International Journal of
Intelligent Systems and
Applications in Engineering

IJISAE

ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

Artificial Bee Colony Algorithm Based Linear Quadratic Optimal Controller Design for a Nonlinear Inverted Pendulum

Baris Ata[1*], Ramazan Coban[2]

Received 13th September 2014, Accepted 08th December 2014

Abstract: This paper presents a linear quadratic optimal controller design for a nonlinear inverted pendulum. Linear Quadratic Regulator (LQR), an optimal control method, is usually used for control of the dynamical systems. Main design parameters in LQR are the weighting matrices; however there is no relevant systematic techniques presented to choose these matrices. Generally, selecting weighting matrices is performed by trial and error method since there is no direct relation between weighting matrices and time domain specifications like overshoot percentage, settling time, and steady state error. Also it is time consuming and highly depends on designer's experience. In this paper LQR is used to control an inverted pendulum as a nonlinear dynamical system and the Artificial Bee Colony (ABC) algorithm is used for selecting weighting matrices to overcome LQR design difficulties. The ABC algorithm is a swarm intelligence based optimization algorithm and it can be used for multivariable function optimization efficiently. The simulation results justify that the ABC algorithm is a very efficient way to determine LQR weighting matrices in comparison with trial and error method.

Keywords: *ABC, LQR, Inverted Pendulum, Optimal Control, Weighting Matrices*

1. Introduction

The inverted pendulum is an unstable and under-actuated system with highly nonlinear dynamics [1]. The control of inverted pendulum is a classic example for design, testing, and comparing of different control techniques as a consequence the control of inverted pendulum has been a research interest in the field of control engineering and the inverted pendulum has been a standard tool in control laboratories for years. Another reason behind the extensive studies of the inverted pendulum is that several important control systems can be modelled with the help of inverted pendulum [2]. Inverted pendulum reveals many interesting system-theoretic properties and its dynamics are fundamental to maintenance balance, such as walking and two-wheeled robots [3], [4].

Optimal control theory is a mathematical optimization method as an extension of the calculus variation and it has numerous applications in control engineering. Determination of the control signals that will cause a process to meet the physical constraints and also maximization or minimization of some performance criteria are the main objectives of optimal control theory [5]. A special case of the general nonlinear optimal control problem where the cost function is a quadratic function and the system dynamics are described by a set of linear differential equations is a linear quadratic optimal control problem. Linear quadratic optimal control can be implemented in numerous control engineering problems, also it provides a basis for many other control techniques and hence it is very important for modern control theory [6]. Linear Quadratic Regulator (LQR) is one of the main solutions for linear quadratic optimal control problem. LQR has a simple process that can achieve the closed loop linear quadratic optimal control with linear state or output feedback [7], [8].

The most challenging part of LQR is selection of suitable weighting matrices which affects the control input [9]. In general, selecting weighting matrices is performed by trial and error method, however, there does not exist a direct connection between weighting matrices and time domain specifications such as overshoot percentage, settling time, and steady state error. There are no relevant systematic techniques for selecting weighting matrices, however; recently a few researchers have proposed artificial intelligence algorithms such as genetic algorithms and particle swarm optimization algorithm for this goal [10], [11]. In addition, the Artificial Bee Colony algorithm, another swarm intelligence optimization algorithm based on the intelligent behaviour of honey bee swarm, can be used to determine LQR weighting matrices.

The control problem is defined as selecting appropriate LQR weighting matrices to stabilize cart position and pole angle of nonlinear inverted pendulum while minimizing settling time, steady state error, and overshoot percentage. In this case study, the ABC algorithm is proposed to determine LQR weighting matrices and simulation results illustrate that proposed method achieves desired control system characteristics and also has a satisfactory control performance.

The rest of this paper is organized as follows. Section II contains the nonlinear mathematical model of the inverted pendulum that is used in this study. In section III, the linear quadratic optimal control problem based on linearized pendulum model is described. Section IV contains an overview of the ABC algorithm. Determination of LQR weighting matrices and simulation results are illustrated in section V followed by conclusion in section VI.

2. The Inverted Pendulum

[1] *Department of Computer Engg. Cukurova University, Adana, Turkey*
[2] *Department of Computer Engg. Cukurova University, Adana, Turkey*
* *Email: bata@cu.edu.tr*

Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

The inverted pendulum system used in this study consists of a motor driven cart and a pendulum hinged to it, as shown in (Figure 1). The inverted pendulum system model used in this paper has been suggested by Ogata [12]. The main aim of the controller is to stabilize the pendulum as to keep pendulum upright position in response to a change in cart position. Designed block diagram of control system is shown in (Figure 2). Linear quadratic optimal control where weighting matrices are selected by the ABC algorithm can be used to determine design variables; integral gain constant K_I and feedback gain matrix K.

Assume that the rod is massless and the pendulum mass is concentrated at the end of the rod. θ is the angel of the rod from vertical line and the control force u is applied to the cart. Also assume that the sampling period T is $0.1\ s$, g is $9.81\ m/s^2$ and the following numerical values for M, m and l:

$$M = 2kg,\ m = 0.1kg,\ l = 0.5m$$

In solving this design problem, we shall define state variables $x_1,\ x_2,\ x_3$ and x_4 as follows:

$$x_1 = \theta,\ x_2 = \dot{\theta},\ x_3 = x,\ x_4 = \dot{x}$$

for nonlinear inverted pendulum model these variables can be written as differential equations as follows: $\dot{x}_1 = x_2$ $\dot{x}_1 = x_2$

$$\dot{x}_1 = x_2 \tag{1}$$

$$\dot{x}_2 = \frac{u + \cos x_1 - (M+m)g\sin x_1 + ml\cos x_1 \sin x_1 x_1 x_2^2}{ml\cos^2 x_1 - (M+m)l} \tag{2}$$

$$\dot{x}_3 = x_4 \tag{3}$$

$$\dot{x}_4 = \frac{u + ml\sin x_1 x_2^2 - mg\cos x_1 \sin x_1}{M + m - m\cos^2 x_1} \tag{4}$$

which are solved using fourth-order Runge-Kutta method in this case study.

