Modeling An Intrusion Detection System Using Data Mining And Genetic Algorithms Based On Fuzzy Logic

G.V.S.N.R.V.Prasad	Y.Dhanalakshmi	Dr.V.Vijaya Kumar	Dr I.Ramesh Babu
Professor	Scholar	Professor &Dean	Professor
Dept of CSE	Dept of CSE	Dept of CSE &IT	Dept of CSE
Gudlavalleru Engg.College	A.N.U	G.I.E.T	A.N.U
Gudlavalleru	Guntur	Rajamandry	Guntur

Summary:

Fuzzy logic based methods together with the techniques from Artificial Intelligence have gained importance. Data mining techniques like clustering techniques, Association rules together with fuzzy logic to model the fuzzy association rules are being used for classifying data. These together with the techniques of genetic algorithms like genetic programming are producing better results. The present paper proposes a model for intrusion detection systems for anomaly detection based on fuzzy association rules which use genetic programming. The model is implemented and tested on sample data with 40 variables and the results are documented in the paper. As the model includes the LGP,MEP and GEP where the three collectively tries to detect the intrusion to a great extent.

Keywords:

Data Mining algorithms, Fuzzy logic, Linear Genetic Programming, Multi Expression Genetic Programming, Gene Expression Programming.

1. Introduction:

As per a survey the number of viruses and intrusions are growing in geometric progression while the number of systems for detecting and correcting them are increasing in arithmetic progression. This amounts to think about new strategies and techniques for countering them. More recently, techniques from the data mining area (mining of association rules and frequency episodes) have been used to mine the normal patterns from audit data. Typically, an IDS uses Boolean logic in determining whether or not an intrusion is detected and the use of fuzzy logic has been investigated as an alternative to Boolean logic in the design and implementation of these systems. Fuzzy logic addresses the formal principles of approximate reasoning. It provides a sound foundation to handle imprecision and vagueness as well as mature inference mechanisms using varying degrees of truth. Because boundaries are not always clearly defined, fuzzy logic can be used to identify complex pattern or behavior variations. This is accomplished by building an intrusion detection system that combines fuzzy logic rules with an expert system in charge of evaluating rule truthfulness.

Data mining techniques have been commonly used to extract patterns from sets of data. Specifically two data mining approaches have been proposed and used for anomaly detection: association rules and frequency episodes.. Association rule algorithm find correlations between features or attributes used to describe a data set. On the other hand, frequency episode techniques are effective at detecting sequential patterns of occurrences in a sequence of events. It is important to note that the use of association rule algorithms to produce rules suitable for anomaly-based and signature-based detection by mining normal and attack network traffic respectively [10]

Genetic algorithms are used to tune the fuzzy membership functions to improve the performance and select the set of features available from the audit data that provide the most information to the data mining component. These algorithms are often used for optimization problems. When using fuzzy logic, it is often difficult for an expert to provide “good” definitions for the membership functions for the fuzzy variables. Those genetic algorithms can be successfully used to tune the membership functions of fuzzy sets used by the intrusion detection system. Each fuzzy membership function can be defined using the standard function representation of fuzzy sets discussed in earlier chapters. Each chromosome for the GA consists of a sequence of these parameters. An initial population of chromosomes is generated randomly where each chromosome represents a possible solution to the problem.

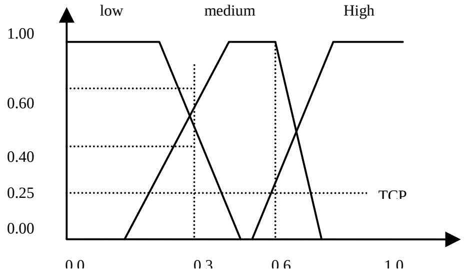

Fig 1..1 TCP linguistc

The Fig 1.1 illustrates the use of fuzzy sets to describe a linguistic variable TCP with domain {0-1} using the terms Low, Medium and High as specified by their respective membership functions. Fuzzy membership functions

