, que pode ser descrito da seguinte forma:

$$= - \sum \log p(f_i | X_i; \theta) + \lambda \|\theta\|^2 \quad (26)$$

φ_i

$=1$

onde n representa o número de amostras de treinamento, φ é o espaço parameter do modelo, que inclui W e b de cada camada, e λ controla a importância do penalidade ou termo de regularização da função de perda.

4. Experimentos

Nesta seção, avaliamos a capacidade de previsão do modelo pro-posed em cinco conjuntos de dados públicos multivariados da série temporal. Por meio da comparação dos modelos de aprendizagem rasa e de aprendizagem

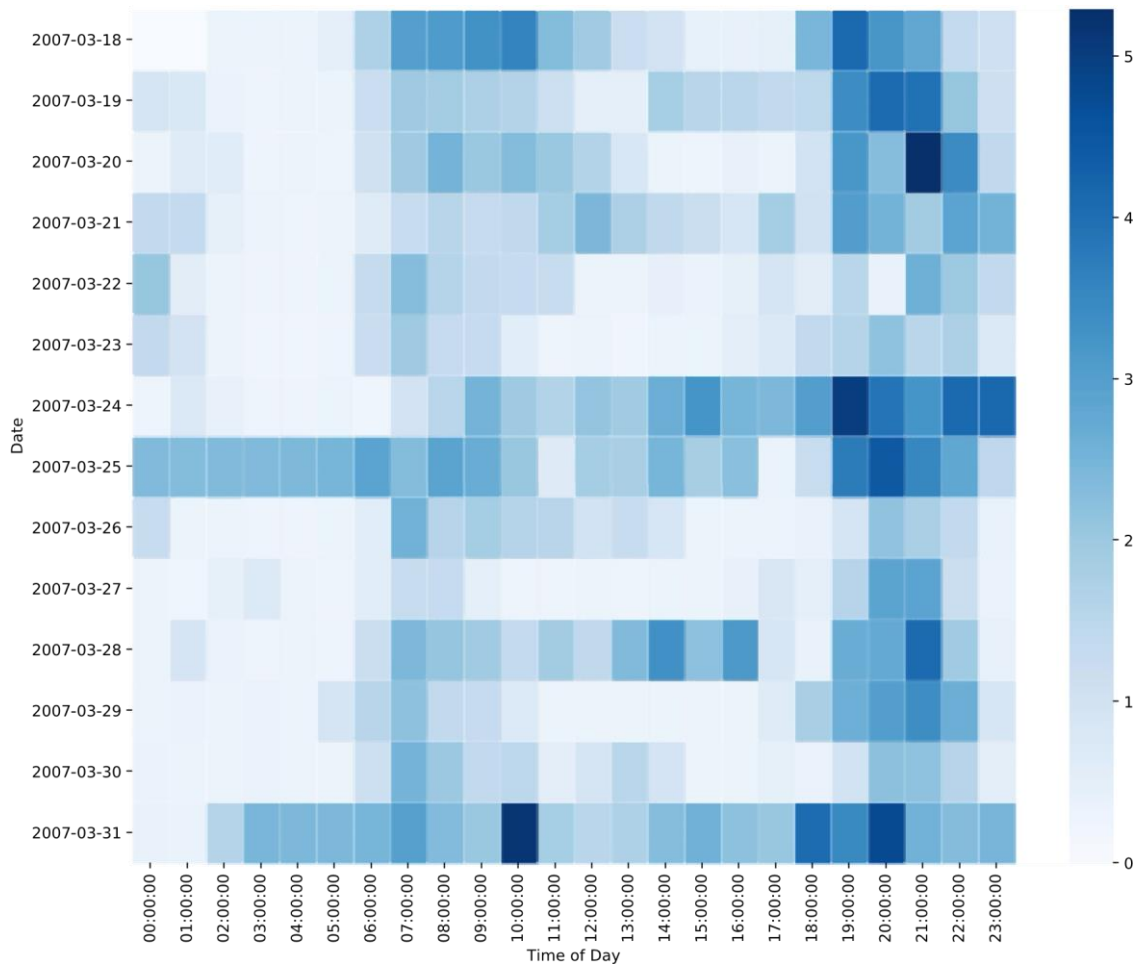


Fig. 3. Visualização do mapa de calor do global_active_power valor de 14 dias (18/03/20 07-31/03/20 07), do conjunto de dados de consumo de energia elétrica individual [40]. profunda, validam-se o desempenho de previsão e eficácia do modelo proposto.

4.1. Conjuntos de dados

Os detalhes dos cinco conjuntos de dados reais da série em tempo multivariada (como mostrado na Tabela 1) são descritos da seguinte forma.

Conjunto de dados de Pequim PM2.5 (Beijing PM25): É um típico conjunto de dados da série tempo mul-tivariate que inclui múltiplas variáveis, por exemplo.

Tabela 1
Detalhes de todos os conjuntos de dados do experimento.

Dataset	Número de instâncias	Número de variáveis	Intervalo de tempo
Pequim PM25	43,824	9	1 hora
Consumo de energia	2,075,259	7	1 min.
Qualidade do ar italiano	9358	15	1 hora
Tráfego de rodovias	37,564	10	15 min.
Baía pems	52,117	14	15 min.

temperatura, pressão, direção do vento, velocidade do vento, condições de neve ou chuva e valor pm2,5 em si, que é publicado no site do repositório de aprendizado de máquina da UCI [40].

Conjunto de dados de consumo de energia elétrica individual (Consumo de Energia Elétrica): O conjunto de dados contém 2.075.259 medidas de consumo de energia elétrica e é coletado com uma taxa amostral de um minuto durante um período de quase 4 anos em uma casa de Sceaux [40].

Conjunto de dados da qualidade do ar (qualidade do ar italiano): O conjunto de dados contém 9358 ocorrências de respostas médias horárias de

uma matriz de 5 sensores químicos de óxido de metal, que estava localizado em um área significativamente poluída da cidade italiana [40].

Highways England Traffic Dataset (Highways Traffic): Este conjunto de dados fornece variáveis médias de tempo de viagem, velocidade de tráfego e fluxo de tráfego para períodos de 15 minutos, que é derivado de altas vias Dados de tráfego da Inglaterra a partir da abertura do governo do Reino Unido [45].

PeMS-Bay: Este conjunto de dados de tráfego é coletado pelo Sistema de Medição de Desempenho das Agências trans-portion da Califórnia (PeMS). Nós

selecione 14 sensores na Área da Baía e colete 6 meses de dados que variam de 01/01/2017 a 31/05/2017 para o experimento [46].

