Research on feature engineering for time series data mining

Lei Li[1], Yihang Ou[1], Yabin Wu[1], Qi Li[1], Daoxin Chen[2]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]CapInfo Company Limited Beijing, China
{leili,ouyihang,wyb,awpboxer}@bupt.edu.cn, chendaoxin@capinfo.com.cn

Abstract: Today Internet and information technology are flourishing rapidly. Data collected from network and mobile devices can bring us huge opportunities to understand some significant characteristics of users and merchants. Time series analyzing is an extremely important topic in data mining that help users and merchants use data to do forecasting. Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) both are excellent algorithms in time series forecasting. Nevertheless, there are few studies focused on feature engineering of LSTM and SVM in time series forecasting. In this paper, we focus on the influence of multi-feature fusion and feature ordering. Our experimental results show that LSTM is more sensitive about feature fusion than SVM, and the position of feature in the feature sequence could affect forecasting results obviously.

Keywords: Feature engineering, Time series, Data mining, Deep learning

1 Introduction

Nowadays, with the rapid development of Internet and information technology, massive data was generated. Thus, data mining become one of the most heated research fields, and it is the computing process of discovering patterns in large data sets involving methods at intersection of machine learning, statistics, and database systems. The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further use [1]. A time series is a series of data points indexed in time order, time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to forecast future values based on previously observed values [2]. Therefore, the combination of time series and data mining means using data mining methods to analyze and forecast time series.

Traditional machine learning model has been applied to time series analyzing for a long time. Support Vector Machine (SVM) which based on statistical learning raised by Vapnik and Cortes in 1990s is a typical example [3]. Jaramillo J et al. [4] use SVM to do forecasting of financial time series. Niu D et al. [5] use SVM in forecasting of short-term power load. Peixian et al. [6] do forecasting of mining subsidence based on SVM. Additionally, with the development of deep learning technology, some deep learning models are applied into the research of time series analyzing gradually. Among deep learning models, recurrent neural network (RNN) import time series into network architecture, and show more powerful adaptation in time series analyzing. RNN comes in many variants, long short-term memory (LSTM) compensates problems in RNN such as vanishing gradient, exploding gradient, and shortage of long time memory ability [7]. Tian et al. [8] and Fu et al. [9] use LSTM neutral network to achieve traffic flow forecasting. Wang et al. [10] use LSTM for earthquake forecasting. Lei Li et al. [11] use LSTM and SVM for consumption forecasting.

Feature engineering is a significant concept in machine learning fields, which could be considered as a job to design feature set for machine learning application [12]. Sandya H B et al. [13] focus on feature extraction and feature engineering of time series forecasting. Overbey L A. [14] uses specific feature extraction technique for structural health monitoring application of time series. However, there are rarely studies focused on feature engineering in time series analyzing neither with traditional machine learning model nor with deep learning model.

In this paper, we do following contributions to these issues.

(1) Considering the influence of feature fusion on two different models. For the SVM algorithm, feature fusion cannot bring about the improvement of forecasting results when tackling the multi-classification problem. For LSTM, feature fusion can improve the forecasting results.

(2) Considering the effects of feature ordering on the two models, and the influence of the position of different length sparse sequences in the feature sequence on the forecasting results, the SVM model is not sensitive about feature ordering, but for the LSTM model, long sparse feature sequence can improve the forecasting results if it was placed in the front of the feature sequence, and the position of short sparse feature sequence in the feature sequence does not affect the forecasting result obviously.

2 System Design

2.1 System Architecture

The system architecture is shown as Figure 1.

978-1-5386-6067-6/18/$31.00 ©2018 IEEE

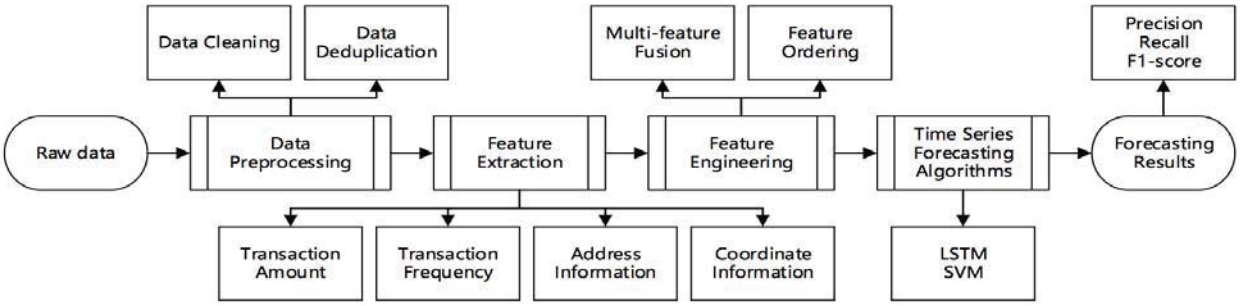

Figure 1 System architecture

2.2 Raw Data

Our raw data is collected from the real users' consumption records, containing 21340966 records of 8992 retail merchants. The consumption time span of the data is from June 26, 2014 to November 10, 2016, a total of 869 days. We select all 8992 merchants and extract their transaction records and geographic information as dataset for our experiment.

