computing@computingonline.net
www.computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
International Journal of Computing

IMPROVING CLASSIFICATION PERFORMANCE OF NEURO-FUZZY CLASSIFIER BY IMPUTING MISSING DATA

Balasaheb Tarle [1], Muddana Akkalaksmi [2]

[1] Research scholar CSE department, GITAM (Deemed to be University),
Hyderabad Campus (TS) India
tarlebs123@gmail.com
[2] Professor CSE department, GITAM (Deemed to be University),
Hyderabad Campus (TS) India
akkalakshmi.muddana@gitam.edu.

Paper history:
Received 05 July 2019
Received in revised form 18 November 2019
Accepted 02 December 2019
Available online 31 December 2019

Keywords:
Classifier;
Imputation;
Neuro-fuzzy Classifier;
Training tuples;
Missing data

Abstract: In medical data classification, if the size of data sets is small and if it contains multiple missing attribute values, in such cases improving classification performance is an important issue. The foremost objective of machine learning research is to improve the classification performance of the classifiers. The number of training instances provided for training must be sufficient in size. In the proposed algorithm, we substitute missing attribute values with attribute available domain values and generate additional training tuples that are in addition to original training tuples. These additional, plus original training samples provide sufficient data samples for learning. The neuro-fuzzy classifier trained on this dataset. The classification performance on test data for the neuro-fuzzy classifier is obtained using the k-fold cross-validation method. The proposed method attains around 2.8% and 3.61% improvement in classification accuracy for this classifier.

1. INTRODUCTION

For various medical data classification problems, Data mining and Machine learning methods are effectively applied [1]. The most critical objective in data mining is to identify the hidden patterns in data and use the acquired knowledge on new cases for classifying the data [2]. Neural networks get the input from training data and adjust the weights mapping input to output that requires at least one tuple for the different cases. Neural networks cannot learn to various situations that are not available in the training data tuples [3]. A sufficient number of training data tuples are needed to increase the classification ability of the classifier. The training dataset may have small data samples from its inception. Another reason for less number of training tuples is the case where the training data may contain data tuples with multiple missing attribute value. And such tuples need to be erased. In this case, extra training tuples added to the original training data set to improve the classification ability of the classifier. Authors [4] proposed the imputation method to produce additional training data tuples and these tuples are added to the original training data samples to generate new training data set. The classification performance of the classifier on this new training data set is improved.

If the number of attributes with missing values in a data instance is more, then we have to delete the data instance. The thumb rule, for instance deletion says that, if a data set has more than 5% missing values those tuples are retained. There are two basic methods for discarding data instance with missing values [6], the ways are complete case analysis and dropping. When these methods are applied, it assumed that deleted cases are a part of the vast dataset and cases are missing completely at random. The deletion of tuples may introduce bias. Subsequently, a small sample size affects the analysis.

Several methods for handling missing values are available in the literature [7] some of the popular

techniques discussed here. First, one is to ignore the tuple, and this method is applied if a tuple contains several attributes with missing values, this method does not provide excellent results. The second method is to fill in the missing value manually. This approach is time-consuming and may not be feasible for an extensive data set with too many missing values. The next methods mentioned in the literature use the attribute mean to fill in the missing value and use the most probable value to fill in the missing value. Some researchers have used regression techniques, inference-based tools using a Bayesian formalism or decision tree induction for replacing the missing value. Missing values are imputed with reasonable probable values; these imputation-based procedures are applied instead of complete deletion. An objective of this method is to use known recognized associations from a valid range of values of the data set [8].

In multiple imputations method, for each incomplete information, multiple simulated values are selected. Then iterative data validation is carried with each simulated value substituted, five imputation copies are generally sufficient for the modest amount of missing data. Despite of these methods some more methods like replacement of missing values with the series mean, by the mean or median of nearby points, or linear interpolation between prior and subsequent known points, interpolating between the adjacent valid values above and below the missing one, or substitution of the linear regression trend value for that point also exists [9].

This paper consists of five sections. Section first covers the brief introduction about missing data and related issues. In section two, a short literature review is presented. Section three includes the problem definition and the proposed algorithm. Section four covers the experimentation and results and the last section is a conclusion.