If displacement of the car is considered as the output of the system then the output equation y becomes

$$y = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \tag{5}$$

After linearizing the nonlinear differential (Equation 1) through (Equation 4) by taking $sin\theta \div \theta$, $cos\theta \div 1$, and $\theta\dot{\theta}^2 \div 0$, the discretized state and output equations of the system for linear quadratic optimal control can be derived as follows:

$$x(k+1) = Gx(k) + Hu(k) \tag{6}$$

$$y(k) = Cx(k) + Du(k) \tag{7}$$

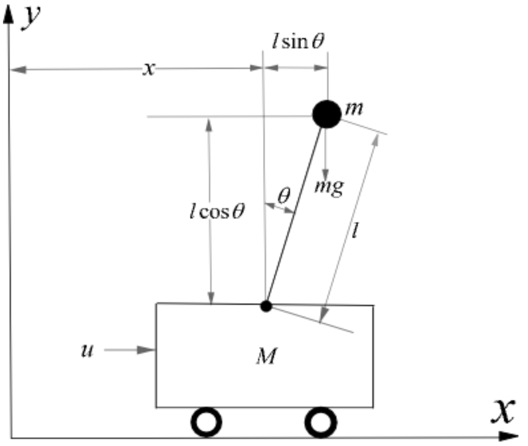


Figure 1. Inverted pendulum system [12]

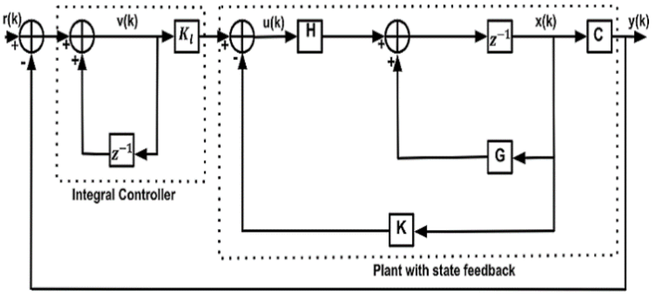


Figure 2. Block diagram of control system

where

$$G = \begin{bmatrix} 1.1048 & 0.1035 & 0 & 0 \\ 2.1316 & 1.1048 & 0 & 0 \\ -0.0025 & -0.0001 & 1 & 0.1 \\ -0.0508 & -0.0025 & 0 & 0 \end{bmatrix}$$

$$H = \begin{bmatrix} -0.0051 \\ -0.1035 \\ 0.0025 \\ 0.0501 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 0 \end{bmatrix}$$

3. Linear Quadratic Optimal Control

The state space representation of a linear time-invariant (LTI) control system can be written as follows [12]:

$$x(k+1) = Gx(k) + Hu(k) \tag{8}$$

$$y(k) = Cx(k) \tag{9}$$

$$v(k) = v(k-1) + r(k) - y(k) \tag{10}$$

$$u = -K(k)x(k) + K_I v(k) = -\begin{bmatrix} K & -K_I \end{bmatrix} \begin{bmatrix} x(k) \\ v(k) \end{bmatrix} \tag{11}$$

where $x(k)$ is state vector (n vector) and $u(k)$ is control vector (r

vector), respectively. G and H are $n \times n$ and $n \times r$ matrices, indicate the constant system. K is state feedback matrix.

At steady-state, the overall system dynamics with constant gain and integral feedback is described by the state equation which is a combination of (Equation 8) through (Equation 11):

$$\begin{bmatrix} x(\mathrm{k}+1) \\ v(k+1) \end{bmatrix} = \begin{bmatrix} G & 0 \\ -CG & 1 \end{bmatrix} \begin{bmatrix} x(\mathrm{k}) \\ v(k) \end{bmatrix} + \begin{bmatrix} H \\ -CH \end{bmatrix} u(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} r(k) \qquad (12)$$

In (Equation 12), it is assumed that the reference is a step change hence $r(k) = r(k+1)$.

Let us define

$$\hat{G} = \begin{bmatrix} G & 0 \\ -CG & 1 \end{bmatrix}$$

$$\hat{H} = \begin{bmatrix} H \\ -CH \end{bmatrix}$$

$$\hat{K} = \begin{bmatrix} K & -K_I \end{bmatrix}$$

$$\hat{x} = \begin{bmatrix} x \\ v \end{bmatrix}$$

$$\hat{D} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Therefore, the last equation can be written as follows:

$$\hat{x}(k+1) = \left(\hat{G} - \hat{H}\hat{K}\right)\hat{x}(k) + \hat{D}r(k) \qquad (13)$$

The (Equation 13) is the state equation of the closed-loop control system. Its output equation is

$$y(k) = \hat{C}\hat{x}(k) + [0]r(k) \qquad (14)$$

where $\hat{C} = [C \quad 0]$.

Linear quadratic optimal control problem may be stated to find the optimal input u sequence that minimizes the quadratic performance index which is defined as:

$$J = \frac{1}{2}\sum_{k=o}^{\infty}\left[x^*(k)Qx(k) + u^*(k)Ru(k)\right] \qquad (15)$$

where Q is positive definite $n \times n$ matrix and R is positive definite $r \times r$ matrix. The notation (*) indicates complex-conjugate transpose of a matrix. Matrices Q and R are selected to weight the relative importance of the performance measures caused by the state vector and control vector, respectively.

The state feedback gain matrix is defined as follows:

$$\hat{K} = (R + \hat{H}^*P\hat{H})^{-1}\hat{H}^*P\hat{G} \qquad (16)$$

which is obtained by solving the following Ricatti equation:

$$P = Q + \hat{G}^*P\hat{G} - \hat{G}^*P\hat{H}(R + \hat{H}^*P\hat{H})^{-1}\hat{H}^*P\hat{G} \qquad (17)$$

By the sense of Liapunov, for a stable matrix $(\hat{G} - \hat{H}\hat{K})$, the matrix P must be a positive definite, or for asymptotic stability a positive semi-definite.

According to main design parameters of the linear quadratic optimal control problem which are weighting matrices Q and R, the quality of the controller design depends on the choice of these matrices. However, there are no relevant systematic techniques to select weighting matrices. This goal is performed by trial and error method in general. Although it depends on the designer’s experience, it is highly time consuming and selected values for weighting matrices cannot establish a direct effect on the desired particular control system specifications.

According to the importance of selecting weighting matrices Q and R, selection of these matrices is performed by the ABC algorithm in this paper.

4. The Artificial Bee Colony Algorithm

The Artificial Bee Colony (ABC) algorithm was introduced by Karaboga in 2005 as a new method which is based on the intelligent behavior of honey bee swarms finding nectar and sharing the information of food resources with each other in the field Swarm Intelligence to solve to optimize numeric benchmark functions [13]. Then it was extended by Karaboga and Basturk and presented to exceed other recognized heuristic methods like Genetic Algorithm as well as Differential Evolution algorithm and Particle Swarm Optimization [14], [15]. The ABC algorithm has the advantages of strong robustness, fast convergence and high flexibility, fewer control parameters and also it can be used for solving multidimensional and multimodal optimization problems [16], [17].