Manuscript received July 5, 2008
Manuscript revised July 20, 2008

determine degrees of membership for each category of term. In Fig 1.1, a TCP value of 0.3 belongs 40% to the low category and 60% to the Medium and a TCP value of 0.6 belongs 100% to Medium and 25% to High. Under this scheme, the TCP value 0.6 is “more important” than the value 0.3 since the sum of its degrees of membership (fuzzy support) is 125% as compared to 100% for a TCP value of 0.3. This shortcoming was eliminated by normalizing the fuzzy terms, ensuring that the fuzzy support for any value totals 100%.

2. Architecture

The Hybrid Fuzzy logic IDS architecture has two modes of operations: rule-generation and detection. When operating in the rule-generation mode, the system processes network data and uses a fuzzy data mining algorithm to generate rules. A subset of the rules produced by the data mining algorithm is used as a model for the input data. The detection mode uses this rule subset for intrusion detection.

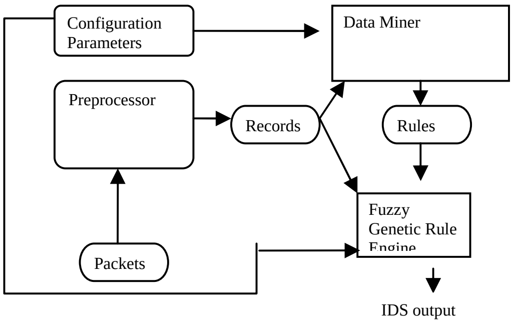

Fig 1.2 System Architecture

2.1 Preprocessor

The preprocessor is responsible for accepting raw packet data and producing records. This component is used in both modes and is capable of reading packets from the wire or a TCP dump file. The output produced by this component consists of records. Records contain aggregate information for a group of packets. Using records and concentrating only on attributes of interest greatly helps in reducing the amount of information to be used by more computationally intensive components of the architecture. Most of the approaches in the literature [7, 10] differ on how those attributes are selected. Here in the approach uses a light weight technique that employs positive and negative examples to identify the subset of attributes that provides the largest information gain in a decision tree. This is done by focusing on the branches of the underlying decision tree that contain the majority of the positive examples. The use of supervised learning in the preprocessor helps to improve the effectiveness of the unsupervised learning algorithm used in the data miner by selecting relevant data subsets.

2.2 Configuration Parameters

Parameter values stored in the configuration file regulate operation of the Data Miner and Fuzzy Inference Engine. The configuration file associates attributes with a term set and describes functions corresponding to the fuzzy membership functions associated with each term. A sample configuration file is shown in Fig.1.3. The structured file identifies the number and names of attributes followed by a description of each attribute. The description includes the type of attribute (binary, categorical, or fuzzy), the term set used to evaluate the attribute, definitions of each element of the term set, and domain information.

For example UDP (second attribute in Fig.1.3) is an attribute of “type” fuzzy that is evaluated using the “terms” BELOW, AVERAGE and ABOVE. RFuzzySet, Trapezoid Fuzzy Set and LFuzzySet are parametric functions that define the three terms (corresponding to names defined in the Fuzzy Jess Library [3]. Finally, range defines the domain for the UDP variable.

Note that the configuration file may be statically specified by system administrators. Alternatively, it may be dynamically generated by the preprocessor in the discovery mode.

2.3 Data Miner

The Data Miner integrates Apriori and Kuok’s algorithms to produce fuzzy rules. With one pass through the records, the fast and efficient algorithm used by the Data Miner extracts rules with sufficient support and confidence.

```
#ICMPUDP
1. attname = ICMP atttype =f termset =
   AVERAGE%ABOVE
2. AVERAGE = RFuzzySet(0.28,0.875)
3. ABOVE = LFuzzySet(0.28,0.917)
4. domain=0.1
5. attname=UDP
   atttype=ftermset=BELOW%AVERAGE%AB
   OVE
6. AVERAGE=trapezoidfuzzyset(0.0,0.051,0.32
   4,0.875)
7. ABOVE=LFuzzySet(0.324,0.917)
  domain=0.1
```

Fig 1.3 Sample Configuration file.

