International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 3, May-Jun-2019, ISSN (Online): 2395-566X

LCHA-SMOTE Machine Learning Approach on Data Imbalance for wireless Sensor Network

Reshma Lakra **Ruchi Dronawat**
reshulakra007@gmail.com dron.ruchi@gmail.com
Department of Computer Science
Sagar Institute of Science and Technology
Bhopal, M.P, 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. After apply our Low Cost High Accurate Synthetic Minority Over-sampling Technique for imbalance to balance data and a random forest algorithm a classification approach for balanced data we achieved a good accuracy rate percentage.**

***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]

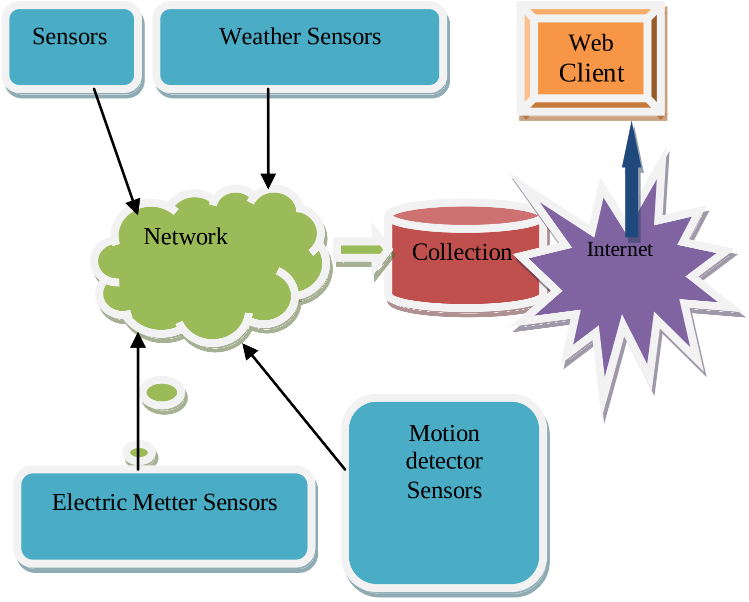


Fig.1 Wireless Sensor Data Collection.

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. 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.

2.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:

3. 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.

4.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.

5. 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.

6. Synthetic minority oversampling technique (SMOTE)- To avoid the over-fitting problem, Chawla et al. (2002) propose the Synthetic Minority Over-sampling Technique (SMOTE). This method is considered a state-of-art technique and works well in various applications. This method generates synthetic data based on the feature space similarities between existing minority instances. In order to create a synthetic instance, it finds the K-nearest neighbors of each minority instance, randomly selects one of them, and then calculate linear interpolations to produce a new minority instance in the neighborhood.

7. 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.

8.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.

9.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. RELATED WORK

Machine learning Approach for Data transmission from the WSN to the end user then becomes much more secure and accurate compared to conventional techniques. [1].A Visualization of the Q-learning method for solutions for their specific application challenges. [2].Modified approach (MSMOTE) for learning from imbalanced data sets, based on the SMOTE algorithm. [3].Modified approach (MSMOTE) for learning from imbalanced data sets, based on the SMOTE algorithm. [4].Study has exposed many facts about the bonding between the protocols and network attacks. [4]. For the learning stage, the proposed method uses the following three steps to learn a class-imbalance oriented model: (1) partitioning the majority class into several clusters using data partition methods such as K-Means, (2) constructing a novel training set using SMOTE on each data set obtained by merging each cluster with the minority class, and (3) learning a classification model on each training set using convention classification learning methods including decision tree, SVM and neural network[3].

III.PROBLEM STATEMENTS

By Above literature we analyzed wireless sensor network generated a huge amount of imbalance data. Wireless sensor data like weather, temperature, motion detector, mobile data extra. Also analyzed different machine learning approaches for improve for imbalanced data, like SMOTE and MSMOTE.

- SMOTE disadvantage with under sampling is that it discards potentially useful data.
- A second reason for using sampling is that many highly skewed data sets are enormous and the size of the training set must be reduced in order for learning to be feasible.