2. RELATED WORKS

In this section, we have focused on some latest developments and relevant information about imputing missing data. Kang and Hyun [10] presented that missing data can decrease the algebraic power of training and can produce partial assessments, leads to unwarranted conclusions. He has handled the various methods of missing data and validated the tools handling missing data. Also presented a comparison of approaches handling the treatment of the missing data. Mosavi et al. [11] proposed an approach for fuzzy classification for missing data. Rough-fuzzy sets are included in logical type neuro-fuzzy systems; subsequently, a rough neuro-fuzzy classifier is derived. The neural network develops an additional purity, and the fuzzy scheme proceeds on the ability of knowledge. When Rough-fuzzy sets are included in NFS output is a rough neuro-fuzzy classifier. Robert K. Nowicki [12], presented a process to execute certain missing data imputation as a statistical method. If the input attributes for learning are numerical, the imputation uses Simpson's fuzzy min-max neural networks.

Shahla and Gerhard [13], proposed a weighted nearest neighbour method for imputing missing attribute values in categorical variables. This method explicitly utilizes the evidence of the relationship between attributes. The imputation error rate is low. M. Albayrak et al. [14], presented a realistic comparison of multiple imputations. The recurrent association of data learning neural network models for estimating missing values. Jin and Dong [15], proposed a new data cleaning method. Three comparative methods are performed to validate the model, and NN is used for pre-processing the training data. The algorithm trains a neural network and is used to create novel training data. The trained system produces several additional training instances are added to the new training dataset. The processed method improves the classification performance of the classifier.

Olanrewaju Akande et al. [16], presented multiple imputations with the mutual method for trade with missing values in numerical records. Missing values are filled with values coming from the predictive model estimation using observed data. It results in multiple, completed versions of the record. Authors also suggested advantages for the regression tree and Bayesian mixture model approaches, making both reasonable default engines for multiple imputations of categorical data. Tarle et al. [17], suggested the fuzzy neural network performs the classification on cleaned data with a correctly reduced feature set. The method has integrated the data cleaning method to improve data quality as a pre-processing method along with a bag of words for feature subset selection. Ezzine and Benhlima [18] presented a comparative analysis for listwise deletion, mean substitution, simple regression, and regression.

Susanti and Aziza [19], suggested handling missing value using DBN. DBN is a beneficial method to maintain the interactions between variables of data. The consequences of the estimate were used to fill missing values in the data. Support Vector Regression system is used for calculating the missing values. It is chosen for its performance as compared to other parallel systems. N. Anindita et al. [20], presented that the hepatitis dataset has an arbitrary arrangement of missing values. This arrangement can be measured by using MCMC and FCS as multiple imputation systems. The research

focused the investigation on equating groups of multiple imputations system and PCA as the happening selection.

S. Azim and S. Aggarwal, [21], suggested the implementation of the 2-stage hybrid model to fill missing values. The proposed algorithm is tested with a simple and complex dataset with varying percentages of missing value and varying value of fuzzifier. The missing data that are not missing completely at random contain non-random elements that may prejudice the results. The deletion of missing data can bring in substantial bias into the results. Also, the reduced sample size may affect the analysis. The Mean-fill approach for finding the estimates of the values is common in missing data imputation.

3. PROPOSED ALGORITHM

In this work problem of insufficient training samples is handled. The training dataset may have insufficient data samples from its inception.

In another case, the training data may contain some data records with multiple missing attribute values. And such records with missing values need to be erased; in these situations, sufficient extra training tuples need can be added to the original training data set to improve the classification ability of the classifier.

In the proposed algorithm, missing data is added in original training tuples explicitly in a random manner. The tuples and features are chosen randomly for the same. Subsequently, missing attribute values are substituted with available domain values of the same attribute, and additional training tuples are generated and validated on the classifier. These tuples are in addition to original training tuples. These generated training samples, along with original training samples, increase the size of the data set and provide sufficient data samples for learning. The neuro-fuzzy classifier is trained on this dataset. The classification performance on test data for the neuro-fuzzy classifier is obtained using the k-fold cross-validation method.

