Anomaly detection using Long Short Term Memory Networks and its applications in Supply Chain Management

Kim Phuc Tran, Hu Du Nguyen, Sébastien Thomassey

Preprints, 9th IFAC Conference on Manufacturing Modelling, Management and Control
Berlin, Germany, August 28-30, 2019

Anomaly detection using Long Short Term Memory Networks and its applications in Supply Chain Management

Kim Phuc Tran * **Huu Du Nguyen** ** **Sébastien Thomassey** ***

* *GEMTEX Laboratory, Ecole Nationale Suprieure des Arts et Industries Textiles, BP 30329, Roubaix 59056, France. (e-mail: kim-phuc.tran@ensait.fr).*
** *Faculty of Information Technology, Vietnam National University of Agriculture, Hanoi, Vietnam (e-mail: nhdu@vnua.edu.vn)*
*** *GEMTEX Laboratory, Ecole Nationale Suprieure des Arts et Industries Textiles, BP 30329, Roubaix 59056, France (e-mail: sebastien.thomassey@ensait.fr)*

Abstract: Anomaly detection has been becoming an important problem in several domains. In this paper, we propose a new method to detect anomalies in time series based on Long Short Term Memory (LSTM) networks. After being trained on normal data, the networks are used to predict interested steps in time series. The difference between the predicted values and observed values is calculated as prediction errors. Then we use a kernel estimator of the quantile function to compute a threshold, which is used to determine anomalous observations. The performance of proposed method is illustrated through an example of anomaly detection of consumer demand in supply chain management. The numerical experiment shows that our approach achieve a higher level of detection accuracy and a lower percentage of false alarm rate compared to the previous One-Class Support Vector Machine method.

Keywords: Long short term memory networks, Anomaly detection, Supply chain management, Quantile kernal estimator, Time series.

1. INTRODUCTION

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. Often, anomalies in data are translated to significant and critical information in a wide variety of applications such as intrusion detection, fraud detection, fault detection, system health monitoring, image processing and sensor networks. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination, or anomalies in credit card transaction data could indicate a credit card or identity theft. Therefore, anomaly detection is an important problem well worth considering. In fact, this problem has been studied by many authors and there is a large number of techniques proposed in the literature to deal with it. A review of novelty detection techniques using neural networks, machine learning and statistical approaches could be seen in Markou and Singh (2003a,b); Hodge and Austin (2004). An extensive survey covering both the anomaly detection techniques and its applications has been conducted in Chandola et al. (2009). Very recently, Mehrotra et al. (2017) provides a comprehensive research on anomaly detection principles and algorithms. In additon, one should consider that control charts in statistics process control (SPC) also belong to anomaly detection techniques. The goal of control charts is similar to that of one-class classification methods as only one class is represented in the training data, which can be used to learn its characteristics as well as to provide a measure of abnormal behavior of new observations. Several types of control charts were presented in Tran (2018).

In the domain of supply chain management (SCM), anomaly detection is also a key factor to make better decisions. An important problem in SCM is to reduce decision cycle times, even though a huge amount of data are being generated at every stage. This overload of data might result in the difficulty of discerning useful signals, which enable meaningful decisions from meaningless ones. By implementing anomaly detection, questionable data are quickly analyzed to determine anomalies or unexpected patterns for making more effective decisions. In the literature, a number of studies focus on abnormal event detection in supply chain based on the radio frequency identification (RFID) technology, see, for example, Shunping and Dong (2014), Angiulli and Masciari (2010), Chen et al. (2017), Garri et al. (2011), Sharma and Singh (2013), and Dolgui and Proth (2012) to name a few. Jandel et al. (2012) investigated the online stability of statistical relational learning (SRL) in the context of supply chain security. Zhao and Zhou (2013) improved the quick outlier detection (QOD) algorithm by clustering based on data streams applied to cold chain logistics. Kraus and Valverde (2014) designed a data warehouse for the detection of fraud in the supply chain by using the Benfoud's law. Roesch and Deusen (2010) suggested a quality control approach for detecting anomalies in analysis of annual inventory data. Two anomaly detection techniques, including statistical-based approach and clustering-based approach, were used to detect outliers in sensor data for real-time monitoring system of perishable supply chain in Alfian et al. (2017).