Selecionamos dois dos conjuntos de dados para exploração visual para dissuadir que as características de distribuição dos conjuntos de dados experiments selecionados não são as mesmas. Exploramos visualmente os dois conjuntos de dados através de um mapa de calor. Um mapa de calor pode exibir visualmente as relações de densidade através de duas dimensões, ou seja, data e ob-served time-step. Ele nos ajuda a obter uma visão instantânea sobre a des-tribuição de dados de séries temporais para cada passo de tempo observado de cada dia. Como mostrado em Figs. 2 e 3, quanto mais escura a cor, maior a densidade de valor da série temporal durante o período de tempo.

Fig. 2 é a visualização do mapa de calor do valor pm2.5 durante um mês (01/12/2014-31/12/2014) do conjunto de dados de qualidade do ar de Pequim, que contém leituras pm2.5 feitas a cada hora. O eixo X mostra passos de tempo observados de cada dia (incluindo 24 pontos) e o eixo Y representa um período de um mês. Como Fig. 2 mostra, a cor da densidade diária do valor PM2.5 é a mais profunda durante a manhã (0:00- 4:00) e à noite (19:00-24:00) período de tempo, que também está em linha com o situação real de poluição do ar. No entanto, durante todo o mês de dezembro, não havia regras óbvias sobre quais dias foram seriamente poluídos e quais dias foram levemente poluídos.

Fig. 3 é a visualização do mapa de calor do valor global de energia ativa durante o meio mês (18/03/20 07-31/03/20 07), que contém uma leitura de consumo de energia elétrica doméstica tomadas a cada minuto [40]. O eixo X mostra etapas de tempo observadas (amostragem a cada uma hora) de cada dia e o eixo Y representa um período de 14 dias. Como pode ser visto de Fig. 3, a cor de densidade diária do valor de consumo de energia elétrica é a mais profunda durante a manhã (6:00-9:00) e à noite (16:00-22:00) período de tempo, que também é em consonância com a situação real da vida doméstica. Além disso, a distribuição de dados de consumo de energia durante meio mês indica que a série de tempo de consumo de energia é mais regular do que o polu aéreo - séries temporentas. O consumo diário não é muito diferente (ex-cept para o fim de semana, por exemplo, dia 21 e dia 25), que é relativamente médio.

4.2. Configuração experimental

Os ambientes de experimento são os seguintes: usamos a biblioteca de aprendizado de máquina de código aberto Scikit-learn e a estrutura de aprendizagem profunda Keras para implementar os modelos de linha de base e nosso modelo. O ambiente experimental de hardware é configurado com CPU Intel (R) E5-2623 3.00 GHz, cada uma das 4 GPUs é 12G NVIDIA Tesla K80C, e a memória é de 128 GB.

No experimento, vários modelos de benchmark são usados para comparar o desempenho de previsão de várias etapas de cinco conjuntos de dados de séries temporal reais, incluindo o típico clássico de aprendizado raso-Modelos SVR (com diferentes funções de kernel), quatro modelos de deep learning de linha de base (RNN, CNN, LSTM, GRU) e o modelo de sequência-a-sequência de linha de base (seq2seq para abreviar) modelo de aprendizagem profunda. As descrições detalhadas dos modelos acima são as seguintes.

ARIMA e VARMA são os modelos estatísticos clássicos para dados de séries temporentas de alyzing e previsão. SVR é um método clássico de kernel, que inclui três tipos de funções de kernel como RBF, POLY e LINEAR.

Tabela 2

Configurações de hiperparâmetro dos modelos experimentais.

Modelo	Parâmetro	Valor
Modelos de aprendizagem profunda da linha de base	Camadas ocultas padrão	1
	Unidades em camadas ocultas	100
	Épocas de treinamento	100
	Abandono escolar	0.3
	Tamanho do lote	96
	Função de perda	Mse
	Otimizador	adagrad
seq2seq modelo e nosso modelo	Camadas ocultas padrão (codificador)	1
	Camadas ocultas padrão (decodificador)	1
	Modo de fusão bidirecional-lstm	soma
	A função ativa da camada de atenção	softmax
	Unidades em camadas ocultas	100
	Épocas de treinamento	100
	Abandono escolar	0.3
	Tamanho do lote	96
	Função de perda	Mse
	Otimizador	adagrad

A CNN é um tipo útil de redes neurais profundas para processamento de imagens e vídeos, que também podem ser usadas para previsão de séries temporais (como cnn unidimensional, ou converter os dados de sequência em uma imagem 2D para processamento).

RNN é uma rede neural profunda clássica. LSTM e GRU são variáveis do modelo RNN clássico, que muitas vezes são usados para lidar com tarefas de aprendizagem sequenciais.

SEQ2SEQ é um modelo clássico de aprendizado profundo sequencial para sequencial de dados.

SEQ2SEQ-BI é uma variante simples do nosso modelo, usando Bi-LSTM como componente codificador.

SEQ2SEQ-ATT é uma variante simples do nosso modelo, usando LSTM com mecanismo de atenção como componente codificador.

Em termos de configurações de parâmetros de modelo, tentamos manter o benchmark consistente para que os mesmos valores da maioria dos parâmetros sejam usados para a linha de base e nosso modelo.

Para os modelos de aprendizagem profunda da linha de base (por exemplo, RNN e LSTM), usamos os mesmos parâmetros de configuração de arquitetura de rede neural (Tabela 2), que usam um oculto camada recorrente com 100 unidades neu-ral, use erro quadrado médio (MSE) como função de perda e use a função Adagrad como otimizador de modelo [47]. O tamanho do lote é de 96 e a taxa de abandono é de 0,3. O processo de treinamento é feito para 100 épocas. Para o seq2seq e nosso modelo, usamos um Bi-LSTM como uma camada oculta do codificador e usamos um LSTM como a camada oculta do decodificador. Seleccionamos a função de soma como o modo de fusão Bi-LSTM e softmax como a função de ativação do temporal

Tabela 3

camada de contexto de atenção. Treinamos o modelo proposto com os mesmos valores parâmetros dos modelos de deep learning da linha de base.

Além disso, todos os conjuntos de dados do experimento são divididos retendo os primeiros 80% de cada conjunto de dados para validação de treinamento e os últimos 20% para testes. Usamos a função min-max para normalizar os dados de entrada para [0, 1]. Finalmente, usamos rmse (Root Mean Square Error) e MAE (Mean Absolute Error) como as métricas de avaliação de erro do modelo.

