# Deep-Learning Based Anomaly Classification in Unstructured Log Data

Jiechao Cheng, Rui Ren

[jetrobert19@gmail.com](mailto:jetrobert19@gmail.com), [renrui@ict.ac.cn](mailto:renrui@ict.ac.cn)

## Abstract

As increasing usage of servers, the threats of anomaly operations to system have also been gone up. Recognizing the anomalous instances is still one of the important works for system anomaly detection and classification.

Anomaly detection and type classification problem for time series is a tedious but mush beneficial task, especially harnessing the massive amount of data generated by variable systems to create value for technology development and business strategies. By using the state-of-art machine learning techniques and deep neural networks, we propose a series of approaches to assist in focusing on detecting the most crucial types of anomalies and gaining critically insights of log information. Here we propose a state-of-the-art recurrent neural network approach and measure the accuracy of our diverse array of classification algorithms. The results reveal the strengths and advantages of our long short-term memory neural network and convolutional neural network for real-world classification tasks of anomaly types. In this article, we employ Long Short Term Memory (LSTM) recurrent neural network and Convolutional neural network (CNN) architectures for anomaly classification models with our system log dataset. Our study shows that deep-learning-based classification models outperform the state-of-the-art machine learning approaches.

**Index Terms**—Anomaly Classification, Machine Learning, Deep Learning, Log Analysis, Long short term memory, Recurrent neural network, Convolutional neural network

## 1 Introduction

During the last decade, a plenty of system serves have become highly popular, and it also induces new challenges that impact operators. The management of system log data and outliers potentially affecting the end users. … For a complex scenario like this, it is of vital importance to effectively detect and classify the occurrence of system anomalies for reducing the loss of the profit.

We propose a simple yet efficient approach to detect and classify system operation anomalies using machine learning and deep learning techniques. Deep learning algorithms build data-driven models from labeled data and make predictions on data which they can learn from. Deep learning provide a more promising alternate for detecting and categorizing log anomalies based on the large set of original features or more relevant set of features for classification process. Deep learning has been largely used in the field of image classification, speech recognition, etc., but not much in log anomalies. Our studies have shown that deep learning algorithms are able to achieve potentially high classification accuracy.

To obtain favorable performances of anomaly classification, we propose a deep learning model based on well-known long short term memory (LSTM) recurrent neural network and deep convolutional neural network. A LSTM recurrent neural network is a classification algorithm that classifies instances by … Deep learning algorithms are like appealing black-box solutions, it’s efficient but very challenging to understand the detailed reasons leading to a particular classification result. In addition, LSTM explicitly show …, the learning algorithm automatically select the most discriminating features. CNN…. Last but not least, previous work [] has shown that deep learning outperforms other machine learning algorithms for the sake of log anomaly classification.

The remainder of this article is organized as follows: Section 2 briefly reviews the related works. Section 3 reveals our deep learning approaches utilized in the rest of the paper. Section 4 describes the proposed anomaly classification experiments including the generation of the semantic datasets used for models training and evaluation purposes. Section 5 presents the discussions of the obtained anomaly classification results. Finally, the last section concludes out work.

Usually, log data is analyzed in order to detect misuses of a system or suspicious events indicating anomalies.

Anomaly discrimination related problems are addressed in a great deal of practical applications, including fraud detection, intrusion detection, system health monitoring as well as event detection in sensor networks. Anomalous items are also referred to as outliers, novelties, noise, deviations and exceptions [1].

In contrast to typical unsupervised anomaly detection, which is often applied on unlabeled data set under the assumption that the majority of the instances are normal, instead, we here take a data set that has been labeled as normal and abnormal into account for supervised anomaly classification tasks with the state-of-the-art deep learning algorithm classifier. In supervised learning, removing the anomalous data from the data set often results in a statistically significant increase in accuracy [2].

There are a diverse array of anomaly detection techniques have been proposed, such as density-based techniques [3, 4, 5], correlation-based outlier detection [6], cluster analysis-based outlier detection [7, 8] and ensemble techniques [9, 10]. When compared across huge data sets and hyper parameters, different methods have little systematic advantages over another in the measurement performance [11].

To classify types of anomalies at scale, we use different combinations of techniques starting with SVM and ending with CNN and LSTM sophisticated deep learning models.

The semantic transformation from a raw unstructured anomaly categorization task to a structured anomaly type classification task requires a solid background knowledge of the dataset, which features and instances are so different from the original raw data, namely the generation of a data view [12].

In fact, many practical anomaly detection problems often require a preprocessing in order to generate the appropriate data to handle with. The final step before the unsupervised anomaly detection algorithm can be applied is normalization. In practical applications, the min-max normalization is often used, every feature is normalized into a [0, 1] interval, so do we in the evaluation in this article

In this paper we present

## 2 Related Work

There has been considerable amount of research about anomaly classification in system log.

