International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

RF-EKHO: Random Forest with Enhanced Krill Herd Optimization Algorithm for proficient Detection of Outliers in Data with High-Dimensions

M Rao Batchanaboyina , Nagaraju Devarakonda

Abstract*: The detection of outliers is a challenging issue in the case of data with high dimensions. It is extensively used in distinct fields of study like social networks, knowledge discovery and statistics. To maintain the network privacy and security in social networks identify structural abnormalities in a constructive way, which are different from the typical behavior of the social network. In this paper, we propose a hybrid model to discover outliers in social networks utilizing Random Forest (RF) and Enhanced Krill Herd Optimization (EKHO) algorithm. The RF is used to enhance the execution and exactness of general procedure and it is a productive classification strategy. The leaves per tree and the trees per the forest are the two parameters of RF. Experimental results shows the efficiency and success of proposed method in terms of accuracy, detection rate, and computational time.

***Index Terms*: Outlier detection, social networks, Random Forest (RF), Enhanced Krill Herd Optimization (EKHO) algorithm.**

I. INTRODUCTION

Outlier detection is an issue known as discovering patterns in information that don't comply with anticipated conduct. The medicine, public health, fraud detection, sports statistics, error detection of measurements, etc, are the major applications of the outlier detection [1]. To detect the outliers there are numerous ways. However, our work is about the detection of outliers with higher dimensional data [2]. The majority of the ongoing works which are intended to discovering outliers formulate implicit presumptions of relatively low dimensionality of the data [3, 4]. Subsequently, for data with high dimensions, the perception of finding important outliers turns out to be significantly progressively complex and non-clear [5].

Generally, detection of outlier is called as intrusion detection, novelty detection [6], and anomaly detection, in fact these procedures have scores of commonalities, and they all attempt to detect unique, distant observations [7]. In modern times the LOF (local outlier factor) algorithm effectively employed for uncovering of outliers [8]. The LOF detects the local outliers of a dataset by using density in which it assigns level of outlierness, called the LOF, to each observation [9]. In LOF, observations with a lesser density when compared to their neighboring points are identified as outliers [10]. Outliers overcome much of the real data; most of research has focused on developing vigorous PCA algorithms that are not overly affected via the occurrence of outliers [11].

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
M Rao Batchanaboyina, Research Scholar, Department of Computer science and Engineering,, Acharya Nagarjuna University, Guntur, India.
Nagaraju Devarakonda, Infirmation Technolgy, Lakireddy Bali Reddy College of Engineering, Mylavaram, India.

II. RELATED WORK: A BRIEF REVIEW

Mohamed Bouguessa has introduced a technique for detection of outliers [15]. In their work an outliers can be detected by utilizing numerical space, categorical space and mixed-attribute space techniques. At last he finished up his work by giving the result of dealing with outliers in single-type feature data with no feature alteration. Xiaowu Deng Et Al. [16] were cooperated for the detection of outliers. They utilized S-SVDD algorithm and R-SVDD algorithm for their work. Use of minimum covariance determinant (MCD) estimator for the detection of outlier gives the consequence of presenting regularized MCD and furthermore setting the regularization parameters to detect the outliers was exhibited by Virgile Fritsch Et Al. [17]. Amid the time of 2017, Huawen Liu Et Al. [18] have introduced a low-rank approximation and local projection-based outlier detection strategies to detect the outlier with high dimensional data. Additionally they present another procedure that was utilized for the abuse of local neighborhood information of an observation to determine whether it was an outlier or not. Randomized methodology for the independent outlier show and randomized robust PCA for the column-sparse outlier demonstrate gives the aftereffect of provably retrieve the right subspace with computational and test complexity which relies upon the size of data dealt with the coherency parameters. It was cleared the time of 2016 by the creators Mostafa Rahmani and George k. Atia [19]. Shu Wu And Shengrui Wang [20] cooperated for research the arrangements of outlier detection. They introduced, information theory based step-by-step (ITB-SS) and single pass (ITB-SP) techniques. This trial indicates both of the algorithm utilized here can manage data sets by an open number of objects and features. The creators [21] utilized hybrid evolutionary technique for the detection of outlier.