4.3. Análise de resultados

Nesta seção, resumimos os resultados de extensos experimentos para analisar o desempenho de previsão dos modelos de linha de base e do nosso modelo. A Tabela 3 mostra os resultados de experimentos de predição de várias etapas para cinco conjuntos de dados reais de séries temporais multivariadas e fornece uma análise comparativa da RMSE média e da RMSE média MAE dos modelos de linha de base e nosso modelo MTSMFF. O erro do modelo é a média dos erros de previsão (RMSE e MAE) para cada passo da 1ª etapa até a 6ª etapa (t-1-t-6). Como mostrado na Tabela 3, os resultados experimentais mostram que nosso modelo tem o menor erro de previsão em todos os conjuntos de dados, e significativamente im - prova o precisão de previsão em comparação com os métodos tradicionais. Análises mais aprofundadas mostram que a RMSE e a MAE dos modelos de rede neural profunda da linha de base são semelhantes e inferiores às dos modelos de aprendizagem rasos (por exemplo, SVR e ARIMA) na maioria dos casos. Além disso, em alguns casos, descobrimos que o modelo estatístico tradicional e o modelo de aprendizagem profunda da linha de base têm previsão diferente perform-mance em diferentes conjuntos de dados. Por exemplo, o VARMA tem um erro de teste maior do que o LSTM no experimento no conjunto de dados PM25 de Pequim. No entanto, o resultado é o oposto no experimento no conjunto de dados de tráfego de rodovias.

Em seguida, analisamos os resultados de ex-periments de previsão de várias etapas com base em dois conjuntos de dados (Pequim PM2.5 e Consumo de Energia). Como mostrado na Tabela 4, em comparação com os modelos de linha de base, o modelo proposto neste artigo tem melhor desempenho de predição, não importa sob a condição de curto prazo previsão de passo de tempo

ou condição de previsão de passo de tempo de longo prazo. Verifica-se que quando o tamanho da previsão para a frente é 1, o erro do nosso modelo no conjunto de dados de Pequim PM2.5 é reduzido para 23,52. Quando o tamanho da previsão para a frente é 6, o erro do modelo é de 46,15 e ainda é mantido ao mínimo. Além disso, nosso modelo tem o menor erro de teste em outro conjunto de dados de consumo de energia em comparação com outros métodos

de linha de base. Resultados experimentais mostram que nosso modelo im-prova a precisão de previsão de várias etapas em comparação com modelos de linha de base, que desempenho também é melhor do que o clássico modelo de codificador: decodificador de aprendizagem profunda (SEQ2SEQ) e seus modelos variantes

Uma comparação do erro de previsão média (RMSE e MAE) do nosso modelo MTSMFF com outros modelos de linha de base em cinco conjuntos de dados

Modelos	Média modelpredição erro em diferente Datasets									
	Pequim PM25		Poder consumo		Italianaer qualidad		Estradas tráfego		Baía pems	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ARIMA	53.41	41.06	0.87	0.69	59.07	50.66	19.92	19.04	5.44	4.37
CERTO	51.78	38.95	0.79	0.66	58.44	47.32	15.10	10.79	5.42	4.21
SVR-LINEAR	88.35	71.48	1.45	1.23	72.39	65.48	36.01	28.85	6.51	5.53
SVR-RBF	57.64	43.47	0.76	0.71	59.02	48.49	25.20	19.52	5.59	4.43
RNN	44.42	27.99	0.46	0.23	53.35	40.16	17.27	12.29	3.95	1.98
CNN	42.98	26.89	0.46	0.24	54.13	41.07	17.05	12.11	3.86	1.94
LSTM	43.12	27.04	0.44	0.23	55.19	41.60	15.82	10.85	3.87	1.96
GROU	43.04	26.95	0.47	0.25	55.02	41.36	16.64	11.67	3.84	1.92
SEQ2SEQ	42.53	26.42	0.46	0.24	54.89	41.25	17.80	12.45	3.85	1.95
SEQ2SEQ-BI	42.29	26.31	0.44	0.23	54.45	41.21	16.33	12.01	3.77	1.86
SEQ2SEQ-ATT	42.15	26.29	0.44	0.23	54.22	41.18	16.27	11.95	3.76	1.84
MTSMFF(Nossa)	39.83	25.45	0.42	0.21	53.28	40.25	15.07	10.38	3.09	1.72

Nota: O erro do modelo é a média dos erros de previsão (RMSE e MAE) para cada passo do tempo da 1ª etapa até a 6ª etapa (t1-t6). E os principais parâmetros dos modelos de aprendizagem profunda são definidos como épocas-100, tamanho do lote-96, abandono-0,3, otimizador-Adagrad, função de perda-MSE.

Tabela 4

Uma comparação da RMSE do nosso modelo MTSMFF com outros modelos de linha de base para a tarefa de previsão multivariada de séries temporais de várias etapas.

	Modelos RMSE											
	Pequim PM25 dadopôr						Poder Consumo dadopôr					
	t + 1	t + 2	t + 3	t + 4	t + 5	t + 6	t + 1	t + 2	t + 3	t + 4	t + 5	t + 6
ARIMA	24.21	41.17	56.43	62.96	65.28	70.38	0.75	0.82	0.89	0.90	0.94	0.96
CERTO	24.10	40.05	55.27	60.36	62.46	68.47	0.71	0.73	0.79	0.82	0.84	0.85
SVR-LINEAR	38.73	47.39	66.3	163.5	120.34	93.89	0.97	1.44	2.52	0.88	1.49	1.40
SVR-RBF	45.02	49.97	55.29	61.66	64.52	69.41	0.67	0.74	0.78	0.77	0.78	0.83
RNN	24.93	35.23	43.27	50.11	54.49	58.48	0.29	0.41	0.48	0.50	0.54	0.57
CNN	24.57	34.26	41.98	48.02	52.68	56.37	0.28	0.40	0.46	0.51	0.53	0.56
LSTM	24.16	34.35	42.06	48.14	53.04	57.01	0.27	0.38	0.44	0.48	0.52	0.54
GROU	23.99	34.34	41.95	48.08	52.94	56.97	0.29	0.40	0.48	0.52	0.54	0.57
SEQ2SEQ	24.42	33.36	41.83	48.07	52.17	55.34	0.29	0.40	0.46	0.50	0.53	0.56
SEQ2SEQ-BI	23.86	33.42	42.05	48.13	52.11	54.20	0.28	0.39	0.46	0.48	0.51	0.55
SEQ2SEQ-ATT	23.94	33.87	42.33	48.09	51.25	53.41	0.28	0.40	0.45	0.47	0.50	0.54
MTSMFF(Nossa)	23.52	33.06	41.09	47.12	48.05	46.15	0.26	0.37	0.42	0.45	0.50	0.52

Nota: O tamanho da previsão de várias etapas é de 6, e os principais parâmetros dos modelos de aprendizagem profunda são definidos como épocas-100, tamanho do lote-96, abandono-0,3, otimizador-Adagrad, função de perda-MSE. Os parâmetros SVR são definidos como: C-2.0, epsilon-0.1, gama-0.5.