It is commonly accepted that deep learning algorithms are well-suited for classification.

Sequences and time series data usually need different algorithms to detect anomalies [13]. P. Fiadino et al. [19] reported statistical detection and diagnosis of anomalies. Lazarevic et al. [14] compared LOF, k-NN, PCA and unsupervised SVM for intrusion detection. Ding et al. [15] studied SVDD, a k-NN classifier, k-means and a GMM for detecting anomalies. Amer et al. [16] proposed One-class Support Vector Machines for anomaly detection. The local density cluster-based outlier factor (LDCOF) [18] detect anomalies by estimating the clusters’ densities assuming a spherical distribution of the cluster members. Sub-space clustering approaches [20] have also been used in anomaly classification.

Combining with multiple anomaly detection algorithms, outlier ensembles boost their joint detection performance [17].

## 3 Method

We applied semantic data, derived from real system operation traces as suggested in []. The data in such format allows analyze the real-time system server operation with a large number of operational anomalies efficiently, moreover, it protects the sensitive information of system services. The procedure of generating semantic data and preprocessing are illustrated as following.

A label can be used as a result indicating whether an instance is an anomaly or not.

Let’s review discriminative algorithms from the perspective of application to finding various types of anomalies. The most suitable type of neural network working with time series is LSTM (Long Short-Term Memory) recurrent neural network, if properly built, it allows you to model the most sophisticated dependencies.

## 4 Experiments Settings

**Dataset**

Using a larger set of log data as input for a deep learning model is not always the best choice, as this increases the dimension of the input data, introducing sparsity issues. Therefore, it may negatively impact classification results. Meanwhile, irrelevant or redundant features will bring more noise to the overall process, thus models will obtain inferior performances.

We total include 14 different variations of anomalies, each one is assigned a class.

We construct a fully labeled dataset spanning a period of a couple of months with consecutive days. The construction of the semantic data set is conceived with the objective of fundamentally maintaining the underlying structural characteristics of the raw temporal operation data as much as possible. The transformation procedure is described as follows. The first step of the construction procedure consists of manually labeling. Then, we transform the textual information into structured representation.

We divide this vector into m blocks, each one corresponding to a two-minute interval….

The data set gained in this way retains certain features of real log data. It keeps the time-series variations of system operation, also, it maintains the differentiation among a variety of anomaly types.

In order to categorize the anomalies, we take the distributions of anomalies across the types described in table 1.

**Modeling**

During several months of trials, we successfully classify anomalies with high accuracy. We aim at modeling...

**Techniques**

Using more features increases the dimensionality of the feature space, usually bring in undesirable effects like sparsity, and some redundant or irrelevant features may diminish performance of models in classification.

In this section we describe the proposed anomaly classification approach based on deep learning, focusing on the principal features as input. Deep learning is widely employed lately as it is very efficient in a large number of scenarios, especially for huge amount of high-dimensional datasets.

Besides convolutional neural network approach, we consider deep learning approach in our work. In addition, ... It is clear that the selection of features for classification tasks plays a major role in its empirical performance. LSTM approach generally consider the temporal analysis of certain features, it employs a powerful ... to build appealing ...

Convolutional neural network is composed of multiple layers of neurons, each of them generally represented by a non-linear function [], every neural employs an activation function that maps the weighted inputs to the output that is passed to the following layer. The weights, originally set to random values, are iteratively adjusted during the training phase.

Global accuracy, recall and precision, F1-score are consider here to evaluate the performance of our deep learning classification models. Global accuracy Ai indicates the… Recall Ri means, … Precision Pi is… These four standard metrics are widely used for performance evaluation in classification tasks. Accuracy measures…, precision measures…, recall measures…, whereas F1-score measures..

Figure 2 depicts the performance comparison of the 2 classifiers in the classification of all 14 anomalous types. To decrease the influence the bias might bring in to all the evaluations, we employ 8-fold cross-validation with different random splits of the dataset, which indicates that we train and test our models in 8 different train/test dataset combinations.

There are no particular bias for both classifiers. The LSTM classifier shows a slightly higher variance in the results, which might suggests that the model is slightly less robust and prone to leading to over-fitting problems.

CNN models provide great insights about

## 5 Results and Analysis

We evaluate our proposed deep learning approaches in this section by comparing anomalous instances classification performances achieved by all methods.

Not only the occurrence event is categorized in our experiment, but also the whole duration is classified here. Note that LSTM is meant to be applied in the temporal dataset.

CNN achieves almost perfect classification performance in both cases, even slightly surpassing the LSTM classifier. Figure 2 presents the classification results achieved by CNN on all features. We can conclude that CNN offers an accuracy comparable or even slightly better that that achieved by LSTM in all anomalous types.

The classification performance obtained of type database anomalies is slightly worse than that of type file.

