International Journal of Computer Science
and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 5, Issue 2, Feb – Mar 2016, 5-14
© IASET

International Academy of Science,
Engineering and Technology
IASET Connecting Researchers; Nurturing Innovations

EFFECTIVE ANALYSIS OF LAND SURFACE WATER RESOURCES OF ANDHRA PRADESH WITH ROUGH SET BASED HYBRID DATA MINING TECHNIQUES USING R

S. NAGINI[1], T. V. RAJINIKANTH[2] & B. V. KIRANMAYEE[3]

[1]Department of CSE, VNRVJIET, Hyderabad, India

[2]Departments of CSE, SNIST, Hyderabad, India

[3]Department of CSE, VNRVJIET, Hyderabad, India

ABSTRACT

Agriculture plays an important role in economy of India. More than half of the population in India depends on Agriculture. It provides raw material for many industries. Early more than half of the land mass is used for Agriculture and over the years there is decline in agriculture land. Various factors like urbanization and development results in the growth of Non-Agriculture land year by year. Agriculture is the largest abstractor and prime consumer of ground water resources across globe and hence study of agro-economies that is ground water dependent became widely popular. Agriculture Irrigation, Surface water and Ground water resources are interlinked to each other. Water Usage and Food Production are dependent on each other extensively. Water is the major parameter that controls the crop yield. Many countries agriculture yield depends on rain fall. Many at time's rainfall is not sufficient to crop yields. It made researchers to do rigorous analysis on water resource availability and suggest farmers for its effective utilization. This paper aims at development and application of new Hybrid Data Mining (HDM) Techniques for effective analysis of Land Surface Water Resources (LSWR) like Canals, Tube wells, Tanks and other water resources. Apart from that analysis is also made on Various Agriculture yield's i.e., both for Cereals and Millets namely Kharif, Rabi, Sugarcane, Maize, Ragi, Wheat, Barley, etc., using new Hybrid Data Mining (HDM) techniques. To model the complex logic, Decision Tables (DT) is used. The results were proved to be good when new Rough Set Based Hybrid Data Mining (RSBHDM) Techniques are applied over the refined data sets.

KEYWORDS: Agriculture, Hybrid Data Mining (HDM) Techniques, Décision Tables (DT), Land Surface Water Ressources (LSWR), Rough Set Based Hybrid Data Mining (RSBHDM) Techniques

INTRODUCTION

India's economy depends on Agriculture. Most of the industries depend on agro products for raw material. Water Usage and Food Production [Ref.9] are dependent on each other extensively. Water is the prime factor for controlling the crop yield. Many countries agriculture yield depends on rain fall. Many a time's rainfall is not sufficient to crop yields. To examine the pathways for increasing efficiency and productivity of water use, the yield response of crops to water must be known. Over the last three and half decades, new knowledge has en-lighten processes underlying the relationship between crop yield and water use and technology has improved. Further, novel needs have emerged related to the planning and management of water in agriculture, including those arising from climate change. The interactions between surface water, agricultural irrigation [Ref.10] and ground water resources are often very close such that active cross-sector dialogue and combined vision is essential to promote sustainable water and land management. Precision Agriculture [Ref.1] deals with small variations in crop production based on the observation of factors like growth, assessment and a timely response. Site Specific Crop Management is related to Precision Algorithm as it involves with Spatial Location, Crop, Climate and Decision Support systems.

