Quality of Service Forecasting with LSTM Neural Network

International Symposium on Integrated Network Management (IM 2019)
May 11, 2019

Tomas Jirsik, Stepan Trcka, Pavel Celeda
Institute of Computer Science, Masaryk University, Brno
Quality of Service Forecasting

what is it good for?

Quality of Service

- Abstract term used for comparing services
- Derived from measurable QoS attributes
- QoS Attributes
  - Application response time
  - Network response time

Applications

- Recommending systems for Web Pages

Forecasting

- Updates from service providers are sparse
Challenges
what do we research

How can be QoS attributes collected?
- Increase the frequency of the QoS attributes updates

How can we use Long Short-Term Memory Neural Network for QoS forecasting?
- How to create LSTM NN model?

What method should we use for QoS attribute forecasting?
- Forecast precision
- Estimation time
Centralized QoS Attribute Collection
how to collect up-to-date data

IP flow network monitoring

- Passive approach to network traffic observation
Centralized QoS Attribute Collection
how to collect up-to-date data

Next-generation IP flow network monitoring

- Bi-flows
- Application layer information

IP flow monitoring for QoS Attributes collection

- Attributes
  - Round trip time
  - Number peers/users
  - Transport size
  - Application response time

- Passive, continuous observation
  - Observation point location makes the difference
Evaluated Forecasting Methods

three approaches to time series forecasting
ARIMA(p,d,q)
autoregression and moving average in one package

Auto-Regression

- evolving variable of interest is regressed on its own lagged (i.e., prior) values

Moving Average

- regression error is a linear combination of error terms whose values occurred at various times in the past

Integrated

- transformation applied to timeseries in order to make it stationary

\[
(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q) \varepsilon_t
\]
Holt-Winters
seasonality included

Model

\[ L_t = \alpha (y_t - I_{t-p}) + (1 - \alpha)(L_{t-1} + T_{t-1}) \]
\[ T_t = \beta (L_t - L_{t-1}) + (1 - \beta)T_{t-1} \]
\[ I_t = \gamma (y_t - L_{t-1} - T_{t-1}) + (1 - \gamma)I_{t-p} \]

Prediction

\[ \hat{y}_t(k) = L_t + kT_t + I_{t-p+1} + (k-1)\text{mod}L \]

Parameters

- Speed of learning/forgetting
Long Short Term Memory Neural Network

recurrent neural network

Recurrent Neural Networks

- Text processing - understanding of the words based on the meaning of the previous ones.
- Classification events in the movie – previous events are necessary for reasoning
- Excellent for modelling sequences
Long Short Term Memory Neural Network
recurrent neural network

Long Short Term Memory

- the context is more “far” in history
- specific function to determine what to remember
- gates
  - Forget
  - Input
  - Output
Methodology

how do we make the comparison
Dataset
real-world data shows the real performance

Two monitored services
- Access portal to information resources at university (libraries, datasets collections, ...)
- Web presentation of the Faculty of Science

Observation period
- one month in 2018

Two granularities
- 5 minute => 8928 observations
- 1 hour => 744 observation

Missing values
### Dataset

Real-world data shows the real performance.

<table>
<thead>
<tr>
<th>QoS Attribute</th>
<th>Measured Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of concurrent users (USR)</td>
<td>count</td>
</tr>
<tr>
<td>Application response time (s) (ART)</td>
<td>min, max, <strong>avg</strong>, p50, p90, <strong>p99</strong></td>
</tr>
<tr>
<td>Transaction count (TC)</td>
<td><strong>count</strong></td>
</tr>
<tr>
<td>Network transport time (s) (NTT)</td>
<td>min, max, <strong>avg</strong>, sum</td>
</tr>
<tr>
<td>Transport size (s) (TS)</td>
<td>min, max, <strong>avg</strong>, <strong>sum</strong></td>
</tr>
</tbody>
</table>
Dataset
real-world data shows the real performance
Forecast
there is not only one forecast