3.1 ARCHITECTURE OF NEURO-FUZZY CLASSIFIER

3.1.1 NEURO-FUZZY CLASSIFIER WITH BOW

The data is classified using the Neuro-Fuzzy classifier. The extracted features are given as the input to the Neuro-Fuzzy Classifier for classifying all the given data. The Neuro-fuzzy system has a three-layered architectural design; following diagram in fig. 1 presents the basic structure of the Neuro-fuzzy classifier system. Neuro-Fuzzy classifier is a fuzzy-based system that is trained by a learning algorithm derived from Neural Networks [22]. The learning algorithm only performs on the local information and provides the local modifications in the fuzzy system. In general, a Neuro-Fuzzy system generates compelling solutions instead of using the system components individually [23]. The steps used in the Neuro-Fuzzy classifier are explained in the following section.

3.1.2 FUZZIFICATION

The input values are the extracted features or attributes are acknowledged by the structure as the feedback, and then these feedback attribute values are fuzzifier based on the membership functions (MF). The MF is providing the membership to each feature to various classes. It is used to extracted features from unseen and inter-related data, according we have to get the additional accuracy of the sorting stage spending Neuro-fuzzy Structure.

Here, the π–type membership function is used to classify the data. The π–type MF has fuzzifier a factor that can be adjusted compared to the necessity of the problem. This controls the simplification capability by choosing a correct value of the fuzzifier a factor and provides more contribute for arrangement the data. The steepness of the Gaussian function is well-ordered by changing the fuzzifier value. The membership function after the Fuzzification process is expressed in the membership matrix. The complete rows and columns in the membership matrix are cascaded and to translate it into a vector. This created vector is set as the input to the neural network.

3.1.3 NEURAL NETWORK

This stage, we have used Feed Forward Multi-layer Perception classifier, it has 3-layers such as an input, unseen, and output layer. The overall amount of input nodes of the neural network is equal to the creation of the number of attributes and modules classes. The total number of output nodes from the neural network is the same as that of the number of classes [24]. The whole number of hidden nodes is equivalent to the square root of the product, of the number of input and output nodes [25].

3.1.4 DEFUZZIFICATION

It is the method of translating the amounts of membership of output stated attribute inside their unwritten positions into strong statistical values, based on the output nodes of the neural network are carried out with defuzzification. A detail of working of Neuro-Fuzzy Classifier algorithm is available in [17].

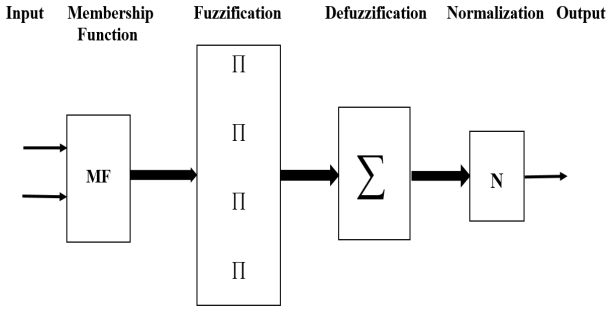

Figure 1 – Architecture of Neuro-Fuzzy Classifier [17]

In this paper, a multiple imputation based method is used to produce added training tuples. These training tuples are then added to the original data set, to form a new training dataset. On this new data set the neural network is trained. This classifier has good classification ability. The details are as follows.

Let D_N be a set of all on hands n training tuples. Let the training tuple denoted by t. Let the attributes denoted by F_N, and the class label denoted by C_N. It is necessary to insert missing feature values randomly in several data tuples if necessary. Let F_N be the feature with missing values and let D_M be a set of tuples that contains a set of tuples with missing features values and usual (non-missing features) data tuples. Compute domain feature values denoted by DV for all features F_N having missing values. Generally, many imputations perform three to five value imputations. In this case, the number of imputations is based on possible rang/domain value for a feature and possible combinations of these domain values. Thus it is not restricted to 3 to 5. Thus with all permutations and combinations of all missing values of attributes are used to build new data set D_{PC}, that data generated with permutations and combinations of domain attribute values for multiple attributes. The ratio μ of original tuples in training data to additional training instances created and added to training data, here the proportion for this ratio is 30% to 50%, depending on the size of the dataset and in this way the dimension of the data set is optimized.

For each feature having values, insert missing feature values by i^{th} domain values. Construct all probable permutations and combinations using existing domain set values for missing features in tuples. Repeat this process for all tuples having missing values. In this way, multiple tuples with imputed features are constructed.

