Indonesian Journal of Electrical Engineering and Computer Science
Vol. 20, No. 3, December 2020, pp. 1435~1443
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v20.i3.pp1435-1443

Outlier detection in WSN by entropy based machine learning approach

Manmohan Singh Yadav, Shish Ahamad
Department of Computer Science & Engineering Integral University, India

Article Info

Article history:

Received Mar 1, 2020
Revised Jun 3, 2020
Accepted Jun 28, 2020

Keywords:

Ensemble
Entropy
K-nearest neighbor
Outlier
Sensor
SVM

ABSTRACT

Environmental disasters like flooding, earthquake, epidemics etc. cause's significant catastrophic effects on population of all over the world. Wireless sensor network (WSN) based techniques have become significantly popular in susceptibility modelling of such challenging disaster due to their greater strength and efficiency in the prediction of such threats occurring enormously day by day. This paper demonstrates the multiple machine learning-based approach to predict outlier in sensor data records with the use of bagging, boosting, random subspace, SVM and KNN based frameworks for outlier prediction using a Wireless sensor network data records. First of all the algorithm follows the pre processing of the database taken from records of 14 sensor motes with presence of outlier due to intrusion. Subsequently the segmented database is created from sensor pairs. Finally, the data entropy is calculated and used as a feature to determine the presence of outlier used different approach. Results show that the KNN model has the highest prediction capability for outlier assessment.

Corresponding Author:

Manmohan Singh Yadav,
Department of Computer Science & Engineering,
Integral University, Lucknow, U.P., India.
E-mail: man_mohan_100@rediffmail.com

1. INTRODUCTION

Outlier due to intrusion is an extreme problem in protection and a top hassle of safety breach, because a single example of Outlier can borrow or delete information from computer and network system in a few seconds. Outlier can also harm machine hardware. Furthermore, Outlier can cause large losses leading to data inferiority in cyber warfare. Therefore, Outlier detection is important and its prevention is necessary. Different Outlier detection techniques are to be had, but their accuracy stays as difficulty; accuracy relies upon on detection and false alarm rate. The problem on accuracy wishes to be addressed to lessen the false alarms rate and to boom the detection rate. This belief changed into the impetus for this research work. Thus, support vector machine (SVM) [1], random forest (RF), KNN etc. are implemented in this work; those techniques were verified powerful in their functionality to deal with the category trouble. Outlier detection mechanisms are proven on a standard dataset. This work used the NSLWSN real time dataset, that's a stepped forward shape and is considered a benchmark within the assessment of outlier detection techniques [2, 3]. Today everything is based on data era of information technology. Day by day the data is getting costlier than the gold but any data has value if it is free of errors otherwise the analysis which is drawn from these data for the purpose of commercial and social welfare may result in false predictions. Hence the modern application like IOT, cloud, WSN, prediction ,estimation etc in defence biomedical, space technology are handicapped without the support of intermediate methods that find and removes the error, noise, disturbance, anomaly etc present in the data. That is why several predictions related to weather monitoring, stock market rates, launching of spacecrafts, diagnosis of disease goes through failures because of the small amount of distortions present in the data records. This paper focus on the problem related to challenges that are being faced today due to outliers in the data records.

Many studies and works are reporting contribution that shows interesting enhancements in terms of the classification performance accuracy. For instance the applications related to tree ensemble process casting to transformed susceptible classifiers into the better robust ones. In this approach, each tree is grown randomly by some training set [4]. Freund brought a boosting set of rules named as Ada boost, which he described as [5]: "deterministic rules sets". In [6] randomness changed into again used to develop the trees, the split became described at every node with the aid of attempting to find the high-quality random choice of capabilities inside the training set. How brought the random subspace, wherein he randomly selects a subset of vectors of features to grow each tree [7]. Dietrich delivered the random split selection in which at each node; a cut up is randomly decided on among great splits. K For these techniques, and like bagging, a random vector sampled to develop a tree is completely independent from the previous vectors, but is generated with the equal distribution. Random cut up selection and introducing random noise into the out puts both gave better outcomes than bagging. Nevertheless, the algorithms implementing approaches of re-weighting the training set, including Ada boost [8], outperform those two techniques. Therefore, Breiman [9] combined the strengths of the methods targeted above into the random forest algorithm. In this method, people are randomly selected from the training set with replacement. At each node, a cut up is chosen by using decreasing the dispersion generated with the aid of the preceding step and therefore decreasing the error fee [10]. The rest of the paper is prepared as specific under. The related work is presented in Section II. The proposed model of Outlier detection to which different machine learning techniques are applied is described in Section III. The implementation and results are discussed in Section IV. The paper is concluded in Section V, which provides a summary and directions for future work.

