International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Framework for Enhancing the Performance of Classification by RCOS and HiForest

Lingam Sunitha , M. Bal Raju

Abstract: ***This framework includes two novel approaches to choose the outlier from various datasets. First one being Relative Cosine-based Outlier Score (RCOS).It's proposed to measure the deviation score of the objects in which each single attribute deviation is calculated and multiplied to get the entire object deviation. Initially we set the threshold. If the calculated score is greater than the threshold, then the instance is considered as an outlier. These are identified and removed since outliers are not required for classification. Now, the remaining normal objects are subjected to different methods of classification. The second method is Hybrid Isolation Forest (HiForest). It is an enhanced version of isolation forest. Similar to method outliers are identified and removed. An experimental analysis is performed on synthetic real time data sets considered from weka and UCI repository. Classification models are built and the generated results are tabulated and accuracy is recorded. The results obtained by the above methods are compared and graphs are plotted for visualization.***

Key terms**: *classification, deviation , HiForest , RCOS***

I. INTRODUCTION

In one of the most recognized work related to outlier detection and removal [1], the authors suggest that first data is divided into different cluster groups based on similarity ,object of same cluster are more similar , the object which are not fall into any of predefined clusters those may be considered as the outliers cannot be considered as noise. Outlier detection and removal techniques have been a significant area of interest amongst the researchers. A remarkable research in this data mining has been made by Hodge and Austin in [2]. clustering techniques [3], multi layer perceptron approach by Zeng and Martinez according to him misclassified instances are assumed as outliers Our work does not focus on a single algorithm, but rather examines the effects of instances that should be mis-classified in a broader context. Another robust technique of identifying outliers efficiently on very large data sets based on distance measured are given in [4-8]. There a Preprocessing technique to fill Missing Values Using a mathematical model Lagrange Interpolation Technique[9] when attributes in relation then we can apply this. In this author differentiated noise and outlier noisy is an error type mismatch wrongly entered during data entry , those will removed simply.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Lingam Sunitha*, Assistant Professor, Department of CSE, KoneruLakshmaiah Education Foundation, Deemed to be University, Hyderabad, (Telangana), India.
M. Bal Raju, Professor Department of CSE Swami Vivekananda Institute of Technology ,Secunderabad, (Telangana), India.

But outlier are some important object those should be further analyzed [10], There are number of basic mathematical and statistical method for Identification of outliers. Dissimilarity measure calculated by Manhattan and Euclidian distances are stored in dissimilarity matrix .if dissimilarity is zero means those objects are same else outlier based on score. And also there are some other methods well suited for single attribute in Data Mining like Inter Quartile Range ,box plots in Real-Time Data[11] .Another set of outlier detection methods is founded based on clustering techniques [12-13]

II. PROPOSED WORK

There are several ways of detecting outliers, but still researches are working on it especially on real time data sets and its applications in day to day life. Methods of outlier detection, classification, clustering vary with the application and hence has always been a constantly evolving research topic. The current work is also one such attempt. Two methods have been designed. First one of them is Gaussian distribution based method which is purely statistical based method. Deviations of all the attributes are measured and final score is calculated and a best permissible score is identified. Second method is based on decision tree. In the current method number of trees are considered and for each tree a score is calculated. An isolated forest is so developed to produce better results.

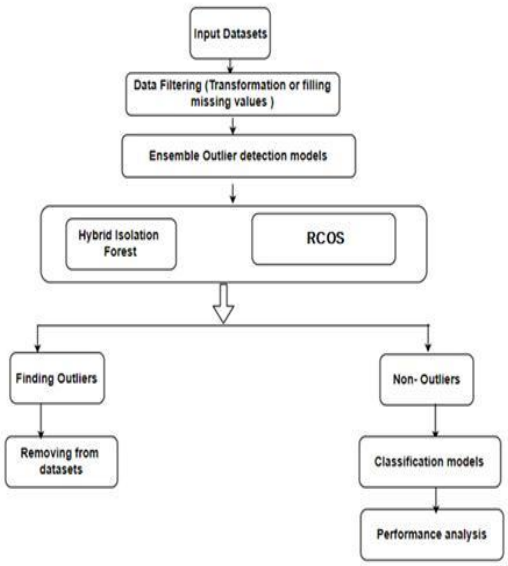

Fig 1. Proposed frame work

2.1: RCOS algorithm:

The training dataset is read and preprocessing begins. This preprocessing [14][15] is done using several methods. Since null values are not supported we need to fill these. This can be done in many ways like mean, median and mode replacement. Here we use mean replacement to fill the null values. For each and every attribute in a data set mean and standard deviation are computed. A naval formula has been derived to calculate score of each instance. This is known as RCOS (Relative Cosine Outlier Score). Initially a threshold is selected. If the RCOS is greater than threshold then its an outlier, else normal. Identified outliers are removed using sorting technique since classification need not be applied on abnormal objects. Normal objects are further investigated. Classification methods are applied on these normal objects. Accuracy is calculated. Repeat the process by changing threshold. The threshold which gives higher accuracy is chosen as final threshold for that method for that dataset. Classification methods applied here are LibSVM [16], RandomForest[17], J48[18], IBk [19], BayesNet [20][21] .

Algorithm1:RCOS

Input: Dataset D, F(D): Features list, T=0.1(threshold)

Output: Outliers O and Normal N.

Procedure:

1. Read Training dataset D
2. To each feature in the features space F(D).
3. repeat
4. Find mean and standard deviation of each feature and compute probability to predict the instance as outlier or not.
5. $$P(F[i][j]) = \frac{1}{\sqrt{2.\pi}}\{\frac{|(F[i][j]) - \mu[F[i]].\cos((F[i][j]) - \mu[F[i]])|}{\sigma}\}$$

Where P(F[i]) represents the probability of each feature i.

F[i][j] represents the instance value at ith feature and jth instance.

Done

6. Sort all the feature probabilities and represent as S(F).
7. Marking outliers
8. For i=1:T
9. Do
10. Mark F[i][j] instance as outlier .
11. Done
12. For i=T:N
13. Do
14. Training data T'=add(F[i][j])
15. Done
16. Apply Classification models on T'.

2.2 Isolation Forest

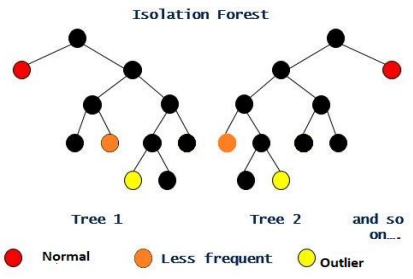

Fig 2 :.Isolation forest

An object which shows deviation from expected is termed an outlier. An outlier is different from other objects of its set and are treated as anomalies. Various techniques are employed for anomaly detection, few of them based of distance, distribution and density. These outliers Analyzed further and it is called as outlier mining. This detection becomes useful in tracking strange events[22]. These outliers can lead to further explorations (good) or are just be meaningless disturbances (noise-bad). Excluding these outliers thus becomes a crucial task for a data scientist. One of the methods involves distance. In this method data set is divided into groups or clusters. Mean value is calculated for each group. A threshold value is assumed at the beginning and a score is calculated which is the distance of an object from mean of each group. The object is treated to belong to the group or is considered as a nearest group to the object when the score is low . If this minimum score is greater than threshold then the object is an outlier, else is a normal object. Another method which can be used for outlier detection using less space and can be performed in linear time is Isolation Forest Method [23]. This method uses decision trees where every object is treated as an individual unit which can be isolated. One attribute of the data set is randomly chosen from the set of attributes. One value is chosen in the range of the values as a splitting value. A decision tree [24][25]is used to represent each trial and possible outcomes. The forest is an average collection of these decision trees. A node in this tree has two children depicting the possibilities of the trial. By comparing an object with the splitting value, two children are assigned to a node. Further these comparison is performed on its child nodes. Each leaf node has a length which is its distance from the root of the tree it belongs to. The leaf node is assigned a score (usually lies between 0 to 1) based on this length which is used to classify objects as outliers. A key value is chosen in this range and the nodes with score greater than this score are termed as outliers. The current employed method for outlier detection is a modified version of the existing isolation forest method and is named as Hybrid Isolation forest.