The ratio $\mu = p/q*100;$ where p is the no. of examples in training data and q be additional training examples created and extra to training data. In this research we have used μ = 30% for small-sized data sets, i.e. data set size up to 100 data examples; and μ = 50% for reasonably sized data sets, i.e. data set size from 100 to 500 data examples by using our own thumb rule, and it worked well.

The proposed Imputation algorithm:

1. Input Data set D_N
2. {
3. Add the missing values to some tuples randomly, explicitly to built D_M.
4. Calculate (DV, (F_N))
5. ∀ (D_M, F_N) Substitute (DV)
6. Repeat steps 4 to 5 for all tuples.
7. }
8. Induct Hypothesis (D_N)
9. ∀ $D_{PC,}$ Test((Hypothesis (D_N), $D_{PC,}$))
{
10. If Correctly Classified keep tuple in D_{PC}
}
Else delete the tuple from D_{PC}
$D_I = D_{PC}$
14. Compute duplicates (D_I, D_N)
15 }
16. Delete duplicates
17. $D_{NI} = D_N + D_I$
18. {
19. Induct Hypothesis (D_{NI})
20. Calculate classification performance.
21. }
22. Stop.
End

Figure 2 – The proposed Imputation algorithm

The data set D_N is applied for training the neuro-fuzzy classifier. This neuro-fuzzy classifier is applied for the correctness of imputed data tuples. The imputed data tuples D_{PC} are applied as test data on the said classifier. The test tuple that is correctly classified is a correctly imputed instance otherwise needs to be deleted. The correctly imputed tuples are part of D_I. Compare D_N, and D_I to find out duplicate tuples constructed in D_I. Duplicate tuple needs to be deleted. Merge D_I in $D_{N,}$ and thus new set is D_{N1}. The model is trained on new data set $D_{N1,}$ and the model executes with enhanced classification performance.

4. EXPERIMENTATION AND PERFORMANCE EVALUATION

The UCI repository data sets Australian, Breast, Lymph, Shuttle and Weather [26] were used for conducting the experiments. In MATLAB the proposed system is implemented. The missing data was explicitly added to few data tuples. The imputing was done following the proposed algorithm. The datasets D_N, D_{N1}, and D_M, were

applied to train the said classifiers. Readings of classification accuracy for these classifiers were acquired using the k-fold cross-validation method.

The accuracy is also obtained with 2 imputation methods first one is substitute missing attribute values with mode and second one substitute with most probable value [2]. This is done to compare the performance of the proposed method with existing techniques. Table 1 presents a comparison of the proposed method with existing techniques. Let A_N be classification accuracy of the classifier using the original data set (D_N) as a training data. Similarly, A_{N1} denotes accuracy on imputed data set D_{N1} and A_M on a dataset containing missing attribute values D_M. The classification performances obtained with existing methods are denoted by A_{Mode} and A_{MP}, that accuracy with mode and most probable value methods of imputation. It can be observed that the proposed method is moderately better than the existing techniques.

Table 1. Presenting Comparison of proposed Method with Existing Techniques

Data Sets	A_N	A_M	A_{N1}	A_{Mode}	A_{MP}
Australian	97.98	98.28	98.70	97.52	97.24
Breast	95.14	96.87	98.89	97.16	97.86
Lymph	98.67	98.00	99.21	97.92	98.61
Shuttle	92.05	88.89	99.21	88.24	90.90
Weather	90.48	88.24	92.31	81.82	81.81
Average	**94.86**	**94.06**	**97.66**	**92.53**	**93.28**

Table 2 presents the results obtained using the method as mentioned above for the proposed algorithm. The table also presents I_{N1} and I_M denote the enhancement in classification accuracy. I_{N1} is an improvement in accuracy in A_{N1} with comparison to A_N. That is the improvement in accuracy with the original data set (D_N) as a training data to imputed data set D_{N1}. Similarly, I_M is calculated. The improvement in the Accuracy calculated by following equation 1and 2.

$$I_{N1} = A_{N1} - A_N, \quad (1)$$

$$I_M = A_{N1} - A_M, \quad (2)$$

where, I_A – improved accuracy.