2. RELATED WORK

Outlier detection is very important part for safety tools having advanced security uses, outlier detection process schemes, outlier prevention systems, and defence applications. Many strategies are used, but their overall performance still possesses multiple limitations. Outlier detection relies upon accuracy and it can be enhanced to decrease the false alarms detection rate. To improve overall performance, multilayer tree based approach, support vector machine (SVM), and other strategies applied in this current work. Such strategies have bounded applications and aren't versatile for large data sets in big system and network data. The outlier detection is applied in studying large traffic data networks hence an efficient category approach is necessary to overcome the challenging issue. This problem is taken into consideration on this paper. Well-known device learning techniques, particularly, SVM, random forest, and extreme learning machine (ELM) are carried out. These strategies are famous because of their functionality in category. The know-how discovery and data mining statistics set is used. The results indicate that present approach outperforms other processes [11]. In a paper random forests for operating devices diagnostics within the presence of a variable wide variety of functions is demonstrated. Wireless sensor network are very helpful to clear many hassle but more subjected to flaws. It is observed that diagnostics at the sink level is important to quantity and to furnish capabilities and that some politics like scheduling or facts aggregation may be developed throughout the network. This paper exposed that random forests are relevant on this context, because of their flexibilityand robustness [12].

Another research work offers a singular hybrid prediction technique, specifically, self-tuning least squares support vector machine (STLSSVM). It is a hybrid technique that makes use of LS-SVM as a supervised-gaining knowledge of-based totally predictor to build a correct input-output courting of the dataset. Prediction accuracy of the ST-LSSVM is compared to other device getting to know methods, particularly, LS-SVM and BPNN in terms of coefficient of correlation (R), mean absolute error (MAE), and root mean square errors (RMSE). Comparisons showed that the ST-LSSVM accomplished higher than LS-SVM, BPNN, and NN in terms of R, RMSE, and MAE [13]. ML class algorithms inclusive of K-nearest neighbour, tree, SVM and Naive Bayesian are performed using real time dataset. Bagging, boosting, and random forest are applied to generate prediction model. The accuracy is then tabulated. Boosting ensemble has the high accuracy in this article [14].

A paper was proposed on an ensemble framework to diagnose disease by means of optimally employing a couple of classifiers primarily based on bagging and random subspace techniques. The proposed framework combines seven of the maximum appropriate and heterogeneous data mining strategies, each with a separate set of suitable functions. The framework is designed as it should be by using deciding on, for each sub-dataset, the maximum suitable feature set and the most correct classifie [15]. The Support vector Machine (SVM) is a famous type approach. However, beginners who aren't acquainted with SVM often get unsatisfactory consequences since they miss some clean however vast steps. In this guide, we recommend an easy system which usually gives affordable outcomes [16].

In Machine Learning, a records set is imbalanced whilst the magnificence proportions are fairly skewed. Imbalanced records sets arise automatically in lots of utility domains and pose a challenge to traditional classifiers. We recommend a new method to constructing ensembles of classifiers for two-class imbalanced statistics units, known as Random Balance. Each member of the Random Balance ensemble is trained with information sampled from the training set and augmented by using synthetic times received using SMOTE [17]. The novelty within the method is that the proportions of the training for each ensemble member are chosen randomly. The instinct at the back of the approach is that the proposed variety heuristic will make certain that the ensemble carries classifiers that are specialized for exclusive running factors on the ROC area, thereby main to large AUC compared to different ensembles of classifiers. Experiments have been performed to test the Random Balance method by using itself, and additionally in combination with popular ensemble techniques. As a result, we propose a new ensemble advent method known as RB-Boost which combines Random Balance with AdaBoost.M2. This combination involves enforcing random elegance proportions in addition to example re-weighting. Experiments with 86 imbalanced records sets from two widely recognized repositories reveal the gain of the Random Balance approach [18, 19].