2.4 Rules

Rules are expressed as a logic implication $p \rightarrow q$ where p is antecedent and q is the consequence. Both p and q are assumed to be in conjunctive normal form, where aa_i, ca_j and cat_{attr} denote an antecedent attribute, a consequent attribute and an attribute category respectively. A typical rule looks like this:

If aa_0 is cat aa_0 $\wedge$ aa_1 is cat aa_1
$\wedge \dots \wedge$ aa_m is cat aa_m
Then ca_0 is cat ca_0 $\wedge$ ca_1 is
cat ca_1 … $\wedge$ ca_n is cat ca_n

The following three conditions hold for each rule:

1) $cat_{aai} \in TERMS(aa_i)0 \le i \le m$

2) $cat_{caj} \in TERMS(ca_j)0 \le j \le n$ and

3) $\{aa_0, aa_1 ... aa_m\} \cap \{ca_0, ca_1 ... a_n\} = \theta$

2.5 Fuzzy Inference Engine

The inference engine [7] makes use of Fuzzy Jess to evaluate fuzzy logic rules. Fuzzy Jess is a rule based expert system shell that integrates fuzzy logic components of the FuzzyJ Toolkit with Jess [9]. The three inputs to the Fuzzy Inference Engine are

i. the configuration parameters that FuzzyJess uses to define the Fuzzy Variables,
ii. the rules produced by the Data Miner that must be defined within the FuzzyJess environment and
iii. The records, which are asserted as facts in Fuzzy Jess. Fuzzy Jess can be configured to use Mamdani or Larsen inference mechanisms to compute the firing strength of each rule applied to each fact.

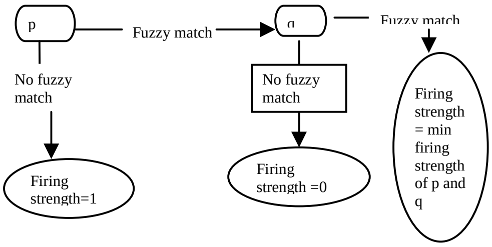

Fig 1.4 Analysis of fuzzy rules

The evaluation of rules begins with the analysis of the antecedent P. (see fig. 1.4). The following two cases are considered for the antecedent p.

- p does not have a fuzzy match so the rule does not apply to the record
- p does have a fuzzy match and the analysis of the consequent q begins

Note that a fuzzy match occurs when the truth value of the predicate is greater than zero. Similarly, the following two cases are considered for the consequence q:

- q does not have a fuzzy match and the firing strength of the rule is zero.
- q has a fuzzy match and the firing strength is determined using Mamdani inference mechanism

Fuzzy rules, as produced by the data mining algorithm, model a behavior represented by the data set employed to run the algorithm. The output of the Fuzzy Inference Engine is the firing strength of each rule for a given fact. This firing strength determines whether or not the fact satisfies the modeled behavior. Firing strengths that have a value close to one indicate that observed behavior closely follows the model behavior, but when several facts register firing strengths at or close to zero for a given rule, then it is likely that a deviation from the model (an attack) has been detected.

2.6 Genetic Algorithms

The goal is to increase the similarity of rules mined from data without intrusions and the reference rule set while decreasing the similarity of rules mined from intrusion data and the reference rule set [2]. A fitness function is defined for the GA which rewards a high similarity of normal data and reference data. A genetic algorithm works by slowly "evolving" a population of chromosomes that represent better and better solutions to the problem.

