Analysis of Various Crop Yields in Different Spatial Locations of Karimnagar District in AP

Ch. Mallikarjuna Rao[1], Dr. A. Ananda Rao[2], N. Madhusudhana Reddy[3]

[1] Department of Computer Science and Engineering, Associate Professor , GRIET, Hyderabad

[2] Department of Computer Science and Engineering ,Professor, JNTUA, Anantapuramu

[3] Department of Computer Science and Engineering , Associate Professor, SDITW, Nandyal

Abstract

The size of spatial data is growing day by day. Spatial data mining is the technique by which one can able to extract interesting and potentially useful spatial information which is hidden in spatial databases. Efficient Spatial Data techniques are required to extract useful information from spatial data sets for effective decision making purpose and are used by various organizations. Presently the Spatial data mining techniques like Classification, Clustering and Association are used. Effective analysis was done using the hybrid data mining techniques by mixing both clustering and classification techniques. In this paper yields of various crops in different spatial locations of Karimnagar district were taken for study by applying the hybrid data mining technique.

Keywords: *Spatial data mining, hybrid data mining technique, Effective Analysis, Decision making, spatial locations.*

1. Introduction

Agriculture yield growth is highly unpredictable in India in spite of growth in Technology and irrigation methods [11]. Karimnagar District is a district in northern Andhra Pradesh, India. The city of Karimnagar is the district headquarters, and has a population of about 3 lakh. It is the 4th biggest city in the Telangana region of Andhra Pradesh. The district has two municipal corporations at Karimnagar and Ramagundam. Karimnagar district occupies an area of 11,823 square kilometers (4,565 sq mi). It borders Adilabad District in the north, Maharashtra and Chhattisgarh in the northeast, Warangal District in the south, Medak District in the southwest and Nizamabad District in the west. Karimnagar district has 57 Mandals namely Bejjanki, Bheemadevarpalle, Boinpalli, Chandurthi, Chigurumamidi, etc, ending with Yellareddipet. Major Crops of this Karimnagar district are Rice, Jowar, Cotton, Turmeric, Maize, Arhar, Chilllies, Sugar Cane and Sesame. Analysis need to be done on the agriculture data sets which requires classical data mining techniques apart from statistical techniques. The data mining techniques[12]like classification, clustering and association are required to apply on the realistic data sets for analysis and draw effective conclusions on the agriculture crop yields [8, 9, and 10] of various seasons like Kharif and Rabi. The following existing techniques are discussed along with proposed hybrid approach.

2. Literature Survey

2.1 K-Means clustering Algorithm:

The k-means algorithm [3] [Hartigan& Wong 1979] is the well known clustering technique used in scientific and industrial applications. Its name comes from centroid which is the mean of c of k clusters C. This technique is not suitable for categorical attributes. It is more suitable for numerical attributes. K-means [1] algorithm uses squared error criteria and is extensively used algorithm. The data is partitioned into K clusters (C1; C2;:::; CK), using this algorithm which are represented by their centers or means. The mean of all the instances belonging to that cluster gives the center of each cluster.

The pseudo-code of the K-means algorithm is given by Fig.1. The algorithm begins with randomly selected initial set of cluster centers. Every instance is assigned to its closest cluster center for each iteration based on Euclidean distance calculated between the two. Again re-calculate the cluster centers. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows:

1. Each instance d_i is assigned to its closest cluster center.

2. Each cluster center Cj is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of instances to clusters. Another stopping criterion is that if it exceeds a pre-defined number of iterations.

Input: S (instance set), K (no. of clusters)
Output: clusters
1: Initialize K cluster centers.
2: while termination condition is not satisfied do
3: Assign instances to the nearest cluster center.
4: Update cluster centers based on the assignment.
5: end while
Figure1. K-means Algorithm

This algorithm is considered as a gradient-decent procedure. In which it begins with random selection of an initial set number of K cluster-centers and iteratively updates by which error function will decrease. A rigorous proof of the finite convergence of the K-means type algorithms is given in [4, 5]. O (NKT) is the order of time complexity where N is the total number of objects, K is the number of clusters and T is the number of iterations.

K-means algorithm is popular because of linear time complexity. Even if large number of instances is considered this technique is computationally attractive. In the Achilles heel of the K-means algorithm selection of the initial partition is required. It is highly sensitive to initial partition selection, which in turn may make the difference between global and local minimum.

