T.C.
BAHCESEHIR UNIVERSITY

The Graduate
School of Natural and Applied Sciences
Computer Engineering

# Spam E-Mail Detection and Filtering Based on an Evolutionary Soft Computing Model Using Neuro-Fuzzy Classifiers and Genetic Algorithms

Master of Science Thesis

Altan PARLAK

Supervisor: Assoc. Prof. Adem KARAHOCA

İstanbul 2010

**T.C**
**BAHCESEHIR UNIVERSITY**
**The Graduate**
**School of Natural and Applied Sciences**
**Computer Engineering**

Name of the thesis: Spam E-Mail Detection and Filtering Based on An Evolutionary Soft Computing Model Using Neuro-Fuzzy Classifiers and Genetic Algorithms
Name/Last Name of the Student: Altan Parlak
Date of Thesis Defense:

The thesis has been approved by the Institute of Science.

Asst.Prof.Dr. Tunç BOZBURA
Director

Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

| Examining Committee Members<br>Title Name and Surname | Signature |
|---|---|
| Assoc.Prof.Dr. Adem KARAHOCA | ---------- |
| Asst.Prof.Dr. Yalçın Çekiç | ---------- |
| Asst.Prof.Dr. M.Alper Tunga | ---------- |

## ACKNOWLEDGEMENTS

This thesis is dedicated to **my parents** for their patience and understanding during my master's study and thesis work.

I would like to express my gratitude to **Assoc. Prof. Dr. Adem Karahoca**, for not only being such a great supervisor but also encouraging and challenging me throughout my graduate program.

## ABSTRACT

Spam E-Mail Detection and Filtering Based on an Evolutionary Soft Computing Model Using Neuro-Fuzzy Classifiers and Genetic Algorithms

Parlak, Altan

M.S. Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Adem Karahoca

January 2010, 35 Pages

Spam mail, common problem for all email users, is getting more popular everyday. Concept drift, reactive creative adversaries makes it difficult to filter spams with basic methodologies. The change in the spam email requires learning based spam filtering. In this thesis literature for the proposed methods are investigated for the spam filtering. The most successful filtering methods are the combinational filtering methods. This thesis proposes a new method for the spam filtering using a combination of Adaptive Neuro-Fuzzy Inference System (ANFIS) and Genetic Algorithms (GA) for the tuning of the rule base. This study also gives brief explanations about spam, spam types, used spam filtering techniques and introduces ANFIS and Genetic Algorithms. The last part compares the results of the NEFCLASS and the proposed method and gives the results for the spam dataset used in this study.



**Keywords**: Spam, Learning Based Spam Filtering Methods, Adaptive Neuro-Fuzzy Inference System(ANFIS), Genetic Algorithms(GA)

# ÖZET

Sinirsel Bulanık Sınıflayıcı ve Genetik Algoritma Kullanarak Evrimsel Yapay Zeka Modeli İle Spam E-posta Tanıma ve Filtreleme Algoritmaları

Parlak, Altan

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Doç. Dr. Adem Karahoca

Ocak 2010, 35 Sayfa

Spam eposta, tüm email kullanıcılar için ortak problem, popülerliğini sürekli arttırmakta. Degişen içerik ve yaratıcı yöntemler basit yöntemlerle spam filtrelemeyi güçleştiriyor. Spam epostalardaki bu hızlı değişim filtrelemede yapay zeka uygulamalarını zorunlu kılıyor. Bu çalışmada spam eposta tanımı, spam çeşitleri ve daha önce kullanılmış olan filtreleme yöntemleri kısaca açıklayarak ANFIS ve Genetik Algoritmaların tanımını vermektedir. Bu çalışmada daha önce kullanılan spam filtreleme yöntemleri incelendi ve ANFIS ile Genetik Algoritmaların birlikte kullanıldığı bir model ele alınarak bir sistem geliştirilmek istendi. Son bölünde ise NEFCLASS ve geliştirilen sistemler karşılaştılaştırıldı.

**Anahtar Kelimeler**: Spam, Öğrenmeye Dayalı Spam Filtreleme Metodları, Adaptive Neuro-Fuzzy Inference System(ANFIS), Genetic Algorithms(GA)

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| Adaptive Neuro-Fuzzy Inference System | : | ANFIS |
| Genetic Algorithms | : | GA |
| Artificial Neural Network | : | ANN |
| Support Vector Machines | : | SVM |
| False Positive | : | FP |
| False Negative | : | FN |

# 1. INTRODUCTION

Spam is a major problem in the Internet area. Although there is not a unique description for spam (also called junk mail) and how it differs from legitimate mail (also called non-spam or genuine mail). The shortest popular definition characterizes spam as “unsolicited bulk email” (UBE) (Androutsopoulos, Koutsias, Chandrinos & Spyropoulos, 2000) or sometimes the word commercial is added (UCE). According to the TREC Spam Track, spam is “unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user” (Cormack, & Lynam, 2005).

Spam is also defined as the following in antispam sites (Mueller): Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Most spam is commercial advertising, often for dubious products, get-rich-quick schemes, or quasi-legal services. Spam costs the sender very little to send -- most of the costs are paid for by the recipient or the carriers rather than by the sender.

There are two main types of spams, and they have different effects on Internet users, usenet spam and email spam. Cancellable Usenet spam is a single message sent to 20 or more Usenet newsgroups. Usenet spam is aimed at "lurkers", people who read newsgroups but rarely or never post and give their address away. Usenet spam robs users of the utility of the newsgroups by overwhelming them with a barrage of advertising or other irrelevant posts. Furthermore, Usenet spam subverts the ability of system administrators and owners to manage the topics they accept on their systems.

Email spam targets individual users with direct mail messages. Email spam lists are often created by scanning Usenet postings, stealing Internet mailing lists, or searching the Web for addresses. Email spams typically cost users money out-of-pocket to receive. Many people - anyone with measured phone service - read or receive their mail while the meter is running, so to speak. Spam costs them additional money. On top of that, it costs money for ISPs and online services to transmit spam, and these costs are transmitted directly to subscribers.

One particularly nasty variant of email spam is sending spam to mailing lists (public or private email discussion forums). Because many mailing lists limit activity to their subscribers, spammers will use automated tools to subscribe to as many mailing lists as

possible, so that they can grab the lists of addresses, or use the mailing list as a direct target for their attacks.

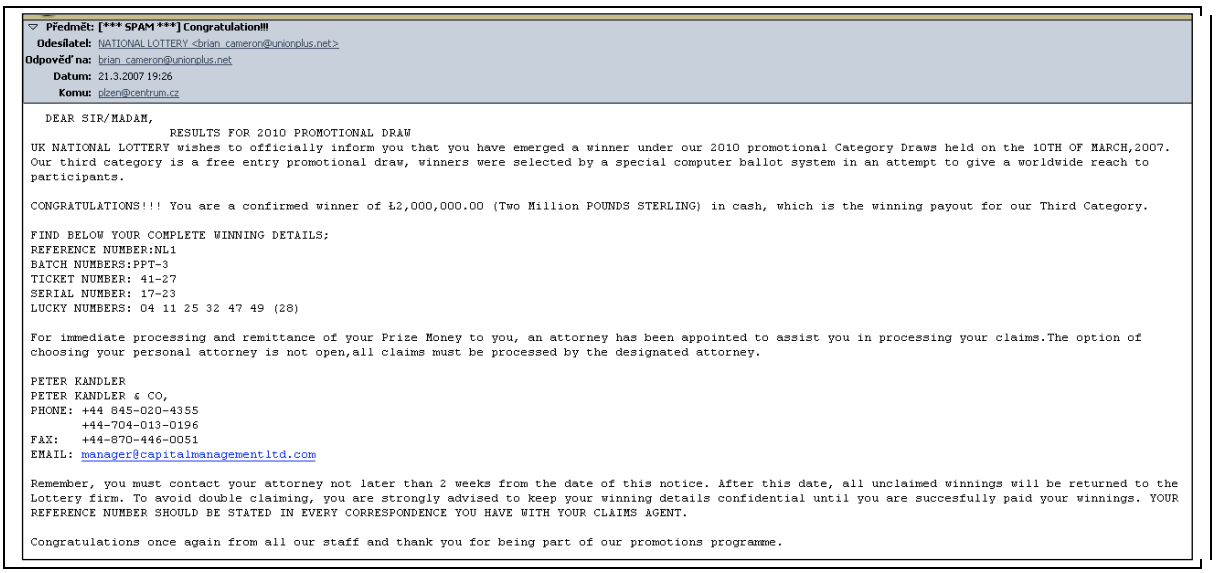
Spam mail becomes a major issue as spam emails constitute approximately 80% of the received emails. Spam causes financial lose, storage space problems, computational power and productive time consumption in deleting emails. Spam emails also may cause legal problems as non legitimate advertisements. The Ferris Research Analyzer Information Service estimates the total worldwide financial losses caused by spam in 2005 as $50 billion (FerrisResearch, 2005).

Due to the negative effects of the spam emails, it is a hot issue for detecting and filtering the unsolicited emails.

Recently, according to study on Anti-spam Strategies in Companies by Siponen and Stucke (2006), filtering is the most commonly used method and it will remain most commonly used method in the near future. The spam filtering is predicted to be an important practical application based on machine learning techniques. This will allow identifying the new types of spams without human intervention.

There are many approaches for spam detection and filtering. The spammers' creativity results in new spam emails that break filter rules. Therefore learning based adaptive detection becomes a key issue to cope with spam.

The combination of the learning based adaptive detection systems filters out the spam emails better. The main aim of this work is to generate a low error rate using combination of Adaptive Neuro-Fuzzy Inference System with Genetic Algorithm where Genetic Algorithm tunes the fuzzy rule base.



**Předmět:** [*** SPAM ***] Congratulation!!!
**Odesílatel:** NATIONAL LOTTERY <brian_cameron@unionplus.net>
**Odpověď na:** brian_cameron@unionplus.net
**Datum:** 21.3.2007 19:26
**Komu:** plzen@centrum.cz

DEAR SIR/MADAM,

RESULTS FOR 2010 PROMOTIONAL DRAW

UK NATIONAL LOTTERY wishes to officially inform you that you have emerged a winner under our 2010 promotional Category Draws held on the 10TH OF MARCH,2007. Our third category is a free entry promotional draw, winners were selected by a special computer ballot system in an attempt to give a worldwide reach to participants.

CONGRATULATIONS!!! You are a confirmed winner of £2,000,000.00 (Two Million POUNDS STERLING) in cash, which is the winning payout for our Third Category.

FIND BELOW YOUR COMPLETE WINNING DETAILS;
REFERENCE NUMBER:NL1
BATCH NUMBERS:PPT-3
TICKET NUMBER: 41-27
SERIAL NUMBER: 17-23
LUCKY NUMBERS: 04 11 25 32 47 49 (28)

For immediate processing and remittance of your Prize Money to you, an attorney has been appointed to assist you in processing your claims.The option of choosing your personal attorney is not open,all claims must be processed by the designated attorney.