As discussed in Bontemps et al. (2016), most of the current studies on anomaly detection do not consider the previous or recent events in detecting the new incoming outlier, i.e., they are based purely on the learning of normally and anomaly behaviors. Using Long Short Term Memory (LSTM) networks is then proposed to deal with challenges associated with temporally dependent anomaly detection problems, due to their ability to maintain long term memories. The method is well-known as a powerful technique to learn the long-term dependencies and represent the relation between current events and previous events effectively. In the literature, most of appliactions of LSTM networks are in predictive analytics. For example, a stacked LSTM network has been presented for anomaly prediction of space shuttle and engine in Malhotra et al. (2015). The bi-directional LSTM is applied to capture long-term dependence for machining tool wear prediction in Zhao et al. (2017). Wu et al. (2018) suggested to use a vanilla LSTM to estimate the remaining useful life of an aircraft turbofan engine under complex operating conditions and strong background noise. In the filed of SCM, Bousqaoui et al. (2017) proposed a LSTM based prediction moddel for demand forecasting. The authors also provided a synthesis of some applications of machine learning and neural networks in supply chains. A brief review of the studies related to applying LSTM to time series and anomaly detection issues was presented in Bontemps et al. (2016). Malhotra et al. (2015) suggested to use stacked LSTM networks for anomaly detection in time series. However, in this study, the authors used an assumption of multivariate Gaussian distribution for error vectors, which is not really practical. The goal of this paper is to avoid this drawback by using a quantile kernal estimator to define a threshold which is needed for detecting anomalies. We consider a scenario in supply chain management to demonstrate the efficiency of our proposed method.

The rest of the paper is organized as follows: Section 2 briefly describes the LSTM network. The approach for detecting anomaly in time series based on LSTM network and the quantile kernal estimator is presented in Section 3. Section 4 shows the performance of proposed method via an example of detecting anomalies in consumer demands. Some concluding remarks are given in Section 5.

2. LONG SHORT TERM MEMORY NETWORKS

LTSM network is a special type of artificial neural networks that accounts for the long-term dependencies between data nodes. It can be seen as a variant of recurrent neural networks, a very power tool in deep learning. Each LTSM has a form of a chain of repeating modules of neural networks to process a variable-length sequence $(x_1, x_2, .., x_t, ..)$ by adding new information into a single memory slot. Figure 1 displays the structure and the operational principle of a typical LTSM module. Each module consists of a three control gates and four neural network layers interacting in a a special way. The horizontal line running through the top of the module is called cell state. It runs straight down the entire chain, maintaining the sequential information in an inner state and allowing the LSTM to persist the knowledge accrued from subsequent time steps. The gates represent the ways to screen information through; they controll the cell state by deciding the extent to which old information should be erased ("forget gate"), new information should be memorized ("input gate"), and current contents should be exposed ("output gate"). Each gate is composed out of a sigmoid neural net layer and a pointwise multiplication operation. The sigmoid layers output numbers in the interval $[0, 1]$, representing portion of input information should be let through. A "zero" output means letting nothing through while a "one" represents passing everything. The LSTM module works as follows. As soon as obtaining the new information x_t in state t, the first step of LSTM is to decide what old information should be forgotten by outputing a number within $[0, 1]$ from forget gate, say f_t with

$$f_t = \sigma_1(W_f.[h_{t-1}, x_t] + b_f), \tag{1}$$

where h_{t-1} is the output in state $t-1$, W_f and b_f is the weight matrices and the bias of forget gate. The second step is to process new information before storing in cell state. In this step, the values updated in cell state, say i_t, is determined in sigmoid layer σ_2 (input gate) along with a vector of candidate values $\tilde{C}_t$ generated by a tanh layer at the same time. The old cell state and new information are then updated into new cell state C_t by

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{2}$$

in which i_t and $\tilde{C}_t$ are defined by

$$i_t = \sigma_2(W_i.[h_{t-1}, x_t] + b_i), \tag{3}$$

$$\tilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c), \tag{4}$$

where (W_i, b_i) and (W_c, b_c) are the weight matrices and the biases of input gate and memory cell state, respectively. From this updated state, the last step is to output values of an LSTM cell. The cell state C_t is put into a tanh layer to scale down the vector state to a number between -1 and 1 while the output gate o_t determines parts of the cell state being outputed throug the number within $[0, 1]$ computing from the sigmoid layer σ_3. The value h_t is then regulated and filtered before outputing by multiplying these two numbers as

$$h_t = o_t * tanh(C_t), \tag{5}$$

in which o_t is defined by

$$o_t = \sigma_3(W_o.[h_{t-1}, x_t] + b_o), \tag{6}$$

where W_o and b_o are the weight matrix and the bias of output gate. There are also various variants of LSTM suggested by different authors. A direct comparision of popular variants of LSTM made by Greff et al. (2017) showed that these varians are almost the same; a few among them are more efficient than others but only in some specific problems.