1. Proposed system

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 Low Cast High Accurate Synthetic Minority Over-sampling Technique LCHA-SMOTE.
- Finally we will compare with different algorithm for imbalanced data. And also check their accuracy.

2. Algorithm

Algorithm in pseudo-code

LCHA- SMOTE(D_{maj} , D_{min}, N, ki, A_{trunc}, B_{def} , C_{sel})

3.Symbol Information

- D_{maj} : Set of majority class samples.
- D_{min}: Set of minority class samples.
- N: Total number of synthetic samples to be generated.
- ki: Number of nearest neighbors.
- A_{trunc}: Truncation factor with $-1 \leq A_{trunc} \leq 1$.
- B_{def} : Deformation factor with $0 \leq B_{def} \leq 1$.
- C_{sel}: Neighbor selection strategy with C_{sel} ∈ n {minority, majority, combined}.

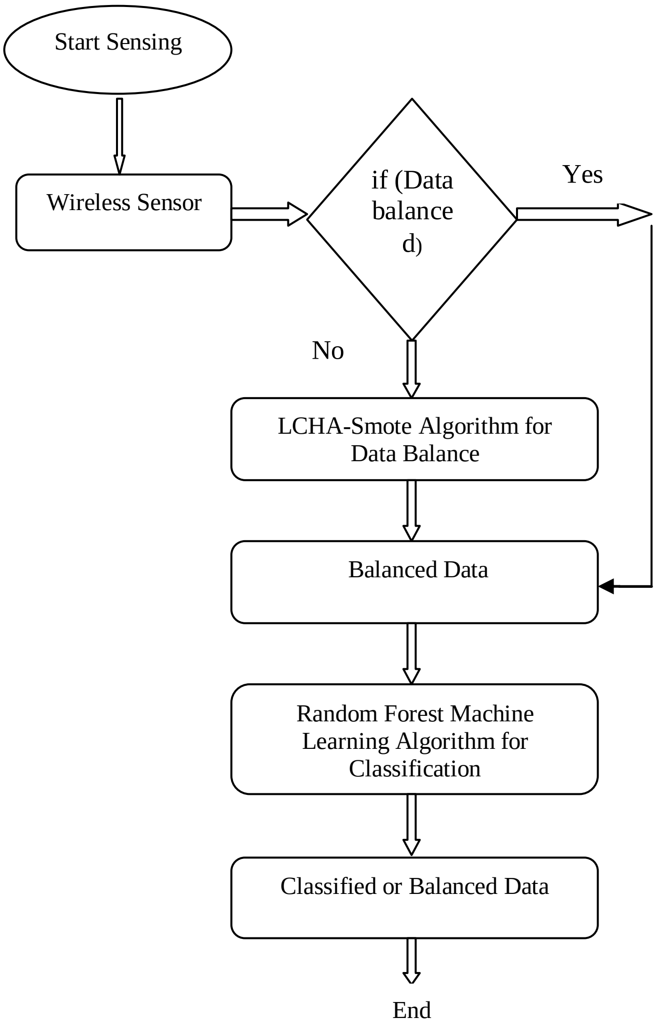

Start Sensing
Wireless Sensor
if (Data balance d)
Yes
No
LCHA-Smote Algorithm for Data Balance
Balanced Data
Random Forest Machine Learning Algorithm for Classification
Classified or Balanced Data
End

Fig. 2 Flowchart of Algorithm.

4. Procedure begin

- Shuffle D_{min}
- Repeat until N minority instances are selected, each multiple times if necessary, in the order that appear in D_{min}:
- Let xcenter ∈ Dmin the selected minority class instance of F components.
- If Csel = minority:
- Calculate the ki nearest neighbors of P_{center} from Dmin. Let $P_{surface}$ one of them, randomly selected.
- Calculate a radius from the relation R ← d(P_{center}, $P_{surface}$).
- Else if Csel = majority:
- Calculate $P_{surface}$ as the nearest neighbor of P_{center} from D_{maj}.
- Calculate a radius from the relation R ← d(P_{center}, $P_{surface}$).
- Else if Csel = combined:
- Calculate the k nearest neighbors of P_{center} from D_{min}. Let xmin one of them, randomly selected.