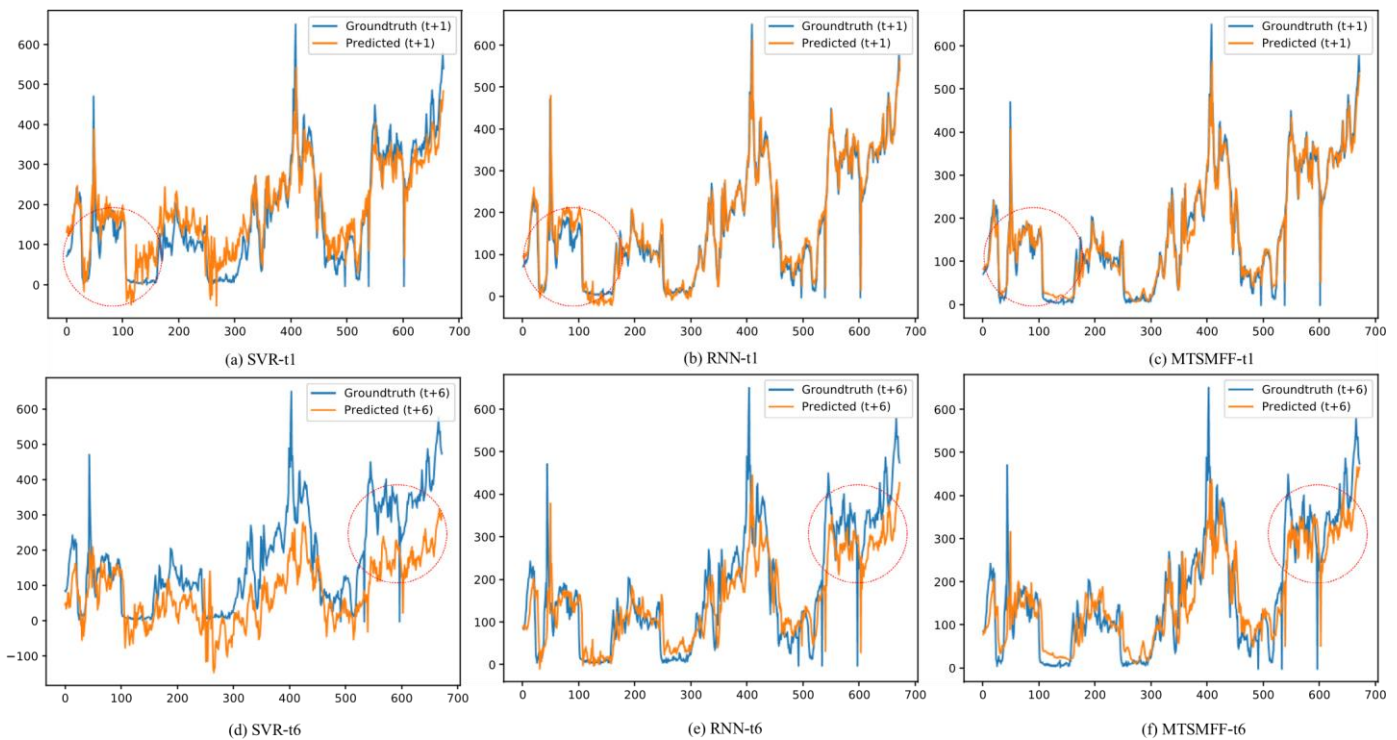


Fig. 4. No experimento sobre o conjunto de dados pm2.5 de Pequim, uma comparação do valor da verdade terrestre e do valor previsto de vários passos para a frente de diferentes métodos sob diferentes tamanhos de previsão (um passo de tempo vs. 6 passos do tempo) condições. a Modelo SVR para previsão de um passo para frente; b Modelo RNN para previsão de um passo para frente; c Nosso modelo MTSMFF para uma previsão de avanço pontual; d Modelo SVR para previsão de 6 passos para frente; e Modelo RNN para previsão de 6 passos para a frente; f Nosso modelo MTSMFF para previsão de 6 passos para a frente.

(SEQ2SEQ-BI e SEQ2SEQ-ATT). Além disso, a capacidade de previsão do nosso modelo pode manter-se estável em diferentes conjuntos de dados.

Também descobrimos um fenômeno comum através de experimentos: Quando o passo do tempo de previsão aumenta, a capacidade de previsão de todos os modelos diminui gradualmente, o que mostra que o erro cumulativo do modelo de previsão de várias etapas está aumentando (com exceção do SVR-LINEAR, que o desempenho da previsão é instável). Isso também reflete os problemas reais enfrentados pelos modelos de previsão de séries temporais em várias etapas. Quanto mais tempo o tempo pisar no futuro, mais difícil é prever. Além disso, com o crescimento do passo de tempo de previsão, nosso modelo tem uma vantagem de desempenho maior do que os modelos de linha de base em experimentos de conjunto de dados PM2.5 de Pequim. Os resultados experimentais mostram que nosso modelo pode manter o menor erro de previsão em comparação com os métodos básicos sob a condição de previsão de passo de tempo de curto e longo prazo.