This algorithm is applicable only when the mean is defined. It is sensitive to noisy data and outliers. In this method, the drawback is the number of clusters needs to be defined in advance. For Numerical attributes Euclidean distance measure is used as similarity measure where as for categorical attributes it is the number of mismatches between objects and the cluster prototypes.

K-medoid or PAM [2] is another partitioning algorithm in which SSE is minimized. This algorithm is very similar to the K-means algorithm. It differs from the latter mainly in its representation of the different clusters. The central object of every cluster represents the respective clusters. Both methods require the user initially to specify K, the number of clusters. On behalf of SSE error criteria other method can be used. Estivill-Castro (2000) analyzed the total absolute error criterion. Namely, instead of summing up the squared error, he suggests to summing up the absolute error. This method requires more computational effort although it is superior in regard to robustness. The objective function is the sum of squares of errors between the centroids and their respective points which is defined as the total intra-cluster variance.

It is derived from the probability frame work and is used extensively in statistics. It is optimized in two ways. The first one is similar to EM algorithm one of the clustering technique. It is in done two-steps. First time assign all the points to their nearest centroids. For the newly formed groups again recalculate the centroids. This process of calculation continues till the stopping condition is reached. This approach is known as Forgys algorithm [6] and has many advantages:

- It easily works with any -norm p L
- It allows straightforward parallelization [5]
- It does not depend on the order.

2.2. J48 algorithm:

J48 algorithm [7] is the optimization of C4.5 algorithm and it is an improvised version of it. Decision tree is its output. It has a root node, along with intermediate nodes apart from leaf nodes. The root node is found based on selection of attribute using Ranker algorithm and evaluation of Info gain. Except Root node and leaf nodes, every other node in the tree consists of decision and which leads to our result. This tree divides the given space of a data set into areas of mutually exclusive. In this its data points are described by every area will have a label, a value or an action. The criterion of splitting is used to find which attribute makes the best split on the portion of tree of the training data set which in turn reaches to a particular node. Decision tree is formed by using the children attribute of the data set.

Proposed Approach

The original data set was subjected to preprocessing techniques and converted in to the required format. The missing values are replaced by mean values. Then apply the data mining technique namely k-means clustering algorithm on the data set and then we get the new data set namely clustered data set. The classification technique J48 was applied on that clustered data set which results in hybrid model with Decision Tree.

Implementation of proposed approach

A data set on Yields Khariff of the Year 2010-2011 of Karimnagar district was analyzed with k-means (or simple k-means) clustering data mining technique. It has 5 attributes namely crop, Normal_area, Area_Sown, Productivity and Production. The results were shown below. In this we considered mandal as Class variable. The whole data was clustered in to five clusters namely Cluster 0 cluster1, cluster2, cluster3 and cluster 4. This has 12 instances. All the instances were clustered and in that 3(25%) instances were under cluster 0, 2(17%) instances were under cluster1, 2(17%) instances were under cluster2, 3(25%) instances were under cluster3 and the other remaining 2(17%) instances are under cluster 4. The clustered bar chart graph is shown on Fig.1. The clustered centroids are shown in Table 1. In this, we have used Euclidean Distance as the similarity measure. Mandal attribute was considered as class to cluster Evaluation on training data. Maximum number of Iterations is 500 with seed 10 and number of Execution slots is 1. Clustering is

an Unsupervised one and here missing values are replaced with mean and mode apart from Discretization with 10 bins. Here full training set as the clustering model. Number of iterations is 3 and the Missing values are globally replaced with mean/ mode. Within cluster sum of squared errors is 8.388004016031848and the Test mode is evaluated on training data. The time taken to build model (full training data) is 0.02 seconds.

The Pictorial representation of Tree structure obtained from J48 classification Algorithm after simple K-means clustering technique was applied. Initially simple K-means algorithm was applied on the preprocessed data set. After that decision tree algorithm J48 classification technique was applied on the clustered data sets which results a hybrid technique. The result was summarized as follows. It is with Minimum confidence is 0.25 and binary tree classifier. It has 12 numbers of instances with 7 attributes. The attributes are Instance_number, crop, Normal_area, Area_Sown, Productivity, Production and Cluster. The test mode is with full training set as classifier model. The resulted J48 pruned tree is given below.