In the ABC algorithm, the colony of artificial bees contains three groups of bees: employed bees, onlooker bees and scout bees. An employed bee memorizes the quality of the food source and finds a food source by modifying this information. Employed bees share the food source information with other bees on the dance area. Onlooker bees watch the dance of employed bees within the hive and find the food sources using the information provided by employed bees. Scout bees search new food sources around the hive randomly. Both onlookers and scouts are also called unemployed bees. The number of employed bees is equal to the number of food sources since each employed bee is associated with one and only one food source. The position of a food source means a possible solution to the problem and the nectar amount of a food source corresponds to the fitness of the associated solution.

The general scheme of the ABC algorithm contains four phase; initialization phase, employed bees phase, onlooker bees phase, and scout bees phase. Detailed pseudo code of the ABC algorithm is as follows [18]:

```
1.  Initialize the population of solutions
2.  Evaluate the population
3.  cycle=1
4.  repeat
5.  Produce new solutions (food source positions) $v_{ij}$ in the neighborhood of $x_{ij}$ for the employed bees using the formula $v_{ij} = x_{ij} + \Phi_{ij}(x_{ij} - x_{kj})$(k is a solution in the neighborhood of $i$ ,$\Phi$ is a random number in the range [-1,1] ) and evaluate them
6.  Apply the greedy selection process between $x_i$ and $v_i$
7.  Calculate the probability values $P_i$ for the solutions $x_i$ by means of their fitness values using the following equation
```

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \qquad (18)$$

In order to calculate the fitness values of solutions we employed the following equation

$$fit_i = \begin{cases} \dfrac{1}{1+f_i} & if\ f_i \geq 0 \\ 1+abs\left(f_i\right) & if\ f_i < 0 \end{cases} \qquad (19)$$

Normalize P_i values into [0, 1]

8. Produce the new solutions (new positions) v_i for the onlookers from the solutions x_i, selected depending on P_i, and evaluate them
9. Apply the greedy selection process for the onlookers between x_i andv_i
10. Determine the abandoned solution (source), if exists, and replace it with a new randomly produced solution x_i for the scout using the following equation

$$x_{ij} = min_j + rand\left(0,1\right) \times \left(\max_j - min_j\right) \qquad (20)$$

11. Memorize the best food source position (solution) achieved so far
12. cycle=cycle+1
13. until cycle= Maximum Cycle Number (MCN)

5. Simulation Results

The importance and difficulty of selecting weighting matrices was mentioned above. The matrices Q and R were chosen by trial and error method as follows in [12]:

$$Q = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \ R = \left[1\right]$$

Based on these matrices, feedback gain matrix K and integral gain constant K_I are determined as follows:

$$K = \left[-64.9346 \ \ -14.4819 \ \ -10.8475 \ \ -9.2871\right]$$

$$K_I = \text{-}0.5189$$

The goal of this simulation is to reduce the settling time $\left(t_s\right)$ of unit-step response $y(\mathrm{k})$ (the cart position) without an overshoot $\left(os\right)$ or with a minimum overshoot also minimize steady-state error$\left(e_{ss}\right)$. The objective weighting method where multiple objective functions are combined into one objective function f_{sum} can be used for multi-objective optimization [19]. The objective function defined as:

$$f_{sum} = K_1 t_s + K_2 os + K_3 e_{ss} \qquad (21)$$

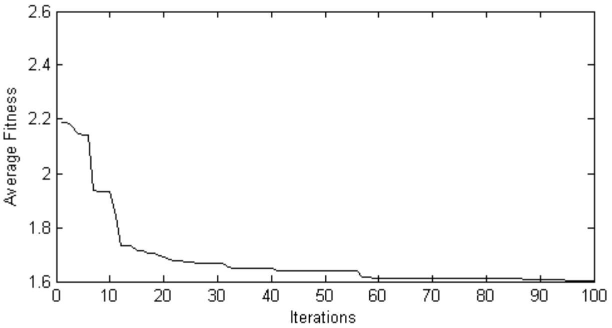


Figure 3. Average fitness during the ABC convergence

where K_1, K_2 and K_3 are weight coefficients of the fitness functions and their values were chosen as 1.0.

The ABC algorithm was employed to select best Q and R matrices that minimize f_{sum} . The results of applying the ABC algorithm to the problem are summarized as follows:

The parameters of the ABC algorithm are set in the range [0.1 100], colony size=20 and max cycle=100. The average fitness values during the ABC algorithm running is shown in (Figure 3). The weighting matrices obtained by ABC algorithm are:

$$Q = \begin{bmatrix} 82.4186 & 34.8695 & 50.4278 & 17.4539 & 66.1159 \\ 34.8695 & 24.7816 & 13.6630 & 3.5719 & 25.2813 \\ 50.4278 & 13.6630 & 51.4173 & 10.8109 & 39.9392 \\ 17.4539 & 3.5719 & 10.8109 & 11.4281 & 18.6369 \\ 66.1159 & 25.2813 & 39.9392 & 18.6369 & 60.7356 \end{bmatrix}$$

$$R = \left[0.1000\right]$$

Using the Q and R matrices obtained by the ABC algorithm the matrix P is calculated by (Equation 17) as follows:

$$P = \begin{bmatrix} 17385 & 3737.4 & 15220 & 6395.5 & -2403.4 \\ 3737.4 & 823.83 & 3315 & 1381.7 & -507.19 \\ 15220 & 3315 & 15439 & 5924.9 & -2660.2 \\ 6395.5 & 1381.7 & 5924.9 & 2436.7 & -952.22 \\ -2403.4 & -507.19 & -2660.2 & -952.22 & 644.42 \end{bmatrix}$$

Moreover, based on these matrices, feedback gain matrix $K$and integral gain constant K_I are determined as follows:

$$K = \left[\text{-}137.7804 \ \ \text{-}31.1832 \ \ \text{-}68.8161 \ \ \text{-}35.4986\right]$$

$$K_I = \text{-}6.9144$$

Table 1. Performance Results

	Trial and Error [12]	***The Proposed***
$t_s\left(s\right)$	4.8987	1.4963
$os\left(\%\right)$	0	0.1038
e_{ss}	2.3938×10^{-004}	5.3479×10^{-009}
f_{sum}	4.8989	1.6002

