A Novel Approach for Text Classification

[1]S.Shanawaz Basha, [2]L.Sunitha Rani

[1,2]Dept. of CSE, Jawaharlal Nehru Technological University Anantapur, Kurnool, AP, India

Abstract

Text Classification (TC) is the process of associating text documents with the classes considered most appropriate, thereby distinguishing topics such as particle physics from optical physics. A lot of research work has been done in this field but there is a need to categorize a collection of text documents into mutually exclusive categories by extracting the concepts or features using supervised learning paradigm and different classification algorithms. In this paper, a new Fuzzy Similarity Based Concept Mining Model (FSCMM) is proposed to classify a set of text documents into pre - defined Category Groups (CG) by providing them training and preparing on the sentence, document and integrated corpora levels along with feature reduction, ambiguity removal on each level to achieve high system performance. Fuzzy Feature Category Similarity Analyzer (FFCSA) is used to analyze each extracted feature of Integrated Corpora Feature Vector (ICFV) with the corresponding categories or classes. This model uses Support Vector Machine Classifier (SVMC) to classify correctly the training data patterns into two groups; i. e., + 1 and – 1, thereby producing accurate and correct results. The proposed model works efficiently and effectively with great performance and high - accuracy results.

Keywords

Text Classification, Natural Language Processing,Feature Extraction, Concept Mining, Fuzzy Similarity Analyzer, Dimensionality Reduction

I. Introduction

Large amounts of data has been collected and stored in large data bases by database technologies and data collection techniques. For some applications only a small amount of the data in the databases is needed. This data is called knowledge or information. Data mining is the process of extracting knowledge from these large databases [1-2]. Data mining is also called knowledge discovery in databases or KDD process.

Although there have been many studies of data mining in relational and transaction databases [1,3], data mining is in great demand in other applicative databases, including spatial databases, temporal databases, object-oriented databases, multimedia databases, etc.

A. Classification

An object can be classified using its attributes. Each classified object is assigned a class. Classification is the process of finding a set of rules to determine the class of an object.

B. Clustering

Clustering means it is the process of grouping the database items in to clusters. All the members of the cluster has similar features. Members belong to different clusters has dissimilar features.

C. Feature Extraction

In text classification, the dimensionality of the feature vector is usually huge. Even more, there is the problem of Curse of Dimensionality, in which the large collection of features takes very much dimension in terms of execution time and storage requirements. This is considered as one of the problems of Vector Space Model (VSM) where all the features are represented as a vector of n - dimensional data. Here, n represents the total number of features of the document. This features set is huge and high dimensional.

There are two popular methods for feature reduction: Feature Selection and Feature Extraction. In feature selection methods, a subset of the original feature set is obtained to make the new feature set, which is further used for the text classification tasks with the use of Information Gain [5]. In feature extraction methods, the original feature set is converted into a different reduced feature set by a projecting process. So, the number of features is reduced and overall system performance is improved [6].

Feature extraction approaches are more effective than feature selection techniques but are more computationally expensive. Therefore, development of scalable and efficient feature extraction algorithms is highly demanded to deal with high-dimensional document feature sets. Both feature reduction approaches are applied before document classification tasks are performed.

D. Similarity Measure

In recent years, fuzzy logic [1-2, 4] has become an upcoming and demanding field of text classification. It has its strong base of calculating membership degree, fuzzy relations, fuzzy association, fuzzy production rules, fuzzy k-means, fuzzy c-means and many more concerns. As such, a great research work has been done on the fuzzy similarity and its classifiers for text categorization.

The categorizer based on fuzzy similarity methodology is used to create categories with a basis on the similarity of textual terms [4]. It improves the issues of linguistic ambiguities present in the classification of texts. So, it creates the categories through an analysis of the degree of similarity of the text documents that are to be classified. The similarity measure is used to match these documents with pre-defined categories [4-15]. Therefore, the document feature matrix is formed to check that a document satisfies how many defined features of the reduced feature set and categorized into which category or class [4, 6-7]. The fuzzy similarity measure can be used to compute such different matrices.

II. Related Works

Al-Mubaid and Umair used distributional clustering to generate an efficient representation of documents and applied a learning logic approach for training text classifiers. The Agglomerative Information Bottleneck approach was proposed by Tishby et al. The divisive information-theoretic feature clustering algorithm was proposed by Dhillon et al. which is an information-theoretic feature clustering approach, and is more effective than other feature clustering methods.