2.3 Hybrid Isolation forest (HiForest) Algorithm(D ,n, t, s)

Input: Dataset D, t number of trees , n sub sample size , S Threshold

Output: Outliers O and Normal N.

Procedure:

1. Read Training dataset D
2. Randomly select subsets of data from D as RS[].
3. To each subset in RS[]
4. Do
5. Build Isolation tree to each subset by randomly split a randomly selected features in RS[i].
6. To each instance in RS[i]
7. Do
8. Compute instance distribution score to each class by considering the average path of the ITree as
$$k(n) = 2*(\log(n-1) + \eta) - (2*(n-1)/n)$$
9. $$Score_i = 1 - \frac{\tan(avgpathlength)}{\log(k(subsamplesize:n))}$$

10. Here, the instances that has the longer average path are marked as outliers.
11. If(Score$_i$>=T) Then
12. Mark instance as outlier
13. Else
14. Mark instance as normal
15. Done
16. Done
17. Apply Classification models on normal instances as training data.

III. EXPERIMENTAL RESULTS

4.1 Data sets : In this frame work six synthetic real time data sets collected from UCI Repository[26] and weka [27] tool first algorithm RCOS is applied on six data sets . IForest algorithm is applied on two data sets.

i. **ILPD (Indian Liver Patient Dataset) Data Set :** It consists This of 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkpho and 583 instances , It is medical data set.

ii. **Heart Disease Data Set:** This dataset gives risk factors for heart disease age, sex, chest pain type (4 values) , resting blood pressure, serum cholesterol in mg/dl . fasting blood sugar > 120 mg/dl ,resting electrocardiographic results (values 0,1,2) . maximum heart rate achieved exercise induced angina,. oldpeak = ST depression induced by exercise relative to rest

the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy

. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect ,13 attributes and 270 instances

iii. Diabetes Data set: number of rows 768 , 9 columns and 2 classes .

iv. Hepatitis data set : It is a real life multivariate data and 155 objects and 19 features