PETER KANDLER
PETER KANDLER & CO,
PHONE: +44 845-020-4355
+44-704-013-0196
FAX: +44-870-446-0051
EMAIL: manager@capitalmanagementltd.com

Remember, you must contact your attorney not later than 2 weeks from the date of this notice. After this date, all unclaimed winnings will be returned to the Lottery firm. To avoid double claiming, you are strongly advised to keep your winning details confidential until you are succesfully paid your winnings. YOUR REFERENCE NUMBER SHOULD BE STATED IN EVERY CORRESPONDENCE YOU HAVE WITH YOUR CLAIMS AGENT.

Congratulations once again from all our staff and thank you for being part of our promotions programme.

Source: The Fight against Spam - A Machine Learning Approach (Jezek & Hynek, 2007)

**Figure 1.1 – Example of a spam email**

## 2. CURRENTLY EXISTING SPAM TYPES

As mentioned in introduction spam does not have an exact explanation and classification, however spam emails are mainly classified into four types by Karel Jezek (2007).

### 2.1 Stock Spam, Pump and Dump

The term "pump and dump" on the Internet represents unsolicited mail offers of very inexpensive goods (typically below $1), urging mail recipients to quick purchase. This evokes massive demand for goods which have already been sold in most cases. Nonetheless, the price of the goods is gradually increased ("pumped").

This type of unsolicited mail often includes links to small or non-existing companies, as it is almost impossible to track any information on the company making the attractive deal. In some cases, "pump and dump" spam is designed to hurt the good name of an existing company, as the consequences of illegal business deals are borne by the actual company, not the spammers.

### 2.2 Phishing

Phishing is used for messages designed to elicit personal data (such as bank account numbers, credit card numbers, passwords, etc.) from email recipients. The term is derived from "fishing", which is exactly what spammers do – distribute "bait" and wait to see what happens. Spammers commonly use exploits such as using the company's image, inserting links to the real company site, or using email that appears to be from the spoofed company.

### 2.3 Image-Based Spam

Tricks used to distribute unsolicited mail get more and more sophisticated. The best way to get around statistical text filters is to use images instead of text. Image handling is quite difficult for antispam software, regardless of the actual image form – plain text converted into an image, various interference items on the background, use of animations, etc. Although use of images for spamming is not a new concept, it is definitely gaining popularity. According to various studies, approximately one-third of all unsolicited mail was represented by image-based spam at the end of 2006. It seems that spammers are quite content with the hit rate of their messages, and keep converting all their text-based mails into images.

### 2.4 Text Spam

Text spam is just unsolicited commercial mail distributed in textual form. Typical features of the text spam are listed below (please note that the majority of these features are language-independent):

HTML text contained in message body,

- High proportion of capital letters (usually more than 30%),
- Exclamation mark(s) in the message subject,
- Instructions on how to unregister from the distribution list,
- Instruction to click on a link,
- Text lines longer than 200 characters,
- High priority assigned to the message,
- Nonsense date of sending (such as 1st January 1970),
- Disclosed message sender,
- More (or disclosed) message recipients.

## 3. LEARNING-BASED METHODS OF SPAM FILTERING

The most popular method for anti-spam technique is spam filtering according to the study of Mikko Siponen and Carl Stucke (2006). Spam filtering classifies the messages into spam and legitimate email. Existing filtering algorithms have quite affective results even close to 90% accuracy and it was found that integrating different learning algorithms actually seems to be a promising way (the evaluation performed by Lai & Tsai, 2004).

Spam filtering is an application which implements a function with binary output, spam or legitimate. Machine learning classification techniques are the main type for the spam filters. In the learning based techniques filtering function input is the message, and parameter vector is the result of a training dataset. However, there are some drawbacks caused by the dataset. Fawcett (2003) states that like most text classification domains, spam presents the problem of a skewed class distribution, i.e., the proportion of spam to legitimate email is uneven. There are no generally agreed upon class priors for this problem. Gomez Hidalgo (2002) points out that the proportion of spam messages reported in research datasets varies considerably, from 16.6% to 88.2%. There are other drawbacks such as unequal and uncertain FP and FN error costs, disjunctive and concept drifting, and reactive creative adversaries (Blanzieri & Bryl, 2008).

For all algorithms there is a problem for determining a reasonable trade-off between errors; classifying spam mail as legitimate and classifying legitimate email as spam. While classifying spam mail as legitimate bothers the end user, classifying a legitimate email as spam results in a valuable data loss.

This trade-off issue is discussed in game theory, training techniques for low false positive and user defined parameters. (Androutsopoulos, Magirou & Vassilakis, 2005; Yih, Goodman & Hulten, 2006; Michelakis, Androutsopoulos, Paliouras, Sakkis & Stamatopoulos, 2004).

There are many methods proposed for the spam filtering. The starting point of the filtering was based on predefined keywords or sender information (blacklist) to detect spam. In time predefined keyword based filters begun to be replaced by learning based approaches like Naïve Bayesian. On the contrary blacklists and whitelists are still in use as part of complex anti-spam solutions as in filtron (Michelakis, Androutsopoulos, Paliouras, Sakkis &

Stamatopoulos, 2004). Moreover there are spammer lists exist in public registers. Another method is the greylist, which is temporary marking an email as spam and unblocks if the email is sent again and sender is not added to the blacklist during this interval. The main idea here is that spam mails generally do not repeat themselves and if they did they are marked as spam in the period between two posts.

### 3.1 Naïve Bayes

Naïve Bayes Classifier is the mainly used classifier in spam filtering (Pantel & Lin, 1998; Sahami, Dumais, Heckerman & Horvitz, 1998). After Paul Graham's 'A plan for spam' (Graham) article it becomes widely known method. This can be mainly classified as a learning based keyword filter when used for the text content. Bayesian method uses d dimensional x vectors to classify the email as spam or legitimate. Here d is the independent features of x, used for estimating the probabilities the email classification. Several variants of Naïve Bayes were applied to spam filtering, an overview and comparison of them can be found in the article by Metsis et al. (Metsis, Androutsopoulos & Paliouras, 2006).

### 3.2 k-Nearest Neighbor

The k-Nearest Neighbor (k-NN) classifier was proposed for spam filtering by Androutsopoulos et al. (2000), which investigates the performance of two machine learning algorithms in the context of anti-spam filtering. In k-NN the decision is made as follows: k nearest training samples are selected using a predefined similarity function, and then the message x is labeled as belonging to the same class as the majority among this k samples.

### 3.3 Support Vector Machines

Another classifier proposed for spam filtering is Support Vector Machine (SVM) (Islam, Chowdhury & Zhou, 2005). This model combines both linear and nonlinear SVM techniques where linear SVM performs better for text based spam classification that share similar characteristics. The proposed model considers both text and image based email messages for classification by selecting an appropriate kernel function for information transformation. Given the training samples and a predefined transformation, which maps the features to a transformed feature space, the classifier separates the samples of the two classes with a hyper

plane in the transformed feature space, building a decision rule. SVM was proposed in particular to classify the vectors of features extracted from images (Aradhye, Myers & Herson, 2005).

### 3.4 Term Frequency-Inverse Document Frequency

The name Term Frequency-Inverse Document Frequency (TF-IDF) actually applies to a term weighting scheme. Weight of each term (token) is calculated by multiplying the occurrence of the term with the log of total messages per messages including given term. This scheme can be combined with the Rocchio algorithm, a detailed description of which can be found in the paper by Joachims (1997). Such combination results in a quite accurate classifier (Drucker, Wu & Vapnik, 1999), which is sometimes also referred to as TF-IDF in the literature.

### 3.5 Boosting

Boosting is a general name for the algorithms based on the idea of combining many hypotheses (for example one-level decision trees). At each stage of the classification procedure a weak (not very accurate) learner is trained, and its output is used to re-weight the data for the future stages: greater weight is assigned to the samples which are misclassified. For spam filtering boosting was proposed by Carreras and Marquez (2001).

### 3.6 Chi By Degrees Of Freedom

Chi by degrees of freedom is proposed for spam filtering by O'Brien and Vogel (2003). This method is usually used for document authorship identification. Messages are represented in terms of character or word N-grams. The idea of the method is to compare the similarity of a new message to the labeled messages using the chi-by-degrees-of-freedom test, which is calculated by dividing the value of the $X^2$ test by the number of degrees of freedom. 'Chi by degrees of Freedom' has the advantage of providing significance measures, which will help to reduce false positives.

### 3.7 Smoothed N-gram Language Models

Word n-gram model is a hidden Markov model which computes, for each sentence of the

speech, a probability of being produced by one author or by the other one. This probability is computed from the probability of a given word coming up next, depending on the prior *n-1* words. The training corpus allows us to build a model for each author from the frequencies of *n*-word sequences. Each model is then applied to each sentence of the test corpus. The recognized author is the one that corresponds to the higher probability of the sentence. Then results are smoothed in order to obtain sentence sequences of the same author by modifying the author of isolated sentences (smoothing across sequences). Medlock (2006) used smoothed higher-order N-gram models.

### 3.8 Neural Network Based Approach

James Clark et al. (2003) designed a 3 layers back propagation (BP) neural networks. It was shown that a BP network with information gain (IG) has rather good effect of identifying spare email in their experiment. Most current anti-spam techniques filter out junk emails based on words of email subjects and body messages. Hu (2008) proposes a better method based on words and behavior based characteristics for judging spam and proves that the Complex Valued Neural Network anti-spare email filter has better performance than simple BP neural network based approaches. Combination of unrelated features may be used for spam detection and possibly results in a better solution. There are also combined methods which may result in a better accuracy. An example is the combination of Neural Recognition and Genetic Features Selection.

### 3.9 Neural Recognition and Genetic Features Selection

Gavrilis (2006) presents a two-step feature selection method that uses term entropy to select a subset of the original features in the first step and genetic selection in the second step. A Radial Basis Function Network (RBF network) is used for the classification with 20 features as inputs producing a 3.27% classification error. The achievement of the proposed method is 96-97% average accuracy when using only 20 features out of 15000.

### 3.10 Artificial Neural Networks (ANN) and Bayesian Networks

Levent Ozgur (2004) proposes anti-spam filtering methods for agglutinative languages. The methods are dynamic and are based on Artificial Neural Networks (ANN) and Bayesian

Networks combined together. The algorithms have two main components. The first one deals with the morphology of the words and the second one classifies the e-mails by using the roots of the words extracted by the morphological analysis. The experiment results show that up to 90% success rate is achieved.