Cluster=cluster0:Soya bean(3.0/2.0)
Cluster=cluster1:Greengram(2.0/1.0)
Cluster = cluster2: Maize (2.0/1.0)
Cluster = cluster3: Paddy (3.0/2.0)
Cluster=cluster4:Turmeric (2.0/1.0)

The number of leaves is 5 and size of the tree is 6. The time taken to build model is 0.02 seconds.

The Detailed confusion matrix is shown in Table 3. Almost all classifiers are classified accurately. Among these classifiers Redgram, cotton, chillies, sugarcane and castor are not classified accurately where as others are classified accurately. The J48 pruned tree is shown in Fig.2. Correctly Classified Instances is 41.6667%. In-correctly Classified Instances is 58.3333 %. The Kappa statistic value is 0.3636. It means the instances are classified correctly to some extent. The Mean absolute error is 0.0972where as Root mean squared error is 0.2205. Relative absolute error is 63.6364% and Root relative squared error is 79.7724 %. Coverage of cases (0.95level) is 100%. Mean rel. region size (0.95level) is 20.833%. In the confusion matrix each column represents instances in the predicted class and each row indicates instances in actual class. These mining techniques were applied on various data sets. Fig.3 Shows J48 pruned tree for Kharif 2010 clustered data set which consists of GPS co-ordinates latitude and longitude as its attributes apart from other attributes.

3. Results & Analysis:

Initially k-means clustering technique was applied on Karimnagar Kharif Yields 2010-11 data set for analysis. The cluster centroid shown by Table-1 shows that

Cluster 0: It was found that Soyabean is the crop with Normal_area centroid as 2960.6667 and has lowest Area_Sown1765.3333. It has Productivity 1833.3333 and has Production 4556.3333.

Cluster 1: It was found that GreenGram is the crop with Normal_area centroid as 10914 and has Area_Sownis8822.5. It has lowest Productivity 542.5 and has Production 4766.5.

Cluster 2: It was found that Maize is the crop with Normal_area centroid as 46390 and has Area_Sownis29436.5. It has highest Productivity 43675 and has Production 158452.5.

Cluster 3: It was found that Paddy is the crop with highest Normal_area centroid as 247514 and has highest Area_Sownis302506. It has Productivity 5356.6667 and has highest Production 517616.7.

Cluster 4: It was found that Turmeric is the crop with Normal_area centroid as 21395 and has Area_Sown is 9370. It has second highest Productivity 7727.5 and has second highest Production 134728.05.

Detailed Accuracy by Class is given by Table 2 in that Maize, Greengram and Turmeric has highest TP rate as 1and F-Measure values as 0.667. Paddy and Soyabean also has TP rate as 1 and F-Measure values are 0.500. The above values states that they are classified accurately.

4. Conclusion

There is a correlation between Rabi yields 2010 & Rabi yields 2011 data and its correlation value is 0.98056243 which is shown in the graph Fig.4. In this graph types of crops were taken along x-axis and production in Tonnes was taken along y-axis. From the graph it is revealed that Paddy yield has reduced from 2010 to 201 where as Maize crop has increased slightly from 2010 to 2011 in Rabi season. There is a correlation between Kharif yields 2010 &Kharif yields 2011 data and its correlation value is 0.893580148 which is shown in the graph Fig.5.In this graph, types of crops were taken along x-axis and production in Tonnes was taken along y-axis. From the graph it is revealed that Paddy & Cotton yield were reduced from 2010 to 2011 and sugarcane and turmeric were increased from 2010 to 2011 in Kharif season. Interesting measures were found when Association rule Apriori was applied on the clustered spatial data set and the rules were shown in fig.6 and fig.7. It was found that Area_Sown, Normal_area, geo_latitude and Cluster2 were associated in Kharif 2010 clustered data set. It was found that Normal_area, Area_Sown and Productivity were associated in Kharif Yield 2010 cluster data set. Initially discretization was done on the clustered data set as a hybrid technique later apriori association rule was applied.

Future scope of this hybrid approach can be extended to various agricultural spatial locations and also to various agricultural yields for effective analysis and useful conclusions.