Para a comparação do desempenho de previsão dos métodos de aprendizagem profunda e dos métodos de aprendizagem rasos, encontramos um fenômeno interessante, ou seja, para o fundição de passos de curto prazo de séries temporais multivariadas, o método de aprendizagem profunda da linha de base não tem nenhuma vantagem óbvia (mesmo obter menor desempenho) em comparação com o desempenho do aprendizado raso método. Por exemplo, como mostrado na Tabela 4, no caso de pré-candidatura de passo único ($t-1$), o erro de previsão do VARMA é menor que o do LSTM e Modelos SEQ2SEQ. No entanto, no caso da previsão de passo de tempo de longo prazo (por exemplo, de $t-4$ a $t-6$), descobrimos que o desempenho de previsão do aprendizado profundo da linha de base método excede o método de aprendizagem rasa. Mas o método proposto

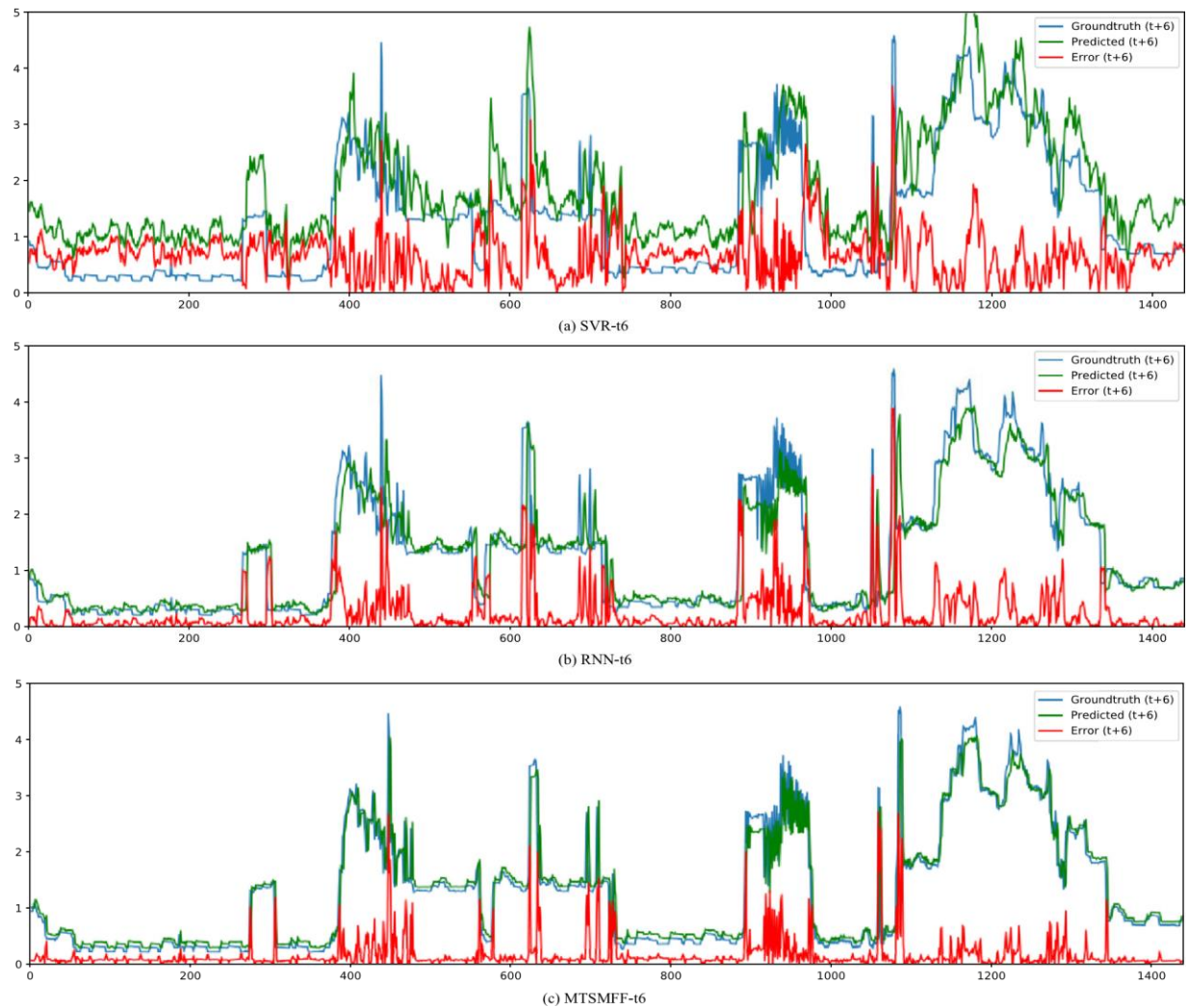


Fig. 5. Uma comparação do valor da verdade do solo e seis etapas ($t+6$) previram o valor de consumo de energia de diferentes métodos durante 24 horas no consumo individual de energia elétrica doméstica conjunto de dados, e o erro é o valor absoluto do valor previsto menos o valor da verdade do solo. a Modelo SVR-RBF para previsão de seis etapas para a frente; b Modelo RNN para previsão de seis etapas; c Nosso modelo MTSMFF para previsão de seis etapas para a frente.

deve evitar esse problema, e os dados experimentais mostram que a capacidade de previsão do MTSMFF é melhor do que os métodos de linha de base em diferentes condições de previsão de passos de tempo.

Além disso, os resultados experimentais mostram que o desempenho de previsão em várias etapas de séries temporais multivariáveis baseadas em modelos SVR-LINEAR é instável. Como mostra a Tabela 4, no experimento do conjunto de dados Power Consumption, o desempenho de previsão do SVR-LINEAR tem grandes flutuações, por exemplo, o desempenho de previsão do primeiro o passo do tempo é muito bom, enquanto o desempenho de previsão do segundo passo de tempo é muito ruim. Em comparação com os modelos de aprendizagem rasos, os modelos de aprendizagem profunda são mais estáveis e não há flutuação violenta, embora o desempenho de previsão seja gradualmente reduzido à medida que o previsão para a frente o passo do tempo aumenta.

Analisamos ainda o desempenho de previsão em várias etapas do modelo proposto através de um display gráfico visual no conjunto de dados pm2.5 de Pequim. Fig. 4 (a)–(f) mostra o valor da verdade do solo e as curvas de valor previstas de diferentes modelos (SVR, RNN e nosso modelo MTSMFF) em diferentes etapas de tempo de previsão (um passo de tempo vs. seis vezes passo). Como mostrado no desvio entre o valor previsto e o valor da verdade terrestre nesses números, no caso da previsão de etapa única, as diferenças entre três modelos não são muito grandes. Podemos ver desvios mais óbvios sob a condição de previsão de várias etapas. Nosso modelo tem a melhor capacidade de previsão em condições de previsão de etapa única e várias etapas.