**Table3.1- Individual Classifier Performance Over Spam Experiments**

| Classifier | Detection Rate | False Positive Rate | Gain |
|---|---|---|---|
| NBayes(non-content) | 88% | 3.8% | 79.8% |
| Ngram | 75% | 4.0% | 72.2% |
| TextClassifier | 90% | 5.0% | 70.0% |
| Pgram | 90% | 5.5% | 77.2% |
| TF-IDF | 74% | 4.2% | 61.5% |
| Limited Ngram | 66% | 5.0% | 61.4% |
| URL | 55% | 10% | 32.0% |

Source: Behavior-based Email Analysis with Application to Spam Detection, Shlomo Hershkop, Doctorate Thesis Columbia University

## 4. DEFINITION OF ANFIS AND GENETIC ALGORITHMS

Proposed method tunes ANFIS parameters utilizing the Genetic Algorithm. Description for the ANFIS and Genetic Algorithms are given in the next sections.

### 4.1 ANFIS Structure

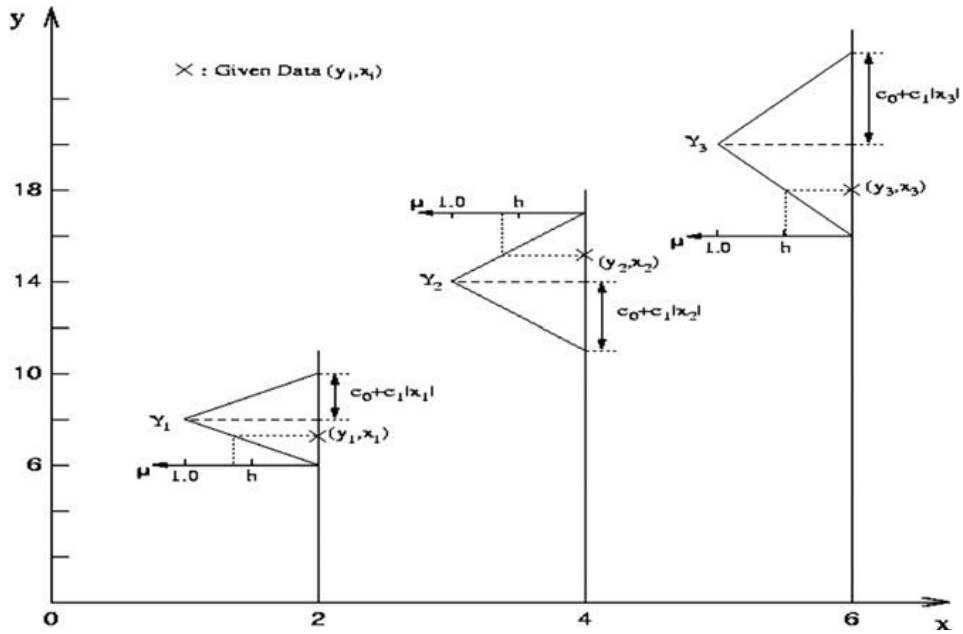
Adaptive Neuro-Fuzzy Inference System is the combination of Artificial Neural Network (ANN) and Fuzzy Inference System (FIS).

*ANFIS is a neuro-fuzzy system developed by Jang. It has a feed-forward neural network structure where each layer is a neuro-fuzzy system component (Fig. 3). It simulates TSK (Takagi–Sugeno–Kang) fuzzy rule of type-3 where the consequent part of the rule is a linear combination of input variables and a constant. The final output of the system is the weighted average of each rule's output. The form of the type-3 rule simulated in the system is as follows:*

*IF* $x_1$ *is* $A_1$ *AND* $x_2$ *is* $A_2$ *AND . . . AND* $x_p$ *is* $A_p$

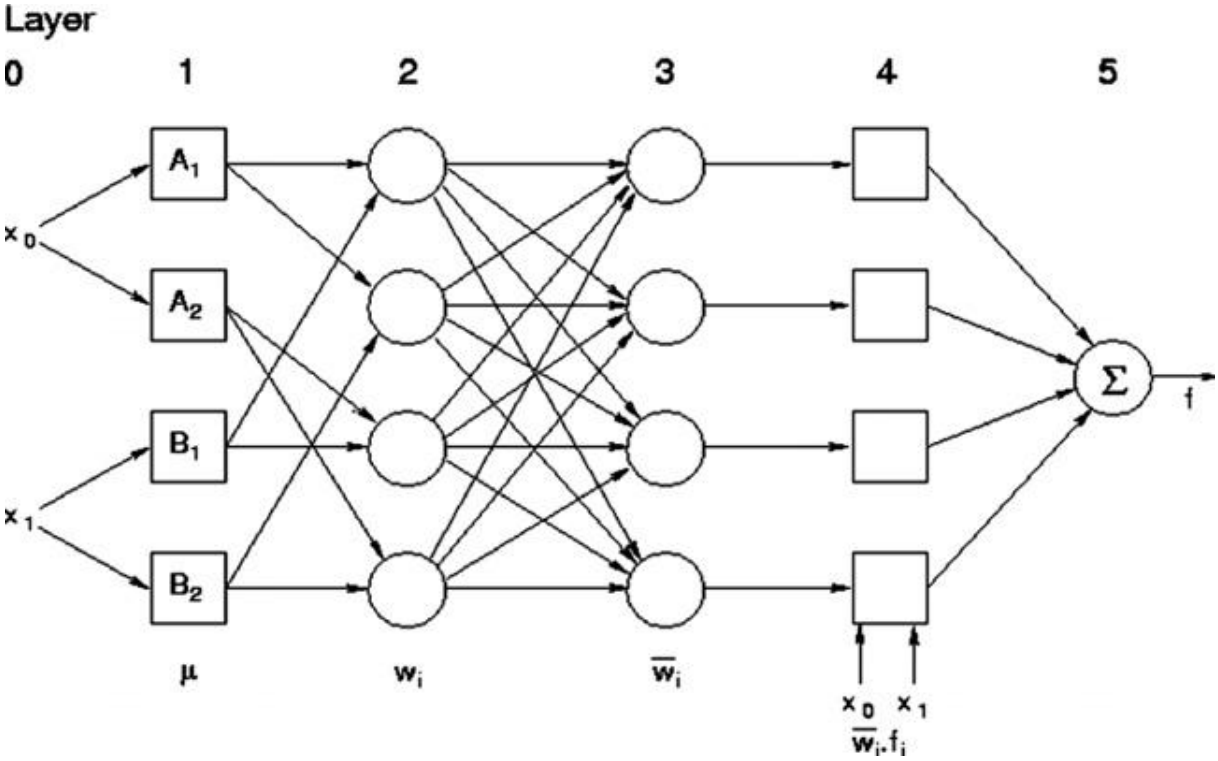
*THEN* $y = c_0 + c_1x_1 + c_2x_2 + \ldots + c_px_p$

*where* $x_1$ *and* $x_2$ *are the input variables,* $A_1$ *and* $A_2$ *are the membership functions, y is the output variable, and* $c_0$, $c_1$, *and* $c_2$ *are the consequent parameters ( Erdem Buyukbingol, 2007 Elsevier Ltd. doi:10.1016/j.bmc.2007.03.065) (Blanzieri & Bryl, 2008).*




Source: Adaptive neuro-fuzzy inference system (ANFIS): A new approach to predictive modeling in QSAR applications: A study of neuro-fuzzy modeling of PCP-based NMDA receptor antagonists (Blanzieri & Bryl, 2008).

**Figure 4.1 - A fuzzy linear regression model**




Source: Adaptive neuro-fuzzy inference system (ANFIS): A new approach to predictive modeling in QSAR applications: A study of neuro-fuzzy modeling of PCP-based NMDA receptor antagonists (Blanzieri & Bryl, 2008).

**Figure 4.2 - Basic ANFIS structure for 6 layers 2 inputs**

### 4.2 Genetic Algorithms

The genetic algorithm is a probabilistic search algorithm that iteratively transforms a set (called a population) of mathematical objects (typically fixed-length binary character strings), each with an associated fitness value, into a new population of offspring objects using the Darwinian principle of natural selection and using operations that are patterned after naturally occurring genetic operations, such as crossover (sexual recombination) and mutation. (Eiben, 1994)

Genetic algorithms have three stochastic operators; selection, crossover and mutation.

Selection replicates the most successful solutions found in a population at a rate proportional to their relative quality. The fittest population is chosen using predefined fitness function. Better individuals are preferred in selection but best is not always picked and worst is not

necessarily excluded.

Crossover is the exchange parts in populations for generating new members. Many crossover techniques exist for organisms which use different data structures to store themselves.

There are three main techniques for crossover:

One-Point Crossover

Two-Point Crossover

Cut and Splice Crossover

Mutation is the change of a part of member randomly. This change is used to maintain genetic diversity from one generation to the next.

The algorithm stops when one of the stopping criteria is reached. This can be the maximum number of population is reached or the improvement between generations is below a threshold.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | * | | | * | | |
| **Before:** | **(5** | **8** | **7** | **2** | **1** | **6** | **3** | **4)** |
| **After:** | **(5** | **8** | **6** | **2** | **1** | **7** | **3** | **4)** |

**Figure 4.3 - An example of mutation**

## 5. DISCUSSIONS

This section compares findings of the different neuro-fuzzy methods, NEFCLASS, ANFIS and GAANFIS.

NEFCLASS is the neuro-fuzzy classification technique that determines fuzzy rules and learns shapes of membership functions. Nefclass (prepared by Detlef Nauck and Ulrike Nauck) was tested with the dataset which is created in HP laboratories using an email account. Data has 57 attributes and a binary output giving spam or legitimate with 4109 instances. NEFCLASS program is trained with a subset of this dataset and NEFCLASS generated 10 rules.