Table 2. Enhancement in classification accuracy with the proposed Method

Data Sets	A_N	A_M	A_{N1}	I_{N1}	I_M
Australian	97.98	98.28	98.70	0.72	0.43
Breast	95.14	96.87	98.89	3.75	2.02
Lymph	98.67	98.00	99.21	0.54	1.21
Shuttle	92.05	88.89	99.21	7.16	10.32
Weather	90.48	88.24	92.31	1.83	4.07
Average	94.86%	94.06%	97.66%	2.8%	3.61%

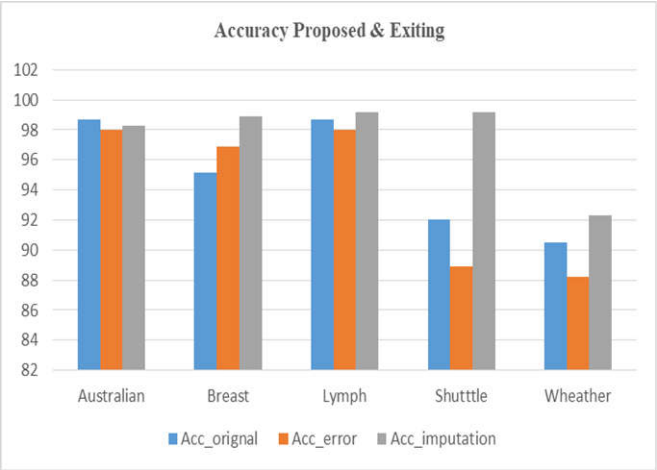

Figure 3 – Graph for classification accuracy comparison of the original and proposed Method

It can be observed that average improvement in the accuracy of the neuro-fuzzy classifier by the proposed method is around 2.8% and 3.61% in comparison with the original data set as new tuples are introduced in the original data set. Thus it improves the classification performance of the neuro-fuzzy classifier.

5. CONCLUSION

The proposed algorithm creates additional data tuples using domain-based multiple imputation methods and adds these tuples in original available training data. It enhances the classification ability of the classifiers. This proposed method utilizes a set of domain values for imputations of feature values. The correctness of the imputed tuple is verified on the classifier. The proposed method significantly enhances the classification performance of the classifiers. This technique is more suitable for small to medium data sets. The data imputation helps in the evolving enhanced and more accurate classifiers. The suggested method presents improved classification performance. The proposed method attains around 3.61% improvement in classification accuracy on a fuzzy neural network.

ACKNOWLEDGEMENT

Thanks to CSE Department staff of GITAM Hyderabad for their valuable guidance and providing resources.

6. REFERENCES

[1] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, Edition 2012.

[2] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan

Kaufmann Publishers Inc. San Francisco, USA, 2011.

[3] B. Tarle, R. Tajanpure, S. Jena, "Medical data classification using different optimization techniques: A survey," International Journal of Research in Engineering and Technology (IJRET), vol. 5, Special Issue 5, ICIAC 2016, pp. 101-108, May 2016.

[4] D.V. Patil, R.S. Bichkar, "Improving generalization ability of classifier with multiple imputation techniques," *ICIP 2012, Communications in Computer and Information Science*, vol. 292, Springer, Berlin, Heidelberg, pp. 309-317, 2012.

[5] R. W. Krause, M. Huisman, C. Steglich and T. A. Sniiders, "Missing network data a comparison of different imputation methods," *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, 2018, pp. 159-163.

[6] A. M. Kalteh and P. Hjorth, "Imputation of missing values in the precipitation-run process database," *Journal of Hydrology Research*, vol. 40, issue 4, pp. 420-432, 2009.

[7] Jaemun Sim, Jonathan Sangyun Lee, and Ohbyung Kwon, "Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications," *Mathematical Problems in Engineering*, vol. 2015, pp. 1-14, 2015.

[8] P. V. de Campos Souza, L. C. B. Torres, A. J. Guimaraes, V. S. Araujo, V. J. S. Araujo, and T. S. Rezende, "Self-organized direction aware for regularized fuzzy neural networks" *Evolving Systems*, pp. 1–15, 2019. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/s12530-019-09278-5

[9] C. de Bodt, D. Mulders, M. Verleysen and J. A. Lee, "Nonlinear dimensionality reduction with missing data using parametric multiple imputations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1166-1179, 2019.