Widespread use of spatial databases and tremendous growth in spatial datasets leaves scope for discovery of spatial knowledge through automated techniques. Guting [5]. Knowledge exploration became vital in spatial databases since vast spatial datasets are obtained from X-ray crystallography, satellite images or from any other automatic equipment. Shashi Shekar ET. al. [6, 7] proposed Spatial Data Mining as a process of mining interesting and unknown knowledge that produces potentially useful patterns. Extracting useful and interesting patterns from spatial datasets involves more complexity (because of spatial data types, spatial relationships and the spatial data correlations) when compared with the extraction of interesting patterns from categorical and traditional numeric data. Efficient tools [7] for extracting meaningful patterns from geospatial data becomes very important for all the organizations that make decisions based on wide spatial datasets, including National Cancer Institute , NASA, National Imagery and Mapping Agency and the USDOT. Such organizations work on application domains like environmental management, ecology, safety of public, field of earth science, transportation, epidemiology, climatology and agriculture. In decision trees, the data is denoted in a hierarchical tree, where each leaf refers to a particular concept and contains a mathematical description of the given concept. Naohisa Koide, Amor V.M Ines,Andrew W.Robertson, David G.De and Jian-Hua Qian [Ref.8] Witt in the reference paper made an attempt to predict the rice production using seasonal climate forecasts. They found that the area harvested positively correlates with rainfall during the preceding dry season, whereas yield positively correlated with the rainfall from June to September and negatively correlated with the rainfall from October to December based on the harvested year respectively. Data mining [Ref.11, 12, 13] is the process of applying various methods that enables in discovering useful hidden patterns from the large data sets. Data mining fills up the gap from artificial intelligence and applied statistics to the databases by exploiting the storage and indexed features that could enable discovery of important patterns through various learning algorithms more efficiently.

LTERATURE SURVEY

The raw data sets have lots of missing attribute values. Handling of these missing attribute values [Ref.2] is a challenging task for present day data analyst. Polish computer scientist Zdzisalw I. Pawlak, said that rough set is an approximation formally applied on crisp set as set of pairs that give the upper and lower approximation of the original set. Rough set theory [Ref.15] deals with imperfect knowledge e.g., being fuzzy sets, evidence theory etc.,. Imperfect knowledge became a crucial issue for scientists in computer field, specifically in the Artificial Intelligence domain. Wide range of problem solving techniques is available to understand and analyze the imperfect knowledge. Fuzzy set theory is one of the most successful approaches proposed by Zadeh [Ref.16]. Future prediction was made on missing attribute values of the data set using Rough set Approach, here missing values are assigned the most common values of the attribute domain, and the obtained results are compared with random tree classifier the case where the missing attribute values are ignored. Performance of the Random tree classifier in terms of results was found to be the best. Rough Set [Ref.3, 4] concept is used as mathematical basis for the Decision Table Theory. A decision table is a system defined by S = (CA, DA, V, U, f) where CA, DA the condition and decision attributes respectively and there are finite sets such that $CA \cup DA \neq 0$, $CA \cap DA = 0$. V is the domain of attributes, f is a decision function and U is a non-empty finite set. A decision table is the best tool used in performing requirement management and testing. While dealing with complex rules of business a well structured exercise to formulate requirements becomes mandatory. Decision tables are used to model complicated logic. In machine learning and statistics, discretization is process of partitioning or conversion of continuous attributes, variables or features to discretized or nominal variables / features / attributes/intervals. Naïve Bayes [Ref.14] Algorithm Performance increased significantly by entropy based Discrediting. Continuous values are discretized during the learning process. Discretization is classified in to three categories like Local vs Global, Unsupervised vs supervised and Dynamic vs Static.

The most widely used K-means clustering technique can be successfully applied in domains such as gene expression, relational databases and decision support. The drawback of K-means is in the choice of centroid locations randomly at the beginning of the algorithm, mapping of numbers to variables and finally the unknown number of clusters 'K'. The impact of first drawback can be assessed through specific initialization or multiple runs methods. However, the specific initialization methods are not better than random centroids. There is a possible solution to the second drawback categorical parameters can be handled using matching dissimilarity measure. The third drawback can be addressed through the use of cluster validity indices. As many other data mining algorithms, K-means reliability reduces when working on high-dimensional data because datasets are almost too sparse. In data classification, a model or description for each class in a dataset is developed; the leaf should be derived from information other than TC. Where TC is a test chosen from mixed cases, based on a single attribute, that has one or more mutually exclusive outcomes {OT_1, $OT_{2, ., . . . ,}$ OT_n}, TC is partitioned into subsets $TC_{1,}$ TC_2, . . . , TC_n , where TC_i contains all the cases in TC having outcome OT_i of the chosen test. The decision tree for TC consists of a decision node that identifies the corresponding test, and one branch for each possible outcome. The same building model of tree is recursively applied on subsets of training cases. **I**n recent years, Support Vector Machines (SVM) [Ref.17] with nonlinear or linear kernel has become one of the most vital learning algorithms for regression and for classification that are the fundamental tasks in data mining. Based on the use of kernel mapping, variants of SVMs have successfully merged flexible and effective nonlinear models. A Support Vector Machine (SVM) [Ref.18] does classification by building an *N*-dimensional hyper plane that derives two categories of data that are optimally separated. SVM models are tightly coupled with neural networks. A two-layer perception neural network can be equated to an SVM model that uses sigmoid kernel function. An association rule defines relationships between set of objects taken from dataset. For a set of transactions that involve a set of items, an association rule expression takes the form X1 →Y1, where X1 and Y1 are two items. The inference from such a rule is that transactions that contain X1 will also tend to contain Y1. An example, "40%"of transactions that contain item1 will also tend to contain item 2; 3% of all transactions contain both item1 and item 2. Here 40% is called the confidence and 3% is called support. The problem discovers all association rules that equates to minimum confidence and minimum support values specified by user as input.