3. METHODOLOGY

3.1. Ensemble of classifier

In ensemble method consist of aggregation of machine learning algorithms used to get better predictive result than obtained from available individual learning algorithms. An ensemble is a supervised way gaining knowledge by algorithm [20].

3.1.1. Bagging

Bootstrap aggregating or bagging is making use of the identical learning algorithm to train each learner on a one-of-a-kind set of data. N' subsets of data are drawn randomly with the replacement from the training data N. The N' subset of records are chosen in parallel. Each of the N' subsets used to train a model M. Test data X is applied to every of the M models for predicting Y.

3.1.2. Boosting

Boosting process is similar to bagging but the iterations applied to be sequential, and every time new classifier applied to get better accuracy of prediction of the previous generation classifier.

3.1.3. Random Subspace

Random subspace approach referred to a feature bagging. It possesses an ensemble learning knowledge that attempts to lessen the correlation among estimators by means of associating them on random samples. The random subspace includes functions like "attributes", "predictors", "unbiased variables" sampled randomly with substitute [21]. An ensemble of models is constructed in this method by following rules:

a) Let number of training points are N and the features within training set are D.
b) Say L is number of individual models for constructed ensemble.
c) For every model l, select nl such that (nl < N) represents number of input points.
d) For every model l, create a training dataset such that dl features within D with desired replacement and then start the train of the model.

Finally use these ensemble model as unseen point, integrate all outputs of individual models. Use random subspace ensembles (Subspace) for better accuracy.

3.2. K-nearest neighbours

This algorithm is based on distance-based classifiers scheme. The class label of a new data is equal to the class of the nearest neighbour found using specific distance formulae. Heterogeneous Euclidean-Overlap Metric (HEOM) is applied for distance measure to get the K-nearest neighbours [22]. Here is step described that are used in K-nearest neighbours (KNN) algorithm:

a) Determine parameter K = number of nearest neighb₀urs
b) Calculate the distance between the query-instance and all the training samples
c) Sort the distance and determine nearest neighbours based on the K-th minimum distance
d) Gather the category of the nearest neighbours
e) Use simple majority of the category of nearest neighbours as the prediction value of the Query instance.

3.3. Support vector machine

Support Vector Machines are found to be very beneficial approach in automated classification process. It has less difficult to apply than Neural Networks process learning. It offers a cookbook approach that helps to provide reasonable results. The users do not need to recognize the underlying concept in the back of SVM. It includes setting apart of records into training and testing out sets. The training set samples must include one target value (known as class labels) and numerous attributes (known as features or observed variables). The SVM provide a model (based totally at the training data) which predicts the test data of the test data given best the check data attributes [23,24].

3.3.1. SVM classifier design

a) After getting the entropy and anomaly values SVM function applied on data variables
b) Use the classifier function for training dataset
   SVMStruct = svmtrain(Entrpall',(Anomalyall>0))
   Group = svmclassify(SVMStruct,Entrpall')
c) c) Find the percent of SVM model accuracy
   Percent training accuracy=100-∑ (output predicted-Anamoly value>0)/total length of anomaly value)*100
d) d) Use the classifier function for testing dataset
   Group = svmclassify(SVMStruct,Entrpall');
e) e) Find the percent of SVM model accuracy for testing dataset
   Percent testing accuracy =100-∑ (output predicted-Anomaly value>0)/total length of anomaly value)*100

4. PROPOSED ALGORITHM FOR OUTLIER DETECTION USING DIFFERENT CLASSIFIER

4.1. Load & process the data variables

a) Data sample time is 0.5 sec., segment length taken of 50 samples .There are fourteen sensors hence 14x13=182 sensor pair id are labelled. Each pair id data bears 3127 sample for record time half hour [21].
b) Sensor data Z 182x3127 and motion data M1 x 3127 (which is as outlier [0 or 1]) is imported to algorithm.
c) Different distance formulae are considered named as‘cityblock', 'chebychev', 'mahalanobis', 'minkowski', 'euclidean', 'seuclidean', 'spearman', 'cosine', 'hamming', 'jaccard'.

4.2. Select sensor pair

Pair id of two sensors communicate with each other is selected randomly to make a training data.

4.3. Data segmentation

a) Dataset is broken in segment of 50 samples (total segment are 62).
b) Motion dataset (outlier data) is also segmented.
c) Total segment per sensor pair =floor [(total sample)/ (segment length)] i.e. 62.
d) Entropy is calculated of each segment.