[10] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anaesthesiology*, vol. 64, issue 5, pp. 402-406, 2013.

[11] M.R. Mosavi, A. Ayatollahi and S. Afrakhteh, "An efficient method for classifying motor imagery using CPSO-trained ANFIS prediction," Evolving Systems, pp. 1-18, 2019.

[12] R. K. Nowicki, "On classification with missing data using rough–neuro-fuzzy systems," *Int. J. Appl. Math. Computer Science*, vol. 20, no. 1, pp. 55–67, 2010.

[13] S. Faisal and G. Tutz, "Nearest neighbor imputation for categorical data by weighting of attributes," arXiv: 1710.01011v1 [stat.ME] 3 Oct 2017.

[14] M. Albayrak, K. Turhan and B. Kurt, "A missing data imputation approach using clustering and maximum likelihood estimation," *Proceedings of the Medical Technologies National Congress*, Trabzon, 2017, pp. 1-4.

[15] X. Ma, Y. Jin, and Q. Dong, "A generalized dynamic fuzzy neural network based on singular spectrum analysis optimized by brain storm optimization for short-term wind speed forecasting," *Applied Soft Computing*, vol. 54, pp. 296–312, 2017.

[16] O. Akande, F. Li & J. Reiter, "An empirical comparison of multiple imputation methods for categorical data," *The American Statistician*, vol. 71, no. 2, pp. 162-170, 2017.

[17] B. Tarle, Ch. Sanjay, S. Jena, "Integrating multiple methods to enhance medical data classification," *Journal Evolving Systems*, Publisher Springer Berlin Heidelberg, pp. 1-10, 2019.

[18] Ezzine and L. Benhlima, "A study of handling missing data methods for big data," *Proceedings of the IEEE 5th International Congress on Information Science and Technology CIST*, Marrakech, 2018, pp. 498-501.

[19] S. P. Susanti and F. N. Azizah, "Imputation of missing value using dynamic Bayesian network for multivariate time series data," *Proceedings of the International Conference on Data and Software Engineering*, 2017, pp. 1-5.

[20] N. Anindita, H. A. Nugroho and T. B. Adji, "A combination of multiple imputations and principal component analysis to handle missing value with the arbitrary pattern", *Proceedings of the 7th International Annual Engineering Seminar (AES)*, Yogyakarta, 2017, pp. 1-5.

[21] S. Azim and S. Aggarwal, "Using fuzzy c means and multi-layer perceptron for data imputation: Simple v/s complex dataset," *Proceedings of the 3rd International Conference on Recent Advances in Information Technology (RAIT)*, Dhanbad, 2016, pp. 197-202.

[22] Q. H. Do and J.-F. Chen, "A neuro-fuzzy approach in the classification of students academic performance," *Computational Intelligence and Neuroscience*, vol. 2013, pp. 1-7, 2013.

[23] M. Juhola, H. Joutsijoki, H. Aalto, and T. P. Hirvonen, "On classification in the case of a medical data set with a complicated

distribution," *Elsevier Applied Computing and Informatics*, vol. 10, no. 2, pp. 52-67, 2014.

[24] M. B. Gorzałczany, and F. Rudziński, "Interpretable and accurate medical data classification – a multi-objective genetic-fuzzy optimization approach," *Expert Systems with Applications*, pp. 1-17, 2016.

[25] Lin, J., Li, N., Alam, M.A. and Yuqing Ma l,. "Data-driven missing data imputation in cluster monitoring system based on deep neural network". *Applied Intelligence, pp,1-18,*2019. doi:10.1007/s10489-019-01560-y

[26] D. Dua, and C. Graff, *UCI Machine Learning Repository*, Irvine, the University of California, 2019. [Online]. Available at http://archive.ics.uci.edu/ml.

Balasaheb Tarle *is presently working as Associate Professor of Computer Engineering department in MVPS's KBTCOE Nasik. He is doing PhD. from GITAM University, Hyderabad. His research area is Data Mining, and he has published more than 12 papers in international journals and conferences.*

Dr. Muddana Akkalakshmi *is a Professor in the Department of Computer Science and Engineering at GITAM (Deemed to be University). Her areas of interest include Artificial Intelligence and Security. She gives training to undergraduate, post-graduate students and guides research scholars in these areas.*