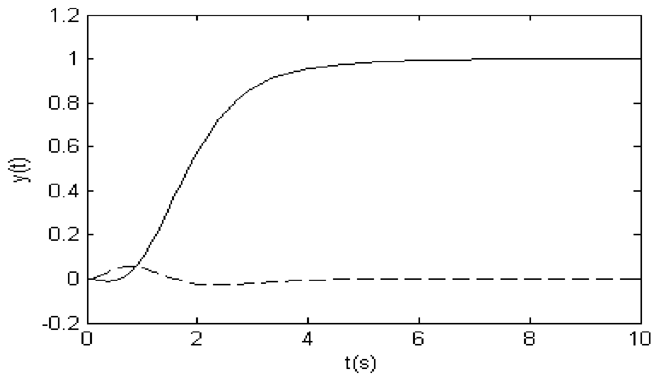


Figure 4. Cart Position and Pendulum Angle for Trial and Error Method

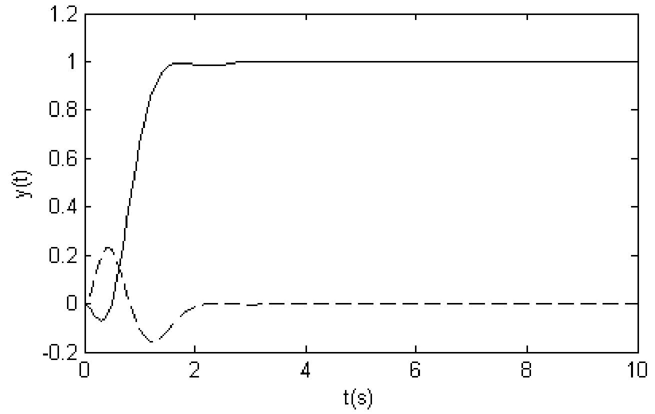


Figure 5. Cart Position and Pendulum Angle for Proposed Method

Since the matrix P is positive definite, the closed-loop control system is stable. That is, all eigenvalues of $\left(\hat{G}-\hat{H}\hat{K}\right)$ lie inside of the unit circle in the following:

$$z_1 = 0.2077$$

$$z_2 = 0.77628 + 0.28782\text{i}$$

$$z_3 = 0.77628 - 0.28782\text{i}$$

$$z_4 = 0.75904$$

$$z_5 = 0.72792$$

The performance results are presented in (Table 1). Also plots of position of the cart and pendulum angels as unit-step response of designed system has been is in (Figure 4) and (Figure 5) for both control systems. Cart position is indicated by solid line and pendulum angle is indicated by dotted line in these figures. Both simulations are performed on the nonlinear pendulum model given by (Equation 1) through (Equation 5).

6. Conclusion

In this paper, the ABC algorithm based linear optimal controller design for nonlinear inverted pendulum has been presented. The ABC algorithm has been employed to determine linear quadratic optimal controller weighting matrices. Using the ABC algorithm has been proved to be effective and feasible to select weighting matrices for nonlinear pendulum controller design more than trial and error method. Also it has been shown that it can optimize multiple time domain control system specifications such as settling time, overshot and steady state error.

References

[1] K. J. Aström and K. Furuta, "Swinging up a pendulum by energy control," Automatica, vol. 36, no. 2, pp. 287-295, 2000.

[2] J. Anderson, "Learning to control of an inverted pendulum using neural networks," IEEE Control Systems Magazine, vol. 9, no. 3, pp. 31-37, 1989.

[3] A. Kuo, "The six determinants of gaint and the inverted pendulum analogy: Adynamic walking perspective," Human Movement, vol. 26, pp. 617-656, 2007.

[4] S. Jeong and T. Takahashi, "Wheeled inverted pendulum type asistant robot: inverted mobile, standing and sitting motions," in IEEE/RSJ International Conferance on Intelligent Robots and Systems, 2007, 2007.

[5] D. E. Kirk, Optimal control theory: an introduction, New Jersey: Prentice-Hall, Inc., 1970.

[6] M. Athans, "The status of optimal control theory and applications for for deterministic systems," IEEE Transactions on Automatic Control, vol. 11, no. 3, pp. 580-596, 1966.

[7] R. E. Kalman, "When is a linear control system optimal?," Journal of Basic Engineering, vol. 86, pp. 51-56, 1964.

[8] H. Kwakernaak and R. Sivan, Linear optimal control systems, New York: Wiley-Interscience, 1972.

[9] J. V. D. F. Neto, I. S. Abreu and F. N. Da Silva, "Neural-genetic synthesis for state-space controllers based on linear quadratic regulator design for eigenstructure assignment," Trans. Sys. Man Cyber. Part B, vol. 40, no. 2, pp. 266-285, 2010.

[10] C. P. Bottura, J. V. da Fonseca Neto, "Parallel eigenstructure assignment via LQR design and genetic algorithms," in Proceedings of the American Control Conference 1999. Vol. 4. , San Diago, 1999.

[11] S. Mobayen, A. Rabiei, M. Moradi and B. Mohammady, "Linear quadratic optimal control system design using particle swarm optimization algorithm," International Journal of the Physical Sciences, vol. 6(30), pp. 6958-6966, 2011.

[12] K. Ogata, "Quadratic Optimal Control Systems," in Discrete-Time Control Systems 2nd Edition, Englewood Cliffs, NJ, Prentice-Hall, 1995, pp. 566-633.

[13] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Erciyes University, Kayseri, Turkey, 2005.

[14] D. Karaboga and B. Basturk, "An artificial bee colony (ABC) algorithm for numeric function optimization.," in IEEE Symp. Swarm Intelligence, Indianapolis, 2006.

[15] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," Applied Soft Computing, vol. 8, no. 1, pp. 687-697, 2008.

[16] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," Journal of Global Optimization, vol. 39, no. 3, pp. 459-471, 2007.

[17] D. Karaboga and B. Akay, "A comparative study of Artificial Bee Colony algorithm," Applied Mathematics and Computation, vol. 214, no. 1, p. 108–132, 2009

[18] "Detailed Pseudocode of the ABC Algorithm," 14 October 2008. [Online]. Available: http://mf.erciyes.edu.tr/abc/pub/PsuedoCode.pdf. [Accessed 1 March 2014].

[19] R. Coban , "A fuzzy controller design for nuclear research reactors using the particle swarm optimization algorithm," Nuclear Engineering and Design, vol. 241, no. 5, pp. 1899-1908, 2011.