PROPOSED APPROACH

In the proposed approach initially the Land Surface water resource raw data set is pre-processed for removal of redundancy, filling of missing attribute values with suitable mean values, etc., and molded into required format. Then apply Rough set based hybridization of Data Mining (RSBHDM) Techniques on the preprocessed Land Surface water resource data sets. The results thus obtained were analyzed effectively using Hybrid Data Mining Techniques. It has proved there is a substantial progress in performance.

IMPLEMENTATION OF PROPOSED METHODOLOGY

The implementation Procedure is shown in the diagram Figure 1. Initially the raw data set is Pre-processed and converted in to the required format. That pre-processed data set is converted in to Rough Set Based Decision Table (RSBDT) using R. The decision Table and then Discretized and the resultant data set is named as Rough set based Descriptive Decision table (RSBDDT). Info Gain Attribute Evaluation procedure is applied along with Ranker Algorithm is applied and attributes selection was done. This concept finds the value of an attribute by measuring information gain for a given class. The RSBDDT is named as dec2.csv Table subjected to three cases

Case I: When two classifier techniques SVM and J48 are applied on the dec2.csv directly without dividing it into Training and Testing data sets as shown in Table 1 and the corresponding graph is shown in Figure 2.

Case II: Initially the dec2.csv is divided in to two data sets namely Training and testing data sets which are Non-clustered and their performances are noted and shown in the Table 2 and the graph is shown in Figure 2.

CASE III: Initially the dec2.csv is subjected Simple K-Means Clustering Algorithm with 5 clusters. The clusters are Cluster0, Cluster1, Cluster2, Cluster3 and Cluster4. The resultant Clustered data set is divided in to two data sets namely Training and testing data sets with 80% and 20% respectively. Then the classifiers SVM and J48 were applied and their performances are shown in the Table 2 and graph is shown in the Figure 2. The Association Algorithm named Apriori is applied on the Resultant Classifier Data sets namely SVM Classifier data set and decision Tree J48 classifier data set. Twenty rules were generated which were shown in the Figure 4.

RESULTS AND ANALYSIS

The results have been analyzed and summarized in TABLE 1, TABLE 2, Figure 2 and Figure 3. In TABLE 1 it is found that for the parameters time to build and time to test, J48 is better than SVM. The performance parameters correctly classified instances and Kappa Statistic values are observed, SVM is proved to be better than J48 Classifier on Non-clustered data set.

TABLE 2 shows that for the parameters time to test and time to build, SVM is relatively poor than J48. The performance parameters are correctly classified instances and further Root Mean Square and Kappa Statistic are observed, here SVM is proved to be better than J48 Classifier on Non-clustered data set for trained data set. Coming to the Test data set also SVM is proved to be better than J48 classifier.