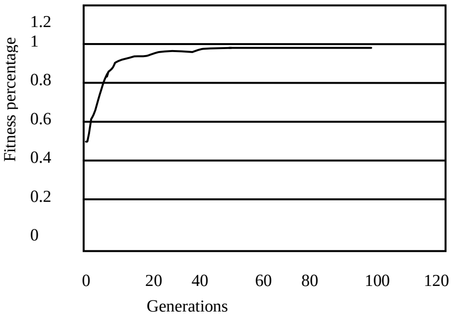

Fig 1.5 the evolution process of the fitness of the population, including the fitness of the most fit individual, the fitness of the least fit individual and the average fitness of the whole population

Fig 1.5 shows how the value of the fitness function changes as the GA progresses. The top line represents the fitness of the best individual in the population. Always retain the best individual from one generation to the next, so the fitness value of the best individual in the population never decreases. The middle line, showing the average fitness of the population, demonstrates that the overall fitness of the population continues to increase until it reaches a plateau. The lower line, the fitness of the least fit individual, demonstrates that continue to introduce variation into the population using the genetic operators of mutation and crossover [2].

3. Proposed Model of Intrusion detection system

In this paper we propose a model of an intrusion detection system using fuzzy logic and genetic algorithm based on data mining techniques. This proposed work is based on the evolutionary design of intrusion detection systems. The idea is to use genetic programming techniques, namely, Linear Genetic Programming (LGP), Multi Expression genetic Programming (MEP) and Gene Expression Programming(GEP).

Genetic programming techniques provide a framework for automatically creating a working computer program for a high level statement of the problem. This is achieved using genetic algorithm techniques of reading a population of computer programs. A population of intrusion detection programs is first initialized new solutions are created using the genetic algorithm techniques of mutation, crossover and reproduction operators. The fitness function of the resulting solutions are evaluated depending on the fitness value. suitable strategy selected and applied to find the solutions that go into the next generation.

Linear Genetic Program is a variant of genetic programming which acts on linear chromosomes. The basic unit of evolution here is a machine code instruction that runs on the floating point processor unit. An LGP individual is represented by a variable length sequence of simple language instructions like C language. The various LGP search parameters the mutation, crossover and the reproduction frequencies. Unlike in GP a MEP chromosome encodes several expressions [5, 8]. The best of the encoder solutions are chosen to represent chromosome by supplying the fitness of the chromosome. MEP genes are represented by substrings of a variable length. However, the number of genes per chromosome is constant and this defines the length of the chromosome. Each gene encodes a terminal or a function symbol. A gene that encodes a function includes pointers towards the function arguments. Function arguments always have indices of lower values than the position of the function itself in the chromosome. The maximum number of symbols in the MEP chromosome is given by

$$NS = (n+1) * (NG -1) +1$$

Where NS denotes the number of symbols, n is the number of arguments of the function with the maximum number of arguments and NG denotes the number of genes. The maximum number of effective symbols is achieved when each gene encodes a function symbol with the maximum number of arguments. The minimum number of effective symbols is equal to the number of genes and it is achieved when all genes encode terminal symbols only.

The individuals of GEP [4, 1] are encoded as linear chromosomes which are translated into expression trees. The linear chromosomes are called genotype and the expression trees are called phenotypes and they are different entities. However, they work together. The main items in GEP are basically two. They are the chromosomes and expression trees. The expression trees are the expressions of the genetic information encoded in the chromosomes. GEP uses linear chromosomes that store expressions in breadth first form. A GEP gene is a string of terminal and functional symbols and is composed of a head and a tail. The head contains both the symbols while the tail contains terminal symbols only. For each problem the length h of the head is chosen by the user and length t of the tail is given by

$$t = (n-1) * h+1$$

Where n is the number of arguments of the function with more arguments.

The initial population is randomly generated. The following steps are repeated until a termination criteria is reached

a). A fixed number of the best individuals enter the next generation.

b)The mating pool is filled by using binary tournament selection. The individuals from the mating pool are randomly paired and recombined. Two off springs are obtained by recombining two parents and they are mutated to enter the next generation.