- Calculate the euclidean distance dmin ← d(P_{center}, P_{min}).
- Calculate P_{maj} as the nearest neighbor of P_{center} from D_{maj}.
- Calculate the euclidean distance dmaj ← d(P_{center}, P_{maj}).
- Calculate a radius from the relation R ← d(P_{center}, $P_{surface}$) where $P_{surface}$ ← $\text{argmin}_{Pmin,Pmaj}$ (dmin, dmaj).
- Generate a synthetic sample P_{gen} ← hyperball(center=0,radius=1)().
- Transform1 the synthetic sample by P_{gen} ← truncate(xgen, xcenter, xsurface, αtrunc).
- Transform2 the synthetic sample by P_{gen} ← deform(P_{gen}, P_{center}, $P_{surface}$, Bdef).
- Transform3 the synthetic sample by P_{gen} ← translate(P_{gen}, P_{center}, R).
- Add the sample P_{gen} to the set of generated samples D_{gen}.

Output

The set Dgen of generated synthetic examples.

IV.RESULT ANALYSIS

For implementation we used python IDE and NSL-KDD dataset is used.

Table 1 Overview of Nsl-Kdd Dataset

Dataset	Normal
NSL-KDD Train	67343
NSL-KDD Test	9711

1. Confusion matrix with Accuracy

The confusion matrix is applied to evaluate the performance and effectiveness of the proposed G network and the original dataset, NSL-KDD. For this purpose, the Accuracy Rate (AR), False Positive Rate (FPR), True Positive Rate (TPR/Recall), and F-measure (F1) are applied and computed by equations 1, 2, 3, and 4 respectively. In the equations, TP, TN, FP, and FN denote number of true positive, true negative, false positive, and false negative cases respectively.

Figure shows the confusion matrix between the testing target output and the predicted output for the generated data from the G. The G network achieved a better binary distribution while improving the accuracy and decreasing the classification error. In addition, TN and FP are two main scriteria for evaluating the performance of the G network data compared to the NSL-KDD dataset results.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

$$\textbf{FPR} = \frac{FP}{FP+TN}$$

$$\textbf{TPR} = \frac{TP}{TP+FN}$$

Error Rate = 1-AR

Table 5.2 CONFUSION MATRIX ORIGINAL DATASET (NSL-KDD)

	TP 9610 42.6%	FN 1044 4.6%	0
ACTUAL CLASS	FP 3223 14.3%	TN 8667 38.4	1
	0	1	

PREDICTED CLASS

Table 5.3 CONFUSION MATRIX OF LCHA-SMOTE

	TP 11480 48.4%	FN 348 3.2%	0
ACTUAL CLASS	FP 1580 7.6%	TN 9140 38.3	1
	0	1	

PREDICTED CLASS

TABLE 5.4 THE COMPARISONS OF FALSE POSITIVE AND RATE

Method	FP	FPR
Original Data [32]	14.3%	0.27
Artificial Neural Network (ANN) [39]	17.4%	0.31
Existent	10.9%	0.21
Proposed Model	7.6%	0.14

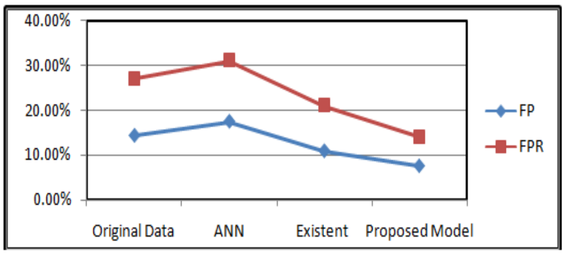
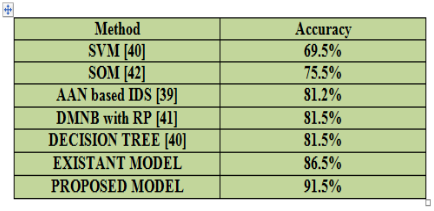