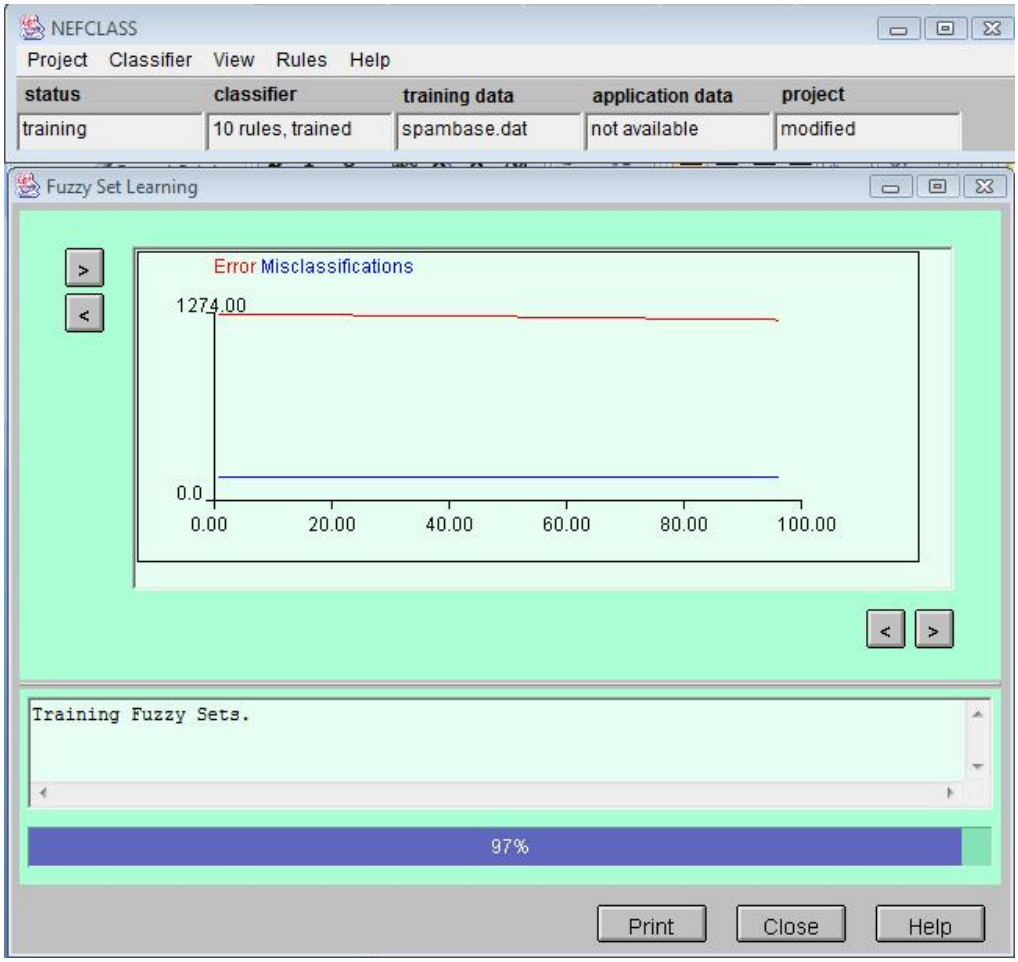



**Figure 5.1 – NEFCLASS output during training process**

NEFCLASS trained using 29 lines training data and 20 lines test data. The error rate for the NEFCLASS is 0.7000.

ANFIS algorithm is the second method after NEFCLASS. Utilizing the standard ANFIS algorithm with a subset of the spam dataset is tested. Training dataset has four inputs and one output with 40 instances. Anfis generated following system.

**Table 5.1 : ANFIS information**

| **ANFIS info:** | |
|---|---|
| Number of nodes | 193 |
| Number of linear parameters | 81 |
| Number of nonlinear parameters | 36 |
| Total number of parameters | 117 |
| Number of training data pairs | 40 |
| Number of checking data pairs | 0 |
| Number of fuzzy rules | 81 |

Numbers of linear parameters are computed by selecting three fuzzy sets for each input. This results in eighty-one for the four inputs and three memberships for each.

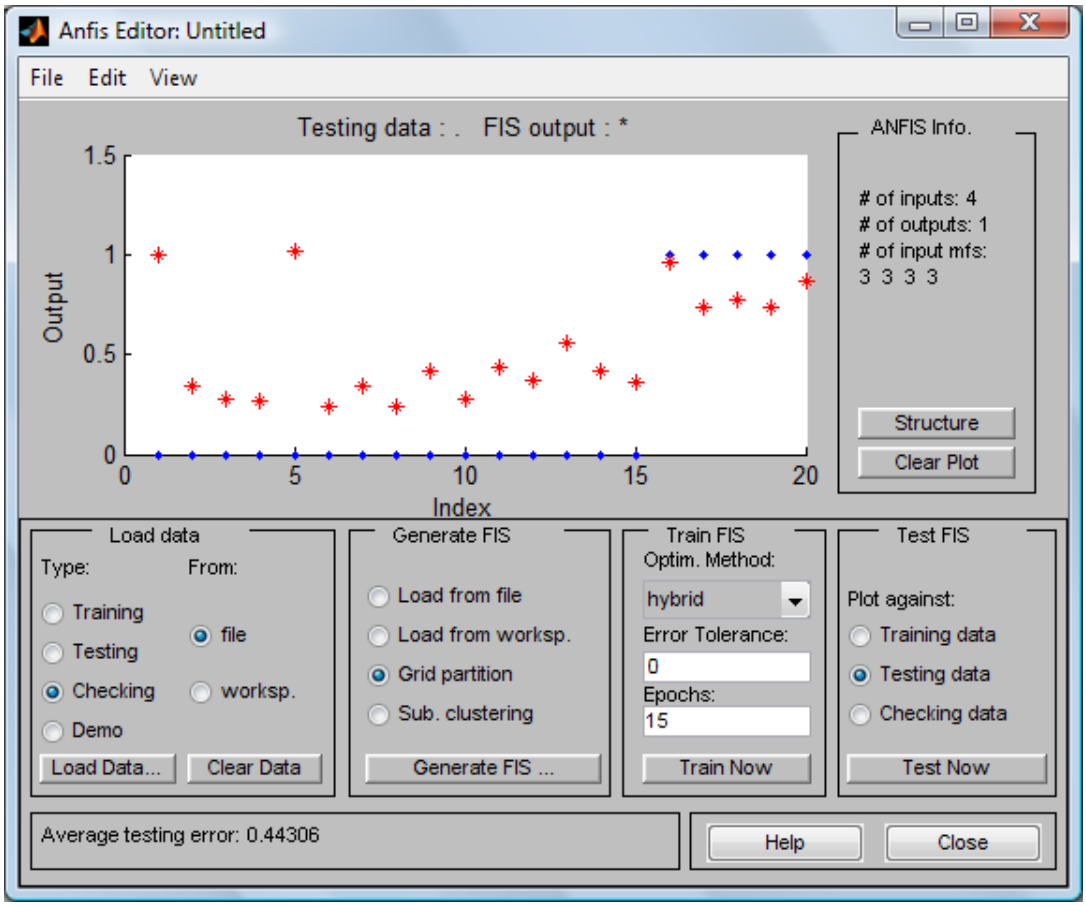



**Figure 5.2 – ANFIS output for sub dataset**

The resulting Average testing Mean Square Error is 0.44306.

Root Mean Square Error is defined as follows. N is the number of data, $e_i$ is the difference between the real value and expected value for ith entry:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}.$$

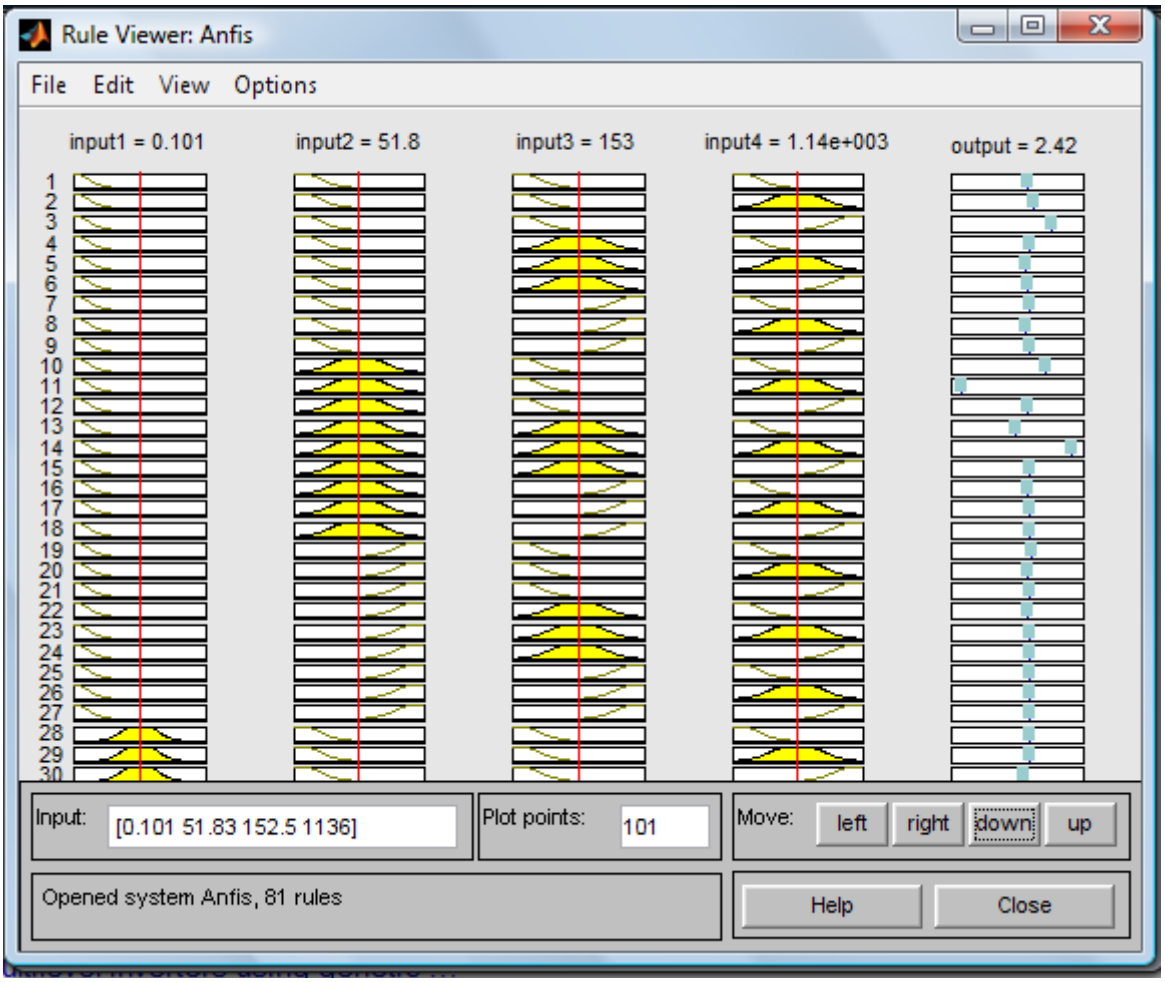



**Figure 5.3 – ANFIS rules for subdataset**

ANFIS generated 81 rules for four inputs and three membership functions for each input. Input of an excessive number would impair the transparency of the system and increases the complexity of the system, therefore increases the computational time. Therefore input selection is necessary for the computational time but selected features should represent the dataset. The purpose of the input selection is as follows:

- Remove noise or irrelevant inputs.
- Remove inputs that depend on other inputs.
- Ensure model is more concise and transparent.
- Reduce the time required for model construction.