Classifier on clustered data set that has been trained , it is interesting that on the test data J48 result has improved a lot because of clustered data set. Apart from that one more fact is observed that the Apriori Association rules generated are same when applied on both SVM and J48 Classifier data sets. These association rules are shown in Figure 4. A resultant data set is formed when discretization technique is applied on the Pre-processed data set without subjecting it to Decision Table conversion. The Association rules are not generated when the Apriori Algorithm is applied on the resultant discretized data set which implies that influence of Decision Table generation is required. So the decision table is also important for generating Association rules. In graphs represented by Figure 2 and Figure 3 performance parameters are taken on X-axis and performance values are taken on Y-axis.

CONCLUSIONS

The classifier results on Clustered test data set are good when compared to the classifier results on Non-clustered test data set. There is not a much of difference for SVM classifier results of test and Trained clustered data set. Whereas as for the J48 classifier results there are 100% improvement on Test data when compared with trained clustered data set.

Apart from that it is also found that the Hybridization Data mining techniques proved to be good when compared non hybridized Data mining techniques. Rough set based Discretized Decision table data sets also enhanced the performance of the Hybrid Data Mining Techniques. The results in Figure 2 and Figure 3 shows that Root relative Squared error, Mean Absolute error, Root relative mean squared error and Relative Absolute error are zero values for SVM where as for J48 shows that they are non zeros. The Association rules are stating that there is a dependency among various attribute of land surface water resources like Canals, Tube wells, Wells, Tanks, Gross Area irrigated from Wells, Wells and Tanks, Other Wells.

FIGUREURES AND TABLES

- TUBE.WELLS==> WELLS.TANK
- WELLS.TANK==> TUBE.WELLS
- GAIFW.TANK==> WELLS.TANK
- WELLS.TANK==> GAIFW.TANK
- GAIFW.TUBE.WELLS==> WELLS.TANK
- WELLS.TANK==> GAIFW.TUBE.WELLS
- GAIFW.TANK==> TUBE.WELLS
- TUBE.WELLS==> GAIFW.TANK
- GAIFW.TUBE.WELLS==> TUBE.WELLS
- TUBE.WELLS==> GAIFW.TUBE.WELLS
- GAIFW.Other.wells==> OTHER.WELLS
- OTHER.WELLS==> GAIFW.Other.wells
- GAIFW.TUBE.WELLS==> GAIFW.TANK
- GAIFW.TANK==> GAIFW.TUBE.WELLS
- TUBE.WELLS, GAIFW.TANK==> WELLS.TANK
- WELLS.TANK, GAIFW.TANK==> TUBE.WELLS
- WELLS.TANK, TUBE.WELLS==> GAIFW.TANK
- GAIFW.TANK==> WELLS.TANK, TUBE.WELLS
- TUBE.WELLS==> WELLS.TANK, GAIFW.TANK
- WELLS.TANK==> TUBE.WELLS, GAIFW.TANK

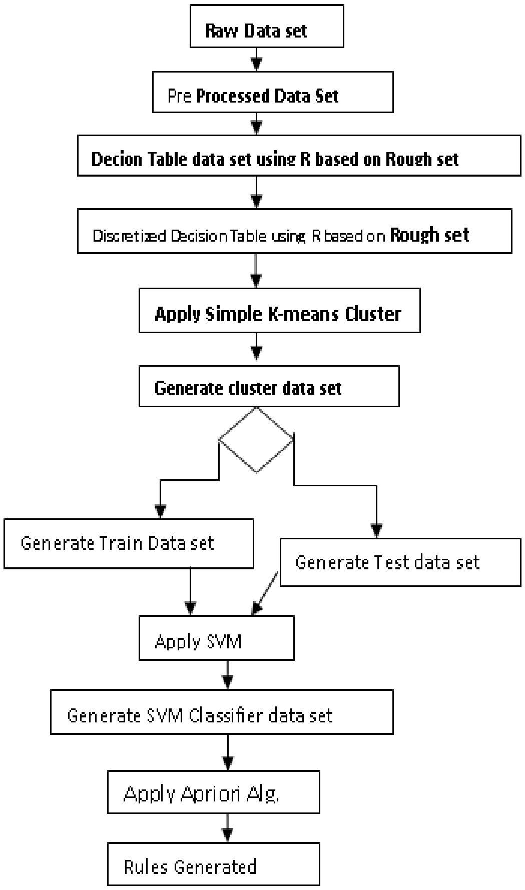
Raw Data set
Pre Processed Data Set
Decion Table data set using R based on Rough set
Discretized Decision Table using R based on Rough set
Apply Simple K-means Cluster
Generate cluster data set
Generate Train Data set
Generate Test data set
Apply SVM
Generate SVM Classifier data set
Apply Apriori Alg.
Rules Generated