Their investigation gives the after effect of exhibited technique to discover the outliers by observing the density distributions of projections from the data. Kim et al. [22] present two algorithms for their work on outlier detection in the time of 2011. After fulfillment of their work the outcome will be in a positive way i.e., the exhibited algorithms were utilized for the reduction of LOF calculation and furthermore it gives predictable and acceptable approximation errors.

III. PROPOSED METHODOLOGY

In this section, we propose a hybrid model which detects outliers in social networks using Random Forest and Enhanced Krill Herd Optimization algorithm. The novel and ensemble machine learning procedure is RF. However, when contrasted the RF shows a great deal of focal points and that of other modeling approach inside the classification. The RF can deal with both discrete and continuous variables which is the fundamental favorable circumstances. Moreover, RF run productively and quickly, and does not over fit as a classifier when taking care of expansive social datasets. The two hyper-parameters of RF are the leaves per tree (at each node splits in the subset) and the trees in the forest. In order to guarantee precise outliers detection in social networks, an optimal quantity of leaves per tree and quantity of trees are selected. The EKHO algorithm can be used to optimize the RF by finding the finest number of trees and leaves per tree in the forest. Thus, optimization is used to enhance the RF execution that implies less error rate for outlier detection.

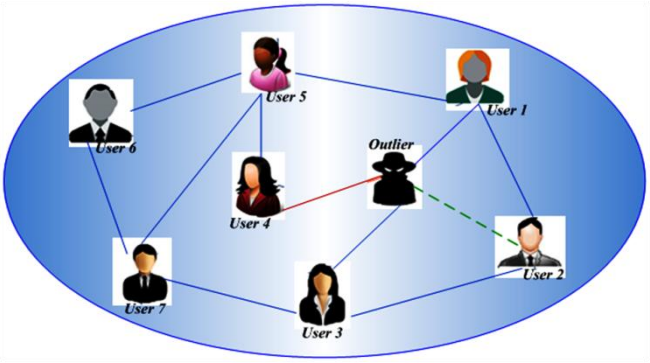

Fig I: Unusual Behavior in the Social Network

Fig I shows unusual behavior of social network. The main objective is to highlight the unusual behavior by identifying unexpected relations assigned or deleted to users in a social network. From the Fig, the outlier deletes the relationship with user 2 which can be represented by the green color dotted line, while red line shows that the outlier assigns a new relationship with user 4.

A. Proposed Hybrid Model for Outlier Detection

i. Random Forest classifier

The productive machine learning approach is RF technique, which integrates two hyper-parameters as the number of trees and number of leaves per tree in the forest (number of splits in the subset at each node). From the social network dataset samples these techniques are obtained and randomly choose at each node. In the meantime two complexities can be noted. The training samples are randomly selected to choose the best sequence in the principal differentiate. Next in the second differentiation, in the forests every one of the trees is maximal trees, because there is no clipping technique is used. In RF classification technique, initially we consider the training set as T_s and from the training set q_s features are extracted. $T_s = \{(A_1, B_1), (A_2, B_2), (A_3, B_3), \ldots, (A_m, B_m)\}$ Where, the input variable is denoted as *Am* and the output class/variable is represented as B_m. Generally, the RF classifier is the combine of a number of decision trees. At first we have considered '*M'* number of trees in the forest in order to make the decision tree. It creates a combination of $M = \{T_{s_1}, T_{s_2}, T_{s_3}, \ldots, T_{s_m}\}$ and each of these is named samples of bootstrap. Therefore, in each bootstrap samples (T_{si}) the one tree (L_i) is generated. Through each tree some input variables $d = \{A_1, A_2, A_3, \ldots, A_m\}$ are going in RF classification and generate one for each tree('M' output) that is represented by $c = \{B_1, B_2, B_3, \ldots, B_m\}$. Here, on this set the final classification is a greater partial vote.