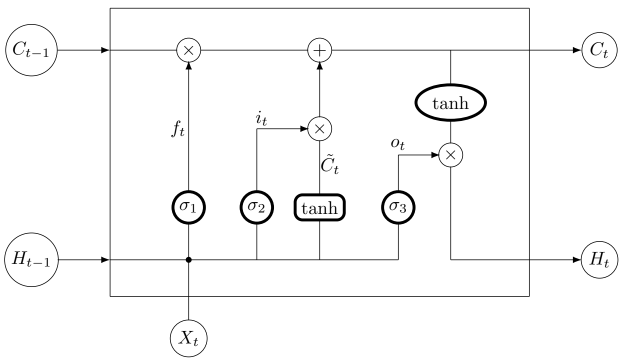

Fig. 1. A module of LSTM network

3. ANOMALY DETECTION USING LSTM AND KERNEL ESTIMATOR OF THE QUANTILE FUNCTION

In most of practical situations, the number of instances of normal behaviors usually overwhelms the number of unexpected cases. The main idea of using LSTM in anomaly detection is to model normal samples by adapting its weights to represent best the training data, and then use the prediction errors to indentify anomalies.

Let $X = \{x_1, x_2, ..., x_{n+1}\}$ denote a time series where each point x_i could be a multi-dimensional vector. This series is considered as a normal sequence. A subset of X is drawn to become a normal validation set, i.e., $Y = \{y_1, y_2, ..., y_m\} \subset X$ $(m \leq n+1)$. The LSTM network is learned using this training data, which consist of the input variables X and its label Y. The prediction model obtained after training is used to compute prediction errors from the time series. Take a simple case as an example. With the input x_{i-1}, the model can predict the next value $\hat{x_i}$ in the time series. That means in this case we obtain a set of error vectors $\{e_1, e_2, ..., e_n\}$ where $e_i = \|\hat{x}_{i+1} - x_{i+1}\|, i = 1, \ldots, n$. The anomaly detection is then based on these prediction errors. In Malhotra et al. (2015), the authors supposed that these error vectors follow a multivariate Gaussian distribution and used the maximum likelihood estimation to define a threshold indicating abnormalities. However, in many practical situations, it seems to be hard to hold this assumption as the distribution of prediction errors is often unknown. We avoid setting any assumption about the distribution of prediction errors by applying a non-parametric method. In particular, we apply a control chart based method using kernel quantile estimator (Sheather and Marron (1990)) to estimate the threshold τ over the set of error vectors. Suppose that the error vectors are reorganized corresponding to the order statistics, i.e., $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(m)}$ (general case). Let K be a density function symmetric about zero. The value of the threshold τ is set to be the kernel quantile estimator calculated by:

$$\tau_p = \sum_{i=1}^{m} \left[\int_{\frac{i-1}{m}}^{\frac{i}{m}} \frac{1}{h} K(\frac{t-p}{h}) dt \right] e_{(i)}, \tag{7}$$

where the bandwith $h > 0$ is to control the smoothness of the estimator for a given sample of size m, and p is a predefined value, $0 < p < 1$. Several versions for the approximation of τ_p were provided in Sheather and Marron (1990). The author also pointed out that the choice of K does not have much effects on the estimation performance. In this paper, we apply a standard Gaussian kernel for K, that is

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}). \tag{8}$$

In contrast to K, the bandwith has a siggnificant influence on the smoothness of density estimate. An asymptotical optimized bandwidth is applied in Sheather and Marron (1990) as:

$$h_{opt} = \left(\frac{p(1-p)}{m+1} \right)^{\frac{1}{2}}. \tag{9}$$

The value of threshold τ_p is used to detect anomalous sample in the time series. In particular, at the new time t, a prediction error is calcluated by $e_t = \|x_t - \hat{x}_t\|$ where x_t is a new observation and $\hat{x}_t$ is a predicion of x_t which is predicted using the trained model. If $e_t > \tau_p$, x_t is classified as "anomalous" and vice versa.