International Journal of
Intelligent Systems and
Applications in Engineering

ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

An Efficient Document Categorization Approach for Turkish Based Texts

Sevinç İlhan Omurca*[1], Semih Baş[2], Ekin Ekinci[1]

Received 14th September 2014, Accepted 18th December 2014

Abstract: Since, it is infeasible to classify all the documents with human effort due to the rapid and uncontrollable growth in textual data, automatic methods have been approached in order to organize the data. Therefore a support vector machine (SVM) classifier is used for text categorization in this study. In text categorization applications, the text representation process could take a huge computation time on weighting the huge size of terms. So far, lexicons that contain less number of terms are used for the solution in the literature. However it has been observed that these kinds of solutions reduce the accuracy of the text classification. In this paper, the term-document matrix is constructed as user dependent according to the purpose of classification. Since the number of terms is still relatively large, we used a hash table for efficient search of terms. Hereby an efficient and rapid TF-IDF method is introduced to construct a weight-matrix to represent the term-document relations and a study concerning classification of the documents in Turkish based news and Turkish columnists is conducted. With the proposed study, the computational time that is required for term-weighting process is reduced substantially; also 99% accuracy is achieved in determination of the news categories and 98% accuracy is achieved in detection of the columnists.

Keywords: *Document categorization, SVM, TF-IDF, User dependent term selecting, Hash table.*

1. Introduction

Due to the rapid and uncontrollable growth in textual data, especially with the domain World Wide Web (www), it is infeasible to manually classify the huge size of documents with high-dimensional text features, so the automatic methods for organizing the data are needed. Text classification is the task of assigning the documents to a set of predefined classes based on their contents. Classification of web pages, filtering of spam e-mails, categorization of topics, retrieving user reviews, author recognition are some popular application areas of text classification.

There are certainly a broad range of machine learning methods available for text classification problems in the literature. The most popular ones include regression models, probabilistic Bayesian models, decision trees, decision rule learners, K-nearest neighbors (KNN), computing with words, association rule mining and SVM. Among these methods, SVM achieves superior results in text classification and pattern recognition problems [1]. (Fabrizio Sebastiani, 2005) also emphasized SVM classifier in his review paper of text categorization because of its best performance in comparative text categorization experiments so far.

Here some of the approaches and techniques have been applied recently in the field of text classification are referred. (Zhang et al; 2008) investigate the effectiveness of using multi-words for text classification with SVM and also effectiveness of linear kernel and polynomial kernel in SVM comparatively. (Li et al; 2011) proposed a hybrid algorithm that combines SVM and KNN and overcomes the drawbacks of sensitive to noises of SVM and low efficiency of KNN. (Sun et al; 2009) realized a comparative study on the effectiveness of strategies addressing imbalanced text classification using SVM and make a survey on the techniques proposed for imbalanced classification. (Miao et al; 2009) proposed a hybrid algorithm which is based on variable precision rough set and KNN to overcome their weaknesses. (Shi et al; 2011) studied semi-supervised text classification; they tried to learn from positive data without negative data and also with the help of unlabeled data. They use SVM, Naive Bayes and Rocchio as classifiers to construct a set of classifier. (Mitra et al; 2007) proposed a least square support vector machine (LS-SVM) that classifies noisy document titles and the proposed system was compared with KNN and Naive Bayes. It was observed that LS-SVM with LSI based classifying agents improves text classifying performance significantly. (Lo, 2008) proposed an auto mechanism to classify customer messages based on the techniques of text mining such as dictionary approach or TF-IDF and SVM then exceeded 83-89% success in classifying. (Rajan et al; 2009) proposed an ANN model for the classification of Tamil language documents and the model achieved 93.33% accuracy. (Zhang et al; 2013) used Rough Set which is based on Rough Set decision making approach for classifying texts which are not easily classified with classical methods. They used CEI for performance evaluation. (Adeva et al; 2014) studied with SVM, Naïve Bayes, KNN and Rocchio for medical-domain texts. They combined these algorithms with 7 different feature selection algorithms and different number of features and used 3 different document sections. (Lee et al; 2012) proposed a new approach, called as Euclidean-SVM. In training phase they used SVM and in classification phase they used Euclidean distance function instead of optimal hyper-plane.

Due to the literature, there are only a few text classification approaches that have been applied in Turkish documents.

[1] *Kocaeli University, Faculty of Engineering, Computer Engineering Department Umuttepe Campus, Kocaeli – 41380, Turkey*

[2] *Tubitak Marmara Research Center Technology Free Zone, IBTECH, Kocaeli – 41470, Turkey*

* *Corresponding Author: Email: silhan@kocaeli.edu.tr*

Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

(Kılıçaslan et al; 2009) explored machine learning models such as Navie Bayes, KNN, decision trees, SVM and voted perceptron for pronoun resolution in Turkish. (Çıtlık and Güngör, 2008) employed SVM and boosting classifiers in spam filtering and achieved high accuracies. (Özyurt and Köse, 2010) studied chat mining. They used Naive Bayes, KNN and SVM to classify Turkish chat conversation texts and achieved 90% accuracy in determination of subject. (Özgür et al; 2004) proposed an anti spam filtering based on Artificial Neural Networks and Bayesian Networks. They tested the system with 750 e-mails and achieved 90% accuracy. (Alparslan et al; 2011) proposed a hybrid system for document classification that considers SVM and adaptive neuro-fuzzy classifier and 96.67% accuracy was achieved. (Uysal and Gunal, 2014) proposed to show impact of preprocessing on text classification. To this end, they used SVM to classify Turkish and English news and e-mails.

In this paper, we have applied a supervised machine learning method in order to classify the Turkish news and also predict the columnists of newspaper articles. There are not many work have been done in Turkish news classification or author detection. (Türkoğlu et al; 2007) identified the author of an unauthorized document by using n-grams and determined the most success classifiers were SVM and Multi Layer Perceptron (MLP). An average accuracy of 88.9% was achieved by SVM. In the current method by using the TF-IDF term weighting method and SVM classifier, a success rate of 96.4% and a lower time complexity are obtained. Thus, it is concluded that great time savings are possible without decreasing the accuracy level.

Our study has two main phases, the first one is text representation phase that is realized by TF-IDF method and the second one is the text classification phase that is realized by a SVM classifier. In text representation process, the huge size of terms entire dataset namely huge amount of feature set causes huge computation time on weighting these terms [21]. In the text representation phase of our application, unlike from the other applications in the literature, the words that are inefficient for classification are subtracted to reduce the term space. The subtracting process is realized by the user due to the characteristics and purposes of classification task. Namely, if sentiment classification of the textual data will be realized then the verbs would be so important even the proper names would not. On the other hand, if the category of documents will be estimated then the proper names would be so important. Or generally, the conjunctions have less importance in text mining independently from the classification task. These kinds of determinations about the term selecting process must be done with the expert persons on the classification tasks.