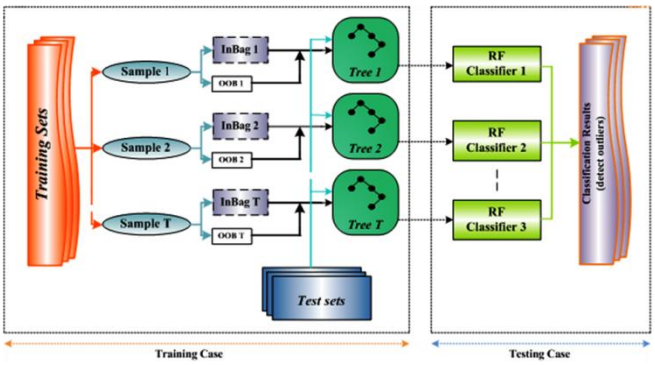

Fig II: Structure of the Random Forests classifier

In the proposed technique, for the increment of classification accuracy the variable importance is assessed. Using the following equation the variable importance is calculated.

$$v_i^{(tr)} = \left[\frac{\sum_{xA \in \varphi^{c(tr)}} i(i_B = c_A^{tr}) - \sum_{xA \in \varphi^{c(tr)}} i(i_B = c_{A,nz}^{tr})}{\varphi^{c(tr)}} \right] \times \frac{1}{t_T} \quad (1)$$

Where, t_T is denoted as the aggregate number of trees, OOB estimates for a unconventional tree is denoted as $\varphi^{c(tr)}$, the classes of RF classifier is indicated as $c_{A,nz}^{tr}$ and c_A^{tr}. For each sample, is sample value, in the forest the quantity of samples per leaves in the tree and quantity of samples per tree is denoted as A and *B* . Generally the exactness is diminished, in the above equation; if the variable importance reduces the precision is expanded. In the training samples, some variables are known as Out-Of-Bag (OOB) and the remaining are called in-bag variables. By contrasting the classification error and OOB term of the variable importance is distinguished. In every node the quantities of splits are randomly chosen to make the binary rule. The mean square error (MSE) is made for each split in the tree to choose the best split. Thus, the leaves per tree are known as best split determination. For expanding the classification accuracy the variable importance measure is valuable.

The fundamental goal of our research is outlier's detection in the dataset if more number of outliers presents then the accuracy is decreased. To find the final predicted tree for all values in the regression tree finally the average is calculated.

ii. Enhanced Krill Herd Optimization algorithm

In this section, the EKHO procedure is used to optimize the RF by finding the finest amount of trees and leaves per tree in the forest. Here, with the optimum number under fitting and the over fitting is identified by looking at the quantity of trees, and quantity of leaves measure per tree in the forest. The under fitting condition may rise, if count of the trees and the count of leaves per tree in the forest is less than the optimal, otherwise the over fitting can occur. In our research, the best quantity of trees and leaves count per trees is chosen based on the RMSE error. Generally, induced motion, foraging motion and random diffusion are the three phases of KHO [23] algorithm. Here, with the help of mutation and crossover updating process the searching behavior of KHO is enhanced, so it is named as EKHO. The procedure of EKHO is as follows.

Initially the input variables and samples of the RF classifier are initialized. Then evaluate the fitness function based on equation (2). After initialization, the variance and best fitting are calculated using the motion calculation of the trees. The motions equations like induced motion, foraging motion and random diffusion are given below.

$$fitness_{function} = \min\{RMSE\} \quad (2)$$

$$M_{ij} = \left[\frac{M_i - M_j}{M^{worst} - M^{best}}\right] \quad (3)$$

$$F_i = \beta_i v_f + F_i^{old} \omega_f \quad (4)$$

$$d_i = d^{\max}\delta \quad (5)$$

Where, the random directional vector is denoted as δ, the maximum diffusion speed is represented as d^{max}, the inertia weight of the foraging motion in the range [0, 1] is ω_f, foraging speed is v_f , the effect of the best fitness of the I^{th} krill is β_i , the best and the worst fitness values of the krill individuals are M^{worst} and M^{best}.

For detecting the outliers the distance between the tress is needed. The distance between two trees is given by following equation. $$D_{s,i} = \frac{1}{5n}\sum_{j=1}^{n}\left\|x_i - x_j\right\| \quad (6)$$

Here, the number of the krill individuals is represented as n and the sensing distance for the i^{th} krill individual is denoted as $D_{s,i}$.