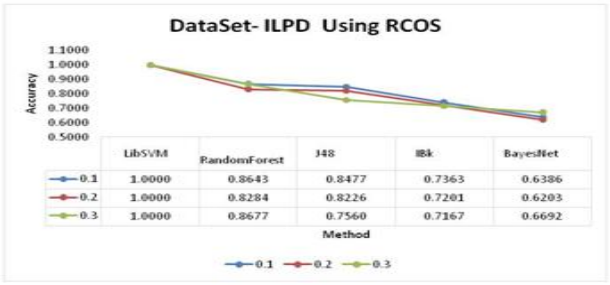

Fig 3. comarision of classification methods on ILPD using RCOS

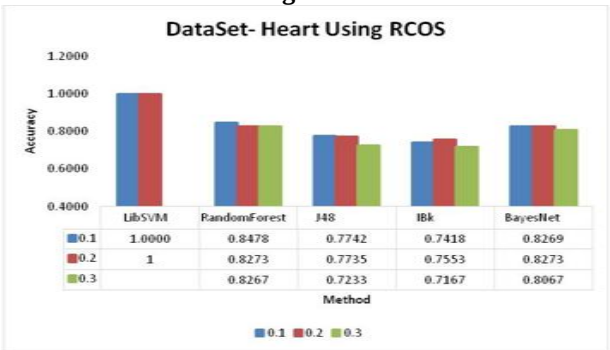

Fig 4 .Accuracy of classification algorithms using RCOS on Heart data set

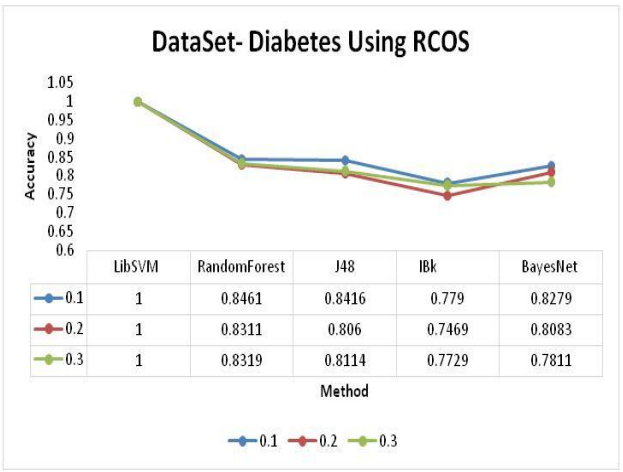

Fig 5 .Accuracy of classification algorithms using RCOS on diabets data set

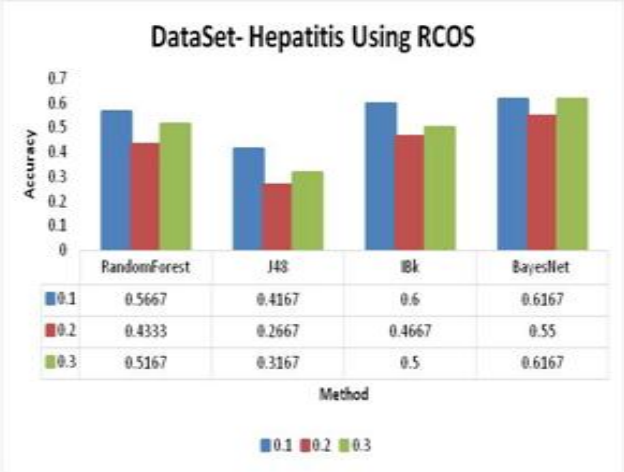

Fig 6 .Accuracy of classification algorithms using RCOS

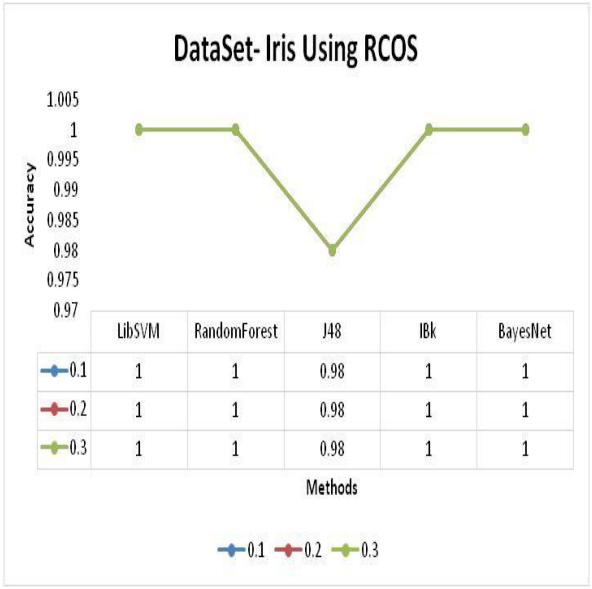

Fig 7 .Accuracy of classification algorithms using RCOS on Iris data set

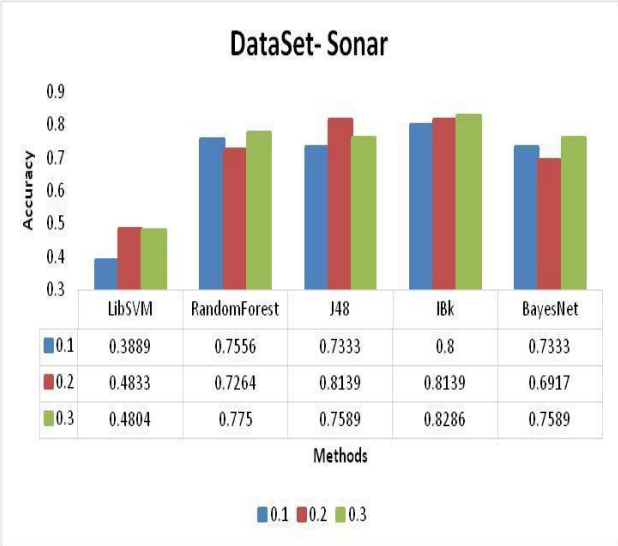

Fig 8 .Accuracy of classification algorithms using RCOS on sonar data set

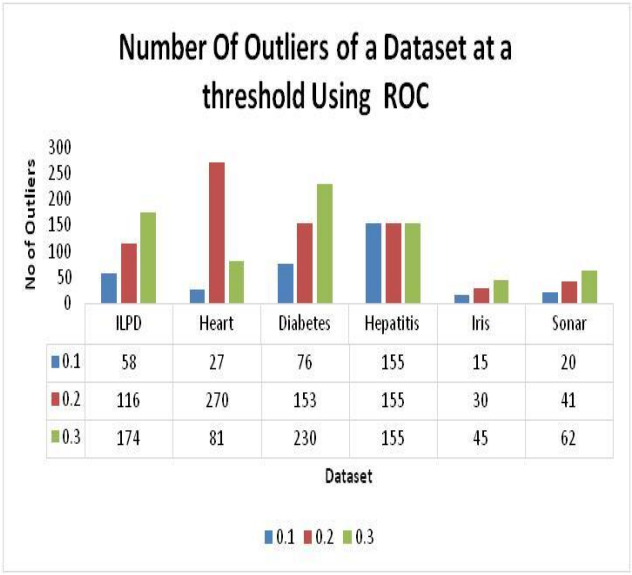

Fig 9 .Number of outliers on at different threshold

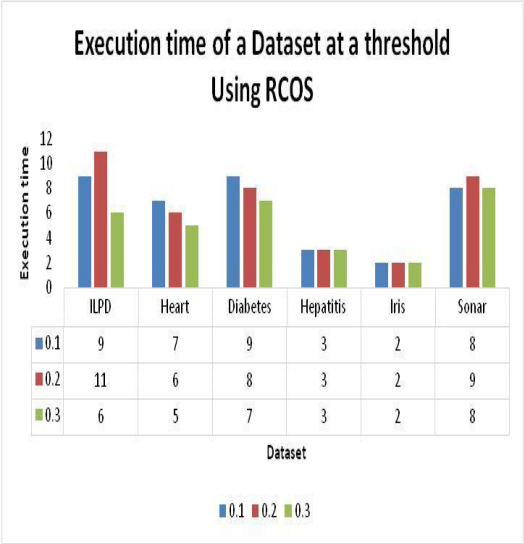

Fig 10 :Execution time of classification algorithms using RCOS on various data sets