Além disso, também analisamos a previsão de várias etapas do modelo proposto através de um display gráfico visual no experimento de conjunto de dados de consumo de energia ao longo do curso de 24 h (1440 pontos de passo de tempo). Fig. 5 (a)–(c) mostra o valor da verdade do solo e as curvas de valor de consumo de energia previstas de diferentes modelos (SVR, RNN e nosso modelo MTSMFF) em seis etapas de previsão condições ($t+6$), onde x -coordenada indica o tempo de observação passo, e o erro é o valor absoluto do valor previsto menos o valor da verdade do solo. Como mostrado no desvio (linha de erro) entre o valor previsto e o valor da verdade terrestre nessas figuras, podemos ver desvios mais óbvios de modelos de linha de base sob a condição de previsão de várias etapas, e a capacidade de previsão do nosso modelo é a melhor. Em resumo, os resultados experimentais comprovam que, em diferentes condições, o desempenho de previsão em várias etapas do modelo proposto é mais estável e eficaz do que os outros modelos de linha de base.

Finalmente, comparamos o desempenho de previsão de diferentes modelos em condições de entrada univariadas e condições de entrada multivariadas da série tempo. Tomando o conjunto de dados pm25 de Pequim, por exemplo, em Tabela 5

Tabela 5
Comparação de erro de teste (RMSE) entre nosso modelo MTSMFF e outros modelos de linha de base sobre condição de entrada de séries temporais univariadas e condição de entrada de séries temporais multivariadas.

Modelos	Pequim PM25 Dataset		PeMS BaíaDataset	
	Multivariado		Multivariado	
Univariate				Univariate
ARIMA	53.41	–	5.44	–

CERTO	–	51.78	–	5.42
SVR-LIN	85.42	88.35	5.65	6.51
SVR-RBF	53.80	57.64	5.11	5.59
RNN	45.09	44.42	4.12	3.95
CNN	43.15	42.98	4.05	3.86
LSTM	44.63	43.12	3.94	3.87
GROU	44.62	43.04	3.92	3.84
SEQ2SEQ	44.56	42.53	3.97	3.85
SEQ2SEQ-BI	43.27	42.29	3.86	3.77
SEQ2SEQ-ATT	43.14	42.15	3.82	3.76
MTSMFF	43.09	39.83	3.28	3.09

Declaração de contribuição de autoria da CRediT

Shengdong Du: Metodologia, Investigação, Software, Redação - rascunho original. **Tianrui Li:** Supervisão, Conceituação, Re-fontes, Redação - revisão & edição, Administração de Projetos. **Yan Yang:** Escrevendo - revisão e edição. **Shi-Jinn Horng:** Supervisão, Escrita - revisão e edição.

Confirmações

Esta pesquisa foi parcialmente apoiada pela [National Natural Science Foundation of China](#) (Nos. 61773324 e 61976247), o "Centro de Inovação do Sistema Cibernético físico" da Featured Areas Research Center Program no âmbito do Projeto Sprout de Ensino Superior pelo Ministério da Educação (MOE) em Taiwan e MOST em 106-2221-E-011-149-MY2 e 108-2218-E-011-006.

Referências

- [1] J.G. De Gooijer, R.J. Hyndman, 25 anos de previsão da série temporal, *Int. J. Forecast.* 22 (3) (2006) 443–473.
- [2] D. Xia, B. Wang, H. Li, et al., Um modelo distribuído de peso espacial-temporal no MapReduce para previsão de fluxo de tráfego de curto prazo, *Neurocomputing* 179 (2016) 246–263.
- [3] Qi Z., Wang T., Song G., et al. Aprendizado profundo do ar: Interpolação, previsão e análise de recursos da qualidade do ar de grãos finos. *IEEE Trans. O Knowl. Data Eng.*, doi: 10. 1109/TKDE.2018.2823740, 2018
- [4] S. Ahmad, A. Lavin, S. Purdy, et al., Detecção de anomalias não supervisionadas em tempo real para dados de streaming, *Neurocomputing* 262 (2017) 134–147.
- [5] L. Guo, N. Li, F. Jia, Y. Lei, J. Lin, et al., Um indicador de saúde baseado em rede neural recorrente para a previsão de vida útil remanescente dos rolamentos, *Neurocomputing* 240 (2017) 98–109.
- [6] G.E.P. Box, D.A. Pierce, Distribuição de autocorrelações residuais em modelos de séries de tempo móveis integradas, *J. Stat. Assoc.* 65 (332) (1970) 1509–1526.
- [7] P.F. Pai, K.P. Lin, C.S. Lin, et al., Previsão da série time por um modelo de regressão vetorial de suporte sazonal, *Exp. Syst. Appl.* 37 (6) (2010) 4261–4265.
- [8] S.H. Park, J.H. Lee, J.W. Song, et al., Previsão de mudança de direção para séries temporais financeiras usando o modelo oculto Markov, em: *Proceedings of the International Conference on Rough Sets and Knowledge Technology*, Berlin, Heidelberg, Springer, 2009, pp. 184–191.
- [9] G.P. Zhang, Série temporal prevendo usando um modelo híbrido de ARIMA e rede neural, *Neurocomputing* 50 (2003) 159–175.
- [10] T. Fu, Uma revisão sobre a mineração de dados de séries tempo, *Eng. Appl. O Artif. Intell.* 24 (1) (2011) 164–181.
- [11] G.P. Zhang, M. Qi, Previsão de rede neural para tempo sazonal e de tendência series, *Eur. J. Oper. Res.* 160 (2) (2005) 501–514.
- [12] N.I. Sapankevych, R. Sankar, Previsão da série time usando suporte vetor ma-chines: uma pesquisa, *IEEE Comput. Intell. Mag.* 4 (2) (2009) 24–38.
- [13] B.S. Freeman, G. Taylor, B. Gharabaghi, et al., Previsão de séries de tempo de qualidade do ar usando deep learning, *J. Manag. de Resíduos de Ar. Assoc.* 68 (8) (2018) 866–886.
- [14] Y. Lv, Y. Duan, W. Kang, et al., Previsão de fluxo de tráfego com big data: uma abordagem de aprendizagem profunda, *IEEE Trans. Intell. Transp. Syst.* 16 (2) (2015) 865–873.
- [15] J. Zhang, Y. Zheng, D. Qi, Redes residuais profundas e temporais para a previsão de fluxos de multidões em toda a cidade, em: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 1655–1661.
- [16] A. Krizhevsky, eu. Sutskever, E. Hinton G, Classificação Imagenet com redes neurais contra volucionais profundas, em: *Proceedings of the International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [17] J. Schmidhuber, Aprendizado profundo em redes neurais: uma visão geral, *Neural Netw.* 61 (2015) 85–117.
- [18] A. Karpathy, F.F. Li, Alinhamentos visuais-semânticos profundos para geração de descriptions de imagem, em: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [19] S. Venugopalan, M. Rohrbach, J. Donahue, et al., Sequência para sequência-vídeo para texto, em: *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4534–4542.
- [20] K. Cho, B. Van Merriënboer, C. Gulcehre, et al., Aprendendo representações de frases usando rnn codificador-decodificador para tradução estatística de máquina, em: *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, 2014, Pp. 1724–1734.
- [21] Gamboa J. C. B. Aprendizado profundo para análise de séries temporentas. *arXiv pré-impressão arXiv: 1701.01887*, 2017.

