ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

Feature Extraction Methods for Time Series Functions using Machine Learning

Ankit Narendrakumar Soni

Department of Information Technology, Campbellsville University, Kentucky

ABSTRACT : This paper expects to utilize different machine learning calculations and investigate the impact between various calculations and multi-include in the time arrangement. The open utilization records comprise the time arrangement as the exploration object. We separate utilization mark, recurrence also, different highlights. Besides, we use support vector machine (SVM), long short-term memory (LSTM) also, different calculations to foresee the data's utilization conduct. Moreover, we have likewise executed multi-feature combination and multi-calculation combination with LSTM also, SVM. In the long run, the trial results show that LSTM calculations are worthwhile in the forecast at the point when the information is inadequate. In the other hand, the SVM is gainful when the information is more plentiful. What's more, LSTM-SVM combination model has points of interest on the removing highlights of LSTM and the order of SVM. As a rule, LSTM-SVM is generally exceptional in the forecast.

KEYWORDS: SVM, LSTM, RNN

I. INTRODUCTION

With the quick improvement of Internet and data innovation, enormous information is developing. Along these lines, information mining gets one of the most significant research to handle these days. Time arrangement investigation is a technique to investigate all the data contained in the time arrangement, to watch, to gauge and to examine the factual normality during the time spent long-term change in such a lot of genuine information [1]. The blend of time arrangement and information mining can investigate the evolving laws of marvels and take proactive control of the activities that have not occurred. As right on time as 1927, the British analyst Yule put forward the Autoregressive (AR) model [2] used to anticipate the law of market changes. In 1931, Walker built up the Moving Average (MA) model. Afterwards, he consolidated these two models to build up an Autoregressive Moving Average (ARMA) model [3]. Based on past exploration, Box and Jenkins proposed the autoregressive incorporated moving regular (ARIMA) model, which is of incredible centrality for current time arrangement examination and forecast. These models are known as time arrangement prescient investigations of traditional strategies. With the ongoing quick turn of events and general use of machine learning and neural network, their mix with time arrangement information mining has gotten a hot issue. Pothuganti et al. [6] use ARIMA, Support Vector Machine(SVM) and the Recurrent Neural Networks (RNN) model to anticipate on various time arrangement information sets and to think about the impacts of the different models under various assignments. Use Long Short-Term Memory (LSTM) nonpartisan system to accomplish traffic stream gauge. Wang et al. use LSTM for quake forecastuse RNN to build a structure about space-time determining for the time arrangement framed via air toxins. Utilizes SVM to anticipate the money related time arrangement. From the above investigations, there are not many examinations centre on the impact of time arrangement include building on the forecast impact of different calculations. Along these lines, we plan to take a shot at the impact of different highlights in various calculations for the time arrangement, including the delegate SVM and LSTM calculations[4]. Through the multi-include combination, we examine the impact of the highlights. Moreover, we consolidate the two calculations of LSTM and SVM.

II. SYSTEM DESIGN

Raw Data:

Our crude information isgathered from the genuine data's' utilization records, containing 21340966 records of 538052 data. The utilization time length of the information is from June 26, 2014, to November 10, 2016, an aggregate of 869

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

days. We haphazardly select 50000 data's and extricate their utilization records as the informational collection for our experimentation.

Pre-processing:

We separate all records, actualize information cleaning and joining, at that point, sparing the necessary highlights.

Feature Extraction

We isolate 869 days' information into 124 weeks which implies time arrangement incorporates 124 measurements. At that point, we separate the accompanying highlights in this time arrangement[5]. Utilization mark: It alludes to whether the data expends inside seven days, on the off chance that utilization happens, at that point recording as 1, in any case, recording it as 0. Utilization recurrence: It is the tally of every data's utilization times each week. Utilization sum: It is the quantity of one's everything utilization cash inside seven days.

Model Algorithms

As referenced before, this paper centres around machine learning and profound learning calculations; thus, the accompanying three calculations are picked for research.

SVM:

SVM is a directed learning model whose quintessence is a paired characterization model. Its fundamental model is characterized in the component space on the biggest direct classifier [7].

RNN:

Compared with the customary neural network, RNN has included a cyclic structure, which can keep the tirelessmemory of data.

LSTM:

LSTM is a variation of RNN. LSTM accomplishes the reason for keeping up the constancy of data through the control of the "door" inside the neuron and stays away from the issue of long haul reliance in RNN[6].

Integration of SVM and LSTM:
Propose a progression of LSTM-SVM structures for highlight extraction and arrangement, and compelling in the subject arrangement and incongruity identification. In this paper, we attempt to consolidate SVM with LSTM in the arrangement, utilizing the consequences of SVM as the contribution to LSTM and the consequences of LSTM as the yield, which is called SVM-LSTM[8]. Essentially, we trade the situation of SVM also, LSTM, at that point, we structure the LSTM-SVM.