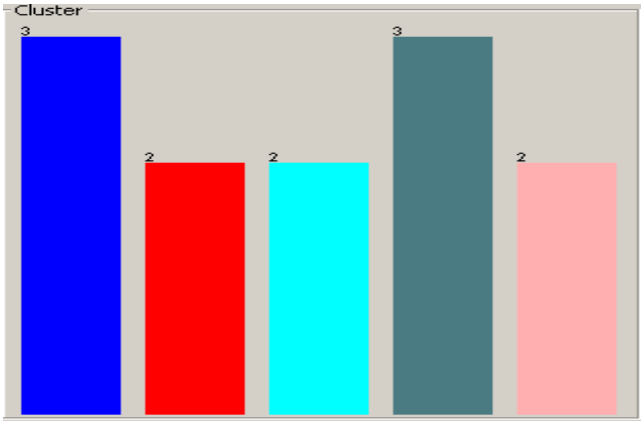

Fig. 1: Clusters Bar chart graph

Table 1: Cluster Centroids

Attribute	Full Data (12)	Cluster# 0 (3)	1 (2)	2 (2)	3 (3)	4 (2)
Crop		Paddy	Soyabean	Greengram	Maize	Paddy	Turmeric
Normal_area	75735.1667	2960.6667	10914	46390	247514	21395
Area_Sown	84006	1765.3333	8822.5	29436.5	302506	9370
Productivity	10455	1833.3333	542.5	43675	5356.6667	7727.5
Production	180201.1	4556.3333	4766.5	158452.5	517616.7	134728.05

Table 2: Detailed Accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.182	0.333	1.000	0.500	0.522	0.909	0.333	Paddy
1.000	0.091	0.500	1.000	0.667	0.674	0.955	0.500	Maize
1.000	0.091	0.500	1.000	0.667	0.674	0.955	0.500	Greengram
0.000	0.000	0.000	0.000	0.000	0.000	0.955	0.500	Redgram
1.000	0.182	0.333	1.000	0.500	0.522	0.909	0.333	Soyabean
0.000	0.000	0.000	0.000	0.000	0.000	0.909	0.333	Cotton
0.000	0.000	0.000	0.000	0.000	0.000	0.909	0.333	Chillies
0.000	0.000	0.000	0.000	0.000	0.000	0.955	0.500	Sugarcane
1.000	0.091	0.500	1.000	0.667	0.674	0.955	0.500	Turmeric
0.000	0.000	0.000	0.000	0.000	0.000	0.909	0.333	Castor
0.000	0.000	0.000	0.000	0.000	0.000	0.955	0.500	Others
0.000	0.000	0.000	0.000	0.000	0.000	0.909	0.333	Total
0.417	0.053	0.181	0.417	0.250	0.256	0.932	0.417	Weighted Avg.

Table 3: Confusion Matrix

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
1	0	0	0	0	0	0	0	0	0	0	0	a = Paddy
0	1	0	0	0	0	0	0	0	0	0	0	b = Maize
0	0	1	0	0	0	0	0	0	0	0	0	c = Greengram
0	0	1	0	0	0	0	0	0	0	0	0	d = Redgram
0	0	0	0	1	0	0	0	0	0	0	0	e = Soyabean
1	0	0	0	0	0	0	0	0	0	0	0	f = Cotton
0	0	0	0	1	0	0	0	0	0	0	0	g = Chillies
0	1	0	0	0	0	0	0	0	0	0	0	h = Sugarcane
0	0	0	0	0	0	0	0	1	0	0	0	i = Turmeric
0	0	0	0	1	0	0	0	0	0	0	0	j = Castor
0	0	0	0	0	0	0	0	1	0	0	0	k = Others
1	0	0	0	0	0	0	0	0	0	0	0	l = Total

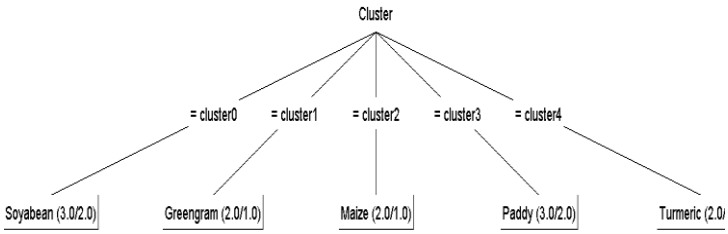

Fig. 2: Decision tree generated by J48 algorithm with cluster as high ranked attribute is crop