2.3 Feature Engineering

In general, through feature exaction, we will get many non-processed features. Besides, there are many missing values in non-processed features needed to be compensated. Thus, we use One-Hot encoding [15] to represent the features. And One-Hot encoding mainly is often used for indicating the state of a state machine. In this task, a one-hot vector is a 1 × N matrix (vector) used to distinguish each feature in a feature set from other features. Then we focus following feature engineering.

2.3.1 Multi-feature Fusion.

This project aims to time series analyzing, multi-feature fusion is combining different features into one sequence.

2.3.2 Feature Ordering.

In general, it is not necessary to think over feature ordering in machine learning tasks, but through a little test, we discovered that different feature ordering in deep learning algorithm taken different effects. These different effects are always caused by following reasons:

(1) Numerical problems: as for input features sequence, if the range of the sequence value fluctuates rapidly, especially when sequence is very sparse, the model will not be learned valid features, and leading to model generalization ability decrease.

(2) Curse of Dimensionality: when dimensions increase, the input spaces size increase exponentially, in higher space, the data become sparse definitely, this also leads to model generalization ability decrease.

2.4 Model Algorithm

As mentioned earlier, this paper focuses on machine learning and deep learning algorithms, hence the following two algorithms are chosen for research.

(1) SVM: SVM is a supervised learning model whose essence is a binary-classification model. Its basic model is in the feature space on the largest linear classifier.

(2) LSTM: LSTM is a variant of RNN. LSTM achieves the purpose of maintaining the persistence of information through the control of the gate inside the neuron, and avoids the problem of long-term dependence in RNN. In our experiments, for all network structures, the number of default hidden unit is 128.

2.5 Evaluation Method

We choose appropriate evaluation index to evaluate experiment results, they are Precision, Recall, F1-score. [16]

3 EXPERIMENTS AND RESULTS

3.1 Pre-Processing

(1) Raw Data processing

Our raw data is collected from the real-life users' consumption records, containing 21340966 records of 8992 retail merchants. The consumption time span of the data is from June 26, 2014 to November 10, 2016, a total of 869 days. We select all 8992 merchants and extract their transaction records and geographic information as the data set for our experimentation. Besides, all these merchants are living in Beijing, so the geographic information includes administrative zones and longitude and latitude coordinate.

(2) Feature extraction and feature engineering

We use original transaction records from 8992 merchants during 869 days and geographic information which includes address information, coordinate information. Then we obtain the following features and in this time series.

(a) Transaction amount (Amount): it refers to the total number of one merchant's sales volume one day, which is an 868 dimensions' vector.

(b) Transaction frequency (Fre): it is the total number of one merchant sales frequency. As same as transaction amount.

(c) Address information (Add): because all of merchants from Beijing, their original geographic information mark

included 2 dimensions represents administrative zoom, and we use One-Hot encoding to convert 2-dimensions feature vector into 345-dimensions feature vector, as long sparse sequence.

(d) Coordinate information (Crd): as the same reason as above, we convert coordinate to a mark which presents the location inside certain loop line of Beijing. This mark includes 5 different values, we use One-Hot encoding to convert it to a 5 dimensions' feature vector, as short sparse sequence.

3.2 Experimental Design

Through data pre-processing and feature engineering, we should set experimental target. we classify these 8992 merchants to 10 categories artificially. We divided all data set which remove 0 into ten parts evenly. Then dataset was classified into 10 classes. We implement experiment for each merchant, using former 868 days' data to predict transaction amount in 869th day, according to above classification to mark label from 1 to 10. We divide the data into training set and testing set as 3:1.

We mainly analyze the following 2 problems through our experiments: 1) What is the influence of adding the different features (long sparse feature vector and short sparse feature vector) on the forecasting results of each algorithm? 2)What is the influence of different feature ordering on each algorithm?

