

K-Fold Cross-Validation Machine Learning Approach on Data Imbalance for Wireless Sensor Data:-A Review

M. Tech. Scholar Mandavi Tripathi

mandavitripathi.2011@gmail.com

Asst. Prof. Ankur Taneja

sisodiya.lakhan53@gmail.com

Asst. Prof. Lakhan Singh

ankurtaneja5@gmail.com

Department of Computer Science & Engineering
SAMCET, Bhopal, MP, India

Abstract - Machine learning is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the pattern or extract information from the data. In that case, we apply machine learning with the abundance of datasets available. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves. Data imbalance problem become greatest issue in for machine learning algorithm. Imbalance problem occur where one of the two classes having more sample than other classes. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The minority samples are those that rarely occur but very important. There are different methods available for classification of imbalance data set which is divided into three main categories, the algorithmic approach, data-pre-processing approach and feature selection approach. We will apply on wireless imbalance data to identify correct information. In this research work systematic study for define which gives the right direction for research in class imbalance problem. By review on K-fold cross validation algorithm we indentify problem as well as we will work on this solution to estimate the skill of the model on new data.

Keywords- Unsupervised learning, Supervised learning, Wireless Sensor, Machine Learning algorithm.

I. INTRODUCTION

Wireless sensor networks (WSNs) have been applied in monitoring systems that are capable of controlling and monitoring various indoor premises. WSN are collections of stand-alone devices which, typically, have one or more sensors (e.g. temperature, light level), some limited processing capability and a wireless interface allowing communication with a base station. As they are usually battery powered, the biggest challenge is to achieve the necessary monitoring whilst using the least amount of power. A Wireless sensor network (WSN) is composed typically of multiple autonomous, tiny, low cost and low power sensor nodes.

These nodes gather data about their environment and collaborate to forward sensed data to centralized backend units called base stations or sinks for further processing. The sensor nodes could be equipped with various types of sensors, such as thermal, acoustic, chemical, pressure, weather, and optical sensors. In particular, WSN designers have to address common issues related to data aggregation, data reliability, localization, node clustering, energy aware routing, events scheduling, fault detection and security [1]. Machine learning (ML) was introduced in the late 1950's as a technique for artificial intelligence (AI). Over time, its focus evolved and shifted more to algorithms which are computationally viable and robust.

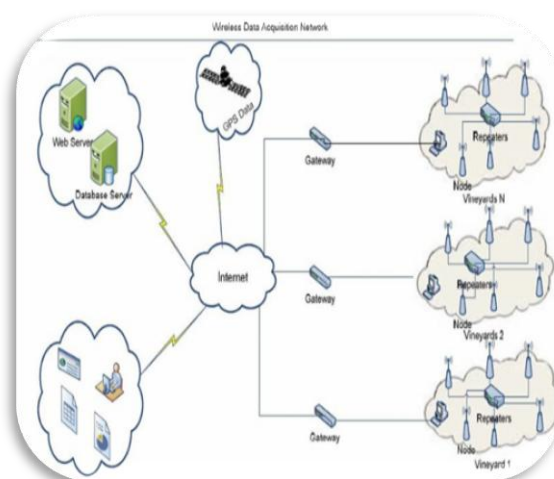


Fig.1 Wireless Sensor Data Network

In the last decade, machine learning techniques have been used extensively for a wide range of tasks including classification, regression and density estimation in a variety of application areas such as bioinformatics, speech recognition, spam detection, computer vision, fraud detection and advertising networks. The algorithms and techniques used come from many diverse fields including statistics, mathematics, neuroscience, and computer science. [1].

1. Approach to handling imbalanced Dataset

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data pre-processing) before providing the data as input to the machine learning algorithm. Re-sampling Techniques Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data pre-processing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application. The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes. Let us look at a few re-sampling techniques:

2. Random Under-Sampling:- A simple under-sampling technique is to under-sample the majority class randomly and uniformly. This can potentially lead to loss of information. But if the examples of the majority class are near to others, this method might yield good results. **Random Over-Sampling:** Random oversampling simply replicates randomly the minority class examples. Random oversampling is known to increase the likelihood of occurring over fitting. On the other hand, the major drawback of Random under sampling is that this method can discard useful data.

3. Cluster-Based Over Sampling:- Cluster-Based Minority Over-Sampling for Imbalanced Datasets. Synthetic over-sampling is a well-known method to solve class imbalance by modifying class distribution and generating synthetic samples. **Informed Over Sampling:** Synthetic Minority Over-sampling Technique This technique is followed to avoid over-fitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created.