Table 1; Accuracy of HiForest for heart data set Comparison Table

Classification Models		Lib SVM	Random Forest	J48	IBk	Bayes Net
Sub ample size S=30	T=15	1.0000	0.8440	0.8242	0.7775	0.8445
	T=20	1.0000	0.8519	0.8114	0.7895	0.8595
	T=25	1.0000	0.8381	0.7943	0.7957	0.8514
	T=30	1.0000	0.8729	0.8314	0.7467	0.8586
Sub Sample size S=40	T=15	1.0000	0.8615	0.7846	0.7308	0.8769
	T=20	1.0000	**0.8802**	**0.8582**	0.7363	0.8341
	T=25	1.0000	0.8679	0.7833	0.7590	0.8449
	T=30	1.0000	0.8212	0.7917	0.7641	0.8526
Sub Sample size S=50	T=15	1.0000	0.8555	0.8227	0.7882	0.8336
	T=20	1.0000	0.8600	0.8600	**0.8100**	**0.8600**
	T=25	1.0000	0.8122	0.7911	0.7400	0.7911
	T=30	1.0000	0.8100	0.7900	0.7500	0.8200

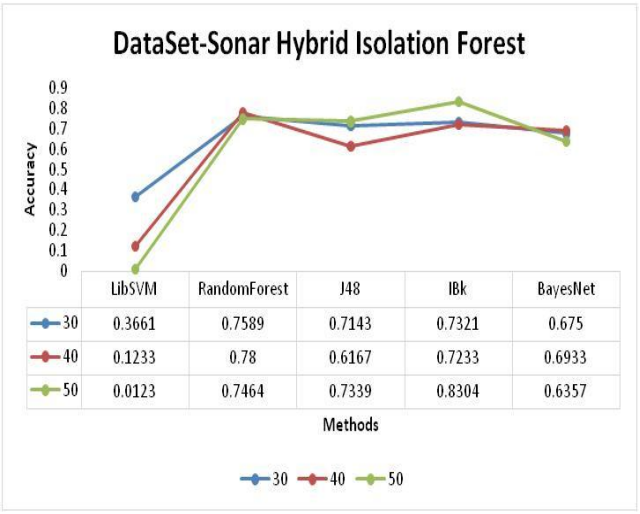

Fig 11 .Accuracy of classification algorithms using Hybrid isolation forest on Heart data set