Feature clustering is an efficient approach for feature reduction which groups all features into some clusters, where features in a cluster are similar to each other. The feature clustering methods proposed in are “hard” clustering methods, where each word of the original features belongs to exactly one word cluster. Therefore each word contributes to the synthesis of only one new feature. Each new feature is obtained by summing up the words belonging to one cluster.

In general, there are two ways of doing feature reduction, feature selection, and feature extraction. By feature selection approaches, a new feature set W1= (w1; w2; . . . ; wk) is obtained, which is a subset of the original feature set W. Then W1 is used as inputs for classification tasks. Information Gain (IG) is frequently employed in the feature selection approach [10]. It measures the reduced uncertainty by an information-theoretic measure and gives each word a weight.

III. Proposed System

The proposed Fuzzy Similarity Based Concept Mining Model (FSCMM) is discussed. This model automatically classifies a set of known text documents into a set of category groups. The model shows that how these documents are trained step by step and classified by the Support Vector Machine Classifier (SVMC). SVMC is further used to classify various new and unknown text documents categorically.

The proposed model is divided into the two phases: Text Learning Phase (TLP) and Text Classification Phase (TCP). TLP performs the learning function on a given set of text documents. It performs the steps of first stage; i. e., Text Document Training Processor (TDTP) and then the steps of second stage; i. e., Fuzzy Feature Category Similarity Analyzer (FFCSA). The TDTP is used to process the text document by converting it into its small and constituent parts or chunks by using NLP concepts at the Sentence, Document and Integrated Corpora Levels. Then, it searches and stores the desired, important and non-redundant concepts by removing stop words, invalid words and extra words. In the next step, it performs word stemming and feature reduction. The result of sentence level preparation is low dimensional Reduced Feature Vector (RFV). Each RFV of a document is sent for document level preparation, so that Integrated Reduced Feature Vector (IRFV) is obtained. To obtain IRFV, all the RFVs are integrated into one. Now, Reduced Feature Frequency Calculator (RFFC) is used to calculate the total frequency of each different word occurred in the document. Finally, all redundant entries of each exclusive word are removed and all the words with their associated frequencies are stored in decreasing order of their frequencies. At the integrated corpora level, the low dimension Integrated Corpora Feature Vector (ICFV) is resulted.

In such a way, feature vectors at each level are made low dimensional by processing and updating step by step. Such functionality helps a lot to search the appropriate concepts with reduced vector length to improve system performance and accuracy.

FFCSA performs similarity measure based analysis for feature pattern (TD – ICFV) using the enriched fuzzy logic base. The membership degree of each feature is associated with it. Therefore, an analysis is performed between every feature of a text document and class.

SVMC is used for the categorization of the text documents. It uses the concept of hyper planes to identify the suitable category. Furthermore, SVMC accuracy is checked by providing some new and unknown text documents to be classified into the respective Category Group (CG). This task is performed in TCP.

The proposed Fuzzy Similarity Based Concept Mining Model (FSCMM) is shown in fig. 1. In the next sections, this model and its components are discussed in detail.

A. Text Learning Phase (TLP)

Consider a set of n text documents,

TD = {TD1, TD2, TD3,…,TDn} (1)

Where TD1, TD2, TD3,…,TDn are the individual and independent text documents which are processed, so that they can be categorized into the required category.

B. Text Document Training Processor (TDTP)

Text Document Training Processor (TDTP) prepares the given text document set TD of n text documents by performing many operations on the sentence, document, and integrated corpora levels. Firstly, each text document TDi, 1 ≤ i ≤ n, is processed at its sentence level. The result of such sentence level preprocessing for all the sentences of TDi is integrated into one, which is further processed and refined to make available for the integrated corpora. Integrated corpora accept and integrate all the refined text documents and perform more processing. Its result is sent to FFCSA.