K-Fold Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods. By using this discover a gentle introduction to the k-fold cross-validation procedure for estimating the skill of machine learning models.

That k-fold cross validation is a procedure used to estimate the skill of the model on new data. There are common tactics that you can use to select the value of k for your dataset. There are commonly used variations on cross-validation such as stratified and repeated that are available in scikit-learn.

4. Algorithmic Ensemble Techniques

The above section, deals with handling imbalanced data by re-sampling original data to provide balanced classes. In this section, we are going to look at an alternate approach

i.e. Modifying existing classification algorithms to make them appropriate for imbalanced data sets. The main objective of ensemble methodology is to improve the performance of single classifiers. The approach involves constructing several two stage classifiers from the original data and then aggregates their predictions. Bagging Based is an abbreviation of Bootstrap Aggregating. The conventional bagging algorithm involves generating 'n' different bootstrap training samples with replacement.

And training the algorithm on each bootstrapped algorithm separately and then aggregating the predictions at the end. Boosting-Based Boost is the first original boosting technique which creates a highly accurate prediction rule by combining many weak and inaccurate rules. Each classifier is serially trained with the goal of correctly classifying examples in every round that were incorrectly classified in the previous round. For a learned classifier to make strong predictions it should follow the following three conditions: The rules should be simple Classifier should have been trained on sufficient number of training examples. The Classifier should have low training error for the training instances

5. NSL-KDD Data Set

We used Network Services Library (NSL) – Knowledge Discovery Data(KDD). The NSL-KDD dataset should be converted to binary, since the neural networks can only process such types of data. Once converted, the dataset can feed into the neural network model as an input layer. NSL-KDD is a data set suggested to solve some of the inherent problems of the KDD'99 data set which are mentioned in.

Although, this new version of the KDD data set still suffers from some of the problems discussed by McHugh and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods. Furthermore, the number of records in the NSL-KDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable.

6. Random Forest

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The term came from random decision forests that were first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features.

II. LITERATURE REVIEW

In this literature a systematic study for define which gives the right direction for research in class imbalance problem.

1 Author: - Tzu-Tsung Wong, Nai-Yu Yang

Problem:- Author, the overlapping of training sets is shown to be irrelevant in determining whether two fold accuracies are dependent or not.

Solution: - The cross validation of non-overlapping training sets can make fold accuracies to be dependent. However, this dependence almost has no impact on estimating the sample variance of fold accuracies, and hence they can generally be assumed to be independent. [1]

2. Author: - Mohammad Abu Alsheikh, Shaowei Lin

Problem: - Wireless sensor networks monitor dynamic environments that change rapidly over time. This dynamic behavior is either caused by external factors or initiated by the system designers themselves.

Solution: - By A Visualization of the Q-learning method for solutions for their specific application challenges. [2]

3. Author: - Shengguo Hu and Yanfeng Liang

Problem: - Not Accurate SMOTE (Synthetic Minority Over-sampling Technique) is specifically designed for learning from imbalanced data sets external factors or initiated by the system designers themselves.

Solution: - By modified approach (MSMOTE) for learning from imbalanced data sets, based on the SMOTE algorithm. [3]

4. Author: - Shengguo Hu and Yanfeng Liang

Problem: - Not Accurate SMOTE (Synthetic Minority Over-sampling Technique) is specifically designed for learning from imbalanced data sets external factors or initiated by the system designers themselves.

Solution: - By modified approach (MSMOTE) for learning from imbalanced data sets, based on the SMOTE algorithm. [4]

5. Author: - L. Dhanabal, Dr. S. P. Shantharajah

Study:- In this paper the NSL-KDD data set is analyzed and used to study the effectiveness of the various classification algorithms in detecting the anomalies in the network traffic patterns.

Solution: - The study has exposed many facts about the bonding between the protocols and network attacks. [4]

6. Author: - Huaping Guo, Jun Zhou

Problem: - Classification of data with imbalanced class distribution has encountered a significant drawback by most conventional classification learning methods which assume a relatively balanced class distribution.

Solution: - For the learning stage, the proposed method uses the following three steps to learn a class-imbalance oriented model:

- Partitioning the majority class into several clusters using data partition methods such as K-Means,

- constructing a novel training set using SMOTE on each data set obtained by merging each cluster with the minority class, and

- learning a classification model on each training set using convention classification learning methods including decision tree, SVM and neural network [3]

In this proposed solution we will reduced data imbalanced problem for wireless sensor network and achieve a accuracy in results steps of proposed algorithm are given blows.