no caso da condição de entrada univariada, apenas a sequência variável única PM2.5 é a entrada no modelo; no caso da condição de entrada de séries temporais multivariadas, a variável PM2.5 em si e as demais variáveis (por exemplo, umidade e velocidade do vento) estão inserindo o modelo em conjunto. Como mostrado na [Tabela 5](#), para o modelo de aprendizagem rasa, a capacidade de previsão em condições univariadas é melhor do que em condições multivariadas, o que indica que é difícil para o modelo de aprendizagem rasa aprender efetivamente as características ocultas de correlação não linear de dados de séries temporais multivariadas. No entanto, para os modelos de deep learning da linha de base e nosso modelo, o desempenho de previsão sob a condição de entrada de séries temporais multivariadas é melhorado em comparação com o previsão de desempenho sob condição de entrada univariada, o que indica que o modelo de aprendizado profundo pode aprender características ocultas de correlação não linear de multivariadas dados da série temporal até certo ponto.

Em suma, através da análise comparativa dos verdadeiros experimentos de conjunto de dados multivariar séries temporais, que valida que nosso modelo tem o melhor desempenho de previsão em comparação com outros métodos de linha de base. A razão pela qual nosso modelo é mais eficaz é porque ele pode aprender representações temporais profundas e características de dependência temporal de longo prazo de dados multivariados de séries temporais, que também pode extrair as características de correlação interdependentes de múltiplas variáveis de dados temporais usando o mecanismo de atenção temporal. O modelo deste artigo pode fornecer uma referência valiosa para a aplicação de sistema de predição inteligente com base em dados de séries temporais multivariadas.

5. Conclusão e trabalho futuro

Neste artigo, propusemos um quadro de aprendizagem profunda de ponta a ponta para a previsão de séries temporais multivariadas, o que aproveita a ideia da estrutura de aprendizagem codificadora-decodificadora com Bi-LSTM e é aumentado com um mecanismo de atenção temporal. É um método de representação em primeiro lugar proposto de dados dinâmicos multivariados de tempo se-ries, que pode aprender conjuntamente o padrão temporal de dependências de longo prazo e não-linear características de correlação de dados temporais multivari-comes. Experimentos em cinco conjuntos de dados de séries temporais multivariadas mostraram que o modelo proposto pode efetivamente prever o valor da série tempo pluceicial em várias etapas, não importa sob condição de passo de tempo de curto prazo ou condições de longo prazo. Além disso, devido ao desempenho de modelos de aprendizagem profunda que geralmente estão relacionados a diferentes parâmetros e condições de conjunto de dados em muitos casos, como melhorar e estabilizar a capacidade de previsão do nosso modelo para melhor suporte à previsão de séries temporais on-line e aplicações de aviso antecipado é o foco de pesquisas futuras.

Declaração de Interesse Concorrente

Os autores declaram que não têm interesses concorrentes ou relações pessoais que possam ter aparecido para influenciar o trabalho relatado neste artigo