Time scale
- Real-time
- Short-term
- Middle-term
- Long-term

Number of forecasted observations
- One-step
- Multi-step

Forecast frequency
- One-time
- Continuous

Our goal
- One step, continuous, real-time/short-term
Models Construction
our approach to estimation

ARIMA(p,d,q)
- Box-Jenkins Methodology
  - Differencing order (Augmented Dickey-Fuller test for stationarity)
  - Autocorrelation plot to determine p,q (AIC if is unclear)
  - Maximum likelihood and Kalman Filter estimation

Holt-winters
- Additive vs multiplicative
- Season length identification (ACF, PACF)
- Parameters estimation (Maximum likelihood)

LSTM NN
- Standardization of time series
- One input, one hidden, one output layer
- MSE – stop loss function
- Stochastic gradient descent optimizer
- Number of iteration determined from learning curve
Models Evaluation
how do we compare

Training and testing dataset
Forecast Precision
- Mean Absolute Percentage Error

\[ MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100 \]

Time complexity
- Time to estimate a model
- 6 AMD Ryzen5 CPUs 3.8GHz, 6GB RAM
Experiment Results

the data reveals the truth
## Models Settings

given by the dataset

### ARIMA

<table>
<thead>
<tr>
<th>QoS Attribute</th>
<th>SERV-1 5 min</th>
<th>SERV-1 1h</th>
<th>SERV-2 5 min</th>
<th>SERV-2 1h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of concurrent users (USR)</td>
<td>(2,0,0)</td>
<td>(2,0,0)</td>
<td>(2,0,0)</td>
<td>(2,0,0)</td>
</tr>
<tr>
<td>Response time - avg (ART-avg)</td>
<td>(2,1,0)</td>
<td>(1,0,0)</td>
<td>(1,0,0)</td>
<td>(1,0,0)</td>
</tr>
<tr>
<td>Response time - p99 (ART-p99)</td>
<td>(1,1,0)</td>
<td>(2,1,0)</td>
<td>(0,0,1)</td>
<td>(1,0,0)</td>
</tr>
<tr>
<td>Transaction count (TC)</td>
<td>(3,0,0)</td>
<td>(2,0,0)</td>
<td>(4,0,0)</td>
<td>(3,0,0)</td>
</tr>
<tr>
<td>Network transport time (NTT)</td>
<td>(2,1,0)</td>
<td>(1,1,0)</td>
<td>(0,0,1)</td>
<td>(1,0,0)</td>
</tr>
<tr>
<td>Transport size (TS)</td>
<td>(3,0,0)</td>
<td>(2,0,0)</td>
<td>(3,0,0)</td>
<td>(3,0,0)</td>
</tr>
</tbody>
</table>

### Holt-Winters

- USR, TC day-night, week pattern
- Season set to 7 days
- Parameter estimation
  - Level – varied over whole interval
  - Trend – no trend identified
  - Season – close to one – recent more weight
Models Settings
given by the dataset

**LSTM NN**

- Two hidden cells
- Number of iterations
  - ART, NTT, TS – rapid drop
  - USR, TC – 1 hour
  - Other linear descend
  - Set to 100