- Start Sensing data from different wireless sensor.
- Collect Data from Wireless sensors and check for balanced or imbalanced.
- If data is balanced we will use directly for Machine Learning classification Approach.
- If data is imbalanced will use a K-fold cross validation algorithm
- Finally we will compare with different algorithm for imbalanced data. And also check their accuracy

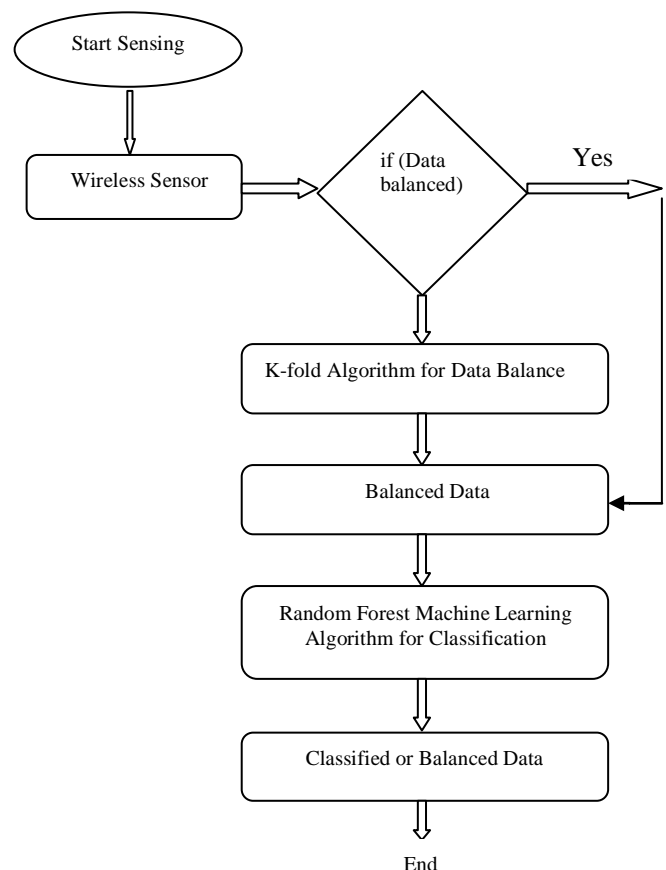


Figure 2 Flowchart of Algorithm

IV. CONCLUSION

In this work we present K-fold cross validation algorithm, a modified technique for learning from different types of imbalanced datasets for improving the performances of model for the minority class. We will use the NSL-KDD

for Experiments and improving the problem on imbalanced data and achieve a accuracy in results.

REFERENCES

- [1] Tzu-Tsung Wong, Nai-Yu Yang "Dependency Analysis of Accuracy Estimates in k-fold Cross Validation" IEEE Transactions On Journal Name, Manuscript Id2017
- [2] Mohammad Abu Alsheikh, Shaowei Lin" Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications" 1553-877X (c) 2013 IEEE.
- [3] Shengguo Hu and Yanfeng Liang "MSMOTE: Improving Classification Performance when Training Data is imbalanced" 978-0-7695-3881-5/09 \$26.00 © 2009 IEEE DOI 10.1109/WCSE.2009.137
- [4] L.Dhanaball, Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification.Algorithms".DOI10.17148/IJARCCE.2015.4696
- [5] K. Bispo, N. Rosa, and P. Cunha, "SITRUS: Semantic Infrastructure for Wireless Sensor Networks," Sensors, vol. 15, no. 11, p. 27436, 2015
- [6] R. Alshinina and K. Elleithy, "Performance and Challenges of Service-Oriented Architecture for Wireless Sensor Networks," Sensors, vol. 17,no. 3, p. 536, 2017
- [7] J. Al-Jaroodi and A. Al-Dhaheri, "Security issues of service-oriented middleware," International Journal of Computer Science and Network Security, vol. 11, no. 1, pp. PP.153-160, 2011.
- [8] X. Chen, K. Makki, K. Yen, and N. Pissinou, "Sensor network security: a survey," IEEE Communications Surveys & Tutorials, vol. 11, no. 2, 02 June 2009 2009.
- [9] L. M. Ibrahim, D. T. Basheer, and M. S. Mahmood, "A comparison study for intrusion database (Kdd99, Nsl-Kdd) based on self organization map (SOM) artificial neural network," Journal of Engineering Science and Technology, vol. 8, no. 1, pp. 107-119, 2013.