4.4. Entropy evaluation

a) Each segment entropy value is input of Classifier predictor algorithm. The equation used for entropy is based on Shannon entropy formulae.
b) Each segment iteratively used to get its entropy value.
c) Segments entropy saved as variable ‘Entrpall’.
d) Outlier data segment are summed and taken as outlier level.
e) We take random pair id and all total 620 data recorded are generated as training database.

4.5. Classifier model development

a) Classifier is trained by using entropy data values as input variable and outlier level as output.
b) Prediction inaccuracy of training & testing data is calculated. Different prediction models at multiple parameters.
c) Find the percentage of prediction accuracy.
d) Find the cross validation accuracy from model testing stage.

5. RESULTS AND DISCUSSION

5.1. About the dataset: Data collection Method Description

The data used in this work is from an experiment at the University of Michigan. It compose of a Sensor kit known as MICA2.Sensors in this kit are Light, Temperature, RH, Pressure, Acceleration etc. This experiment setup consists of 14 sensor nodes randomly deployed inside and outside a lab room. Sensors communicate by broadcasting and the received signal strength (RSS) is recorded as the voltage measured. 14 x 13 i.e. 182 sensor pairs records of RSS measurements over a 30 minute period is taken at a sample of 0.5 sec to get 3191 time samples. During the experiments volunteer walked into and out of the lab to create outlier patterns in the RSS voltage readouts. A web camera also used to record activity. The original raw data is stored in the matrix 'dataLinear' (of size 182 x 3191) in the file 'dataLinear.mat'. The ground truth is recorded in the vector 'motionCode' (of size 1 x 3191) in the file 'motionCode.mat' Datalinear is sensor RSS (volts), Motion is Anomaly (0 or 1), Z is preprocced data. Total segment is 3127/50 i.e. 62, and pair id is any two sensor communicate each other (let pairid=60). At a particular pairid (let pairid=60) dataset is divided into segment (total segment 62. similarly motion data (outlier data) is also broken in equal segments [25].

5.2. Entropy

Entropy as it relates to machine learning, is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information. Flipping a coin is an example of an action that provides information that is random. For a coin that has no affinity for heads or tails, the outcome of any number of tosses is difficult to predict [26, 27]. This is the essence of entropy.

$$\text{Entropy (p)} = -\sum_{K=1}^{C} p\left(\frac{K}{P}\right) * Log\left(\frac{K}{P}\right)) \quad (1)$$

a) After segmentation we find out the segment entropy which is input of different predictor models.
b) Entropy={ e1, e2, e3,……e62}
c) Anomaly data is also segment wise.
d) Find the summation of segment wise anomaly data.
e) Anomaly segment are summed and saved as anomaly level.

Figure (1) represents the block diagram of the outlier detection schemes. In Figure 1 all the sensors collected signals in voltage form as data for processing. To estimate the probability density function (PDF) of a dataset using data-split technique divide data as segments. To estimate the entropy of the dataset. To use different classifier model (KNN, SVM etc.). To use this metric to detect outlier.

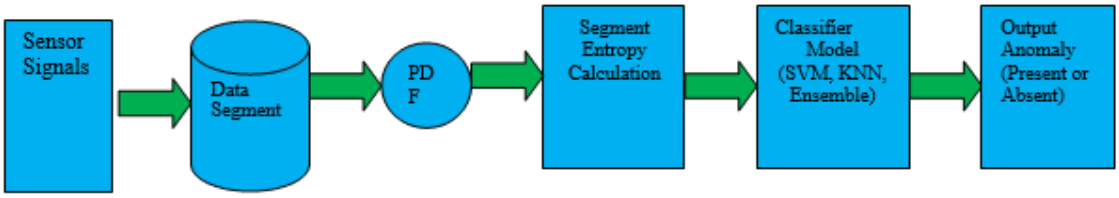

Figure 1. Block diagram representation of anomaly detection model using different classifiers

5.3. Data sets use for training, validation and testing

The original raw data is stored in the matrix 'data Linear' (of size 182 x 3191) in the file dataLinear.mat' .Out of 580762 obtain from different combination of 14 sensor pair id but this valid data is divided 75% data taken for training purpose for develop all the model . Under the training process detection model is develop and error in between actual and calculated output is determined. Validation is performance along with the training using 15% data that is not used under the training purpose. The error in the validation is used to update the model parameters finally if the model gives desired level of occurs then this performance is finally re-evaluated by using 15% remaining testing data.