Figure 6.1 THE COMPARISONS OF FALSE POSITIVE AND RATE

We compared the performance of our approach alongside existing methods that use the NSL-KDD dataset. As shown in Table 5.4 and 5.5, the proposed model achieves significantly better accuracy with a lower error rate. The performance of ML techniques optimized accuracy over the NSL-KDD dataset. For example, the accuracy of support vector machine (SVM) and decision tree are much lower compared to other ML techniques [16]. In [17], the authors introduced Discriminative Multinomial parameter learning using Naïve Bayes (DMNB) with a supervised filter called Random Projection at the second level.

The authors achieved 81.47% accuracy in their system. In [16], the authors implemented self-organizing map (SOM) with a very low accuracy rate. The ANN [17] reported that the accuracy of their model was similar to other ML techniques. Our Proposed model Low Cast High Accurate the accuracy rate is 91.5%. It is batter then above algorithm. In this algorithm we used LCHA-SMOTE algorithm for data balance and after balance data we used a classification approach for classified wireless sensor data in more accurate way and we achieved a good accuracy rate.

TABLE 5.5 THE COMPARISONS OF ACCURACY RATE



Method	Accuracy
SVM [40]	69.5%
SOM [42]	75.5%
AAN based IDS [39]	81.2%
DMNB with RP [41]	81.5%
DECISION TREE [40]	81.5%
EXISTANT MODEL	86.5%
PROPOSED MODEL	91.5%

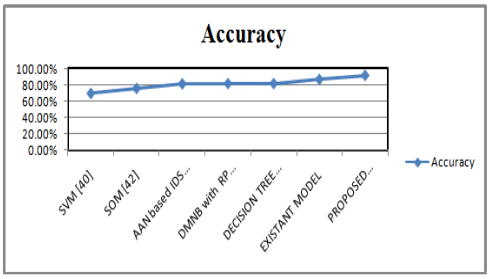


Figure 6.2:-THE COMPARISONS OF ACCURACY RATE

V.CONCLUSION

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 paper systematic study for define which gives the right direction for research in class imbalance problem.

REFERENCES

[1]. Remah Alshinina and Khaled Elleithy “A Highly Accurate Machine Learning Approach for Developing Wireless Sensor Network Middleware”. WTS 2018 1570433702 IEEE.

[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.Dhanabal1, 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 Infrastructurefor 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]. D. Jayalatchumy, P.Thambidurai, “Web Mining Research Issues and Future Directions -A Survey”, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 14, Issue 3, e-ISSN: 2278-0661, p- ISSN: 2278-8727, PP 20-27,(Sep. - Oct. 2013).

[8]. Jay Prakash, RakeshKumar “Web Crawling through Shark-Search using PageRank”,International Conference on Intelligent Computing, Communication & Convergence (ICCC 2015),PP 210-216,(2015).

[9]. Rashmi Rani and Vinod Jain, “Weighted Page Rank using the Rank Improvement”,International Journal of Scientific and Research Publications, Volume 3, Issue 7, 1, ISSN 2250-3153 (7th july,2012).

[10]. Supreetkaur, UsvirKaur “An Optimizing Technique for Weighted Page Rank with K-Means Clustering” IJARCSSE, (2013).

[11]. R. O. Duda, P. E. Hart, D. G. Stork, Pattern classification, John Wiley & Sons, 2012.

[12]. R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of eugenics 7 (2) (1936) 179–188.

[13]. H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, Neural Networks, IEEE Transactions on 17 (1)(2006) 157–165.

[14]. X. Chang, F. Nie, Y. Yang, H. Huang, A convex formulation for semi-supervised multi-label feature selection., in: AAAI, 2014, pp. 1171–1177.

[15]. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 2009, pp. 1-6.

[16]. M. Panda, A. Abraham, and M. R. Patra, "Discriminative multinomial Naïve Bayes for network intrusion detection," in Sixth International Conference on Information Assurance and Security, Atlanta, GA USA, 2010, pp. 5-10: IEEE.