The rest of the paper is organized as follows. Section 2 gives an overview of the TF-IDF and SVM methods. Section 3 discusses the experimental setup. Section 4 shows the results of experiments and section 5 gives the concluding remarks.

2. Brief Overview

2.1. TF-IDF

In text classification problems, for most of the training algorithms, a document should be represented as a vector of numbers. A method called term frequency (TF) and inverse document frequency (IDF) are used to represent text with vector space model. There is an extension of term frequency inverse document frequency (TF-IDF) developed from IDF which is proposed by ([22], [23]) and expresses that a term which appears in many documents is not a good discriminator and should be given less weight than another term which appears in few documents [24]. Intuitively, this method determines how relevant a given word is in a particular document. Terms that are common in a small group of document-set tend to have higher TF-IDF numbers than common terms such as prepositions or articles.

Assume there are N documents in the collection, t_i denotes term i and occurs n_i of documents. Then inverse document frequency is formulized as in (Equation.1).

$$IDF(t_i) = \log \frac{N}{n_i} \tag{1}$$

In text classification models, a text can be defined as a term matrix, $D = \lfloor d_{ij} \rfloor_{m\times n}$, where n denotes the number of documents, m denotes the number of different terms and d_{ij} denotes the weight value of the term t_i in document d_j. TF-IDF method expressed by (Equation.2) is used to compute the term weight values. Where tf_{ij} indicates the frequency of the term i in the document j [25].

$$dij = TF_{ij} \times IDF_i = \frac{TF_{ij} \times \log_2 (N / n_i + 0.01)}{\sqrt{\sum_{j=1}^{m} (TF_{ij} \times \log_2 (N / n_i + 0.01))^2}} \tag{2}$$

TF-IDF formulation is used to measure the discrimination or importance value of a term in the document collection [26]. However there is an important criticism of using this method for text representation. This comes from the huge dimensionality of term-document matrix, resulting in that it causes a huge computation time on weighting these terms [21]. High dimensionality of the feature space is also addressed as the major difficulty of text categorization problems. The classification accuracy is directly connected with how much of the document terms can be reduced without losing useful information in category representation [27]. Consequently a powerful test representation implementation would not only decrease the computational time for the task but also improve the accuracy of the classification task.

2.2. SVM

SVM is a computational learning method uses machine learning theory presented and developed by [28]. (Joachims, 1998) was firstly proposed SVM for text classification tasks and just it is clearly known that, SVM is one of the most important learning algorithms for text classification due to its robustness on high dimensional spaces [30].

In SVM, original input space is mapped into high dimensional feature space and in this space there are many hyper-planes (linear classifiers) that separate the data. The optimal hyper-plane among them that achieves maximum separation is determined by optimization theory to maximize the generalization ability of the classifier [31].

A training data set represented by n-dimensional input $x_i \in R^n$, $i = 1,\ldots,l$ and l is the number of samples that belong to target classes $y_i \in \{1,-1\}$. A hyper plane $f(x)=0$ that separates the data is tried to find.

$$f(x) = w \cdot x + b = \sum_i^n w_i \cdot x_i + b = 0 \tag{3}$$

where $w = (w_1,\ldots,w_n)$ and $b \in R$. The aim is correctly classifying the data and a distinctly separating hyper-plane satisfies these conditions.

$$y_i(x_i \cdot w + b) \geq 1, i = 1, \ldots, l \tag{4}$$

$$f(x) = w \cdot x_i + b \geq 1, y_i = +1 \tag{5}$$

$$f(x) = w \cdot x_i + b \leq 1, y_i = -1 \tag{6}$$

Among all possible hyper-planes SVM selects the optimal separating hyper-plane that creates the maximum margin. The optimal hyper plane can be found by solving a quadratic optimization problem in (Equation 7). ξi is slack variable represents noise and C is error penalty determines the trade-off between model complexity and loss function.

Minimize: $$\phi(w\xi) = \frac{1}{2(w \cdot w)} + C(\Sigma_{i=1}^{l}\xi_i) \tag{7}$$

Subject to: $$y_i(x_i \cdot w + b) \geq 1 - \xi_i, i = 1, \ldots, l \tag{8}$$

For simplification of the calculations, the optimization problem has been converted to Lagrange dual problem with Kuhn-Tucker conditions.

$$\Sigma_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i}\sum_{j}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \tag{9}$$

Subject to: $$\sum_{i=1}^{l}\alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, l \tag{10}$$

$K(x_i, x_j)$ is the inner product $\left\langle \phi(x_i)\phi(x_j) \right\rangle$ in feature space and called as kernel function.

$$K(x_i, x_j) = \left\langle \phi(x_i)\phi(x_j) \right\rangle \tag{11}$$

ϕ is a mapping from X to inner product feature space F. In practice ϕ and F are derived from the definition of kernel function. There are different kernel functions for SVM. (Joachims, 2002) and (Dumais et al; 1998) reported an important finding in text classification that linear SVM performs better than nonlinear SVM so in this paper a linear kernel function is used for SVM. The other common kernel functions are as follows and called Polynomial Kernel, Radial Basis Kernel and Sigmoid Kernel Function respectively.

$$K(x_i, x_j) = \left[\left(x_i \cdot x_j + 1\right)\right]^q \tag{12}$$

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{2\alpha^2}\right) \tag{13}$$

$$K(x_i, x_j) = \tanh\left(v\left(x_i, x_j\right) + C\right) \tag{14}$$

The final mapping function $f(x)$ between the input variable space and the desired output variable can be expressed in terms of the SVs (training examples) as follows:

$$f(x) = \Sigma i, j = 0 \alpha_i y_i K(x_i, x_j) + 1 \tag{15}$$

where x_i, x_j are SVs for class 1 and class 2, respectively [31].

3. Experimental Setup

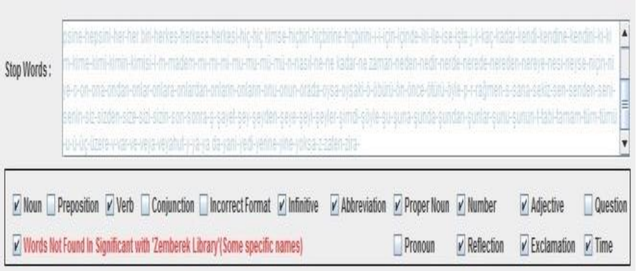
3.1. Text Collection and Pre-processing

The first phase of the text classification model is the pre-processing phase which includes elimination of stop words, determination of the terms and stemming. Stop words elimination filters out the words that are not relevant in the analysis of documents and usually consist of articles, pronouns, prepositions, interjections among others.