Fig III: Flowchart of Enhanced Krill Herd Optimization algorithm

the motion calculation has not achieved the better solution, then solutions are abandon and using the mutation and crossover updation it produce the random number of solution. The mutation and the crossover rate of krill herd are calculated based on the accompanying equations (7) & (8):

$$Crossover_{rate} = \frac{M_{Genes}}{Length_{channel}} \quad (7)$$

$$Mutation_{rate} = \frac{Mutation_p}{Length_{channel}} \quad (8)$$

Where, M_{Genes} demonstrates the number of genes crossover, $Mutation_p$ represents the mutation point and $Length_{channel}$ indicates the length of chromosome. After updating, check the termination criterion if it satisfies the algorithm completes the technique by selecting finest amount of trees and leaves per each tree. Then, selected optimal result of EKHO is given to the Bagger algorithm to train the RF samples and variables. At that point the trained forests samples are given to the classifier to select the class. For high dimensional datasets, the HADOOP framework is used to detect the outliers. Fig III shows the flowchart of proposed EKHO algorithm.

IV. RESULTS AND DISCUSSION

Several experiments have been performed to evaluate the effective performance of the proposed methodology. The performance of proposed method is executed in MATLAB/Simulink platform by utilizing social network datasets. In our research three types of datasets are utilized for outlier detection. They are WikiSigned [27], HEP-PH [28] and Cond-mat [29].

A. Performance Comparison with Other Techniques

The performance of the proposed method is tested by utilizing three social network datasets and compared with various existing classifiers. The evaluation metrics like accuracy, detection rate, computation time and error rate are used to analyze the proposed method performance and compared with Graoui et al. [24],

Heard et al. [25] and Gao et al. [26] existing outlier detection techniques. The performance of outlier detection technique utilizing the metric of detection rate is expressed as follows.

$$D_{rate} = \frac{Correct_{matching}}{Original_{nodes}} \quad (9)$$

Where, the number of original nodes that were attacked is $Original_{nodes}$ and the number of correct matching between pairs of outlier and original node is represented as $Correct_{matching}$.

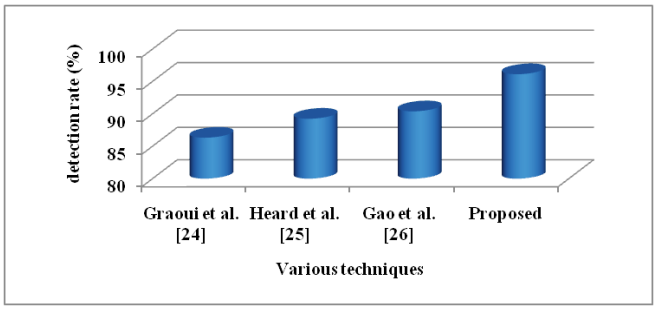

Fig IV: Detection rate comparison of various techniques

Fig IV illustrates comparison of the detection rates of various techniques. From the Fig it is clearly observed that, proposed method which utilized the hybrid model (RF with EKHO) for outlier detection is having better performance of 10.19%, 7.1% 5.96% when compared with Graoui et al. [24], Heard et al. [25] and Gao et al. [26]. In the proposed method the

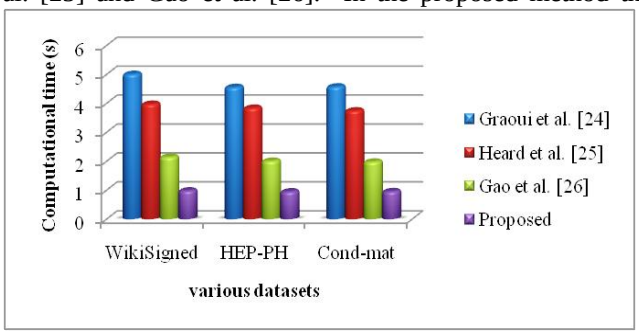

Fig V: Accuracy comparison of various datasets

accuracy comparison of various datasets is shown in Fig V. It is clearly observed that, the proposed method shows better performance for WikiSigned 8.03%, 7.07% and 5.9% , for HEP-PH 10.2%, 6.9% and 6.1%, and Cond-mat 8.5%, 6.4% and 5.7% in terms of accuracy when compared with Graoui et al. [24], Heard et al. [25] and Gao et al. [26].