5.4. Result analysis

Result analysis demonstrates that training gives 86% accuracy with 2 nearest neighbour for seuclidean distance. In testing 78% accuracy is obtained for 3 nearest neighbour with seuclidean distance type. In Figure 2 the receive signal strength in voltage at sensor node is shown at y axis and the x xis

is the time instants in seconds. In Figure 2 when data received from sensor with time 30 minutes and take sample of 5 seconds. It is a plot graph between Received signal strength (RSS) in voltage with respect to time in sec. Figure 3. During the measuring period, students walked into and out of lab at random times, which caused anomaly patterns in the RSSI measurements. Intruder motion (anamoly) representation with respect to time samples (n).Whenever intruder disturbs the sensor waves the intrusion is marked as one otherwise the value is zero. Whenever this figure touches high value the anomaly is embedded in the signal voltage wave form.

In Figure 4 RSS data segments plot for segment length of 50 samples, segment length is equal to any value decided that can cover data variation (let L=50) Total segment= c/L = 3127/50 = 62 ,Pair id= id (identity number)) of any two sensor communicate each other. At a particular pairid (let pairid=60) dataset is divided into segment which is total segment 62 and is stored in another matrix. Similarly motion dataset which is known as anomaly data is also segment into 62 segments. In this figures segment 1 to segment 10 RSS waveform is shown with respect to time.

In Figure 5.a we have shown the calculated value of entropy for each segment of a specific sensor pairid 167, in Figure 5.b for same pairs Anamoly level with respect to all segments for sensor pairid 167.Similar to this figure the entropy value of all the segments of all sensor pairs is calculated. This entropy of segment is taken as input and number of anomaly occurred during that segment is the anomaly level under that segment and shown in Figure 5.b.

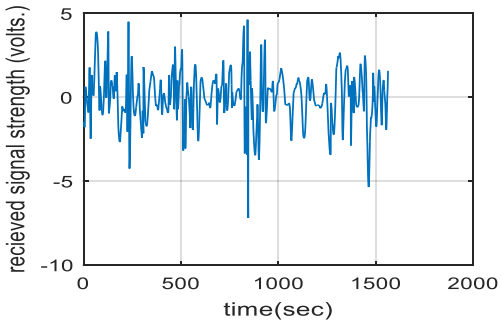

Figure 2. Received signal strength voltage with respect to time (sec)

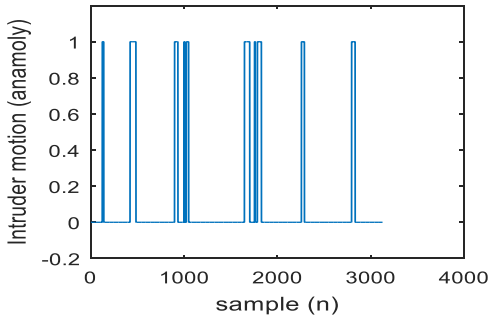

Figure 3. Intruder motion (anamoly) representation with respect to time samples (n)