4. ILLUSTRATIVE EXAMPLE

In this Section, we illustrate the performance of proposed method on a situation of detecting anomalous consumer demands in SCM. Understanding consumer demands is a key factor that ensure the success of each enterprise in nowadays globally competitive market (Thomassey (2010)). Detecting anomalies in customer demands enables suppliers to better manage the supply chain. For example, in some real situations, a manager can recognize the difference between current consumer demands and previous demands (which were assumed to be normal), but he or she is not sure whether this difference is abnormal or not. If the answer is "*yes*", he or she should pay attention on discovering the assignable causes of abnormalities to make reasonable adjustments. Moreover, anomaly detection in customer demands also makes the consumer demand forecasting more efficient as it eliminates errors and anomalies in the data used for forecast models, avoding "*garbage in, garbage out*".

In our experiment, a training data set of 2810 normal samples and a validation data set of 148 normal samples representing the normal consumer demands are simulated. After training with LSTM, the performance of proposed model is evaluated based on 400 normal and abnormal samples of a simulated testing data set. Let TP (True Positive) denote the number of anomalies correctly diagnosed as anomalies, TN (True negative) denote the number of normal events correctly diagnosed as normal, FP (False Positive) denote the number of normal events incorrectly diagnosed as anomalies, and FN (False Negative) denote the number of anomalies incorrectly diagnosed as normal events. The following measures, which have been widely used to analyze the performance of a certain detection method, is applied:

- Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$,
- DR $= \frac{TP}{TP+FN}$,

Method	DR(Recall)	Precision	Accuracy	F-score
LTSM	0.98	0.96	0.94	0.86
OCSVM	0.78	0.93	0.76	0.62

Table 1. Performance metrics of proposed approach

- Precision $=\frac{TP}{TP+FP}$,
- F-score $= 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$.

Among these measures, DR (Recall or sensitivity) represents the true positive rate, Precision represents the probability of predicting a true positive from all positive predictions, Accuracy represents a general anomaly detection performance and F-score provides a configurable harmonic mean between Precision and Recall.

Our computation is performed on a platform with 2.6 GHz Intel(R) Core(TM) i7 and 32GB of RAM. With the quantile $p = 0.99$, we obtain $\tau_p = 0.448$. Figure 2 displays the process of detecting anomalies in the simulated data of consumer demands. The test on the simulated testing data set shows a good performance as presented in Table 1. The same procedure is applied to the traning and testing data set by using the One-Class Support Vector Machine (OCSVM) method suggested in Schölkopf et al. (2001). As can be seen in Table 1, the proposed LSTM based method leads to an Accuracy of 94%, a Precision of 96%, a F-score of 86%, and a Recall of 98%, which are all significantly higher than those corresponding values of $78\%, 93\%, 76\%$ and 62%, respectively, from the OCSVM method. That is to say, our proposed method outperforms the OCSVM based method, ensuring more accurate detection of anomalies in customer demands. These anomalies will be carefully analyzed to find out the assignable causes. This could bring very useful information for managers to make better decisions. Also, the data after being processed anomalies are the valuable data for other tasks like forecasting. Predicting accurately or approximately the information about expected customer demands is a very important problem that have a great influence on SCM, especially on production planning and inventory management. In this context, the LSTM based mehod of forecasting demand for time series will be a powerful tool. For example, the LSTM model could be applied to mitigate the tight assumptions of mathematical models such as assumption of linear mixed effects of customer demands suggested in Stefanescu (2009). This promising application of the LSTM for inventory management is our goal of study in the near future.

5. CONCLUSION

In this paper, a new method to detect anomalies based on a LSTM network has been presented. We have proposed using a quantile kernal estimator to define the threshold for anomaly detection from error vectors. An advantage of this suggestion is that it does not need any assumption about distribution of prediction errors. The efficiency of proposed method has been illustrated through a situation in supply chain management. In gereral, this method could be applied to any scenarios that require detecting anomalies related to time series. As future work we would like to address the anomaly detection problem using autoencoder and control charts for large stream data, and to address the customer demand prediction using LSTM networks.

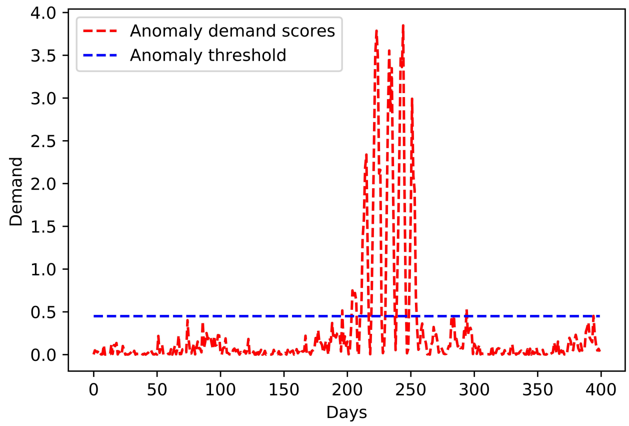

Fig. 2. Anomaly detection in consumer demand

REFERENCES

Alfian, G., Syafrudin, M., and Rhee, J. (2017). Real-time monitoring system using smartphone-based sensors and nosql database for perishable supply chain. *Sustainability*, 9(2073), 1–17.