Figure 1: Data Flow Diagram

Table 1: Performance of Classifiers on Non-Clustered Data Set

Parameters	SVM on Non Clustered Data Set	J48 on Non Clustered Data Set
Time to Build model	0.52 sec	0.06 sec
Time to Test model	0.08 sec	0 sec
Correctly Classified	100 %	57.143%
In Correctly Classified	0.0%	42.86%
Kappa Statistic	1.0	0.5386
Mean Absolute error	0.0	0.0613
Root mean Squared error	0.0	0.176
Relative absolute error	0.0	46.1544%
Root relative squared error	0.0	67.9367%

Table 2: Performance of SVM & J48 Classifiers on Non-Clustered and Clustered Data Sets

Parameters	SVM on Non Clustered Data Set	SVM on Clustered Data Set	J48 on Non Clustered Data Set	J48 on Clustered Data Set
Time to Build model	0.36 sec	0.31sec	0 sec	0
Time to Test model	0 sec	0.02sec	0 sec	0
Correctly Classified TR	100%	100%	45.455%	72.727%
Correctly Classified TE	0%	100%	0%	100%
In Correctly Classified TR	0%	0%	54.546%	27.273%
In Correctly Classified TE	100%	0%	100%	0%
Kappa Statistic TR	1.0	1.0	0.4	0.6163
Kappa Statistic TE	0.0	1.0	0.0	1.0
TR(Mean Absolute error)	0.0	0.0	0.0779	0.1552
TE(Mean Absolute error)	0.1429	0.0	0.1429	0.116
Root mean Squared error	0.0	0.0	0.1974	0.2786
Root mean Squared error	0.378	0.0	0.3021	0.187
Relative absolute error	0.0%	0.0%	59.289%	50.945%
Root relative squared error	0.0%	0.0%	77.201%	71.949%