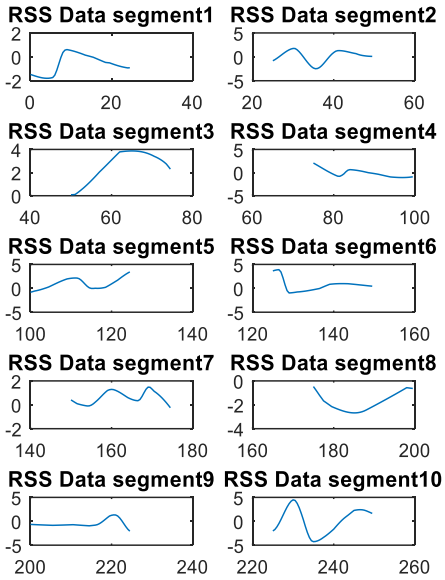

Figure 4. RSS data segments plot for segment length of 50 samples

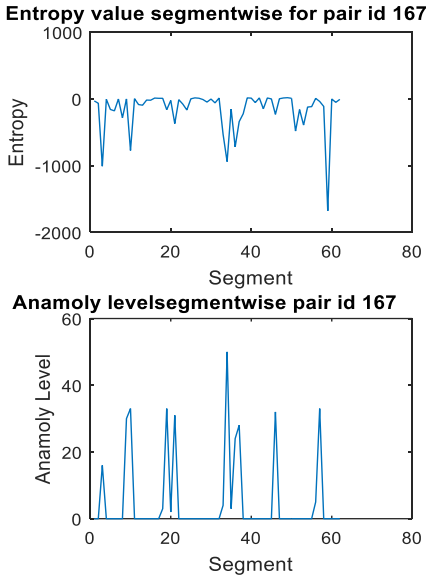

Figure 5. a) Entropy value calculted for segment of sensor pairid 167, b)Anamoly level with respect to all segment for anonymous sensor pairid 167

If data size is increase then accuracy is high with hamming distance & no. of neighbor is 5. The training dataset accuracy=94.34 % and Testing dataset accuracy=89.85%. After the result analysis in training analysis, 76.22% accuracy for SVM model and KNN model 86% accuracy, we conclude that KNN is better than SVM model. In testing analysis, 75.19%, for SVM model and KNN model 78% accuracy, we conclude that KNN is better than SVM model.

Table 1 shows the summary of the results for the KNN scheme at different distance type. Every row of table 1a shows the percent detection accuracy for training and testing process on applying the model development at different attempts. First column shows the best number of neighbour at which highest accuracy in training and testing is observed at specific distance formulae type.

Table 2 shows the similar results for KNN but on considering the large size of database. In thirds case training and testing accuracy is improved but the model development time and memory requirement is heavily increased. Table 3 gives the details of percentage training accuracy for the application of SVM. It is applied for different number of attempts to validate the consistency of performance. It is observed that the training accuracy obtained to be 75 to 78% and the testing accuracy is 73% to 76%.

Table 4 is the comparison table for all the 5 methods applied for developing the anomaly detection scheme. These are KNN, SVM, Boosting, Bagging and Random Subspace classifier based detection methods developed for segment entropy as input parameter. High accuracy is observed for the KNN.

Table 1. KNN result on Training and testing data analysis

(a) Analysis of Training dataset			(b) Analysis of Testing dataset		
No of neighbor	Distance type	Accuracy (%)	No of neighbor	Distance type	Accuracy (%)
2	chebychev	83.39	9	spearman	76.89
2	euclidean	82.26	2	spearman	76.32
2	euclidean	84.68	3	seuclidean	77.74
2	seuclidean	85.65	2	spearman	76.72

Table 2. Observation on Training and Testing process with large dataset

Analysis Training dataset				Analysis on Testing dataset			
Attempt	Nearest neighbor	Distance type	Accuracy (%)	No of attempt	No of neighbor	Distance type	Accuracy (%)
1	6	jaccard	94.34	1	5	Hamming/jaccard	89.8/85.1
2	6	jaccard	94.34	2	5	Hamming/jaccard	89.8/85.1
3	6	jaccard	94.54	3	5	Hamming/jaccard	89.8/85.1
4	6	jaccard	94.34	4	5	Hamming/jaccard	89.8/85.1

Table 3. Observation of Training and Testing process using SVM

No of attempt	%cent of Training dataset	%cent of Testing dataset
1	75	76.77
2	75.6	75.5
3	76	73
4	76	75.6
5	78.5	75.1

Table 4. To compare the entire classifier algorithm results

Classifier	% accuracy of Training dataset	% accuracy of Testing dataset
K- Nearest Neighbour	85.65	77.74
K- Nearest Neighbour with large dataset	94.34	89.85
Support Vector Machines	76.22	75.19
Boosted	77.8	72
Bagged	70.5	56
Subspace	78.5	72

Figure 6 shows the results obtained after applying different anomaly detection scheme using the K-NN, SVM, Boosting, Bagging and subspace classifier. The vertical axis shows the percentage accuracy of detection scheme based on entropy feature. The numeral values are also attached for training and testing scheme. It has been observed in the graph that the KNN scheme gives highest detection accuracy as compared to other methods. It is about 85% and as the training data size is increased it can go up to 94%.