3.3 Experimental result and analysis

3.3.1 Experiment 1. All of four features will be considered in this experiment, we choose two features fused. Meanwhile, we also consider the influence of feature ordering with different algorithms. The forecasting results of each feature group are as showed in Table I:

Table I Experiment 1 Results

	Alogorithm	LSTM			SVM		
Exp	Feature	P	R	F	P	R	F
1.1	Amount-Fre	0.61	0.72	0.66	0.51	0.71	0.59
1.2	Fre-Amount	0.64	0.73	0.67	0.51	0.71	0.59
1.3	Add-Amount	**0.72**	**0.77**	**0.73**	**0.51**	**0.71**	**0.59**
1.4	Amount-Add	0.51	0.71	0.59	0.51	0.71	0.59
1.5	Add-Fre	**0.68**	**0.76**	**0.72**	**0.51**	**0.71**	**0.59**
1.6	Fre-Add	0.51	0.71	0.59	0.51	0.71	0.59
1.7	Crd-Amount	0.64	0.73	0.68	0.51	0.71	0.59
1.8	Amount-Crd	**0.69**	**0.77**	**0.72**	**0.51**	**0.71**	**0.59**
1.9	Crd-Fre	0.68	0.76	0.71	0.51	0.71	0.59
1.10	Fre-Crd	0.66	0.76	0.70	0.51	0.71	0.59

(1) LSTM is more sensitive about feature ordering than SVM. From the results, no matter where the new feature is in feature sequence, the results of SVM are invariable, but the results of LSTM is not identical. So, the SVM is not sensitive to feature ordering.

(2) When long sparse feature placed in the front of the feature sequence, then the forecasting results of LSTM are worse. From experiment 1.3, 1.4, 1.5, 1.6, when adding address information feature into feature sequence and putting it in the front of the feature sequence, these forecasting results are better than single feature with transaction amount or transaction frequency, so the address information indeed contain valid information for forecasting. Through one-hot encoding, address information feature became very sparse, and then putting it at back of feature sequence, these sparse features make model forgotten information which LSTM model had learned from the front of feature sequence. This is why we put the long sparse feature sequence in the front of the feature sequence. But there is no different of results when changing the location of long sparse in feature sequence in SVM. It was known that the forecasting result of address information feature independently is very poor (Precision=0.51, Recall=0.71, F1-score=0.59) and as same as experiment 1.4, 1.6 (address information feature at the back of the feature sequence). Compared with experiment 1.3,1.5, and the forecasting results is better if we put the long sparse feature sequence in the front of feature sequence. Thus, reasonable feature ordering is beneficial for forecasting in LSTM.

(3) The position of short sparse feature sequence in feature sequence will not affect the forecasting results of LSTM seriously. From experiment 1.7, 1.8, 1.9, 1.10, no matter we place coordinate information feature in the front of the feature sequence or place it at the back of feature sequence, it affects the forecasting slightly. Because the length of coordinate information feature is too short (only 5 dimensions), it is hard to affect the forecasting results.

In sum up, forecasting results will be affected by new address information feature and coordinate information feature, then we continue to increase the number of features.

3.3.2 Experiment 2. Based on experiment 1, we choose transaction frequency feature and transaction amount feature fusion as the fundamental feature. Then we consider the influence of the long sparse address information feature sequence and short sparse coordinate information feature sequence and consider feature ordering, and consider the influence of feature ordering among different algorithms. The forecasting results of each feature group are showed in table II:

Table II Experiment 2 Results

	Algorithm	LSTM			SVM		
Exp	Feature	P	R	F	P	R	F
2.1	Fre-Amount-Add	0.51	0.71	0.59	0.51	0.71	0.59
2.2	Fre-add-Amount	0.69	0.75	0.71	0.51	0.71	0.59

2.3	Add-Fre-Amount	**0.70**	**0.76**	**0.71**	**0.51**	**0.71**	**0.59**
2.4	Fre-Amount-Crd	0.69	0.76	0.72	0.51	0.71	0.59
2.5	Fre-Crd-Amount	0.66	0.76	0.70	0.51	0.71	0.59
2.6	Crd-Fre-Amount	**0.70**	**0.75**	**0.71**	**0.51**	**0.71**	**0.59**

(1) SVM is not sensitive to new added feature and feature ordering when the length of sequence over certain range. From experiment 1, 2, all results of SVM experiments are identical. It was argued that the results of SVM are not affected by feature fusion and feature ordering when addressing multi-classification time series forecasting problem.

(2) Address information feature and coordinate information feature contribute to forecasting for LSTM. From experiment 1.1,1.2, 2, no matter to add address information feature or coordinate information feature, the results of which are better than the result of transaction amount feature fused with transaction frequency feature. By this way, the address information and coordinate information are valid for time series forecasting in this data set. However, the results of SVM experiment is identical in experiment 2, so the address information feature and coordinate information feature are not useful for time series forecasting in SVM.