3.1 General approach for intrusion detection

We define the general approach adapted for modeling an intrusion detection system using fuzzy logic and genetic algorithms with the help of data mining. A broad general approach is presented below

Step 1: Identity the three parameters or features of the problem statement.

Step 2: Classify the parameters or features depending on their uncertainty or crisp nature.

Step 3: Once the parameters are classified use fuzzy logic for modeling the uncertain parameters or features.

Step 4: The crisp values can be modeled using statistical distributions depending on their classifications.

Step 5: Normalize these modeled values.

Step 6: Use clustering techniques and association rules for grouping them together suitably.

Step7: Identify a suitable model based on Mathematical technique genetic algorithms or any other mathematical techniques for solving a problem.

Step 8: Solve the problem using the above Mathematical technique.

Five different types of data were chosen with 40 attributes each [6]. The data contain 24 attack types which are classified into four categories. They are Denial of Service (DOS), unauthorized access from a Remote Machine (URM), unauthorized access to Local Super user (ULS) and Probing and Surveillance (PAS).

Denial of service(DOS) is a class of attack where an attacker makes a resource too busy to handle authorized request and in turn deny access to the authorized users. URM is a class of attack where an attacker exploits the vulnerability of the machine by sending packets to the machine, to gain illegal access as a user. In the case of ULS an attacker starts with gaining access to the account of a normal user and then exploits the systems vulnerability. PAS is a class of attack where an attacker scans a network to know the vulnerabilities and exploits them. The 40 variables are given in table 1.1. The variables from 24 to 40 are modeled using normal distribution. The variables 8 and 9 are modeled using fuzzy techniques and the remaining values are taken as they are numerically viable. All the data are normalized between 0 and 1. A clustering algorithm is used for classifying them into five classes namely, NORMAL, PAS, DOS, URM and ULS. There are two phases namely training and testing. In the training phase LGP, MEP and GEP models are constructed using the training data. The test data is then passed through the saved training model to detect intrusions in the testing phase. The various parameter settings for LGP, MEP, and GEP are given in tables 1.2, 1.3 and 1.4 and the results of using all these three are compared in table 1.5&1.6. The true positive rates and false positive rates for are obtained using the formula

True positive rate = (positives correctly
classified)/ (total positives)
and
False positive rate = (total negatives – negatives
incorrectly classified)/ (Total negatives).

The results are shown in table 1.5&1.6

S.No	Variable Name
1	Duration
2	Protocol Type
3	Service
4	Flag
5	Src_bytes
6	Dst_bytes
7	Wrong fragment
8	Urgent
9	Hot
10	Num_failed _logins
11	Logged_ in
12	Num_compromized
13	Root_shell
14	Su_attempted
15	Num_root
16	Num_file_creations
17	Num_shells
18	Num_access_files
19	Num_outbound_cmds
20	Is_host_login
21	Is-guest _login
22	Count
23	Srv_count
24	Serror_rate
25	Srv_serror_rate
26	Rerror_rate
27	Srvr_rerror_rate
28	Same_srv_rate
29	Diff_srv_rate
30	Srv_diff_host_rate
31	Dst_host_count
32	Dst_host_srv_count
33	Dst_host_same_srv_rate
34	Dst_host_diff_srv_rate
35	Dst_host_same_src_port_rate
36	Dst_host_srv_diff_host_Rate
37	Dst_host_serror_rate
38	Dst_host_srv_serror_rate
39	Dst_host_rerror_rate
40	Dst_host_srv_rerror_rate

Table 1.1 Variables taken

3.2 Parameter setting

We have chosen 40 parameters as listed in table 1.1, The population size is taken as 1024 for LGP with maximum number of tournaments 10000. The maximum program size is taken as 256 with number of demes being 10 in each case. The details are given in table 1.2. The parameter settings for MEP and GEP are given in tables 1.3 and 1.4 respectively.