The computational time comparison of various datasets is shown in Fig VI where it is noticed that the computational

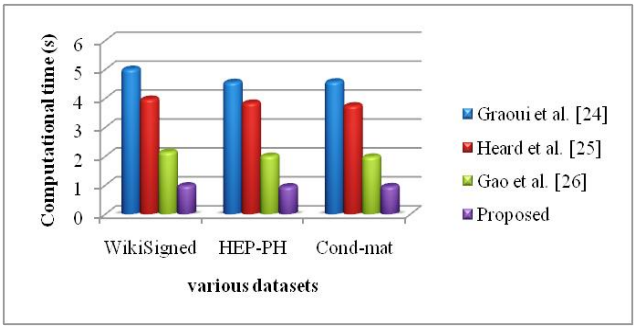

Fig VI: Computational time comparison of various datasets

time of the proposed method is low for WikiSigned 80.3%, 79.2% and 54% , for HEP-PH 79%, 75.1% and 52.5%, and Cond-mat 79%, 74% and 51.7% when compared with Graoui et al. [24], Heard et al. [25] and Gao et al. [26].The HADOOP based classification for high dimensional data is shown in following table 1. Here, the proposed classification approach is compared with the existing classification approaches such as SVM, KNN, and NN. In this paper, the error rate of these three approaches is re-evaluated by utilizing the MATLAB for the results comparison. The proposed classification approach has only 1.80 % of error in the outlier detection. So, the proposed approach generates the less error in the detection of outliers.

Methods	Error rate
SVM	16.13
KNN	8.59
NN	10.43
Proposed Method	1.80

Table I: HADOOP based Classification

V. CONCLUSIONS

In social networks the outlier detection is an important process that aims to automatically identify outlier in network, in the form of unusual behaviors. This paper a hybrid model to detect outliers in social networks utilizing RF and EKHO algorithm is presented. The RF enhances the execution and exactness of general procedure and it is a productive classification strategy. By using EKHO the optimal number of leaves per tree and number of leaves in the forest in RF are selected to detect the outliers. Simulation result shows that our proposed method outperforms better when compared with other state of art approaches. The error rate of proposed method is low when compared with SVM, KNN, and NN classifier approaches. The detection rate of proposed method is better of 10.19%, 7.1% and 5.96% when compared with other three existing research works.