(3) The position of long sparse sequence in feature sequence affect forecasting results of LSTM obviously. From experiment 2.1, 2.2, 2.3, there is different among forecasting results of LSTM by changing address information features' position in feature sequence. Specifically, if address information feature was placed at the back of the feature sequence, the result will be weakened, and we put forward a point that long sparse feature sequence leads to LSTM model forgetting some information that had learned from the front of feature sequence. On the contrary, if address information feature was placed in the front of the feature sequence, then the information from this feature will be learned by the model and do not affecting feature at the back of sequence learned by model. Additionally, if address information feature was placed between transaction frequency feature and transaction amount feature, the results of which is worse than the results that feature placed in front of the feature sequence. Thus, new added long sparse feature sequence will weaken the forecasting ability of LSTM. However, the SVM is not sensitive about this.

(4) The position of short sparse sequence in feature sequence affects forecasting results slightly for LSTM. From experiment 2.4,2.5,2.6, no matter we place coordinate information feature sequence in front of the feature sequence or place it at back of the feature sequence, the difference of results is not obvious. However, when we place it between transaction frequency feature and transaction amount feature, the result is the worst, and the numerical problem obviously. Because transaction frequency feature and transaction amount feature come from the same resource, model will be affected deeply when adding newly short sparse feature between them. Besides, the SVM also not sensitive about newly added short sparse feature.

In sum up, SVM is not sensitive to new added feature and feature ordering, but new added address information is beneficial for forecasting in LSTM. Then we continue to add the number of features in following experiment.

3.3.3 Experiment 3. In this experiment, we consider all four features of time series, and focus on the influence of feature ordering on different algorithms, and conclude the best feature ordering. The forecasting results of each feature group are showed in Table III:

Table III Experiment 3 Results

Algo		LSTM			SVM		
Exp	Feature	P	R	F	P	R	F
3.1	Fre-Amount-Add-Crd	0.51	0.71	0.59	0.51	0.71	0.59
3.2	Add-Crd-Fre-Amount	**0.70**	**0.76**	**0.72**	**0.51**	**0.71**	**0.59**
3.3	Add-Fre-Crd-Amount	**0.70**	**0.76**	**0.72**	**0.51**	**0.71**	**0.59**
3.4	Crd-Add-Fre-Amount	0.68	0.75	0.71	0.51	0.71	0.59
3.5	Crd-Fre-Add-Amount	**0.69**	**0.75**	**0.72**	**0.51**	**0.71**	**0.59**
3.6	Crd-Fre-Amount-Add	0.51	0.71	0.59	0.51	0.71	0.59

(1) The results of SVM are identical from experiment 1,2,3, thus SVM is not sensitive about adding new features and feature ordering in time series forecasting, and there is not exist the best feature ordering for time series forecasting in this dataset.

(2) Information contained in long sparse sequences must be placed in front of the feature sequence so that it can be learned by the model and not influences the model's learning of non-sparse features at the back of the feature sequence in LSTM. From experiment 3.2,3.4,3.5,3.6,2.6, when adding new address information feature (sparse feature with 345 dimensions) and changing its position in feature sequence, the results of forecasting are difference. Specifically, when long sparse feature is placed in front of the feature sequence, the information of this feature can be learned. However, when it is placed at the back of the feature sequence, not only the information contained in this feature cannot be learned by the model, but also the model's previous learned information will be forgotten, so the forecasting results is declined. Additionally, from experiment 2.6 and 3.2, adding address information feature promotes the Recall 1% and F1-score 1%, thus, address information feature is useful for forecasting.

(3) The position of short sparse sequences in the feature sequence has little influence on the forecasting results for LSTM. From experiment 3.2,3.3,3.4,2.3, by changing the position of coordinate information feature (sparse feature with 5 dimensions), there is little influence on the forecasting results. Specifically, when the short sparse feature sequence is placed at the first of the feature sequence, its precision drops 2% and recall drops 1%, mainly because when the short sparse

sequence is placed at the first of the feature sequence, the latter long sparse feature sequence makes its information "forgotten". Comparing the results of experiments 3.2, 3.3 and 2.3, coordinate information feature do contribution to the forecasting (1% increase in F1-score).

After these three experiments, we get following results.

(1) For the LSTM algorithm, feature fusion improves the forecasting results of time series. From experiment 1,2,3, it was suggested that all the forecasting results are improved when adding the transaction amount feature, transaction frequency feature, address information feature and coordinate information feature respectively. It shows that feature fusion is an effective way for time series forecasting. However, feature fusion is invalid for SVM in time series forecasting.