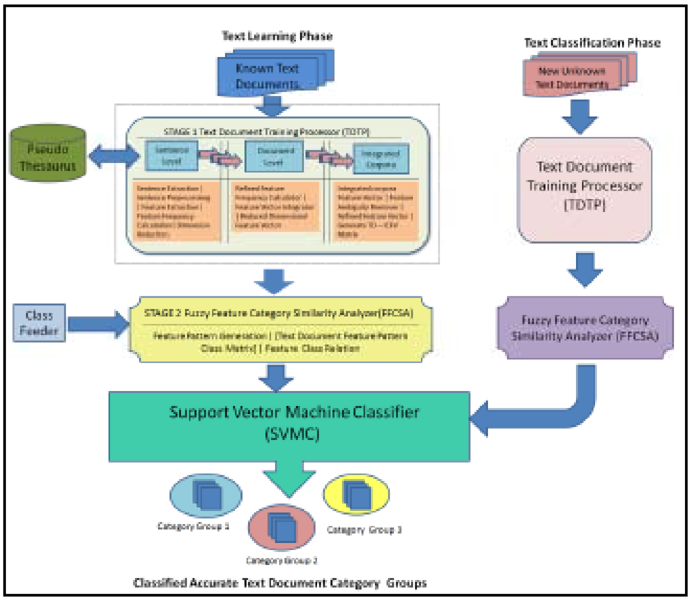

Text Learning Phase
Known Text Documents
Text Classification Phase
New Unknown Text Documents
Pseudo Thesaurus
STAGE 1 Text Document Training Processor (TDTP)
Sentence Level
Document Level
Integrated Corpora
Sentence Extraction | Sentence Preprocessing | Feature Extraction | Feature Frequency Calculation | Dimension Reduction
Refined Feature Frequency Calculator | Feature Vector Integration | Reduced Dimensional Feature Vector
Integrated corpora Feature Vector | Feature Ambiguity Removal | Refined Feature Vector | Generate TD – ICFV Matrix
Text Document Training Processor (TDTP)
Class Feeder
STAGE 2 Fuzzy Feature Category Similarity Analyzer(FFCSA)
Feature Pattern Generation | (Text Document Feature Pattern Class Matrix) | Feature Class Relation
Fuzzy Feature Category Similarity Analyzer (FFCSA)
Support Vector Machine Classifier (SVMC)
Category Group 1
Category Group 2
Category Group 3
Classified Accurate Text Document Category Groups

Fig. 1: The Fuzzy Similarity Based Concept Mining Model (FSCMM).

B. At Sentence Level

A text document TDi is composed of a set of sentences, so consider

TDi = {si1, si2, si3,…,sim} (2)

Where i denotes the text document number and m denotes the total number of sentences in TDi. Sentence Extractor (SE) is used to extract the sentence si1 from TDi. Each sentence has its well-defined and non-overlapping boundaries, which makes the sentence extraction a simple task for SE.

When the sentence is extracted, a verb – argument structure is made for si1. The syntax tree is drawn using the pre-defined syntactic grammar to separate the verbs and the arguments of the sentence. The sentence can be composed of the nouns, proper nouns, prepositions, verbs, adverbs, adjectives, articles, numerals, punctuation and determiners. So, with the construction of the syntax tree, the stop word and other extra terms are removed except the nouns, proper nouns and numerals which are considered as the concepts. To remove the invalid and extra words, the Pseudo Thesaurus is used. It also helps in word stemming.

The next step is to make the Feature Vector (FV) of the sentence sij of text document TDi as

FV = {Fi11, Fi12, Fi13, . . . ,Fi1r} (3)

Where 1 ≤ i ≤ n, 1 ≤ j≤ m, and r depicts the total number of present features in the sij. Feature Frequency Calculator (FFC) calculates the frequency of each different feature occurred in FV. Frequency represents the number of occurrences of a feature in the sentence. So, each different feature is associated with its frequency in the form of a Feature Frequency pair as <Fijk, freq (Fijk)>, where i is the text document number, j is the sentence number of the sentence, k is the feature number, freq () is a function to calculate the frequency of a feature, and $1 \leq k \leq r$.

The next step is to convert the high dimensional FV into low dimensional Reduced Feature Vector (RFV) to reduce the storage and execution time complexities. So, a counter loop is invoked to remove the duplicate or redundant entries of a feature. Therefore, only one instance of each different feature occurred is stored in RFV. It highly reduces the FV dimension and increases the efficiency of the system with good performance. The complete sentence level processing is shown in the fig. 2.