## 6 Conclusion

In this paper we have proposed a deep-learning-based approach for anomaly classification of large scale system operation log data, offering a very powerful and straightforward technique to categorize anomalous instances. We believe this appealing approach is capable of providing high insights for understanding system server operations without disclosing any business sensitive information. By depending on deep learning techniques, we have shown the classification performance of the labeled anomalies in an efficient fashion. In general, CNN approach outperforms the LSTM method. We will explore better deep learning model for anomaly classification in the future work.

This area is still on-going research, and it requires a lot of work to build the model for the time series. Should you succeed, you may achieve outstanding performance results in terms of accuracy.

## Acknowledgments

## References

[1] Victoria J. Hodge, Jim Austin. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review. 2004; 22(2):85-126. doi:10.1007/s10462-004-4304-y.

[2] Smith M. R., Martinez T. Improving classification accuracy by identifying and removing instances that should be misclassified. The 2011 International Joint Conference on Neural Networks. 2011; p. 2690. ISBN 978-1-4244-9635-8. doi:10.1109/IJCNN.2011.6033571.

[3] Knorr E. M., Ng R. T., Tucakov, V. Distance-based outliers: Algorithms and applications. The VLDB Journal the International Journal on Very Large Data Bases. 2000; 8(3-4):237-253. doi:10.1007/s007780050006.

[4] Breunig M. M., Kriegel H.-P., Ng R. T., Sander, J. LOF: Identifying Density-based Local Outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD. 2000; pp. 93–104. ISBN 1-58113-217-4. doi:10.1145/335191.335388.

[5] Schubert E., Zimek A., Kriegel H. P. Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. Data Mining and Knowledge Discovery. 2012; 28: 190–237.

doi:10.1007/s10618-012-0300-z.

[6] Kriegel H. P., Kroger P., Schubert E., Zimek A. Outlier Detection in Arbitrarily Oriented Subspaces. 2012 IEEE 12th International Conference on Data Mining. 2012; p.379. ISBN 978-1-4673-4649-8. doi:10.1109/ICDM.2012.21.

[7] He Z., Xu X., Deng S. Discovering cluster-based local outliers. Pattern Recognition Letters. 2003; 24(9-10):1641-1650. doi:10.1016/S0167-8655(03)00003-5.

[8] Campello R. J. G. B., Moulavi D., Zimek A., Sander J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Transactions on Knowledge Discovery from Data. 2015; 10(1):5:1-51. doi:10.1145/2733381.

[9] Zimek A., Campello R. J. G. B., Sander J. R. Ensembles for unsupervised outlier detection. ACM SIGKDD Explorations Newsletter. 2014; 15:11–22. doi:10.1145/2594473.2594476.

[10] Zimek A., Campello R. J. G. B., Sander J. R. Data perturbation for outlier detection ensembles. Proceedings of the 26th International Conference on Scientific and Statistical Database Management-SSDBM. 2014; p.1. ISBN 978-1-4503-2722-0. doi:10.1145/2618243.2618257.

[11] Campos Guilherme O., Zimek Arthur, Sander Jörg, Campello Ricardo J. G. B., Micenková Barbora, Schubert Erich, Assent Ira, Houle Michael E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery. 2016; 30(4):891. ISSN 1384-5810. doi:10.1007/s10618-015-0444-8.

[12] Goldstein M. In: Markus Hofmann RK, editor. Anomaly Detection. Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC. 2013; p.367-394.

[13] Aggarwal CC. Outlier Analysis. Springer-Verlag. NewYork. 2013.

[14] Lazarevic A, Ertoz L, Kumar V, Ozgur A, Srivastava J. A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. In Proceedings of the Third SIAM International Conference on Data Mining. vol. 3. Siam; 2003. p. 25–36.

[15] Ding X, Li Y, Belatreche A, Maguire LP. An Experimental Evaluation of Novelty Detection Methods. Neurocomputing. 2014; 135:313–327. doi:10.1016/j.neucom.2013.12.002

[16] Amer M, Goldstein M, Abdennadher S. Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description (ODD’13). New York, NY, USA: ACM Press. 2013; p. 8-15.

[17] Zimek A, Campello RJGB, Sander J. Ensembles for Unsupervised Outlier Detection: Challenges and Research Questions a Position Paper. SIGKDD Explor Newsl. 2014; 15(1):11-22. doi: 10.1145/2594473.2594476.

[18] Amer M, Goldstein M. Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner. In: Simon Fischer IM, editor. Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012). Shaker Verlag GmbH; 2012. p. 1-12.

[19] P. Fiadino et al. RCATool - A Framework for Detecting and Diagnosing Anomalies in Cellular Networks. ITC. 2015.

[20] P. Casas et al. MINETRAC: Mining Flows for Unsupervised Analysis & Semi-Supervised Classification. ITC. 2011.