CERTAIN INVESTIGATIONS ON IMPROVING OUTLIER DETECTION ACCURACY IN WIRELESS SENSOR NETWORKS

ABSTRACT

Submitted by

ARUL JOTHI S

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

ANNA UNIVERSITY

CHENNAI 600 025

MAY 2023

ABSTRACT

Nowadays, wireless sensor networks (WSNs) are gaining popularity in a variety of civilian and military applications. A WSN can be viewed as a collection of tiny sensor nodes with network connectivity that communicate and exchange information and data. In WSN, data fusion or data aggregation is carried out to gather data in a cluster head (CH) from other nodes in that cluster. These aggregated data will be forwarded to the base station for analysis. In a lively environment, WSN data that is measured and gathered may be sometimes inaccurate. The issues about data reliability in WSNs imply that the sensor data can be inaccurate, which further affects the quality of the raw data and the aggregated results that are forwarded to the base station for analysis. Additionally, transmitting inaccurate data to the base station consumes needless battery power, reducing the lifespan of the network. Identification of abnormal or inaccurate data that vary intensely from remaining data readings is considered suspicious that needs to be focused on and researched. This issue is addressed as Outlier detection (OD) which performs the classification of normal data from abnormal data. Implementing OD in WSNs aids in removing inaccurate data transmission from CH to the base station which is further considered in this research work.

The gathered sensor data may be imbalanced where the abnormal instances are available in minimum amounts. When dealing with imbalanced data, the OD system can suffer from yielding better detection accuracy and more false positives than false negatives. This challenging task motivated researchers to build an OD model to improve detection accuracy with less computation complexity and a reduced number of false alarms. This attracted researchers to develop an efficient OD model that should be able to classify abnormal instances with high detection accuracy and reduce false alarms. In recent years, unsupervised outlier detection (UOD) using deep learning has proved to improve classification accuracy when dealing with imbalanced data produced by various

real-time applications. Accordingly, the goal of this research is to successfully implement the OD model using deep learning techniques to provide better detection accuracy with reduced false alarms, and computational complexity.

Over the years, many researchers have contributed to research in the field of OD in WSNs. Numerous machine learning (ML) strategies were developed, but it suffers from the issue of computational complexity for high dimensional imbalanced data. Very few studies have focused on how to handle data imbalances in WSNs. Existing OD models were developed using machine learning techniques and recently researchers were motivated by the performance of deep learning techniques in terms of time and detection precision for imbalanced and high dimensional data. Despite existing models proposed using deep learning, there are certain research gaps to be focused on improving the OD model detection accuracy due to WSNs resource restrictions. Furthermore, the autoencoder deep learning models that were considered by researchers suffer from learning the significant features of the data during the training phase of the model and are built with complex architecture resulting in expensive computation complexity. This in turn affects the model's accuracy in classifying normal instances from abnormal data instances and WSNs energy consumption. Also, to handle the problems of manual tweaking of hyperparameters, adopting an algorithm for automated tuning of hyperparameters in a high-dimensional space dataset is a noteworthy research gap. Therefore, this research focuses on constructing an efficient OD model to produce better detection accuracy with minimal energy consumption in WSNs.

The first contribution proposes a Feed-forward Autoencoder Neural Network (FANN) model to detect abnormal instances with better accuracy. The FANN model is trained to learn the significant features of normal instances thereby developing a pattern for normal data behaviour. This process helps in reconstructing the data during the decoding phase and analyzing the FANN model's behaviour by evaluating reconstruction errors. The reconstruction error

is later considered for fixing the classification threshold to categorize the data as normal instances or abnormal instances. This enables the proposed FANN model also acts as a False Positive Reducer intending to reduce false alarms. The model performance is validated using various performance metrics such as Accuracy, Precision, Recall, Specificity, F1 Score, false alarm rate (FAR), and Matthews Correlation Coefficient (MCC). Furthermore, the FANN model is compared with the existing algorithms with a real-time Weather dataset.

The second contribution proposes an automated hyperparameter tuning algorithm using particle swarm optimization and stochastic reasoning (PSOSR) model in the proposed FANN model for OD to enable consistent detection accuracy and avoid model overfitting. The proposed model is contributed to eliminating the bias and time involved in the manual search of the hyperparameters. Experimental results were conducted concerning the detection performance of the proposed PSOSR algorithm with the FANN model considering metrics such as Area Under curve (AUC) score, accuracy, precision, recall, and F1 score. Also, the best hyperparameters of each existing model and proposed PSOSR algorithm are also assessed.

The third contribution proposes a Modified Variational AutoEncoder (ModVAE) model with a rule-based approach to focus on well-known feature classification methods dealing with the hidden relation between the features. The motivation for this contribution is that the variational autoencoder (VAE), a deep learning model, illustrates that it can work for unlabeled data with the benefit of the occasional labels when they are present, in contrast to autoencoders utilized in the FANN model. In the proposed ModVAE with a fuzzy inference system (FIS) for OD, the traditional VAE model is modified to learn the latent vector space with problem-specific classification. The model works in two phases: first, the ModVAE model is trained with normal instances and validated with various abnormality ranges using two thresholds (minimum threshold and maximum threshold) that were fixed with the help of reconstruction error.

These thresholds are used to classify validation data as a normal instance (NI) if the model produces an error below the minimum threshold, an abnormal instance (AI) if the model produces an error above the maximum threshold, and suspicious instances (SI) if the model produces an error in between minimum and maximum threshold. In the second phase, the SI will be fed into FIS which was built with fuzzy rules to produce a confidence level to classify the SI as either NI or AI. The proposed model has been experimented with four benchmark datasets and results are discussed in terms of AUC score, detection accuracy, precision, recall, specificity, FAR, and F1 score. Additionally, the proposed model's detection accuracy performance in terms of rank power (RP) and computation complexity has been compared with existing models.

The research contributions presented in the thesis aid in implementing the OD model in the CH during data aggregation and eliminate inaccurate data transmissions to the base station guaranteeing data integrity in WSN. Also, the experimental analysis help in developing an efficient OD model with reduced false alarms to improve the detection accuracy, and has a positive impact on computational complexity. The contributions will also inspire researchers to ponder on future initiatives to enhance the proposed methods to understand the latent features of the data instances even more effectively with incomplete information to measure data uncertainty.