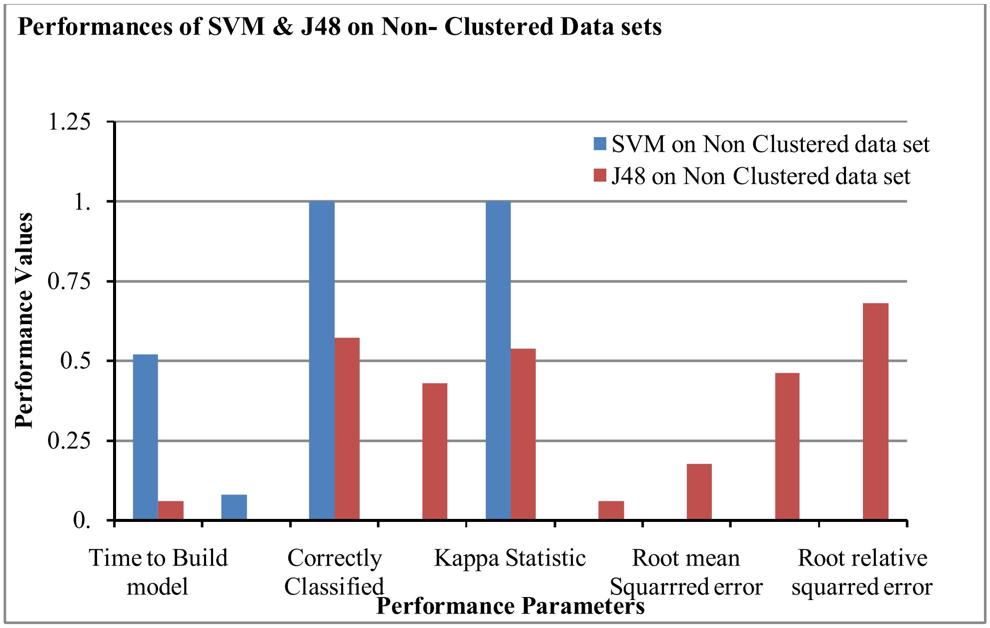

Figure 2: Performances of SVM & J48 on Non-Clustered Data Sets

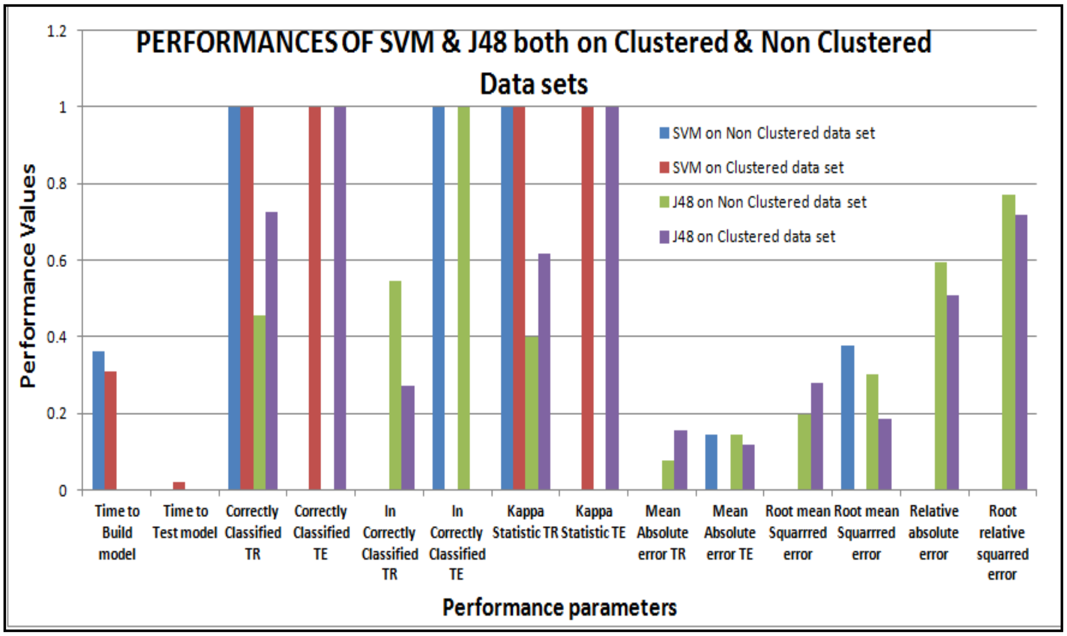

Figure 3: Performances of SVM & J48 both on Clustered & Non-Clustered Data Sets

REFERENCES

1. Sudhanshu Sekhar Panda, Gerrit Hoogenboom and Joel O. Paz, Remote Sensing and Geospatial Technological Applications for Site-specific Management of Fruit and Nut Crops: A Review, *Remote Sensing***2010**,*2*,1973-1997; doi:10.3390/ rs2081973.

2. M. Sandhya, Dr. A. Kangaiammal, Dr. C. Senthamarai, A Comparative Study on Decision Rule Induction for incomplete data using Rough Set and Random Tree Approaches, OSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, PP 06-10, Volume 9, Issue 3(Mar. -Apr. 2013).

3. Zdeislawpawlak Decision Tables- A Rough Set Approach, Department of Complex control Systems, Polish Academy of Sciences, Poland.

4. Rough Sets And Decision Tables, Z. Pawlak, Institute of Computer Science Polish Academy of Sciences.

5. Guting, "An Introduction to Spatial Database Systems", *VLDB 1994*

6. Shashi Shekar, Pusheng Zhang, Ranga Raju Vatsavai, "Research Accomplishments and Issues on Spatial Data Mining"

7. Shashi Shekhar, Pusheng Zhang, Yan Huang, Ranga Raju Vatsavai, Spatial Data Mining

8. Amor V. M. Ines, James W. Hansen, Andrew W. Robertson, Enhancing the utility of daily GCM rainfall for crop yield prediction