Parameter	Value				
	Normal	PAS	DOS	ULS	URM
Tournament size	8	8	8	8	8
Mutation Frequency (%)	85	82	75	86	85
Cross over frequency	75	70	65	75	70

Table 1.2: Parameters for LGP

Parameter	Value				
	Normal	PAS	Dos	ULS	URM
Population size	100	200	250	100	100
Number of generations	30	200	800	20	800
Chromosome length	30	40	40	30	40
Cross over frequency (%)	90	90	80	90	90
Number of mutations per chromosome	3	4	5	3	4

Table 1.3 Parameters used by MEP

Parameter	Value				
	Normal	PAS	DOS	ULS	URM
Population size	100	100	100	100	100
Number of generations	800	500	500	500	500
Number of genes	12	12	14	12	12
Mutation	0.1	0.1	0.1	0.1	0.1
One point cross over	0.4	0.4	0.3	0.4	0.4
Two point cross over	0.4	0.4	0.3	0.4	0.4
Gene recombination	0.2	0.2	0.1	0.1	0.2
Gene transposition	0.2	0.2	0.1	0.1	0.2

Table 1.4 Values of parameters used by GEP

	Classification Accuracy on Test Data Set Percentage				
	Normal	PAS	DOS	ULS	URM
LGP	99.31	99.85	99.95	67.40	99.40
MEP	99.76	95.13	98.76	99.50	99.50
GEP	99. 58	97.85	95.64	99.36	98.85

Table 1.5 Performance comparison between LGP,MEP and GEP

	Best results for training applied for testing					
	LGP		MEP		GEP	
	TP	FP	TP	FP	TP	FP
Normal	0.992	0.997	0.994	0.999	0.994	0.994
PAS	0.544	0.996	0.945	0.982	0.997	0.974
DOS	0.982	0.992	0.985	0.999	0.917	0.946
ULS	0.360	0.997	0.402	0.999	0.435	0.993
URM	0.965	1	0.971	1	0.982	0.982

Table 1.6 Comparison of false alarm rates

4.Conclusions

There are many intrusion detection systems proposed in the literature based on various techniques like cryptographic techniques, Encryption methods etc. In recent times Fuzzy logic based methods together with the techniques from Artificial Intelligence have gained importance. However, none of them is fool-proof and have their advantages and limitations. Data mining techniques like clustering techniques, Association rules together with fuzzy logic to model the fuzzy association rules are being used for actually classifying data. These together with the techniques of genetic algorithms like genetic programming are producing better results. The present paper proposes a model for intrusion detection systems for anomaly detection based on fuzzy association rules which use genetic programming. The model is implemented and tested on sample data with 40 variables and the results are documented in the paper. By the observations it can be concluded that the model on the whole tries to irradiate the intrusions. This is because by including LGP,MEP,GEP in the model. For eg. the LGP is less efficient in controlling unauthorized access to local super user where as MEP is more efficient. Likewise it is observed that for Denial of Service and unauthorized access to local super user using MEP will give FP with less error. So it can be concluded that the drawbacks of one programming in certain aspects can be overcome by including the three in the model to complete intrusion detection system.

5.Directions for future work

According to a survey, the number of viruses is growing in geometric progression while the number of anti-virus solutions is growing in arithmetic progression. This suggests that we need to employ newer strategies to combat threats for any networks. There are many more new approaches being tried by researchers. Some of the promising areas in this direction, especially in the area of intrusion detection are ant colony optimization methods and neural networks.

In ant colony optimization, the processes are defined based on the techniques of ant movements. The propagation algorithms may used for training the system and use that knowledge for detecting the intruders. Further, fuzzy reasoning may be replace with Demster-Shaffer theory where ever applicable.

References

[1] Friedman, J. H., “Multivariate adaptative regression splines”, Annals of statistics, Vol. 19, 1991.