(2) For the LSTM algorithm, feature ordering affects the forecasting of time series. From experiment 1, it has suggested that LSTM algorithm is sensitive for feature ordering, even if the two features exchange their locations of the sequence, their results are not same. Additionally, for the new added features, we consider the long sparse feature and the short sparse feature (both of which through one-hot encoding). As for long sparse feature sequence (345 dimensions), we need to place it in front of the sequence so that it could be learned by model, if it is placed at the back of the sequence, not only making the information placed in front of the sequence "forgotten", but also the information contained in this feature could not be learned by the model. Nevertheless, the location of the short sparse features (5 dimensions) does not have a significant effect on the forecasting. However, changing the location of long sparse feature sequence and short sparse feature sequence could not change the results of SVM in time series forecasting.

4 Conclusions

In this paper, we work on the real user consumption records to extract a variety of features to construct time series. We use SVM and LSTM and combine feature engineering to achieve forecasting. We examine the impact of different feature representation methods, multi-features fusion and feature ordering on various algorithms from different angles respectively. We know that the forecasting results based on different feature presentations are different under different algorithms, feature fusion is inconsistent with the performance of different algorithms, the position of long sparse feature sequence and short sparse feature sequences in the feature sequence has a great impact on the forecasting.

Acknowledgements

This work was supported by National Social Science Foundation of China [grant number 16ZDA055]; National Natural Science Foundation of China [grant numbers 91546121, 71231002]; EU FP7 IRSESMobileCloud Project [grant number 612212]; the 111 Project of China [grant number B08004]; Engineering Research Center of Information Networks, Ministry of Education; Beijing BUPT Information Networks Industry Institute Company Limited; the project of Beijing Institute of Science and Technology Information; the project of CapInfo Company Limited

.References

[1] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. Data mining: concepts and techniques. Elsevier.
[2] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD '03). ACM, New York, NY, USA, 2–11. https: //doi.org/10.1145/882082.882086
[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine Learning 20, 3 (1995), 273–297.
[4] Johana Jaramillo, Juan David Velasquez, and Carlos Jaime Franco. 2017. Research in Financial Time Series Forecasting with SVM: Contributions from Literature. IEEE Latin America Transactions 15, 1 (2017), 145–153.
[5] Dongxiao Niu, Yongli Wang, Chunming Duan, and Mian Xing. 2009. A New Short-term Power Load Forecasting Model Based on Chaotic Time Series and SVM. Journal of Universal Computer Science 15, 13 (2009), 2726–2745.
[6] Peixian Li, Zhixiang Tan, Yah Lili, and Kazhong Deng. 2011. Time series prediction of mining subsidence based on a SVM. International Journal of Mining Science and Technology() 21, 4 (2011), 557–562.
[7] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. Nature 521, 7553 (2015), 436.
[8] Yongxue Tian and Li Pan. 2015. Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network. In IEEE International Conference on Smart City/socialcom/sustaincom. 153–158.
[9] Qianlong Wang, Yifan Guo, Lixing Yu, and Pan Li. 2017. Earthquake Prediction based on Spatio-Temporal Data Mining: An LSTM Network Approach. IEEE Transactions on Emerging Topics in Computing PP, 99 (2017), 1–1.
[10] Rui Fu, Zuo Zhang, and Li Li. 2017. Using LSTM and GRU neural network methods for traffic flow prediction. In Chinese Association of Automation. 324–328.
[11] Li L, Wu Y, Ou Y, et al. Research on machine learning algorithms and feature extraction for time series[C]// IEEE, International Symposium on Personal, Indoor, and Mobile Radio Communications. IEEE, 2017:1-5.
[12] Michael Anderson, Dolan Antenucci, Victor Bittorf, Matthew Burgess, Michael Cafarella, Arun Kumar, Feng Niu, Yongjoo Park, Christopher Ré, and Ce Zhang. 2012. Brainwash: A Data System for Feature Engineering. Cidr (2012).
[13] H. B Sandya, Kumar P Hemanth, Himanshi Budhiraja, and Susham K. Rao. 2013.Fuzzy Rule Based Feature Extraction and Classification of Time Series Signal. (2013).
[14] Lucas A Overbey. 2008. Time series analysis and feature extraction techniques for structural health monitoring applications. (2008).
[15] David Harris and Sarah Harris. 2014. Digital Design and Computer Architecture, Second Edition. Chian Machine Press,. 289–361 pages.
[16] Zhihua Zhou. 2015. Machine Learning. Tsinghua University Press