9. Crop Yield Response To Water, Fao Irrigation And Drainage Paper, ISSN 0254-5284.

10. Ground water Resources and Irrigated Agriculture – making a beneficial relation more sustainable, Perspectives Paper, Global Water Partnership. www.gwp.org

11. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.
12. Tarun Rao, N.Rajasekhar, T.V.Rajinikanth, Sundar K S, Classification of Remote Sensed Data Using Linear Kernel Based Support Vector Machines.
13. Zaibin Liu, Dewu Jin, Qisheng Liu, Prediction of Water Inrush Through Coal Floors Based on Data Mining Classification Technique, Procedia Earth and Planetary Science, Volume 3, 2011, Pages 166-174, ISSN1878-5220, 10.1016/j.proeps .2011. 09. 079.
14. James Dougherty, Ron Kohavi, Mehran sahami, Supervised and Un-Supervised Discretization of Continuous Features. Computer Science department, Stanford University.
15. Zdzislaw Pawlak, Rough Sets University of Information Technology and Management, Poland
16. L. Zadeh: Fuzzy sets, Information and Control, 8, 338-353, 1965
17. VAPNIK, V. Statistical Learning Theory. New York: Wiley, 1998.
18. http://www.dtreg.com/mlfn.htm

APPENDICES

Authors Biography

S.Nagini, Associate Professor in CSE, VNRVJIET, Hyderabad, is currently pursuing Ph.D. degree in Computer Science and Engineering in Acharya Nagarjuna University. She has total teaching experience of 18 years. Her specialized area of research is Data Mining

Dr. T. V. Rajini Kanth has obtained his Ph.D. degree in C.S.E branch from Osmania University, Hyderabad in July, 2008 and M. Tech. (C.S.E.) degree from Osmania University, Hyderabad in January, 2001. His specialization area in research is “Spatial Data Mining”. He obtained his PGDCS degree from HCU, Hyderabad in 1996. He received his M. Sc. (Applied mathematics) degree in the year 1989 from S.V. University, Tirupati as University Ranker.

He is working as a Professor in CSE department, Sreenidhi Institute of Science and Technology (SNIST), Hyderabad. Before that, he worked as a Professor and Head, Department of IT, GRIET, Hyderabad. Before that, he worked as a Professor in CSE department, GRIET, Hyderabad since 2007 November. Prior to that, he worked as Associate Professor in VNRVJIET, Hyderabad. He joined in VNRVJIET in 1996. His total teaching experience is 22 years. His writings have appeared in numerous Professional conferences and Journals (International journals-21, national level-4,). He was an author of a few books, i.e. Artificial Intelligence, etc. His current research area interests include Image processing, Data Warehousing & Mining, Spatial Data Mining, Web Mining, Text Mining and Robotic area, etc. He is presently, guiding research students in the research areas like Spatial Data Mining, Web Mining, Image Processing and Text Mining.

He has conducted two International conferences, namely ICACT-08, and ICACM-11 at GRIET, Hyderabad and also acted as session chair for many conferences. He was the convener for Pragnya-08, a national level technical Fest at GRIET. Apart from that, many National levels ROBO workshops were conducted, namely RoboTrix, eTrix, iTrix, LogiTrix, VisionTrix and Haptic Robotic arm in association with IIT Bombay and Technophilia etc. Many projects were guided and developed as products some of them are Emolument management system, Mobile based automation and gamming projects, journals maintenance project, Web crawler's project, Office automation projects, Vision based Robo projects and stegonography projects, etc., Apart from these presently guiding Ph.D. scholars at various universities namely, JNTUH, JNTUK, JNTUA and NU.

He is currently editor for two journals. He was called for about 50 AICTE sponsored workshops as resource person. Recently the Robo products display at 2nd Annual IT summit was appreciated by Hon. Minister Ponnala Lakshmaiah held at HICC, Hyderabad. He received seminar grants from AICTE and DST organizations. He is currently the Nodel Officer for academics, TEQIP at GRIET.

B.V. Kiranmayee, Associate Professor in CSE, VNRVJIET, and Hyderabad is currently pursuing Ph.D. degree in Computer Science and Engineering in JNTUH University. She has total teaching experience of 18 years. Her specialized area of research is Data Mining.