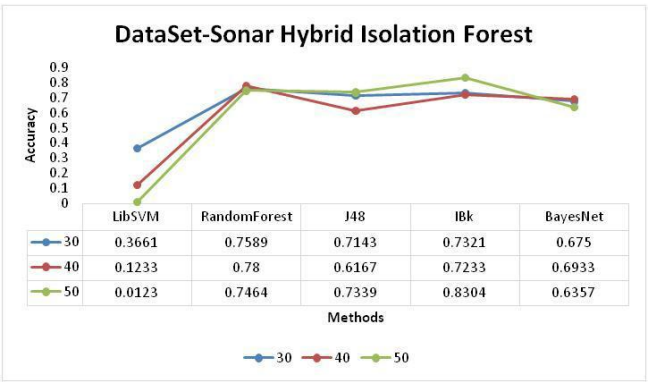

Fig 12 .Accuracy of classification algorithms using hybrid isolation forest on sonar data set

4.2 Comparison of results

From the above results certain inferences can be drawn for RCOS method. For the data set ILPD and diabetes at all the thresholds considered (0.1, 0.2, 0.3) LibSVM is the method with highest accuracy. Similarly for heart data set at 0.1 and 0.2 threshold LibSVM is the method with highest accuracy. But at a threshold of 0.3 LibSVM method cannot be applied and random forest is the preferred method. Bayesnet has the greatest accuracy with respect to Hepatitis dataset at all the considered thresholds. Also, LibSVM method cannot be applied to Hepatitis dataset at any of these thresholds. While all the methods of classification namely LibSVM, randomforest, bayesnet, IBk,J48 are equally accurate when it comes to Iris dataset. For Sonar dataset IBk is the preferred method of classification.

IV. CONCLUSION

RCOS measure, including the expected value and the false alarm probability. The theoretical results suggest parameter settings for practical applications. Simulation results on both synthetic data sets and real-life data sets demonstrated superior performance of our proposed methods on different data set ,There are numbers outlier detection methods are available , there is no common approaches , outlier estimation is specified to problem . In Isolation forest scaling the values in is not required in the vector space. It is an effective method when value distributions cannot be initialized. I t has few parameters , this makes this method fairly robust and easy to optimize. The concept of isolation has not been explored in the current literature and the use of isolation is shown to be highly effective in detecting anomalies with extremely high efficiency. Taking advantage of anomalies' nature of 'few and different', iTree isolates anomalies closer to the root of the tree as compared to normal points. iForest's fast execution with low memory requirement is a direct result of building partial models and requiring only a significantly small sample size as compared to the given training set.

At finally we discuss on performance of isolation forest number of trees 't' controls ensemble size ,we find that the path lengths n=30 ,n=40 , n=50 , the training process will gives the group of trees, the complexity is O(tnlogn).