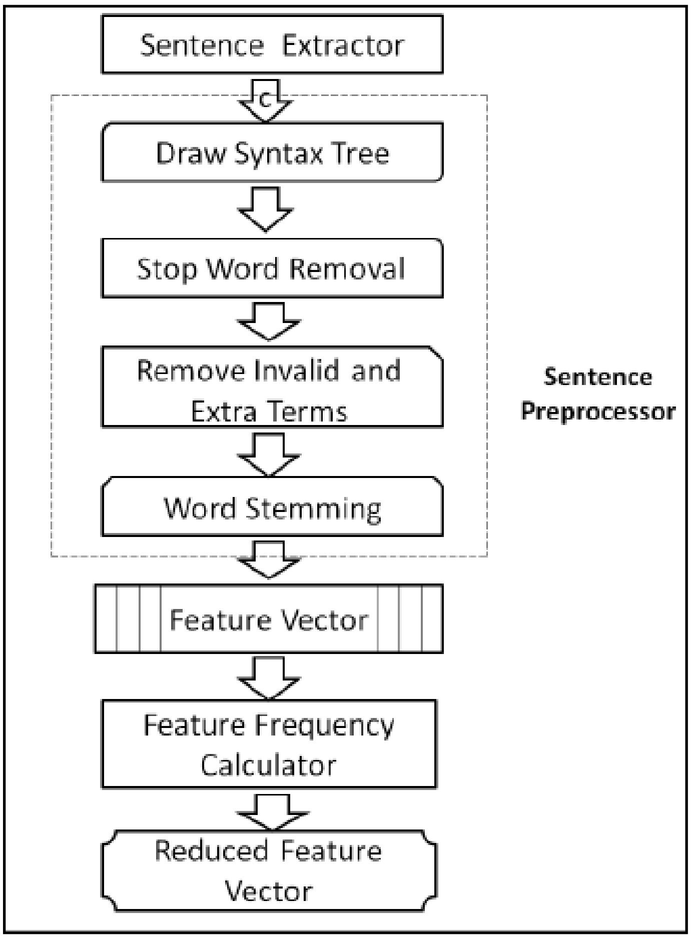


Fig. 2: Text Document Preparation at Sentence Level

.

Such sentence level preprocessing for the TDP is performed for each sentence of each document and then progressing toward the integrated corpora.

C. At Document Level

This step accepts the resultant RFV which is sent for document level pre-processing. Firstly, a counter loop is invoked to match the similar features in each of the RFVj with every other RFVq of a TDi where $1 \leq j, q \leq m, j \neq q$ and $1 \leq i \leq n$. Refined Feature Vector Calculator (RFVC) updates each feature's frequency for those features which are present two or more times in more than two sentences. These updates are done by adding up their frequencies in terms of combined calculated frequency of that feature only. In this way, it updates the count of each different occurred feature with more than one occurrence. The features that have occurred only once in the document will not update their frequency.

The next step is that all RFVs of a TDi are integrated into one as

IRFV = Integrat (RFV1, RFV2,…,RFVm) (4)

Where Integrat () is a function to combine all RFVs.

Now each RFVj is compared with every other RFVq where $1 \leq j, q \leq m$, and $j \neq q$. In this way, each feature of RFVj is compared with the every other feature of the RFVq and thereby, the duplicate and redundant features are removed. The complete procedure on the document level is shown in fig. 3.

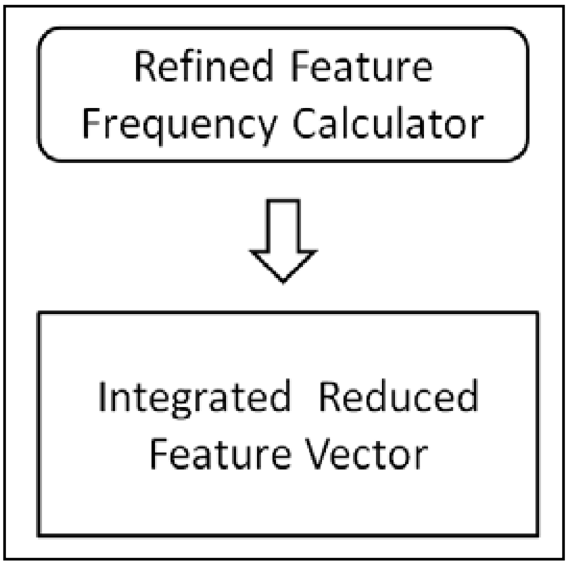

Refined Feature Frequency Calculator
Integrated Reduced Feature Vector

Fig. 3: Text Document Preparation at Document Level.