The most relevant four inputs have been selected and a small subset of the data created as follows for the ANFIS training:

**Table 5.2 : Feature Selected Spam Database Subset**

| Input1 | Input2 | Input3 | Input4 | Output |
|---|---|---|---|---|
| 0 | 1,225 | 3 | 38 | 0 |
| 0 | 1,256 | 5 | 98 | 0 |
| 0 | 1 | 1 | 13 | 0 |
| 0 | 1,489 | 11 | 137 | 0 |
| 0 | 1,22 | 6 | 61 | 0 |
| 0 | 1,72 | 11 | 43 | 0 |
| 0 | 1,488 | 5 | 64 | 0 |
| 0 | 1,2 | 3 | 24 | 0 |
| 0 | 1,372 | 5 | 70 | 0 |
| 0,202 | 3,766 | 43 | 1789 | 0 |
| 0 | 1,312 | 6 | 21 | 1 |
| 0 | 1,243 | 11 | 184 | 1 |
| 0 | 3,728 | 61 | 261 | 1 |
| 0 | 2,083 | 7 | 25 | 1 |
| 0 | 1,971 | 24 | 205 | 1 |
| 0 | 5,659 | 55 | 249 | 1 |
| 0 | 4,652 | 31 | 107 | 1 |
| 0 | 35,461 | 95 | 461 | 1 |
| 0 | 1,32 | 4 | 70 | 1 |
| 0 | 3,509 | 91 | 186 | 1 |
| 0 | 3,833 | 9 | 23 | 1 |

| | | | | |
|---|---|---|---|---|
| 0,059 | 2,569 | 66 | 2259 | 1 |
| 0 | 4,857 | 12 | 34 | 1 |
| 0 | 1,131 | 5 | 69 | 1 |
| 0 | 5,466 | 22 | 82 | 1 |
| 0,059 | 2,565 | 66 | 2258 | 1 |
| 0 | 5,466 | 22 | 82 | 1 |
| 0 | 2,611 | 12 | 47 | 1 |
| 0 | 4 | 11 | 36 | 1 |
| 0 | 2,687 | 66 | 129 | 1 |
| 0,059 | 3,836 | 79 | 211 | 1 |
| 0 | 1,238 | 6 | 78 | 1 |
| 0 | 4,155 | 38 | 507 | 1 |
| 0 | 1,972 | 19 | 146 | 1 |
| 0 | 2,37 | 96 | 588 | 1 |
| 0 | 2,379 | 96 | 583 | 1 |
| 0 | 102,666 | 304 | 308 | 1 |
| 0 | 4,875 | 140 | 195 | 1 |
| 0 | 2,37 | 96 | 588 | 1 |
| 0 | 2,379 | 96 | 583 | 1 |

Table5.2 shows the inputs and outputs for the spam dataset. Output {0} refers to not spam and output {1} refers to spam. Input 1 refers to character '#', value is continuous. Input 2 refers to capital run length average, value is continuous. Input 3 refers to longest capital run length, value is continuous. Input 4 refers to total capital run length and value is continuous.

FIS structure for the ANFIS is as follows. FIS type is Sugeno type fuzzy system. Table 5.2 gives FIS system information.

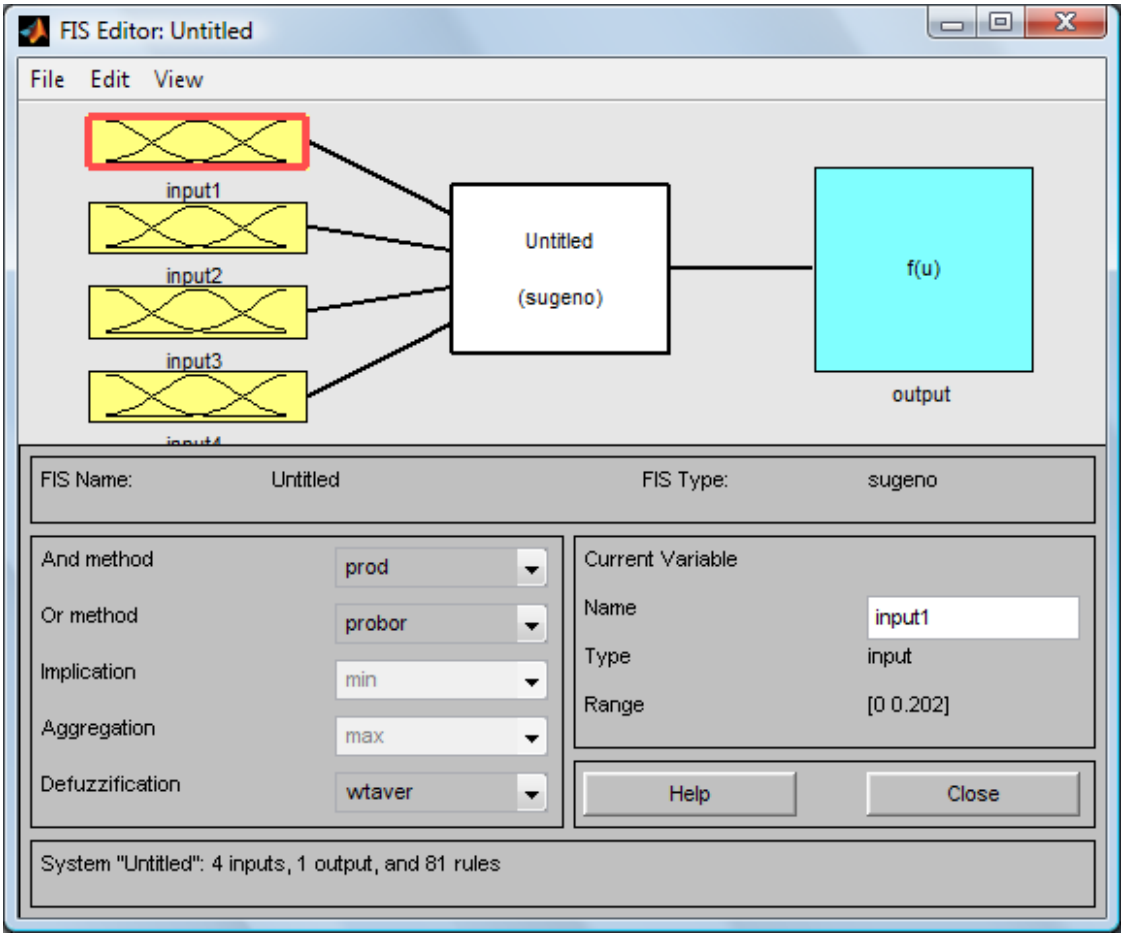



**Figure 5.4 – ANFIS fis editor**

Figure shows the 4 inputs 1 output FIS structure.

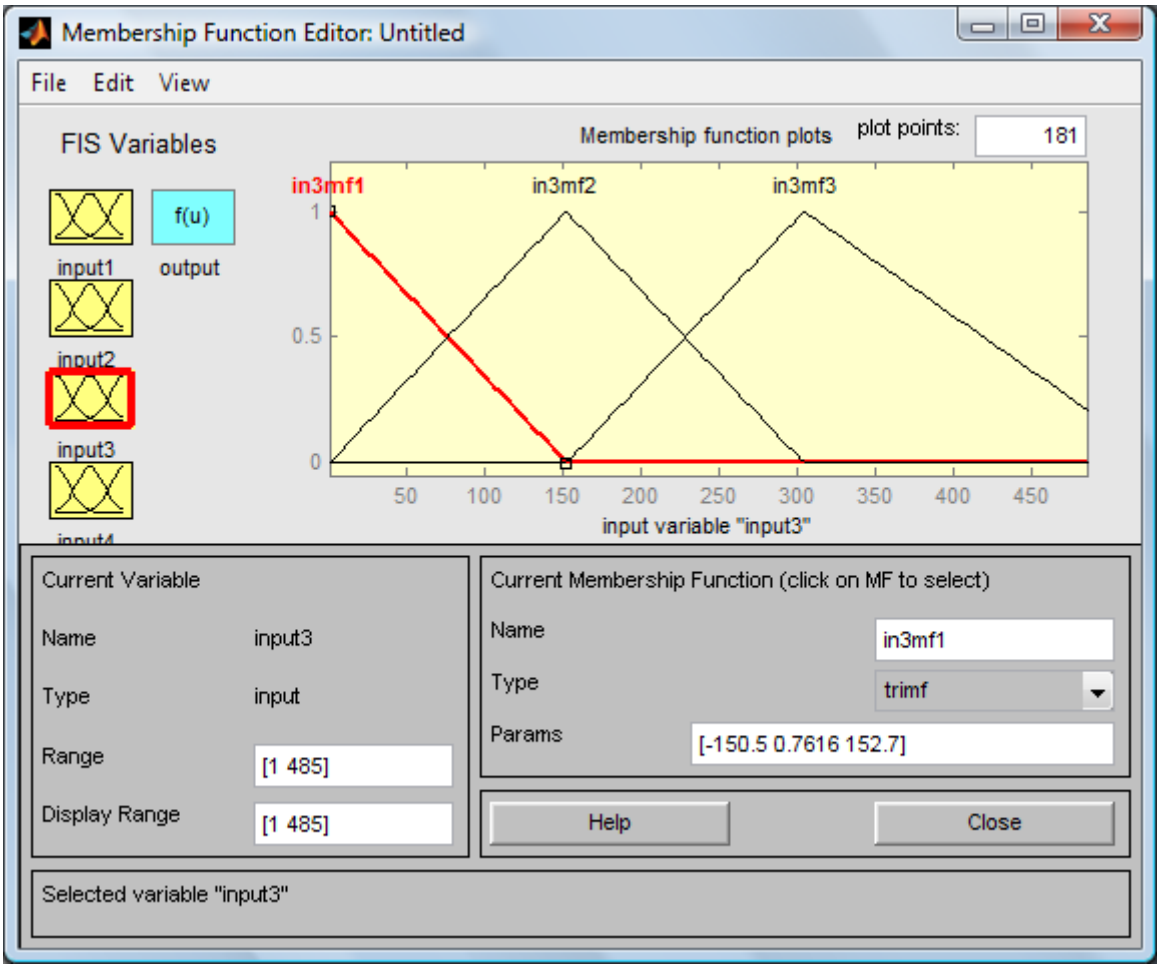



**Figure 5.5 – ANFIS Membership Functions**

The generated FIS structure is created using triangular membership functions. Figure 5.5 shows triangular membership functions. The ANFIS information is given below:

**Table 5.3 : ANFIS Information**

| **ANFIS info:** | |
|---|---|
| Number of nodes | 193 |
| Number of linear parameters | 81 |
| Number of nonlinear parameters | 36 |
| Total number of parameters | 117 |
| Number of training data pairs | 40 |
| Number of checking data pairs | 20 |
| Number of fuzzy rules | 81 |

Anfis information in Table 5.3 shows the information about the training data pairs, checking data pairs, node numbers according to inputs. The training data has 40 entries and 20 pairs checking data is used.

ANFIS was trained using the Sugeno type fuzzy system. System information is following:

**Table 5.4 : ANFIS System Information**

| **NAME** | **VALUE** |
|---|---|
| Type | Sugeno |
| andMethod | Prod |
| orMethod | Probor |
| defuzzMethod | Wtaver |
| impMethod | Prod |
| aggMethod | sum |
| input | [1x4 struct] |
| output | [1x1 struct] |
| rule | [1x81 struct] |

This system has four inputs and an output. System creates eighty-one rules by using all combinations of three fuzzy memberships for each four inputs.