Angiulli, F. and Masciari, E. (2010). Effectively monitoring rfid based systems. In B. Catania, M. Ivanovic, and B. Thalheim (eds.), *Advances in Databases and Information Systems. ADBIS 2010*, volume 6295 of *Springer, Berlin, Heidelberg*, 31–45.

Bontemps, L., McDermott, J., Le-Khac, N.A., et al. (2016). Collective anomaly detection based on long short-term memory recurrent neural networks. In *International Conference on Future Data and Security Engineering*, 141–152. Springer.

Bousqaoui, H., Achchab, S., and Tikito, K. (2017). Machine learning applications in supply chains: long short-term memory for demand forecasting. In *International Conference of Cloud Computing Technologies and Applications*, 301–317. Springer.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 15.

Chen, M., Liu, J., Chen, S., Qiao, Y., and Zheng, Y. (2017). Dbf: A general framework for anomaly detection in rfid systems. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*.

Dolgui, A. and Proth, J.M. (2012). Radio frequency identification (rfid) in supply chain: Technology, applications and concerns. In *Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing*.

Garri, K., Sailhan, F., Bouzefrane, S., and Uy, M. (2011). Anomaly detection in rfid system. *International Journal of Radio Frequency Identification Technology and Applications*, 3(1), 31–46.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10), 2222–2232.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2), 85–126.

Jandel, M., Svenson, P., and Wadstromer, N. (2012). Online learnability of statistical relational learning in anomaly detection. *Proc 15th Int Conf Information Fusion.*

Kraus, C. and Valverde, R. (2014). A data warehouse design for the detection of fraud in the supply chain by using the benford's law. *American Journal of Applied Sciences*, 11(9), 1507–1518.

Malhotra, P., Vig, L., Shroff, G., and Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In *Proceedings*, 89. Presses universitaires de Louvain.

Markou, M. and Singh, S. (2003a). Novelty detection: A review-part 1: Statistical approaches. *Signal Processing*, 83(12), 2481–2497.

Markou, M. and Singh, S. (2003b). Novelty detection: A review-part 1: Statistical approaches. *Signal Processing*, 83(12), 2499–2521.

Mehrotra, K., Mohan, C.K., and Huang, H.M. (2017). *Anomaly Detection Principles and Algorithms.* Springer.

Roesch, F.A. and Deusen, P.C. (2010). Anomaly detection for analysis of annual inventory data: A quality control approach. *Southern Journal of Applied Forestry*, 34(3), 131–137.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7), 1443–1471.

Sharma, M. and Singh, M. (2013). Outlier detection in rfid datasets in supply chain process: A review. *International Journal of Computer Applications*, 65(25), 47–51.

Sheather, S.J. and Marron, J.S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410), 410–416.

Shunping, H. and Dong, W. (2014). Research on supply chain abnormal event detection based on the rfid technology. *Applied Mechanics and Materials*, 513-517, 3309–3312.

Stefanescu, C. (2009). Multivariate customer demand: modeling and estimation from censored sales.

Thomassey, S. (2010). Sales forecasts in clothing industry: The key success factor of the supply chain management. *International Journal of Production Economics*, 128(2), 470–483.

Tran, K. (2018). Designing of run rules t control charts for monitoring changes in the process mean. *Chemometrics and Intelligent Laboratory Systems*, Inpress. doi: 10.1007/s00362-016-0769-4.

Wu, Y., Yuan, M., Dong, S., Lin, L., and Liu, Y. (2018). Remaining useful life estimation of engineered systems using vanilla lstm neural networks. *Neurocomputing*, 275, 167–179.

Zhao, R., Yan, R., Wang, J., and Mao, K. (2017). Learning to monitor machine health with convolutional bidirectional lstm networks. *Sensors*, 17(2), 273.

Zhao, W. and Zhou, W.D.S. (2013). Outlier detection in cold-chain logistics temperature monitoring. *Elektronika ir Elektrotechnika*, 19(3), 65–68.