For the pre-processing step of the application, initially, we eliminate the stop words from the document set directly. The stop word list including about 223 words is obtained from [33]. Thus the computation complexity for multiword representation can be reduced by stop word elimination because they usually have high frequency in documents. Then, the user selected terms are also eliminated from the document set. For the classification experiments in this study, different kinds of terms like noun, preposition, verb, conjunction, incorrect format, infinitive, abbreviation, proper noun, number, adjective, question, pronoun, reflection, exclamation, time and words not found significant with Zemberek Library (some specific shortenings) can be chosen for the elimination. After all, the document list that composed of selected terms is handled to find out the root forms of words by a comprehensive Turkish stemming library Zemberek [34] in order to reduce the number of terms needed to represent the document collection. After the pre-processing step, term-document matrix contains in its cells the importance of terms in the document set have been constructed.

Differently from the current studies on the subject, in this study a user dependent term selection and weighting method is used. The proposed study allows the user to eliminate the words by the term selection among different function words like prepositions, pronouns, conjunctions and also among different content words like nouns, adjectives, verbs, shortenings or proper names. Every different task of text categorization may require different kinds of function and content term analysis. For instance, while classifying the documents that include mathematical information the numbers are so important; however the numbers do not have any importance for sentiment classification. In category determination, the proper names may be so important for defining the magazine category; however in sentiment analysis they are not so important, the adjectives are more important. In author detection, frequently used articles, prepositions may be helpful. In brief the right terms should be chosen for the right task requires a perception of the nature of text categorization task [2]. In this regard, the proposed term selection part is an effective factor for achieving high classification accuracy rates. The main contribution of the study is, with developed software tool the users can chose which kinds of terms must be evaluated and which kind of terms are redundant for text classification process. In other words the chosen terms are not weighted and also evaluated. Thus the term-document matrix space is intelligently decreased with respect to text categorization task.

The stop words and the terms that can be chosen by the users for constructing term-document matrix are shown in (Figure.1).



Stop Words : psine-hepsini-her-her biri-herkes-herkese-herkesi-hiç-hiç kimse-hiçbiri-hiçbirine-hiçbirini-ı-i-için-içinde-iki-ile-ise-işte-j-k-kaç-kadar-kendi-kendine-kendini-ki-ki
m-kime-kimi-kimin-kimisi-l-m-madem-mı-mi-mi-mu-mu-mü-mü-n-nasıl-ne-ne kadar-ne zaman-neden-nedir-nerde-nerede-nereden-nereye-nesi-neyse-niçin-ni
ye-o-on-ona-ondan-onlar-onlara-onlardan-onların-onların-onu-onun-orada-oysa-oysaki-ö-öbürü-ön-önce-ötürü-öyle-p-r-rağmen-s-sana-sekiz-sen-senden-seni-
senin-siz-sizden-size-sizi-sizin-son-sonra-ş-şayet-şey-şeyden-şeye-şeyi-şeyler-şimdi-şöyle-şu-şuna-şunda-şundan-şunlar-şunu-şunun-t-tabi-tamam-tüm-tümü
-u-ü-üç-üzere-v-var-ve-veya-veyahut-y-ya-ya da-yani-yedi-yerine-yine-yoksa-z-zaten-zira-

☑ Noun □ Preposition ☑ Verb □ Conjunction □ Incorrect Format ☑ Infinitive ☑ Abbreviation ☑ Proper Noun ☑ Number ☑ Adjective □ Question

☑ Words Not Found In Significant with 'Zemberek Library'(Some specific names) □ Pronoun ☑ Reflection ☑ Exclamation ☑ Time

Figure 1. Stop words and redundant words for the text analysis application task.

3.2. Term Weighting

In text mining, the term-document matrix is mostly weighted by TF-IDF method. In conventional TF-IDF method, how many times each term appears in document set is calculated. The major difficulty here is the high computational time caused by high dimensionality of the terms. In text mining, supervised linear feature extraction methods may be used to reduce the feature dimensionality [35]. When the relevant literature is analyzed it is seen that, the high computational time problem is usually solved by linear discriminant analysis (LDA) or any of the supervised linear feature extraction methods, in this study, without any need to reducing the term-document matrix dimension, the term frequency determination process is accelerated. This is achieved by combining the proposed user dependent term selection method with hashing method.

Since the number of unique terms in document set is relatively large, a hash table is built and used for efficient searching. The hash table consists of <key, value> pairs that are the unique terms appear in the document set and their appearing frequencies respectively. The keys represent the domain dependent and user selected unique terms; the values represent the number of documents that contain these keys. The TF-IDF values are easily computed by configuring the hash table term-based. When a term is appeared in the first document, it has been added to hash table as a key and also the frequency of it as a value. After that, while the term frequencies are been calculated for the second document, if the same term appears in this document again, that means, this term was already added to hash table and it can be easily reached by the key value. Thus, instead of searching the term in a list structure, it has been reached directly by the generated hash code. The list structure typically indexed with integer numbers, while hash table indexed with a word.

Hash structure can be very efficient for processing large scaled data, because the time to locate a value on a hash table is absolutely independent of its size. The length of the frequency list for each term is the index of the document this term is last occurred in. By this means, it saves us to make unnecessary computation loops on document set such as for a term which is only occurred in the first document. As an example the first document content is like "… the clustering application is …" and the second document is like "... the next word in the next application …" the number of documents is denoted n. Then our hash table structure for this example is as in (Table.1).

Table 1. Hash table structure

Key	Document 1 (Value list)	Document 2 (Value list)	...	Document n (Value list)
clustering	1	0	...	...
application	1	1	...	...
next	0	2	...	...
word	0	1	...	...

4. Classification Results

4.1. Text Collection and Pre-processing

In this study, two Turkish text datasets that taken from a natural language research group of Yıldız Technical University [36] in Turkey are used in order to examine the performance of the proposed document categorization system. The first sample data set contains 10 different columnists for each of 9 different authors. The second data set contains 150 different documents for each of 5 different news groups that have different subject in each such as economy, magazine, medical, politics and sports. In machine learning techniques, the ratio between the training data and the test data is recommended as 75% and 25% of all data respectively [37]. Accordingly, for the first dataset, 63 documents have been used for training and 27 texts for testing. The second dataset is split into 560 texts for training and 190 texts for testing. Thus different parts of the whole data are treated as training and test examples for SVM learning. Once the training phase is completed, the SVM model can be able to classify some unknown text data.