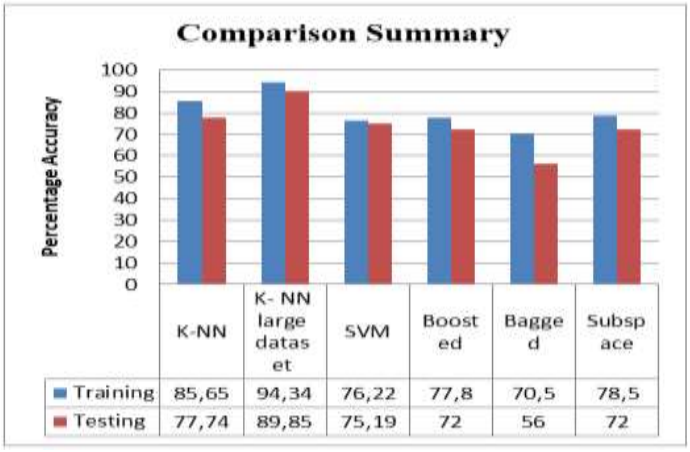

Figure 6.Comparisons of all the classifier algorithm results

6. CONCLUSION

A novel technique is demonstrated on wireless sensor dataset different prediction model learning methods: Based on the idea of varying the parameters of the model attribute classes and applied it to design a new prediction method. Despite the simplicity of development approach multiple methods have validated when compared with each other in terms of the prediction accuracy of outlier in sensor data including segment entropy as input feature specifically devised for data classification.

Study the performance of Boosting, Bagging Tree, KNN, SVM etc in use of wireless sensor network data with outlier characteristics due to intrusion. It has been proven to have a strong influence on performance of classification. In testing analysis, 75.19% accuracy of SVM model and 78% accuracy of KNN model, we conclude that KNN is better than SVM model. The problems which are faced in detecting outlier is due to overlapping, noisy signals [28], small disjoints or borderline values. These problems are minimized with advanced pre-processing techniques. The entropy feature based outlier detection strategy is successfully implemented on data with small disjoints, using different self learning techniques specially KNN and SVM. The ideas in this article may further be extended to multiple-class unbalanced problems with other classifiers. These approaches are presently developed on systematic data records but in future same algorithm may be used to remove the outliers in real time situations with proper hardware setup for the challenging environmental conditions for weather monitoring purpose.

ACKNOWLEDGEMENTS

The present article work was supported by Faculty of Doctoral Studies and Research (DSR) and Department of Computer Science & Engineering, Integral University, Lucknow under the processing of research manuscript communication number (IU/R & D/2019-MCN000773).

REFERENCES

[1] H.Wang, J. Gu, and S.Wang, "An effective intrusion detection framework based on SVM with feature augmentation", *Knowl.-Based Syst.*, vol. 136, pp. 130-139, doi: 10.1016/j.knosys.2017.09.014, Nov. 2017.
[2] S. Teng, N.Wu, H. Zhu, L. Teng, andW. Zhang, "SVM-DT-based adaptive and collaborative intrusion detection", *IEEE/CAA J. Automatica Sinica*,vol. 5, no. 1, pp. 108-118,doi: 10.1109/JAS.2017.7510730, Jan. 2018.
[3] Iftikhar Amad , Mohd. Basheri, Mohd. Javed Iqbal and Aneel Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", *IEEE access*, vol. 6, 2018.
[4] Yali Amit and Donald Geman, "Shape quantization and recognition with randomized trees". *Neural Computation*, vol. 9, pp. 1545-1588, 1997.
[5] Nasir Saeed, Tareq Y. Al-Naffouri, Mohamed-Slim Alouini, "Outlier Detection and Optimal Anchor Placement for 3D Underwater Optical Wireless Sensor Networks Localization", *IEEE*, pp. 0090-6778, 2018.
[6] Leo Breiman, "Using adaptive bagging to debias regressions. Technical report", *Statics Department UCB*, 1999.
[7] Victor Garcia-Font, Carles Garrigues and Helena Rifà-Pous, "A Comparative Study of Anomaly Detection Techniques for Smart City Wireless Sensor Networks",www.mdpi.com/journal/sensors,*Sensors*, vol. 16, p. 868, 2016.
[8] Thomas G, Dietterich. "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization". *Machine Learning*, vol. 40, pp. 139-157, 2000.