HELPFUL HINTS

REFERENCES

1. M. Pardo and T. Hobza, "Outlier detection method in GEEs", Biometrical Journal, vol. 56, no. 5, pp. 838-850, 2014.
2. P. Raña, G. Aneiros and J. Vilar, "Detection of outliers in functional time series", Environmetrics, vol. 26, no. 3, pp. 178-191, 2015.
3. J. Zhang and M. Zulkernine, "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection", 2006 IEEE International Conference on Communications, 2006.
4. S. Barua and R. Alhajj, "High performance computing for spatial outliers detection using parallel wavelet transform", Intelligent Data Analysis, vol. 11, no. 6, pp. 707-730, 2007.
5. A. Ghoting, S. Parthasarathy and M. Otey, "Fast mining of distance-based outliers in high-dimensional datasets", Data Mining and Knowledge Discovery, vol. 16, no. 3, pp. 349-364, 2008.
6. N. Nigam and T. Saxena, "Global High Dimension Outlier Algorithm for Efficient Clustering and Outlier Detection", International Journal of Computer Applications, vol. 131, no. 18, pp. 1-4, 2015.
7. A. Kaur and A. Datta, "Detecting and ranking outliers in high-dimensional data", International Journal of Advances in Engineering Sciences and Applied Mathematics, 2018.
8. F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets", IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 2, pp. 203-215, 2005.
9. G. Kollios, D. Gunopulos, N. Koudas and S. Berchtold, "Efficient biased sampling for approximate clustering and outlier detection in large data sets", IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 5, pp. 1170-1187, 2003.
10. B. Liu and E. Fokoué, "Random Subspace Learning Approach to High-Dimensional Outliers Detection", Open Journal of Statistics, vol. 05, no. 06, pp. 618-630, 2015.
11. D. Gervini, "Outlier detection and trimmed estimation for general functional data", Statistica Sinica, 2012.
12. P. Gogoi, B. Borah, D. Bhattacharyya and J. Kalita, "Outlier Identification using Symmetric Neighborhoods", Procedia Technology, vol. 6, pp. 239-246, 2012.
13. X. Ru, Z. Liu, Z. Huang and W. Jiang, "Normalized residual-based constant false-alarm rate outlier detection", Pattern Recognition Letters, vol. 69, pp. 1-7, 2016.
14. J. Liu and J. Lian, "Outliers Detection of Dam Displacement Monitoring Data Based on Wavelet Transform", Applied Mechanics and Materials, vol. 71-78, pp. 4590-4595, 2011.
15. M. Bouguessa, "A practical outlier detection approach for mixed-attribute data", Expert Systems with Applications, vol. 42, no. 22, pp. 8637-8649, 2015.
16. X. Deng, P. Jiang, X. Peng and C. Mi, "Support high-order tensor data description for outlier detection in high-dimensional big sensor data", Future Generation Computer Systems, vol. 81, pp. 177-187, 2018.
17. V. Fritsch, G. Varoquaux, B. Thyreau, J. Poline and B. Thirion, "Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators", Medical Image Analysis, vol. 16, no. 7, pp. 1359-1370, 2012.
18. H. Garces and D. Sbarbaro, "Outliers Detection in Environmental Monitoring Data", IFAC Proceedings Volumes, vol. 42, no. 23, pp. 330-335, 2009.
19. Rahmani, M. and Atia, G. (2017). Randomized Robust Subspace Recovery and Outlier Detection for High Dimensional Data Matrices. IEEE Transactions on Signal Processing, 65(6), pp.1580-1594.
20. Wu, S. and Wang, S. (2013). Information-Theoretic Outlier Detection for Large- Scale Categorical Data. IEEE Transactions on Knowledge and Data Engineering, 25(3), pp.589-602.
21. Rao, A., Somayajulu, D., Banka, H. and Chaturvedi, R. (2012). Outlier Detection in Microarray Data Using Hybrid Evolutionary Algorithm. Procedia Technology, 6, pp.291-298.
22. Kim, S., Cho, N., Kang, B. and Kang, S. (2011). Fast outlier detection for very large log data. Expert Systems with Applications, 38(8), pp.9587-9596.
23. A. Gandomi and A. Alavi, "Krill herd: A new bio-inspired optimization algorithm", Communications in Nonlinear Science and Numerical Simulation, vol. 17, no. 12, pp. 4831-4845, 2012.
24. E. Graoui, N. Zrira, S. Mekouar, I. Benelallam and E. Bouyakhf, "Outlier and anomalous behavior detection in social networks using constraint programming", 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), 2016.
25. N. Heard, D. Weston, K. Platanioti and D. Hand, "Bayesian anomaly detection methods for social networks", The Annals of Applied Statistics, vol. 4, no. 2, pp. 645-662, 2010.
26. T. Ji, J. Gao and D. Yang, "A Scalable Algorithm for Detecting Community Outliers in Social Networks", Web-Age Information Management, pp. 434-445, 2012.
27. M. Newman, "The structure of scientific collaboration networks", Proceedings of the National Academy of Sciences, vol. 98, no. 2, pp. 404-409, 2001.
28. S. Maniu, T. Abdessalem and B. Cautis, "Casting a web of trust over Wikipedia", Proceedings of the 20th international conference companion on World wide web - WWW '11, 2011.
29. J. Leskovec, J. Kleinberg and C. Faloutsos, "Graph evolution", ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, p. 2-es, 2007.

AUTHORS PROFILE

M Rao Batchanaboyina is research scholar. He is pursuing his Ph.D in Acharya Nagarjuna University (ANU). He completed his M.Tech in the year of 2009 from Information technology from Andhra University (AU). He received his B.Tech in Computer science and Information Technology from JNTU Hyderabad IN 2005. He has 10 years of teaching experience and published few papers in various international journals, National and international conferences. His research interests are Data Mining, Machine learning, soft computing and Information Security.

Dr. Nagaraju Devarakonda the professor and HOD of IT Department in Lakireddy Balireddy College of Engineering, Mylavaram. He was awarded his Ph.D in Computer science & Engineering from the Jawaharlal Nehru Technological University in the year 2014. He completed his Master Degree M.Tech(CSE) From the Jawaharlal Nehru University (JNU), New Delhi in the year 2005.He completed his B.Tech(CSE) from Sri venkateswara University,Tirupathi in the year 2002. He has 16 years of Teaching experience and published papers in various international journals, National and international conferences. His research interests are Data Mining, Machine learning, soft computing and Pattern Recognition.