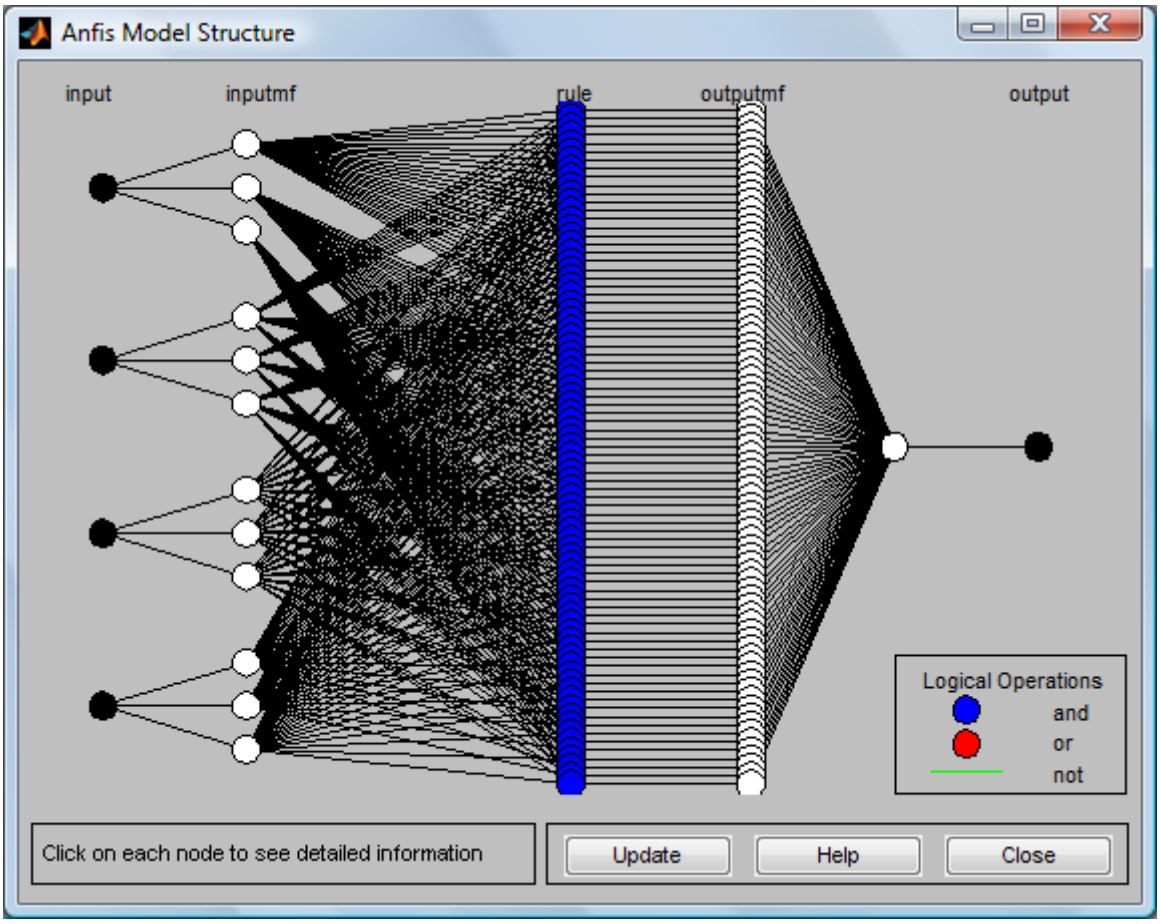



**Figure 5.6 – ANFIS Structure for 4 Input System**

ANFIS structure complexity increases with the input number and membership function numbers. This generates a complexity in training process and increases the learning time.

ANFIS uses back propagation or a combination of least squares estimation and back propagation for membership function parameter estimation. Membership function parameters can be optimized by utilization of Genetic Algorithms. This results in the proposed method for the spam filtering.

ANFIS program is updated using the genetic algorithms to fine tune the membership functions. GAANFIS is the genetic algorithm enhanced anfis program. GAANFIS generates a genetic algorithm based program based on MATLAB fuzzy toolbox and updated the membership function shapes using GA. This program provides a better starting and a quick convergence.

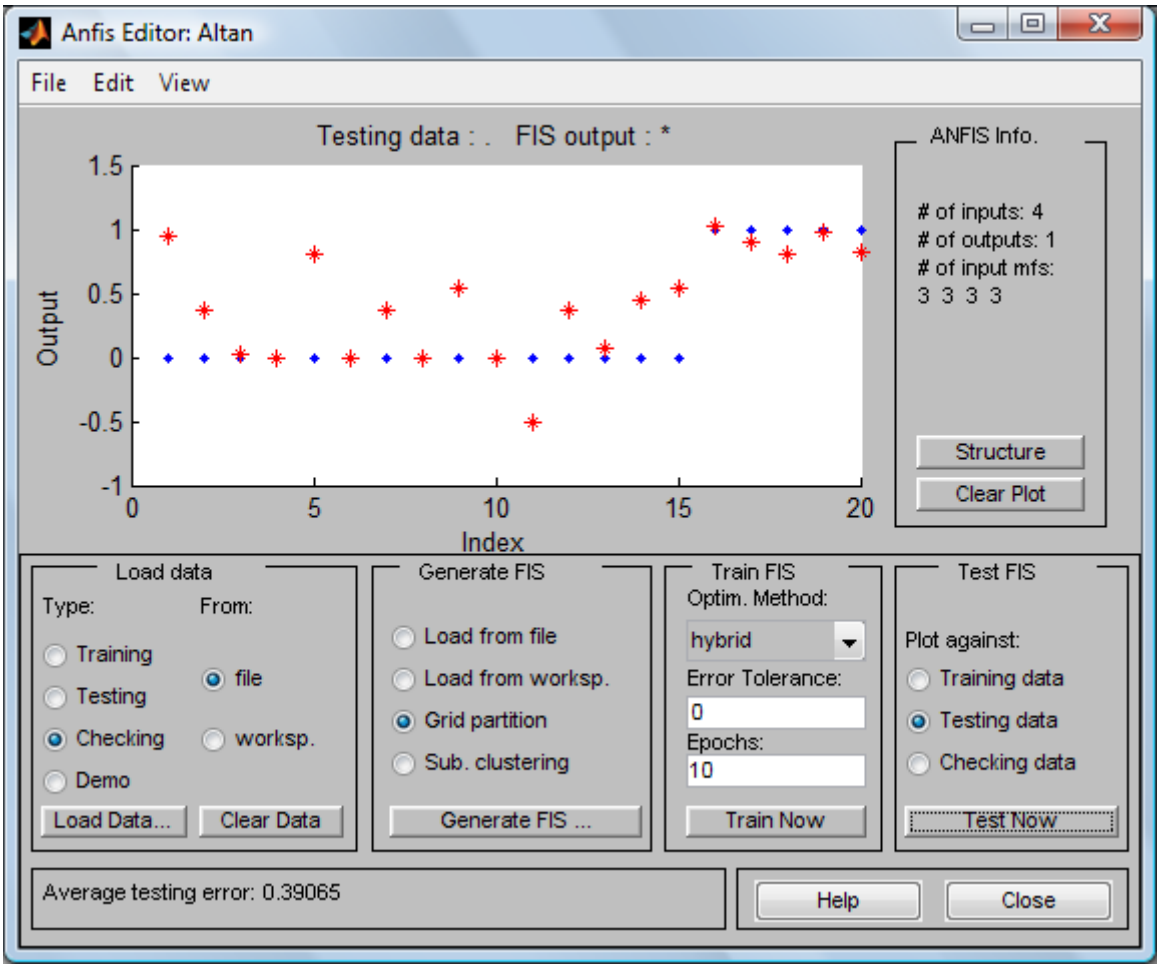



**Figure 5.7 – GAANFIS output for sub-dataset**

Average testing error for GAANFIS is 0.39065.

ANFIS information for the proposed method is as follows:

**Table 5.5 : ANFIS Information**

| **ANFIS info:** | |
|---|---|
| Number of nodes | 193 |
| Number of linear parameters | 81 |
| Number of nonlinear parameters | 36 |
| Total number of parameters | 117 |
| Number of training data pairs | 40 |
| Number of checking data pairs | 20 |
| Number of fuzzy rules | 81 |

Training process for the GAANFIS:

GAANFIS is trained using the same dataset as ANFIS and used Sugeno type fuzzy system.

System information is as follows:

**Table 5.6 : GAANFIS System Information**

| **NAME** | **VALUE** |
|---|---|
| Type | Sugeno |
| andMethod | Prod |
| orMethod | Probor |
| defuzzMethod | Wtaver |
| impMethod | Prod |
| aggMethod | sum |
| input | [1x4 struct] |
| output | [1x1 struct] |
| rule | [1x81 struct] |

GAANFIS generates a similar FIS structure where membership functions are fine tuned using genetic algorithms. GAANFIS FIS structure is given in Figure 5.8.