D. At Integrated Corpora

In integrated corpora, all the IRFVs of n documents are integrated into one. This step is used to calculate and update the final frequency of each different feature occurred in the corpora. Firstly, it removes duplicate and redundant feature entries of IRFV, and then removes all ambiguous words with the help of the pseudo thesaurus. In such a way, Integrated Corpora Feature Vector (ICFV) is generated. A Threshold Value (TV) is defined for ICFV. TV cuts off those features whose total frequency is less than TV. Finally, it reduces the dimension of ICFV.

Therefore, Integrated Corpora Feature Vector (ICFV) is constructed as

ICFV = {F1, F2,…,Fn} (5)

Where F1, F2,…,Fn represent the features.

Such features with their associated frequencies show the statistical similarity between the text documents. Each feature is counted for each document and represented as the feature and its frequency as follows.

TD1 = {<F1, f (F1)>, <F2, f (F2)>,…,<Fn, f (Fn)>} (6)

Where f(Fi) is the function to calculate the frequency of feature Fi in the text document.

The next step is to generate the matrix of TD and ICFV in the given form of table 1. In this matrix, the 0 represents the absence of that feature in TD and the numerical value represents the total number of occurrences of the feature in the TD.

Table 1: TD ICFV Matrix

Text Document	Feature		
	F1	F2	F3
TD1	1	0	1
TD2	1	3	0
TD3	0	2	1
TD4	4	0	0

E. Pseudo Thesaurus

The Pseudo Thesaurus is a predefined English Vocabulary Set which is used to check the invalid words or to remove extra words from a sentence while processing the sentence in the TDTP. It is also used for word stemming so that the exact word can be obtained. For example, consider three different words for the word research - researching, researcher and researches. When the word stemming is performed, research is the final resulting word with the feature frequency counted as 3.

F. Class Feeder (CF)

Text Classification is the process of assigning the name of the class to a particular input, to which it belongs. The classes, from which the classification procedure can choose, can be described in many ways. So classification is considered as the essential part of many problems solving tasks or recognition tasks. Before classification can be done, the desired number of classes must be defined.

G. Fuzzy Fetaure Category Similarity Analyzer (FFCSA)

In FFCSA, firstly the Feature Pattern (FP) is made in the form of the membership degree of each feature with respect to every class. Consider text document set TD of n text documents as per given in equation 1, together with the ICFV F of y features f1, f2,…fy and e classes c1, c2, …, ce. To construct the FP for each feature fk in F, its FP fpi is defined, by

fpi = < fpi1, fpi2, fpi3,…, fpie> (7)
= <μ (fi, c1), μ (fi, c2), μ (fi, c3),…,μ (fi, ce)>

Where,

fi represents the number of occurrences of fi in the text document TDg where 1 ≤ g ≤ n.

μ (fi, ce) is defined as the sum of product of the feature value of fi present in n text documents TD, w. r. t. a column vector and the 1 or 0 as the presence or absence of that feature in class ce / Sum of the feature

$$\mu\ (f_i,\ c_e) = \frac{\sum_g (TDgi).bi}{\sum_g (TDgi)} \quad (8)$$

b_i is represented as

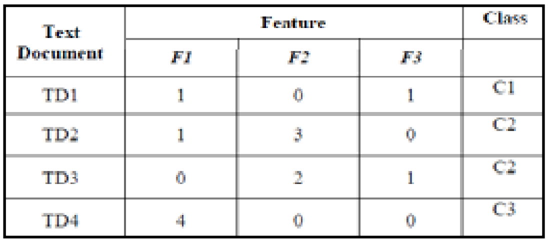
$$b_i = 1,\ \text{if } document\ \epsilon\ class\ c_e \\ = 0,\ otherwise \quad (9)$$

Each text document TD belongs to only one class c. In this way, each class can belong to one or more text documents. A set of n documents and their related categories or classes are represented as an ordered pair as shown

TD = {<TD1, C (TD1)>, <TD2, C (TD2)>,…,<TDm, C (TDm)>} (10)

Where the class of the text document TDi: C(TDi) ϵ C, C (TD) is a categorization function whose domain is TD and range is C. Each document belongs to one of the classes in the C (TD).The resulted text documents that have many features are stored with their relevant classes as shown in the Table 2. They are in the form of <doc no, number of occurrences of each feature, class no>.