![MSE Score vs Number of iterations graph](image)
# Models Comparison

## MAPE performance

<table>
<thead>
<tr>
<th>QoS Attribute</th>
<th>Service</th>
<th>ARIMA 5 min</th>
<th>ARIMA 1 h</th>
<th>Holt-Winters 5 min</th>
<th>Holt-Winters 1 h</th>
<th>LSTM NN 5 min</th>
<th>LSTM NN 1 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of concurrent users (USR)</td>
<td>SERV-1</td>
<td>7.79</td>
<td>13.70</td>
<td>28.09</td>
<td>38.89</td>
<td>2.16</td>
<td>20.27</td>
</tr>
<tr>
<td></td>
<td>SERV-2</td>
<td>5.44</td>
<td>10.02</td>
<td>24.41</td>
<td>32.11</td>
<td>1.61</td>
<td>20.84</td>
</tr>
<tr>
<td>Response time - avg (ART-avg)</td>
<td>SERV-1</td>
<td>119.04</td>
<td>113.01</td>
<td>212.61</td>
<td>141.45</td>
<td>100.99</td>
<td>116.52</td>
</tr>
<tr>
<td></td>
<td>SERV-2</td>
<td>103.44</td>
<td>41.24</td>
<td>66.48</td>
<td>45.08</td>
<td>40.39</td>
<td>30.87</td>
</tr>
<tr>
<td>Response time - p99 (ART-p99)</td>
<td>SERV-1</td>
<td><strong>250.42</strong></td>
<td><strong>110.83</strong></td>
<td>504.83</td>
<td>195.52</td>
<td>497.23</td>
<td>153.43</td>
</tr>
<tr>
<td></td>
<td>SERV-2</td>
<td>205.54</td>
<td>84.62</td>
<td>165.77</td>
<td>126.70</td>
<td><strong>106.58</strong></td>
<td>71.70</td>
</tr>
<tr>
<td>Transaction count (TC)</td>
<td>SERV-1</td>
<td><strong>76.28</strong></td>
<td>50.80</td>
<td>310.98</td>
<td>272.62</td>
<td>252.68</td>
<td>119.89</td>
</tr>
<tr>
<td></td>
<td>SERV-2</td>
<td>36.23</td>
<td>28.75</td>
<td>226.91</td>
<td>198.95</td>
<td><strong>28.07</strong></td>
<td><strong>11.40</strong></td>
</tr>
<tr>
<td>Network transport time (NTT)</td>
<td>SERV-1</td>
<td>288.63</td>
<td>96.53</td>
<td><strong>238.26</strong></td>
<td>99.16</td>
<td>460.22</td>
<td>69.96</td>
</tr>
<tr>
<td></td>
<td>SERV-2</td>
<td><strong>374.92</strong></td>
<td>81.40</td>
<td>394.67</td>
<td>96.04</td>
<td>409.73</td>
<td>34.31</td>
</tr>
<tr>
<td>Transport size (TS)</td>
<td>SERV-1</td>
<td>46.82</td>
<td>25.99</td>
<td>51.79</td>
<td>160.12</td>
<td><strong>46.40</strong></td>
<td>39.72</td>
</tr>
<tr>
<td></td>
<td>SERV-2</td>
<td><strong>112.90</strong></td>
<td>48.84</td>
<td>210.06</td>
<td>154.127</td>
<td>386.56</td>
<td>35.73</td>
</tr>
</tbody>
</table>
### Time Complexity

**how long does it take**

<table>
<thead>
<tr>
<th>Granularity</th>
<th>ARIMA</th>
<th>Holt-Winters</th>
<th>LSTM NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 minutes</td>
<td>574.56 ± 509.06</td>
<td>44.04 ± 4.17</td>
<td>397.48 ± 43.63</td>
</tr>
<tr>
<td>1 hour</td>
<td>30.21 ± 30.42</td>
<td>2.92 ± 0.82</td>
<td>33.70 ± 1.61</td>
</tr>
</tbody>
</table>
Further Notes

what can be improved

Initial weights for LSTM NN

Outliers present

- Use Symmetric Mean Absolute Percentage Error instead MAPE

LSTM Time complexity

- Adam or RMSProp optimizer instead SGD

Data preprocessing
**Summary**

and future work

**Centralized monitoring of QoS**

**Comparison of methods for QoS timeseries forecasting**

- ARIMA vs. Holt-winters vs. LSTM NN
- LSTM NN better for high granular data
- Dataset and experiment released for public

**Future work**

- K-step prediction
- Optimization of LSTM NN performance
- Data preprocessing
Thank you for your attention

Tomas Jirsik et al.
jirsik@ics.muni.cz

https://csirt.muni.cz/
https://github.com/CSIRT-MU/QoSForecastLSTM
@csirtmu