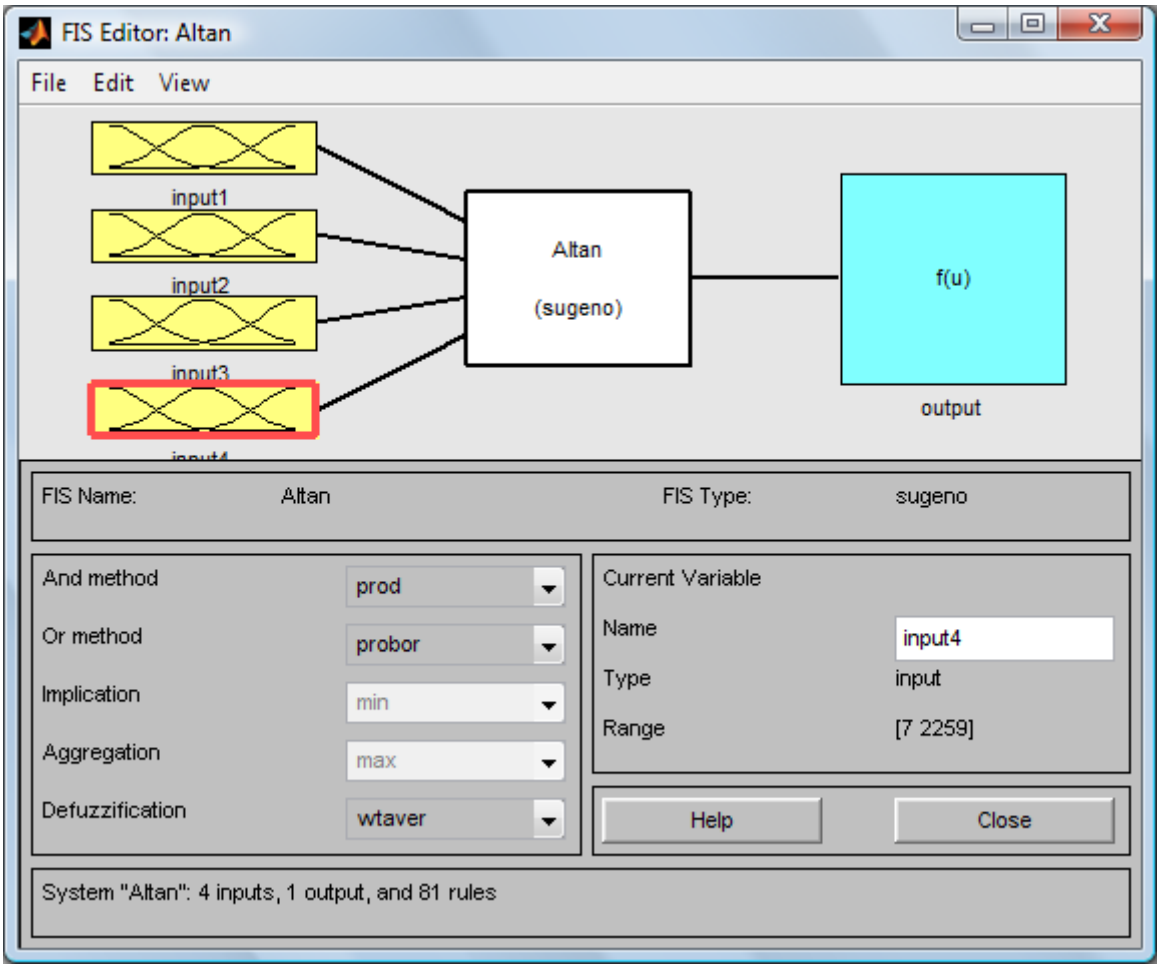



**Figure 5.8 – GAANFIS FIS structure**

Sugeno type system is preferred for anfis due to its advantages; computationally efficient, work well with optimization and adaptive techniques, guaranteed continuity of output surface and well suited to mathematical analysis.

Output surface plot for the input3 and input4 is given in the Figure 5.9.

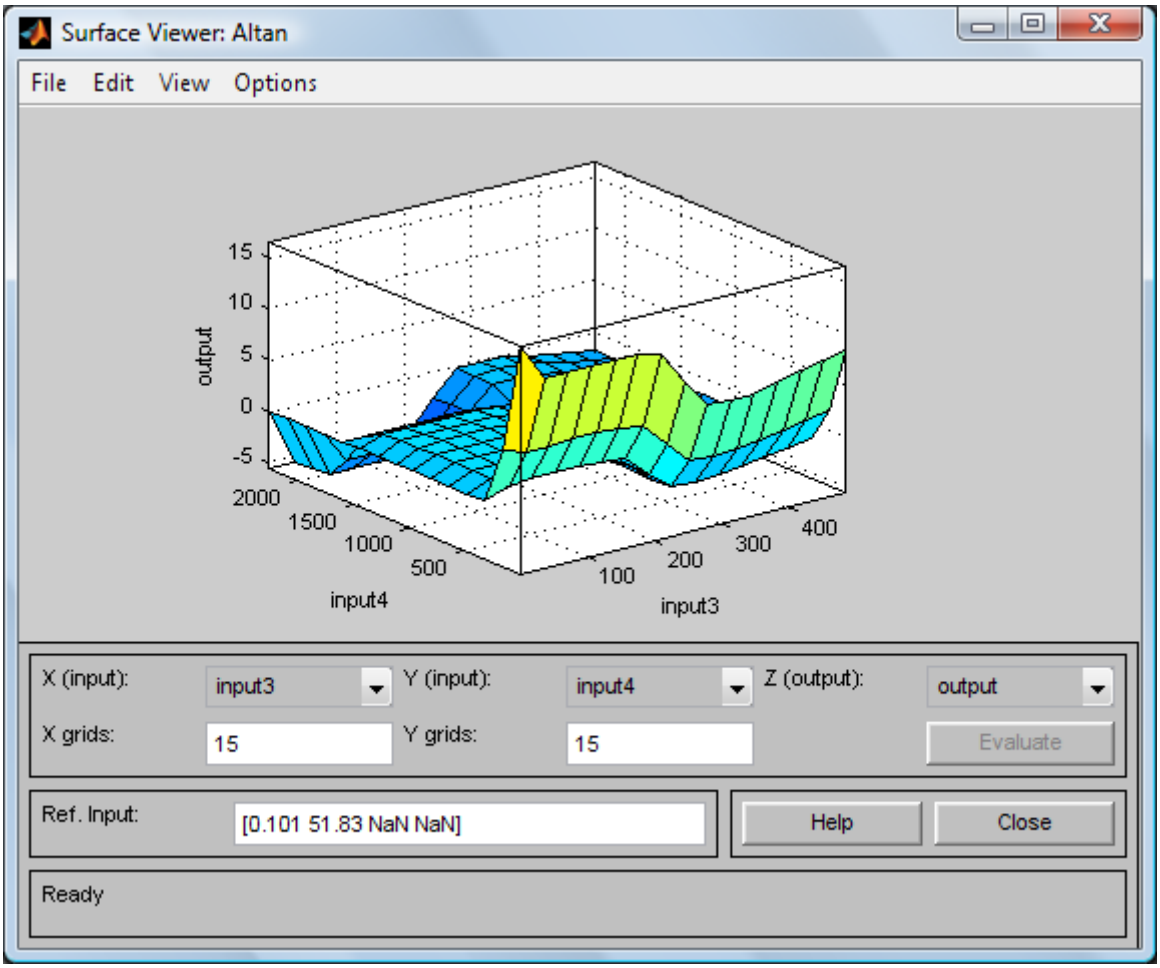



**Figure 5.9 – GAANFIS Rule Surface plot**

ANFIS optimization using GA improves the output error results. In ANFIS test case the fitness value function did not give necessary change for the rule base. The ANFIS program generated only one rule for the 4 inputs 30 entries small dataset.

Error Rates for the system:

**Table 5.7: Error Rates**

| **System** | **Error Rate (Mean Square Error)** | **Input number** |
|---|---|---|
| NEFCLASS | 0.70000 | 57 |
| ANFIS | 0.44306 | 4 |
| GAANFIS | 0.39065 | 4 |

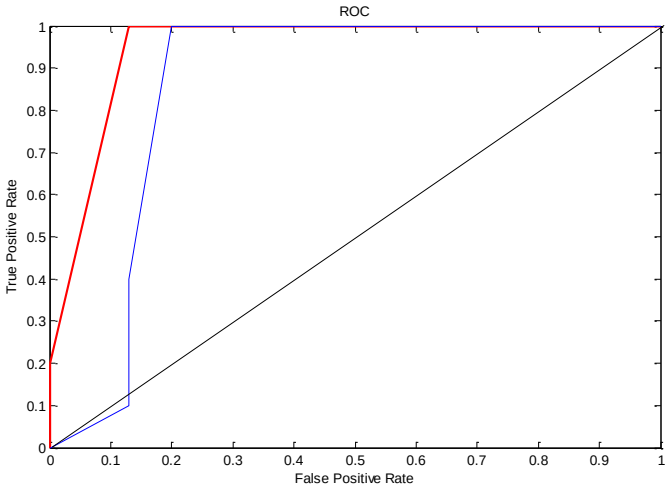



**Figure 5.10 – ROC curves**

Figure 5.10 shows the ROC curves for GAANFIS (red) and ANFIS (blue).

**Table 5.8: Classifier Performance**

| Classifier | Detection Rate | False Positive Rate | Gain |
|---|---|---|---|
| ANFIS | 90% | 2.8 | 75.7% |
| GAANFIS | 92% | 2.4 | 76.3% |
| NEFCLASS | 86% | 4.7 | 73.8% |

GAANFIS resulted in better output for the spam dataset with 20 entries as given in Table 5.7. However complexity of the model increased causing increase in training time.

GAANFIS structure is very slow, and time consumption for training is relatively too high for multiple inputs system. The current computer with 1.73GHz CPU and 1GB RAM is not sufficient to use MATLAB fuzzy toolbox with dataset including 57 inputs and 4109 instances. Computational time is not satisfactory in finding a spam email on the run.
The measurements taken from current computer using different input numbers in dataset with same number of instants and the corresponding figure is below:

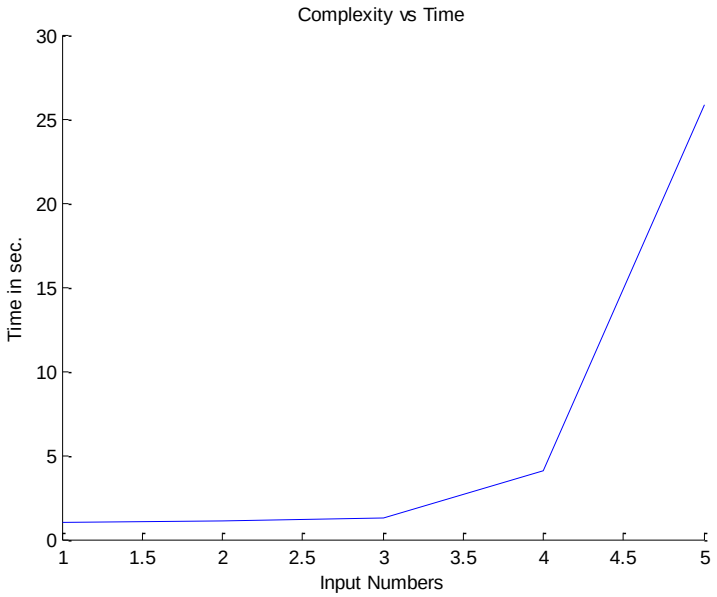



**Figure 5.11 – GAANFIS Model Complexity for Different Datasets**

Measurements for complexity:

**Table 5.9: Complexity vs. Time**

| Input Number | FIS Training | 3 Epocs Training | 10 Epocs Training |
|---|---|---|---|
| 1 | 11.1 sec | 0.5 sec | 1 sec |
| 2 | 11 sec | 0.82 sec | 1.1 sec |
| 3 | 13.5 sec | 1.1 sec | 1.3 sec |
| 4 | 13.6 sec | 1.84 sec | 4.1 sec |
| 5 | 21 sec | 3 sec | 25.8 sec |
| 6 | 27.1 sec | 25 sec | 4.5 mins |

Time measurement for the increasing input values and training period is given in the Table 5.5. Figure 5.11 shows the graphical view of the time for 10 epocs training using GAANFIS for different number of input variables. The measurements are taken for 40 entries datasets. Even for 6 inputs case MATLAB run out of memory and stuck at the training step and complete training in 4.5 minutes. The results show that time for training increases exponentially for input number increase and for constant number of instances.