Evaluation method:
The undertaking of our test framework is mostly to foresee whether the data will expend or not in the next time unit; subsequently, we will give more consideration to the accuracy of the estimate. As indicated by the disarray lattice, the accuracy P is characterized as follows:

$$P = \frac{TP}{TP+FP} \qquad (1)$$

What's more, to precisely depict the rich level of the information, we characterize an information plenitude record K. Assume that a databunch contains N datas, each of which has utilization records for Di (i = 1, 2, ..., N) weeks in 124 weeks. This data bunch information has a rich degree as follows[9]:

$$K = \frac{\sum_{i=1}^{N} C_i}{124*N} \qquad (2)$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

III. RESULTS

Experimental design:

Various datasets have distinctive utilization propensities, subsequently, building a bound together model for all data can't perform adequately for all data. Indeed, this has been affirmed in our unique tests. Like this, thinking about the likenesses among some data utilization practices, we use K-intends to group data and separate data gatherings. Afterwards, we manufacture model for each data gathering to anticipate their utilizations. K-implies is a standard group calculation dependent on distance. We use K-intends to group 50,000 dataset into three various groups. The three bunches are as per the following: Group0 has 29228 dataset, Group1 has 17154 dataset, and Group2 has 3618 dataset. Their information-rich degrees are as per the following: Group0:10.95%, Group1:28.49%, Group2:46.59%. We actualize explore for every data gathering, utilizing previous 123 weeks information to anticipate whether data devours in the 124th week, on the off chance that expends, at that point recording as '1', if not, recording as '0'. We separate the information into preparing set and testing set as 9:1. The quantity of dataset in each data gathering's trying set, and the number of dataset in each bunch has appeared in table I.

Table I: The Number Of Users In Each Cluster

User Group	Group0	Group1	Group2
0	1755	943	127
1	472	713	193
Total	2227	1656	320

We dissect the accompanying three issues through our tests: What is the impact of every calculation on the forecast of time arrangement when the information is scanty what's more, has low information extravagance? What is the impact of including the combination highlights on the forecast aftereffects of each calculation? What is the impact of the combination between calculations and how well they foresee?

Experimental Results And Analysis:

Experiment1 Taking the utilization checks as highlights, we use SVM, LSTM and RNN individually to prepare a model and anticipate for every data gathering. The forecast precisions are as per the following: Group0's data information is scanty and imbalanced. We get high accuracy in the 0 classifications with these calculations, yet SVM can't foresee the one classification which is less extent. LSTM can anticipate the one classification with higher exactness, which shows that LSTM has a conspicuous impact on the lopsided time arrangement. RNN additionally can foresee the one classification, yet RNN can't channel memory data. In this way, the forecast capacity of RNN leaves something to be desired as that of LSTM. The information extravagance of Group1 and Group2 are expanded; in this way, these calculations can foresee both of two sorts' dataset. From the outcome, by expanding the extravagance of the information can improve execution, particularly in type which low extent. Also, RNN's invaluable profit by the information's wealth, so it's exactness is somewhat better than LSTM. From another edge, the Group2's information is rich than the Group1, so the expectation of Group2 is improved. Be that as it may, we see that the information extravagance surpass certain degree will result in RNN somewhat sub-par compared to LSTM. The explanation of this is LSTM utilize the 'entryway' to channel information which is helpful for recall the highlights, and take out repetition. Investigation 1 demonstrates the benefit of LSTM in retaining long time arrangement when the information is meagre, lopsided succession.

Furthermore, RNN isn't as great as LSTM because of long haul reliance. Furthermore, at the point when information is bountiful, SVM is more potent than LSTM and RNN. At that point, we have a go at including new highlights, also, analyze the impact of the multi-highlight combination on each algorithm.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

Table II: Experiment 1 Results

Groups	**Group0**			**Group1**			**Group2**		
	SVM	LSTM	RNN	SVM	LSTM	RNN	SVM	LSTM	RNN
0	0.79	0.75	0.82	0.63	0.63	0.65	0.81	0.59	0.49
1	0.02	0.86	0.47	0.57	0.52	0.57	0.62	0.68	0.70
Total	0.81	0.81	0.75	0.61	0.58	0.59	0.72	0.65	0.63

Experiment 2:

We include the recurrence highlights, utilizing both of the utilization recurrence and the utilization mark as the highlights in this test. The forecast aftereffects of every data bunch are as per the following:After including the recurrence includes, the compelling highlights is additionally enhanced. In Group0's outcomes, SVM can foresee class 1 and accomplish 67% exactness. Including highlights can improve SVM's forecast on the information which is scanty and the class with less extent essentially. LSTM's forecast is better than SVM what's more, RNN. For Group1, it can anticipate the two classes, and keep up steady gauge results. Be that as it may, the forecasts of LSTM have declined. The highlights included may affect LSTM's memory capacity, and there is a certain measure of futile clamour. RNN can't channel the data, the clamour doesn't have an extraordinary effect, and the forecast outcomes are better than that of LSTM. SVM keeps up a high expectation exactness while the information is plentiful. SVM in Group2 is same as in Group1; it shows an expectation advantage when the information is plentiful and accomplishes the highest forecast accuracy in the three calculations. LSTM's exactness is somewhat lower than the SVM. As it were, the point at which the information is more plentiful, LSTM's expectation results can be more unfortunate than that of SVM. RNN's expectation result is irregular; the forecast exactness of class 0 will be 0, which may be brought about by overfitting. In our investigation, the neural arrange has 128 shrouded units, which might be an excessive number of for the information. Henceforth, we attempt to decrease the number of shrouded units to 64 and 32. Every exactness of 64-RNN is: Class0:0.55, Class1:0.69 and total:0.64. Furthermore, each accuracy of 32-RNN is Class0:0.55, Class1:0.66 and total:0.62 as should be obvious that RNN with 128 concealed units in the Group2 has the issue of overfitting. After including the recurrence includes the information wealth increments to a certain degree, which negatively affects RNN. It shows up overfitting quicker than LSTM. In the equivalent information size, LSTM can dodge the issue of overfitting very well through its particular memory. Looking at the aftereffects of examination one and analysis 2, we can see that the expectation exactness of LSTM on the three data bunches doesn't change, furthermore, including the recurrence highlights has little impact on LSTM. Including highlights and improving the important information can improve SVM's expectations on inadequate and uneven time arrangement. Yet, when the information is excessively rich, unreasonable clamour can unfavourably influence the expected consequences of SVM. At the point when the information wealth is low, including highlights can likewise improve RNN's outcome, yet when the information is plentiful, it is conceivable to make RNN show up overfitting if we keep on including highlights.

Table III: Experiment 2 Results

Groups	**Group0**			**Group1**			**Group2**		
	SVM	LSTM	RNN	SVM	LSTM	RNN	SVM	LSTM	RNN
0	0.79	0.79	0.80	0.61	0.62	0.57	0.71	0.57	0.00
1	0.67	0.86	0.60	0.59	0.51	0.65	0.63	0.68	0.61
Total	0.80	0.81	0.75	0.61	0.58	0.61	0.68	0.65	0.37

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 8, August 2018

LSTM-SVM has essential accuracy of 2 outcomes; also, SVM has the most critical accuracy of 5 ones, and they have a similar exactness twice. We can see that both LSTM-SVM and SVM calculations have the best gauge results for most cases. SVM has excellent forecasts when the information is plentiful. Yet, it is regularly not unsurprising for inadequate time arrangement, and SVM will show up overfitting issue when the information is bountiful for some situation. By and large, LSTM-SVM has important focal points over different calculations with scanty information and unequal time arrangement. Additionally, after the information is bountiful, it can likewise adequately anticipate the outcomes. There are no undeniable inadequacies in LSTM-SVM; what's more, it can adequately abstain from overfitting issues.

IV. CONCLUSION

In this paper, we utilize open data utilization records to develop time arrangement, and precise assortment highlights to join with machine learning calculations to accomplish the utilization forecast. We think about the effect of various highlights on different calculations and afterwards, accomplish multi-highlight combination and calculation combination to supplement the shortcoming of every single element and calculation. This investigation gives the premise to study further. Simultaneously, it essential to consider the time arrangement different attributes and utilization of prescient model later on.

REFERENCES

[1]. X. Xu, “Research and application on time series prediction based on the data mining method." PhD dissertation, the China University of Geosciences for Master Degree (Beijing), 2011.
[2]. Karunakar Pothuganti, Aredo Haile, Swathi Pothuganti,” A Comparative Study of Real Time Operating Systems for Embedded Systems” International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2016.
[3]. T. Le, Y. Cai, X. Ma, and L. Wang, “Development and application of time series forecasting,” Ordnance Industry Automation, no. 2, pp. 63–68, 2015.
[4]. Vishal Dineshkumar Soni 2018. Artificial Cognition for Human-robot Interaction. International Journal on Integrated Education. 1, 1 (Dec. 2018), 49-53. DOI:https://doi.org/10.31149/ijie.v1i1.482.
[5]. Y. Tian and L. Pan, “Predicting short-term traffic flow by long short-term memory recurrent neural network,” in Smart City/Social Com/Sustain Com (Smart City), 2015 IEEE International Conference on. IEEE, 2015, pp. 153–158.
[6]. Pothuganti, Karunakar, Jariso, Mesfin, Kale, Pradeep. 2017. A Review on Geo Mapping with Unmanned Aerial Vehicles. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 5, Issue 1, January 2017
[7]. R. Fu, Z. Zhang, and L. Li, “Using lstm and GRU neural network methods for traffic flow prediction,” in Chinese Association of Automation (YAC), Youth Academic Annual Conference of. IEEE, 2016, pp. 324–328.
[8]. Vishal Dineshkumar Soni. (2018). An IoT Based Patient Health Monitoring System. International Journal on Integrated Education, 1(1), 43-48. https://doi.org/10.31149/ijie.v1i1.481
[9]. J. Fan, Q. Li, Y. Zhu, j. Hou, and y. Feng, “A spatiotemporal prediction framework for air pollution based on a deep run,”Science of Surveying and Mapping, pp. 1–16, 2017(07).

.