[2] Susan M.Bridges, Rayford B. Vaughan,“Fuzzy Data mining and Genetic Algorithms Applied to Intrusion Detection”, Conference on National Information Systems Security, Oct. 2000.

[3] Wengdong Wang, Susan M. Bridges“Genetic Algorithm Optimization of Membership Functions for mining Fuzzy Association Rules” Presented at International Conference on information System Fuzzy theory, March 2000.

[4] Ferreira. C, “Genetic representation and Genetic Neutrality in Gene Expression programming”, Advances in complex systems, 5(4):pp. 389-408, 2002

[5] Oltean M and Grosen C, “Evolutionary algorithms using multi-expression programming”, Proc. Of VII European Conf. on Artificial Life, Dortmund, Germany, pp. 651-658, 2003.

[6] Mukkamala S, Sung. A, Abraham A, “ Intrusion Detection using ensemble of soft computing and hard computing paradigms”, J. of Network and Comp. Appl., Elsevier Science, Vol. 28, issue 2, pp. 167-182, 2005.

[7] Aly EI Semary, Jamica Edmonds, Jesus Gonzalez Pino, Mauricio Papa“Implementation of a hybrid intrusion detection system using fuzzyjess” in 7th International Conference on Enterprise Information Systems, (Miami Florida) 2005,pp.390-393.

[8] Peddabachigari. S, Abraham A, Grosan C, Thomas J, “ Modeling Intrusion Detection system using Hybrid Intelligent systems”, J.of Network and Comp. Appl., Elsevier Science,2005.

[9] K.M. Faraoun, and A. Boukelif ”Genetic Programming Approach for Multi-Category Pattern Classification Applied to Network Intrusions Detection”, International Journal of Computational Intelligence Vol.3, No.1 2006, pp.79-90.

[10] Aly EI Semary, Jamica Edmonds, Jesus Gonzalez-Pino, Mauricio Papa “Applying Data mining of Fuzzy Association Rules to Network Intrusion Detection “, IEEE Proc. On Information Assurance, West Point, New York, 2006,pp.100-107.

Prof. G.V.S.N.R.V.Prasad did his MS Software Engineering in BITS Pilani and M.Tech in Computer Science and Technology in Andhra University .He has 15 years of teaching experience . Published 7 Research Papers in Various National and International Conferences He is a member in various Professional Bodies . Presently working as Professor and Head in CSE at Gudlavalleru Engineering College , Gudlavalleru ,A.P.His area of interest is Datamining,Network Security and Image Processing

MrsY.Dhanalakshmi did her MCA,M.Phil 2000 and 2003 respectively.She was awarded Ph.D from Acharya Nagarjuna University. She has published 7 papers in international and national journals. Her area of interest is Data Mining, Network Security.

Dr.V.Vijaya Kumar did his MS Engineering in Computer Science [USSR –TASHKENT STATE UNIVERSITY] and Ph.D in Computer Science . Worked as Associate Professor in Department of CSE and School of Information Technology (SIT) at Jawaharlal Nehru Technological university (JNTU) Hyderabad . Having a total of 13 years of experience. He Published 60 Research Papers in various National and Inter National Conferences/Journals. Guiding 10 Research scholars . He is a Member for various National and Inter National Professional Bodies .Presently working as Dean for CSE & IT at GODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY Rajamandry .

Prof. I.Ramesh Babu . He joined as an Assistant Professor in the Department of Computer Science and Engineering in Acharya Nagarjuna University in 1988,and became a Professor in 2004. He held many positions in Acharya Nagarjuna University as Executive council member, Chairman Board of Studies, Head, Director Computer Centre, Member of academic senate, member of the standing committee of academic senate. He is also a member of Board of Studies for other universities. He has published many research papers in International, national journals and presented papers in international conferences also. His research areas of interest include Image Processing, Computer Graphics, Cryptography, Network Security and Data Mining. He is member of IEEE, CSI, ISTE,IETE, IGISS, Amateur Ham Radio (VU2UZ)