REFERENCES

1. Aggarwal, C.C., Philip, S.Y.: Outlier detection for high dimensional data. ACM Sigmod Record. Vol. 30. No. 2. ACM (2001)
2. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22.2: 85-126 (2004)
3. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. SIGMOD Record, 29(2):93–104, June 2000
4. Knorr, E.M., Raymond, T.N.: A Unified Notion of Outliers: Properties and Computation. KDD. (1997)
5. Knox, E.M., Raymond T.N.: Algorithms for mining distance based outliers in large datasets. Proceedings of the International Conference on Very Large Data Bases. (1998)
6. Fawcett, T., Foster P.: Adaptive fraud detection. Data mining and knowledge discovery 1.3: 291-316 (1997)
7. Williams, G.J., Zhexue, H.: Mining the knowledge mine. Advanced Topics in Artificial Intelligence. Springer Berlin Heidelberg. 340-348 (1997) .
8. Knorr, E.M., Raymond, T.N., Tucakov, V.: Distance-based outliers: algorithms and applications. The VLDB Journal The International Journal on Very Large Data Bases 8.3-4: 237-253 (2000)
9. Data Mining: Estimation of Missing Values Using Lagrange Interpolation Technique, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013,
10. L. Sunitha, A Comparative Study between Noisy Data and Outlier Data in Data Mining, International Journal of Current Engineering and Technology ISSN 2277 - 4106 © 2013 INPRESSCO.
11. Automatic Outlier Identification in Data Mining Using IQR in Real-Time Data,(IJA RCCE) Vol. 3, Issue 6, June 2014.
12. Raymond, T.N, Han, J.: E cient and E ective Clustering Methods for Spatial Data Mining. Proc. of. (1994)
13. Ramaswamy, S., Rastogi, R., Shim, R.: Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record. Vol. 29. No. 2. ACM (2000) .[14] Changming Zhu* Influence of Data Preprocessing , Journal of Computing Science and Engineering, Vol. 10, No. 2, June 2016, pp. 51-57
14. Wesam S. Bhaya Review of Data Preprocessing Techniques in Data Mining*n* Journal of Engineering and Applied Sciences 12(16):4102-4107 · September 2017
15. Durgesh k. Srivastava, 2lekha bhambhu , Data Classification Using Support Vector Machine , Journal of Theoretical and Applied Information Technology Vol12No1/1Vol12No1.pdf.
16. Eesha Goel* , Er. Abhilasha , Random Forest: A Review, International Journal of Advanced Research in Computer Science and Software Engineering Research, Volume 7, Issue 1, January 2017 ISSN: 22778
17. Gaganjot Kaur, Amit Chhabra, Improved J48 Classification Algorithm for the Prediction of Diabetes , *in* International Journal of Computer Applications 98(22):13-17 · July 2014
18. Jasmina Đ. Novakovic, Experimental Study Of Using The K-Nearest Neighbour Classifier With Filter Methods , Conference: COMPUTER SCIENCE AND TECHNOLOGY,June -2016
19. Michal Horný , Bayesian Networks ,Technical Report No. 5 April 18, 2014
20. Developing a Bayes-net based student model for an External Representation Selection Tutor,Conference: Artificial Intelligence in Education - Supporting Learning through Intelligent and Socially Informed Technology, Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED 2005, July 18-22, Amsterdam, The Netherlands
21. Deepika Pahuja , Outlier Detection for Different Applications: Review, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 3, March – 2013 pg no 1-13.
22. Fei Tony Liu Isolation Forest· DOI: 10.1109/ICDM.2008.17 · Source: IEEE Xplore
23. Conference: Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on January 2009 .
24. (Abdul Fattah Mashat IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 10, 2012 17 | P a g e www.ijacsa.thesai.org A Decision Tree Classification Model for University Admission System
25. K. Bache and M. Lichman. UCI machine learning repository, 2013.
26. Jason Brownlee Standard Machine Learning Datasets To Practice in on June 28, 2016 in Weka Machine Learning Last Updated on December 11, 2019

AUTHOR PROFILE

Lingam sunitha received her MCA from Kakatiya University in 1999, and M.Tech (CSE) from JNTU, Hyderabad in 2009. She is now working as Assistant Professor in Department of CSE, Koneru Lakshmaiah Education Foundation, Deemed to beUniversity, Hyderabad, Telangana 500075.
and also pursuing PhD in Computer Science and Engineering from JNTU Hyderabad,Telangana, India. Her area of interest include Data Mining and artificial intelligence and Machine Learning

Dr M.Bal Raju He received both Graduation B.Tech(ECE) and Post GraduationM.Tech (CSE) from Osmania University and PhD from JNTU Hyderabad in 2010.Now he is working as Professor and PrincipalSwami Vivekananda Institute of Technology Secunderabad, India.

His area of interest includes Data Base, Data Mining and image processing; He was published 30 research papers in various National and International Journals. He was attended and presented 10 research papers in National and International conferences.