- [22] T. Kuremoto, S. Kimura, K. Kobayashi, et al., Série time prevendo usando uma rede de crenças profundas com máquinas Boltzmann restritas, *Neurocomputando* 137 (2014) 47-56.
- [23] Y. Tian, K. Zhang, J. Li, et al., Previsão de fluxo de tráfego baseada em LSTM com dados faltantes, *Neurocomputing* 318 (2018) 297-305.
- [24] Y. Bao, T. Xiong, Z. Hu, Previsão da série de tempo de vários passos à frente usando regressão vetorial de suporte multi-ple-output, *Neurocomputing* 129 (2014) 4 82-4 93.
- [25] N.K. Ahmed, A.F. Atiya, N.E. Gayar, et al., Uma comparação empírica dos modelos de aprendizagem ma-chine para previsão de séries temporais, *Econom. Rev.* 29 (5-6) (2010) 594-621.
- [26] M.R. Hassan, B. Nath, M. Kirley, Um modelo de fusão de HMM, ANN e GA para previsão do mercado de ações, *Exp. Syst. Appl.* 33 (1) (2007) 171-180.
- [27] S.B. Taieb, G. Bontempi, A.F. Atiya, et al., Uma revisão e comparação de estratos para a previsão de séries de tempo de vários passos com base na competição de previsão NN5, *Exp. Syst. Appl.* 39 (8) (2012) 7067-7083.
- [28] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, et al., *Análise série time : Previsão e Controle*, John Wiley & Sons, 2015.
- [29] Liu, Z. Wang, X. Liu, et al., Um levantamento de arquiteturas de rede neural profunda e suas aplicações, *Neurocomputing* 234 (2017) 11-26.
- [30] Chambon S., Galtier M. N., Arnal P. J., et al. Uma arquitetura de aprendizagem profunda para classificação do estágio de sono temporal usando séries temporais multivariadas e multimodais. *arXiv: 1707.03321* (2017).
- [31] N. Laptev, J. Yosinski, L.E. Li, et al., Previsão de eventos extremos em séries de tempo com redes neurais na uber, em: *Proceedings of the International Conference on Machine Learning*, 34, 2017, pp. 1-5.
- [32] J. Yang, M.N. Nguyen San, et al., Redes neurais convolucionais profundas em séries temporais multicanais para reconhecimento da atividade humana, em: *Prosseguir as Conferências Conjuntas Internacionais sobre Inteligência Artificial*, 15, 2015, Pp. 3995-4001.
- [33] X. Ding, Y. Zhang, T. Liu, et al., Aprendizado profundo para previsão de ações orientadas a eventos, em: *Proceedings of the International Joint Conferences on Artificial Intelligence*, 2015, pp. 2327-2333.
- [34] Y. Zheng, Q. Liu, E. Chen, et al., Classificação da série time usando redes neurais multi-canais profundas e convolucionais, em: *Proceedings of the International Conference on Web-Age Information Management*, Cham, Springer, 2014, Pp. 298-310.
- [35] S. Yao, S. Hu, Y. Zhao, et al., DeepSense: uma estrutura unificada de deep learning para o processamento de dados de sensoriamento móvel de séries temporais, em: *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 351-360.
- [36] N. Jaitly, Q.V. Le, O. Vinyals, et al., Um modelo de sequência on-line, condicionando parcialmente, em: *Proceedings of the International Conference on Neural Information Processing Systems*, 2016, pp. 5067-5075.
- [37] J.K. Chorowski, D. Bahdanau, D. Serdyuk, et al., Modelos baseados em atenção para reconhecimento de fala, em: *Proceedings of the International Conference on Neural Information Processing Systems*, 2015, pp. 577-585.
- [38] I. Sutskever, O. Vinyals, Q.V. Le, Sequência para sequenciar aprendizado com redes neurais, em: *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 3104-3112.
- [39] Kuznetsov V., Mariet Z. Fundamentos da Modelagem Sequência-para-Sequência para Séries Tempores. *arXiv pré-impressão arXiv:1805.03714*, 2018.
- [40] UCI Machine Learning Repository, 2017 [Online]Disponível: <http://archive.ics.uci.edu/ml/index.php>.
- [41] J. Du Preez, S.F. Witt, Univariate versus multivariada séries temporais : uma aplicação à demanda turística internacional, *Int. J. Previsão*. 19 (3) (2003) 435-451.
- [42] D.E.H. Zhuang, G.C.L. Li, A.K.C. Wong, Descoberta de associações temporais em séries temporais multivariadas, *IEEE Trans. O Knowl. Data Eng.* 26 (12) (2014) 2969-2982.
- [43] S. Hochreiter, J. Schmidhuber, Memorando de longo prazo, *Neural Comput.* 9 (8) (1997) 1735-1780.
- [44] Bahdanau D., Cho K., Bengio Y. Tradução de máquina neural aprendendo conjuntamente a alinhar e traduzir. *arXiv pré-impressão arXiv:1409.0473*, 2014.
- [45] Highway England Traffic Data Set, 2013 [Online] Disponível: <http://data.gov.uk/dataset/highways-england-network-journey-time-and-traffic-flow-data>
- [46] Conjunto de dados de fluxo de tráfego pems, 2017 [online] Disponível: <http://pems.dot.ca.gov/>.
- [47] J. Duchi, E. Hazan, Y. Cantor, Métodos subgradiantes adaptativos para aprendizado online e otimização estocástica, *J. Mach. Aprender. Res.* 12 (2011) 2121-2159.
- [48] Malhotra P., Ramakrishnan A., Anand G., et al. Encoder-decodificador baseado em LSTM para detecção de anomalias multisensor. *arXiv pré-impressão arXiv:1607.00148*, 2016.
- [49] S.H Park, B.D. Kim, C.M. Kang, et al., Previsão sequencial da trajetória ve-ícle via arquitetura de codificador-decodificador LSTM, em: *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2018, pp. 1672-1678.



Shengdong Du recebeu o B.S. e ESM graduado em Com-puter Science pela Universidade de Chongqing em 2004 e 2007, respectivamente. Ele é atualmente ph.D. candidato na Escola de Ciência da Informação e Tecnologia da Universidade de Jiaotong Do Sudoeste. Seus interesses de pesquisa incluem mineração de dados e aprendizado de máquina.



Tianrui Li recebeu b.S., M.S., e Ph.D. Graduados pela Universidade de Jiaotong do Sudoeste, Chengdu, China em 1992, 1995 e 2002, respectivamente. Foi pesquisador de pós-doutorado na SCK •CEN, Bélgica, de 2005 a 2006, e professor visitante na Universidade Hasselt, Bélgica, em 2008, na Universidade de Tecnologia, Sydney, Austrália, em 2009, e a Universidade de Regina, Canadá, 2014. Atualmente é Professor e Diretor do Key Laboratory of Cloud Computing and Intelligent Techniques, Southwest Jiaotong University. Ele é autor ou coautor de mais de 300 artigos de pesquisa em jour-nals e conferências referenciadas. Seus interesses de pesquisa incluem big data, computação em nuvem, mineração de dados, computação granular e conjuntos ásperos. É membro da Receita Federal e membro sênior da ACM e do IEEE.



Yan Yang recebeu o B.S. e ESM Graduado pela Universidade de Ciência e Tecnologia de Huazhong, Wuhan, China, em 1984 e 1987, respectivamente. Ela recebeu Ph.D. graduação pela Southwest Jiaotong University, Chengdu, China em 2007. De 2002 a 2003 e 2004 a 2005, foi bolsista visitante na Universidade de Água-loo, Canadá. Atualmente é Professora e Vice-Reitora da Escola de Ciência da Informação e Tecnologia da Universidade de Jiaotong do Sudoeste, Chengdu, China. Ela é uma líder acadêmica e técnica da província de Sichuan. Seus interesses de pesquisa incluem inteligência artificial, análise e mineração de big data, aprendizado de conjunto, computação em nuvem e serviço. Ela participou de mais de 10 altas...

projetos de nível recentemente, autor e coautoria de mais de 150 artigos em revistas e processos de conferência internacional. Ela também atua como vice-presidente da ACM Chengdu Chapter, um membro distinto da CCF, um membro sênior da CAAI, e um membro do IEEE e ACM.



Shi-Jinn Horng recebeu o B.S. graduação em eletrônica engineering do National Taiwan Institute of Technology, o M.S. graduado em engenharia da informação pela Universidade Central Nacional e doutorado. graduação pela Universidade Nacional Ts-ing Hua, em 1980, 1984 e 1989, respectivamente. Atualmente é presidente do Departamento de Ciência da Computação e Engenharia da Informação da Universidade Nacional de Ciência e Tecnologia de Taiwan. Ele fez mais de 200 artigos de pesquisa e recebeu muitos prêmios. Especialmente, o Distinguished Research Award recebeu do Conselho Nacional de Ciência em Taiwan em 2004. Seus interesses de pesquisa incluem deep learning, recognition biométrico, processamento de imagens e segurança da informação.