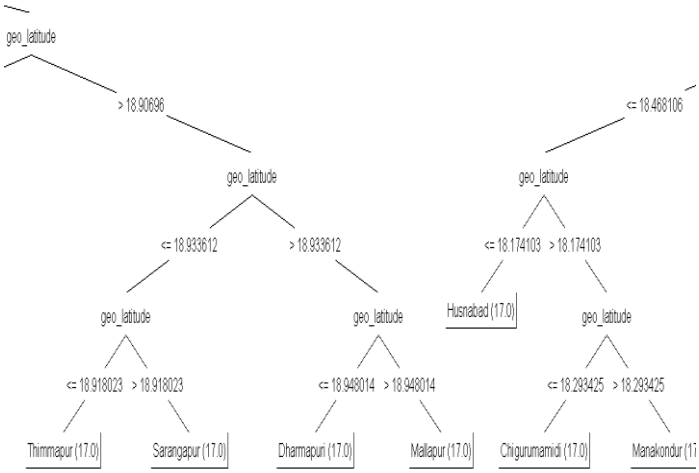

Fig. 3: Decision tree generated by J48 algorithm with cluster as high ranked attribute as mandal for Kharif 2010 cluster data set with latitude and longitude

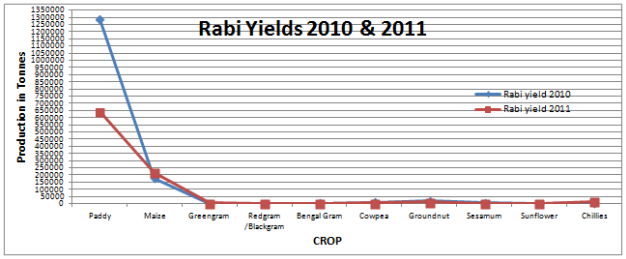

Fig.4: Correlation Graph between Kharif Yields 2010 & 2011

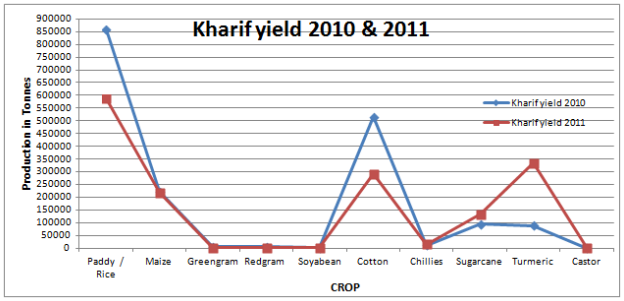

Fig.5: Correlation Graph between Kharif Yields 2010 & 2011

Best rules found Kharif 2010 clustered with latitude and longitude values are considered

1. Area_Sown='(-inf-1013.4]' 247 ==>Normal_area='(-inf-908.3]' 205 <conf:(0.83)> lift:(2.32) lev:(0.12) [116] conv:(3.69)

2. geo_latitude='(18.232532-18.324939]' 153 ==> Cluster=cluster2 106 <conf:(0.69)> lift:(2.65) lev:(0.07) [66] conv:(2.36)

3. Normal_area='(-inf-908.3]' 347 ==>Area_Sown='(-inf-1013.4]' 205 <conf:(0.59)> lift:(2.32) lev:(0.12) [116] conv:(1.81)

4. Cluster=cluster0 205 ==>Normal_area='(-inf-908.3]' 109 <conf:(0.53)> lift:(1.48) lev:(0.04) [35] conv:(1.36)

Best rules found Kharif Yield 2010 clustered

1. Area_Sown='(-inf-50492.7]' 8 ==>Normal_area='(-inf-46026.1]' 8 <conf:(1)> lift:(1.5) lev:(0.22) [2] conv:(2.67)

2. Normal_area='(-inf-46026.1]' 8 ==>Area_Sown='(-inf-50492.7]' 8 <conf:(1)> lift:(1.5) lev:(0.22) [2] conv:(2.67)

3. Area_Sown='(-inf-50492.7]' Productivity='(-inf-8678.5]' 6 ==>Normal_area='(-inf-46026.1]' 6 <conf:(1)> lift:(1.5) lev:(0.17) [2] conv:(2)

4. Normal_area='(-inf-46026.1]' Productivity='(-inf-8678.5]' 6 ==>Area_Sown='(-inf-50492.7]' 6 <conf:(1)> lift:(1.5) lev:(0.17) [2] conv:(2)

References

[1] Pavel Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, Inc. [2] Kaufman, L. and Rousseeuw, P.J., 1987, Clustering by Means of Medoids, In Y. Dodge, editor, Statistical Data Analysis, based on the L1 Norm, pp. 405- 416, Elsevier/North Holland, Amsterdam.