Table 2: Text Document Feature Vector Class Matrix



Text Document	Feature			Class
	F1	F2	F3	
TD1	1	0	1	C1
TD2	1	3	0	C2
TD3	0	2	1	C2
TD4	4	0	0	C3

In such a way, the relation between a feature and a class is made. Sometimes, it is quite possible that one document belongs to two or more classes that concern has to be considered by making the more presences of the text document in the table with the cost of increased complexity, so it is required to check each feature's distribution among them.

H. Support Vector Machine Classifier (SVMC)

The next step is to use the Support Vector Machine Classifier (SVMC). SVMC is a popular and better method than other methods for text categorization. It is a kernel method which finds the maximum margin hyper plane in the feature space paradigm separating the data of training patterns into two groups like Boolean Logic 1 and 0. If any training pattern is not correctly classified by the hyper plane, then the concept of slack measure is used to get rid out of it.

Using this idea, SVMC can only separate apart two classes for h = +1 and h = -1. For e classes, one SVM for each class is created. For the SVM of class cl, 1 ≤ l ≤ e, the training patterns of class cl are for the h = +1 and of other classes are h = -1. The SVMC is then the aggregation of these SVMs.

SVM provides good results then KNN method because it directly divide the training data according to the hyper planes.

I. Text Classification Phase

To check the predictive accuracy of the SVMC, new and unknown text document is used, which is independent of the training text documents and is not used to construct the SVMC. The accuracy of this document is compared with the learned SVMC's class. If the accuracy of the SVMC is acceptable and good, then it can be used further to classify the future unseen text documents for which the class label is not known. Therefore, they can be categorized into the appropriate and a suitable category group.

IV. Conclusion

The proposed FSCMM model is made for text document categorization, it works well with high efficiency and effectiveness. Although this model and methodology seem very complex, yet it achieves the task of text categorization with high performance, and good accuracy and prediction. Feature Reduction is performed on the sentence, document and integrated corpora levels to highly reduce feature vector dimension. Such reduction improves the system performance greatly in terms of space and time. Result shows that the feature reduction reduces the space complexity by20%.

Fuzzy similarity measure and methodology are used to make the matching connections among text documents, feature vectors and pre-defined classes. It provides the mathematical framework for finding out the membership degrees as feature frequency.

References

[1] N. P. Padhy,"Artificial Intelligence and Intelligent Systems", 5th ed., Oxford University Press, 2009.

[2] Eliane Rich, Kevin Knight, Shivashankar B Nair, "Artificial

Intelligence", 3rd ed., Mc Graw Hill, 2010.

[3] Jiawei Han, MicheLine Kamber,"Data Mining: Concepts and Techniques", 2nd ed., Elsevier, 2006.

[4] Marcus Vinicius, C. Guelpeli, Ana Cristina, Bicharra Garcia,"An Analysis of Constructed Categories for Textual Classification Using Fuzzy Similarity and Agglomerative Hierarchical Methods", Third International IEEE Conference Signal-Image Technologies and Internet-Based System, September 2008.

[5] S. Shehata, F. Karray, M. S. Kamel,"An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.

[6] Jung-Yi Jiang, Ren-Jia Liou, Shie-Jue Lee,"A Fuzzy Self Constructing Feature Clustering Algorithm for Text Classification", IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 3, March 2011.

[7] Choochart Haruechaiyasak, Mei-Ling Shyu, Shu-Ching Chen, Xiuqi Li,"Web Document Classification Based on Fuzzy Association", IEEE, 2007.

[8] Ahmad T. Al-Taani, Noor Aldeen K. Al-Awad,"A Comparative Study of Web-pages Classification Methods using Fuzzy Operators.

S.Shanawaz Basha received his M. Tech degree in Computer Science from Jawaharlal Nehru Technological University Anantapur, India in 2011, MCA form Osmaina University, and B.Sc.in Electronics from S.K. University, A.P, India in 2009 and 2006 respectively. He is currently working as Assistant Professor at AVRSVR Engineering College of JNTUA. His current research interest includes Wireless Sensor Networks and networking protocols.

L. Sunitha Rani was born in Dhone, Kurnool Dist Andhra Pradesh, India. She received B.Tech in C.S.E from JNT University, Hyderabad, Andhra Pradesh, India. Presently, she is pursuing M.Tech in C.S.E from Indira Priyadarshini College of Engineering, Nannur, Kurnool Dist, Andhra Pradesh, India. Her Research interest includes Data-warehousing and Data-Mining and clustering.