The proposed model for fine tuning the ANFIS with GA is as follows:

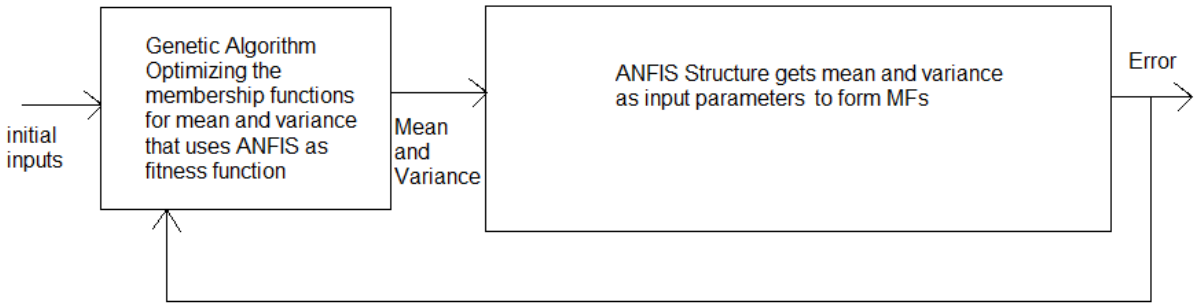



**Figure 5.12 – GAANFIS Model for MF Tuning**

In the ANFIS structure the initial MF is selected as fixed mean and variance. The number of MFs is selected by user where this requires expertise and analysis to identify. Genetic algorithm is well suited for the initial MF generation due to its ability to solve this fast and accurate way. This program requires to run ANFIS for each generations created by Genetic Algorithms. Genetic algorithm population fitness is evaluated by the ANFIS structure.

However this requires a server that is capable of handling this complexity and solution requires a long time using the 57 features dataset.

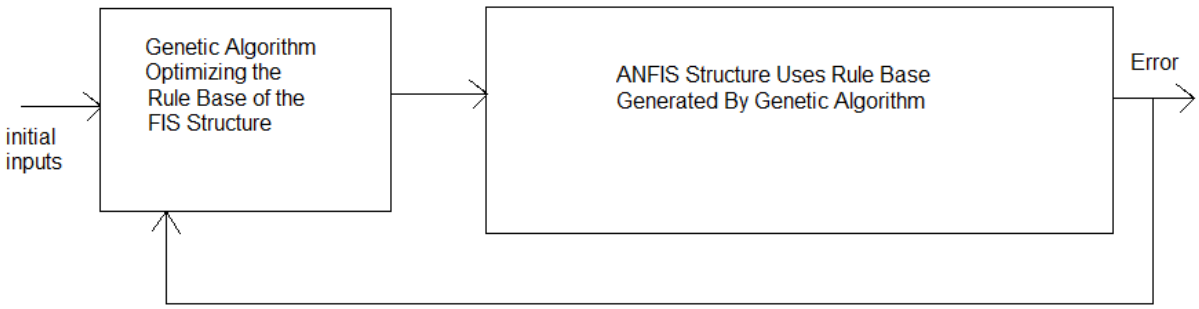



**Figure 5.13 – GAANFIS Model for Tuning Rule Base**

Rule Base Tuning model for the anfis structure proposed for the GA tuned ANFIS model. This model fine tunes the rule base and generates better results than ANFIS structure but when the complexity increases the training time increases exponentially with the increasing complexity. This is not an efficient way of finding the spam mails but can generate accurate results in the long run.

## 6. CONCLUSION

Internet is widely used nowadays and the increase in the spam emails causes time and money loss with disrupted users. Spam emails cause a waste of time and money for the individuals having approximately a hundred spam each day. Spam filtering is an important topic for saving people from unsolicited commercial emails. Spam filtering started as keyword filtering and black-list white-list approach. In the recent more aggressively created spam mails requires faster, self adoptable high accuracy filters.

In this thesis different spam detection and filtering methods are reviewed. The comparisons of the machine learning algorithms are given by Chih-Chin Lai and Ming-Chi Tsai (2004). The main issue in comparison is the weighting problem of the false positives and false negatives. One means a spam passing the filter whereas other means a valuable data loss. This measure depends on the user expectations.

According to the previous researches, combination of different techniques results in a better achievement in spam detection rates. An example is the ANN and Bayesian. The newly proposed system verifies this.

Proposed GA-ANFIS improved the results for the spambase data. However that increased the training period for the ANFIS structure. Since ANFIS on its own works slowly, proposed method makes it even slower while increasing the success of output prediction.

## REFERENCES

Androutsopoulos, I., Koutsias, J., Chandrinos, K.V. & Spyropoulos, C.D., 2000. "An experimental comparison of naive bayesian and keyword-based antispam filtering with personal e-mail messages." In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Athens, Greece, pages 160-167.

Androutsopoulos, I., Magirou, E. & Vassilakis, D. , 2005. "A game theoretic model of spam e-mailing." In Proceedings of Second Conference on Email and Anti-Spam, CEAS'2005.

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. & Stamatopoulos, P., 2000. "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory based approach." In H. Zaragoza, P. Gallinari, and M. Rajman, editors, Proceedings of the Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2000, pages 1–13.

Aradhye, H., Myers, G. & Herson, J. , 2005. "Image analysis for efficient categorization of image-based spam e-mail." In Proceedings of Eighth International Conference on Document Analysis and Recognition, ICDAR 2005, volume 2, pages 914–918. IEEE Computer Society.

Banday, M.T. & Jan, T.R., 2008. "Effectiveness and Limitations of Statistical Spam Filters", 2008 International Conference on "New Trends in Statistics and Optimization" – Imprint, 2009 – arxiv.org

Blanzieri, E., & Bryl, A., 2008. "A survey of learning-based techniques of email spam filtering",2008. Tech. rep. DIT-06-056, University of Trento, Information Engineering and Computer Science Department.

Buyukbingol, E., Sisman, A., Akyildiz, M., Alparslan, F. N. & Adejare A., 2007. "Adaptive neuro-fuzzy inference system (ANFIS): A new approach to predictive modeling in QSAR applications: A study of neuro-fuzzy modeling of PCP-based NMDA receptor antagonists", 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.bmc.2007.03.065

Carreras, X. & Marquez, L. , 2001. “Boosting trees for anti-spam email filtering.” In Proceedings of 4th International Conference on Recent Advances in Natural Language Processing, RANLP-01.

Clark, J., Koprinska, I. & Poon, J., 2003. “A Neural Network Based Approach to Automated E-mail Classification.” In Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI’03)

Cormack, G.. & Lynam, T. , 2005. “Spam corpus creation for TREC.” In Proceedings of Second Conference on Email and Anti-Spam, CEAS’2005.

Drucker, H., Wu, D. & Vapnik, V. , 1999. “Support vector machines for spam categorization.” IEEE Transactions on Neural networks, 10(5):1048–1054.

Eiben, A. E., 1994. "Genetic algorithms with multi-parent recombination". 1994, PPSN III: Proceedings of the International Conference on Evolutionary Computation. The Third Conference on Parallel Problem Solving from Nature: 78–87.

Fawcett, T., 2003. “in vivo spam filtering: A challenge problem for data mining.” KDD Explorations, 5(2):140–148.
http://home.comcast.net/~tom.fawcett/public_html/papers/spam-KDDexp.pdf

FerrisResearch, 2005. “The global economic impact of spam. report #409.”
Available at http://www.ferris.com/2005/02/24/the-global-economic-impact-of-spam-2005/
[Accessed 01 April 2010].

Gavrilis, D., Tsoulos, I.G., & Dermatas, E., 2006. “Neural Recognition and Genetic Features Selection for Robust Detection of E-Mail Spam”, G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 498 – 501.

Graham, P., “A plan for spam.”
Available at http://www.paulgraham.com/spam.html
[Accessed 01 April 2010].

Hidalgo, J. M. Gomez. , 2002. “Evaluating cost-sensitive unsolicited bulk email categorization.” In Proceedings of SAC-02, 17th ACM Symposium on Applied Computing, pages 615-620, Madrid, ES, 2002.

Hu, J., Li, Z., Hu, Z., Yao, D. & Yu J., 2008. “Spam Detection with Complex-Valued Neural Network using Behavior-based Characteristics”, © 2008 IEEE

Islam, Md. R., Chowdhury, M. U. & Zhou, W., 2005, “An Innovative Spam Filtering Model Based on Support Vector Machine”

Jezek, K. & Hynek, J., 2007. “The Fight against Spam - A Machine Learning Approach” In Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria.

Joachims, T. , 1997. “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.” In Douglas H. Fisher, editor, Proceedings of ICML-97, 14th International Conference on Machine Learning, pages 143–151, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

Lai, Chih-Chin & Tsai, Ming-Chi , 2004. “An empirical performance comparison of machine learning methods for spam e-mail categorization.” Hybrid Intelligent Systems, pages 44–48.

Medlock, B. , 2006. “An adaptive approach to spam filtering on a new corpus.” In Proceedings of the Third Conference on Email and Anti-Spam, CEAS’2006.

Metsis, V., Androutsopoulos, I. & Paliouras, G. , 2006. “Spam filtering with naïve bayes? which naive bayes?” In Proceedings of Third Conference on Email and Anti-Spam, CEAS’2006.

Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G.. & Stamatopoulos, P. , 2004. “Filtron: A learning based anti-spam filter.” In Proceedings of the First Conference on Email and Anti-Spam, CEAS’2004.

O'Brien, C. & Vogel, C. , 2003. "Spam filters: bayes vs. chi-squared; letters vs. words." In Proceedings of the 1st international symposium on Information and communication technologies, ISICT '03, pages 291–296, Dublin, Ireland. Trinity College Dublin.

Ozgur, L., Gungor, T. & Gurgen F., 2004. "Adaptive anti-spam filtering for agglutinative languages: A special case for Turkish", 2004 Elsevier B.V. doi:10.1016/j.patrec.2004.07.004

Pantel, P. & Lin, D., 1998. "Spamcop: A spam classification & organization program." In Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05.

Sahami, M., Dumais, S., Heckerman, D. & Horvitz, E. , 1998. "A bayesian approach to filtering junk e-mail." In Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05.

Scott Hazen Mueller, "What is spam"
Available at: http://spam.abuse.net/overview/whatisspam.shtml
[Accessed 01 April 2010].

Siponen, M. & Stucke, C., 2006. "Effective Anti-spam Strategies in Companies: An International Study"

The Spamhaus Project Ltd. "SPAMHAUS. The definition of spam."
Available at http://www.spamhaus.org/definition.html
[Accessed 01 April 2010].

Yih,W., Goodman, J., & Hulten, G. , 2006. "Learning at low positive rates." In Proceedings of the Third Conference on Email and Anti-Spam, CEAS'2006.

# VITAE

**Name Surname** : Altan Parlak

**Address** : NETAS Alemdag Caddesi No171 Umraniye Istanbul

**Birth Place / Year** **:** Eskişehir- 1982

**Languages** : Turkish (native) – English (Fluent) – French (Intermediate)

**Elementary School** : Ulku Ilkokulu – 1993

**High School** : Eskişehir Kılıcoglu Anatolian High School - 2001

**BSc** **:** Middle East Technical University - 2006

**MSc** **:** Bahcesehir University- 2010

**Name of Institute** : Institute of Science

**Name of Program** : Computer Engineering

**Work Experience** **:** Nortel Networks Netas

Software Engineer (December 2006 – Present)