[3] Hartigan, J. A. Clustering algorithms. John Wiley and Sons., 1975.

[4] Selim, S.Z., and Ismail, M.A. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. In IEEE transactions on pattern analysis and machine learning, vol. PAMI-6, no. 1, January, 1984.

[5] Dhillon I. and Modha D., Concept Decomposition for Large Sparse Text Data Using Clustering.Machine Learning.42, pp.143-175. (2001).

[6] FORGY, E. 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. Biometrics, 21, 768-780.

[7] Yugalkumar and G. Sahoo, Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA.

[8] Dr.T. V. RajiniKanth, Ananthoju Vijay Kumar, Estimation of the Influence of Fertilizer Nutrients Consumption on the Wheat Crop yield in India- a Data mining Approach, 30 Dec 2013, Volume 3, Issue 2, Pg.No: 316-320, ISSN: 2249-8958 (Online).

[9] Dr.T. V. RajiniKanth, Ananthoju Vijay Kumar, A Data Mining Approach for the Estimation of Climate Change on the Jowar Crop Yield in India, 25Dec2013,Volume 2 Issue 2, Pg.No:16-20, ISSN: 2319-6378 (Online).

[10] A. Vijay Kumar, Dr. T. V. RajiniKanth "Estimation of the Influential Factors of rice yield in India" 2nd International Conference on Advanced Computing methodologies ICACM-2013, 02-03 Aug 2013, Elsevier Publications, Pg. No: 459-465, ISBN No: 978-93-35107-14-95

[11] Ramesh Chand, S.S. Raju, Instability in Andhra Pradesh agriculture -A Disaggregate Analysis, Agricultural Economics Research Review Vol. 21 July-December 2008 pp283-288.

[12] D. Hand, el al., Principles of Data Mining. Massachusetts: MIT Press, 2001.

AUTHORS

Ch. Mallikarjuna Rao

Received his B.Tech degree in computer Science and engineering from Dr.Baba sahib Ambedkar Marathwada University, Aurangabad, Maharastra in 1998,and M.Tech Degree in Computer Science and Engineering from J.N.T.U Anantapur ,Andhrapradesh in 2007. He is currently pursuing his Ph.D degree from JNTU Ananthapur University, Andhra Pradesh. Currently he is

working as Associate Professor in the department of Computer Science and Engineering of Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.His research interest includes Data bases and data mining.

Dr. Ananda Rao Akepogu

Received his B.Tech degree in ComputerScience&Engineering from University of Hyderabad, Andhra Pradesh, India and M.Tech degree in A.I & Robotics from University of Hyderabad, Andhra Pradesh, India. He received Ph.D degree from Indian Institute of Technology Madras, Chennai, India. He is Professor of ComputerScience& Engineering Department and currently working as Principal of JNTUA College of Engineering, Anantapur, Jawaharlal Nehru Technological University, Andhra Pradesh, India. Dr. Rao published more than 100 publications in various National and International Journals/ Conferences. He received Best Research Paper award for the paper titled "An Approach to Test Case Design for Cost Effective Software Testing" in an International Conference on Software Engineering held at Hong Kong, 18-20 March 2009. He also received Best Educationist Award for outstanding achievements in the field of education by International Institute of Education & Management, New Delhi on 21st Jan. 2012. He bagged Bharat Vidya Shiromani Award from Indian Solidarity Council and Rashtriya Vidya Gaurav Gold Medal Award from International Institute of Education & Management, New Delhi on 19th March, 2012. Dr.Rao got Best Computer Science and Engineering Faculty award from ISTE for the Year 2013. His main research interest includes software engineering and data mining.

N. Madhusudhana Reddy

Received his B.Tech degree in computer Science and engineering from Jawaharlal Nehru Technological University, Hyderabad in 1999, and M.Tech Degree in Computer Science from University of Hyderabad, Hyderabad, Andhra Pradesh in 2002. He is currently pursuing his Ph.D degree from JNTU Ananthapur University, Andhra Pradesh. Currently he is working as Associate Professor in the department of Computer Science and Engineering of Syamaladevi Institute of Technology for women, Nandyal, Andhra Pradesh. His research interest includes Big data, Data mining, and cloud computing.