[9] Y. Freund and R. Schapire. "Experiments with a new boosting algorithm". In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148-156, 1996.
[10] Tin Kam Ho. "The random subspace method for constructing decision forests". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.
[11] W. ElGhazel, C. Guyeux, A. Farhat, M. Hakem, K. Medjaher, N. Zerhouni, and J.M. Bahi, "Random Forests for Industrial Device Functioning Diagnostics Using Wireless Sensor Networks",arXiv:1706.08106v1 [cs.AI], 2017.
[12] Doddy,Prayogo and Yudas , "Optimizing the Prediction Accuracy of Friction Capacity of Driven Piles in Cohesive Soil Using a Novel Self-Tuning Least Squares Support Vector Machine",*HindawiAdvances in Civil Engineering*,, Article ID 6490169, 2018.
[13] B. Emil Richard Singh and E. Sivasankar, "Enhancing Prediction Accuracy of Default of Credit Using Ensemble Techniques",First International Conference on Artificial Intelligenceand Cognitive Computing, *Advances in Intelligent Systems and Computing* 815,https://doi.org/10.1007/978-981-13-1580-0_41, Springer, 2019.
[14] Shaker El-Sappagh , Mohammed Elmogy, Farman Ali, Tamer ABUHMED, S. M. Riazul Islam and Kyung-Sup Kwak, "A Comprehensive Medical Decision–Support Framework Based on a Heterogeneous Ensemble Classifier for Diabetes Prediction" ,*Electronics 2019*, vol. 8, p. 635; doi:10.3390/electronics8060635,2019.
[15] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classi_cation", http://www.csie.ntu.edu.tw/~cjlin ,April 15, 2010.
[16] Jose F. Diez-Pastor , Juan J. Rodriguez,Cesar Garcia-Osorio , Ludmila I. Kuncheva, "Random Balance: Ensembles of variable priors classifiers for imbalanced Data", journal homepage: www.elsevier.com/locate/knosys, 7 May 2015.
[17] J. Stefanowski, "Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data, in: Emerging Paradigms in Machine Learning", *Springer*, pp. 277-306, 2013.
[18] C.E. Brodley, M.A. Friedl, "Identifying mislabeled training data", *J. Artif. Intell. Res*. Vol. 11, pp. 131-167, 1999.
[19] G.M. Weiss, "The impact of small disjuncts on classifier learning, in: Data Mining", *Springer*, pp. 193-226, 2010.
[20] K. Napierała, J. Stefanowski, S. Wilk, "Learning from imbalanced data in presence of noisy and borderline examples, in: Rough Sets and Current Trends in Computing", *Springer*, pp. 158-167, 2010.
[21] T. Jo, N. Japkowicz, "Class imbalances versus small disjuncts", *ACM SIGKDD Explor. Newslett*. vol. 6, no. 1, pp. 40-49,2004.
[22] D. Wilson, "Asymptotic properties of nearest neighbor rules using edited data", *IEEE Trans. Syst. Man Cybern*. vol. 2, no. 3, pp. 408-421,2016.
[23] K. Gowda, G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighbourhood (corresp.)", *IEEE Trans. Inform. Theory*,vol. 25, no. 4, pp. 488-490,1979.
[24] J. Stefanowski, S. Wilk, "Selective pre-processing of imbalanced data for improving classification performance", in: *Data Warehousing and Knowledge Discovery, Springer*, pp. 283-292, 2008.
[25] Manmohan singh yadav, Shish Ahmad, "Outlier detection in Wireless sensor networks data by entropy based K-NN Predictor",*International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN: 2278-3075, vol. 8, no. 12, October 2019.
[26] Dr. Aaisha Makkar, Dr. Neeraj Kumar, Prof. Ahmed Ghoneim, "An Efficient Spam Detection Technique for IoT Devices using Machine Learning" *IEEE Transactions on Industrial Informatics*, DOI 10.1109/TII.2020.2968927,2020.
[27] Nazia Tabassum, Tanvir Ahmed, "A theoretical study on classifier ensemble methods and its applications", *International Conference on Computing for Sustainable Global Development,IEEE*, 978-9-3805-4421-2/16, 2016.
[28] Garcia-Font, V., Garrigues, C., and Rif-Pous, H. "A Comparative study of anomaly detection techniques for smart city wireless sensor networks", *Sensors*, vol. 16, no. 6, p. 868, 2016.