To evaluate the proposed document categorization system, four kinds of classical evaluation measures constantly used in document categorization, precision, recall, F-measure and accuracy are adopted for the experiments. Precision is a measure of the ability of a classification model to present only relevant items. Recall is a measure of the ability of a classification model to present all relevant items. F-measure is the weighted harmonic mean of precision and recall.

4.2. Experimental Results and Analysis

In this study, Java programming language is used to develop a document categorization application. We run the experiments on an Intel Core 2 duo (2.27 GHz) PC with 2GB Ram. First, we examine the proposed document categorization model on author categorization dataset. For author categorization process, the total meaningful words in training text are ranked as 3743 and the training time of the classifier is measured approximately 0.12 seconds. After the training phase, the test examples are uploaded to software and they classified by the SVM model.

For each author in this dataset, the number of training, test and misclassified examples are also shown in (Table.2).

Table 2. Classification performance due to the selected terms

Author	Training	Test	Misclassified
Doğan Hızlan	7	3	0
Erkan Çelebi	7	3	0
Ercan Mumcu	7	3	0
Ertuğrul Özkök	7	3	1
Ertuğrul Sağlam	7	3	0
Fatih Altaylı	7	3	0
Gündüz Tezmen	7	3	0
Pakize Suda	7	3	1
Serdar Turgut	7	3	0

In text mining, the importance of the terms chances due to document categorization task [2]. To demonstrate this point, the SVM classifier runs for two different user selected term sets for weighting. The first set consists of noun, abbreviation, proper noun, adjective, reflection, exclamation and the second set consists of noun, verb, infinitive, abbreviation, proper noun, adjective, reflection, exclamation and time. According to these two sets the classification results are shown in (Table.3).

Table 3. Classification performance due to the selected terms

Selected Terms	Precision %	Recall%	F-measure %	Accuracy%
1st set	91	91	91	98
2nd set	89	92	90	97

In the first case the SVM classifier achieves 98% accuracy by weighting 3743 words, but in the second case, the SVM classifier achieves 97% accuracy by weighting 4358 words. Despite, the number of meaningful terms used for constructing SVM classifier is decreased, the accuracy of classifier is increased in consequence of selecting the right dimensions for the right task. Briefly, the proposed text representation model increases the accuracy of SVM classifier; in addition that it decreases dimension of the term-document matrix and consequently the required classification time.

Secondly, the number of training, test and misclassified examples for each news document in the news-group dataset are shown in (Table.4). Three of the documents are misclassified among 190 test examples. Thus the classification performance is calculated as follows; the precision of the SVM classifier for this dataset is 98.6%, recall is 98%, F-measure is 98% and the accuracy is 99%.

Table 4. Classification results for newsgroup dataset

Category	Training	Test	Misclassified
Ekonomi	112	38	1
Magazin	112	38	2
Sağlık	112	38	0
Siyaset	112	38	1
Spor	112	38	0

There is another critical point of the results of this study. When the test results of the SVM classifier were evaluated, it was observed that the classification error rates are considerable small; what is more, when the misclassified documents were evaluated, it was observed that these documents do not contain sufficient distinguishing words to represent their categories.

Considering the newsgroup dataset will be descriptive for understanding the reasons of misclassifications. This data set contains several documents in five different news groups, such as economy, magazine, medical, politics and sports. When the results in Table 5 are examined, it is seen that one document in economy category and two documents in magazine category are misclassified.

First we evaluate the fifth misclassified document in economy category. It is classified as in medical category (3) by SVM classifier even though it is in economy category (1). Title of this news document is “The ministry of health’s objection towards the cord blood trade” and the document expresses the legal restrictions which have been applied due to limit the cord blood trade. It contains a lot of medical terms and the SVM classifier labeled it as in medical category. At first blush the document seemed misclassified; however, indeed the decision of the classifier is not so wrong.

Other misclassified documents are in magazine category. One of them is labeled as in sports category by SVM classifier even though it is in magazine category. When the content of the news document was evaluated, it was observed that it is about a Turkish basketball player who had played in AEK and Panatinaikos basketball teams. The document refers his attendance to the Olympiads in Athens and his fifteen days basketball camp for the national match. As a result of these reviews, the decision of the classifier is not so wrong. The other misclassified document in magazine category is about the hairs and it is labeled as in medical category.

5. Conclusion

There are two major factors that make the text classification process difficult. The first one is the problem of defining the document feature vector that better distinguish the category to which each document belongs. The second one is the problem of deciding the best learning model as document classifier. In this paper, a new approach for the first issue in Turkish based texts is directly addressed. Then a SVM classifier with a linear kernel function is implemented in order to observe the accuracy of the classification.

In text classification applications, the text representation process causes huge computation time on weighting the huge size of terms. Lexicons that contain less number of terms are used as a usual solution for this problem. In this study, distinctively from the literature, a user dependent term-document matrix is determined for text representation. The terms like noun, adjective, infinitive, verb, abbreviation, proper noun, number, reflection, exclamation, time and words not found significant with Zemberek Library are considered as the terms that are going to be evaluated. Due to the characteristics of the classification model, user can chose any of these terms to construct the representatives of the documents. While the characteristics and the purpose of the classification model changes, the important kind of terms will also be changed. For example, on the one hand the nouns may be more important for news classification model, on the other hand the verbs and adjectives may be more important for sentiment classification. In brief, user can determine the terms that will be used in term-document matrix construction, according to the purpose and characteristics of classification model.

Experiments conducted over author and newsgroup datasets and as a result of these 98% and 99% accuracy are achieved respectively. Increasing number of meaningful terms, which are used for constructing SVM, has caused a decrease in accuracy in determination of columnist. With 4358 meaningful term SVM classifier achieves 97% accuracy for author dataset. Our feature extraction strategy based on hash table consistently improves text classification in the terms of classification accuracy and computational time. As in other studies it has been clearly shown by this study, SVM achieves superior classification accuracy in classification problems.

The future work should be done on the issues of trying different document classification problems and determining the relation between the classification purpose and corresponding important terms. A self learning system should be developed for determining which kind of terms is useful for which kind of classification task.

References

[1] M. A. Kumar, and M. Gopal, “A comparison study on multiple binary-class SVM methods for unilabel text categorization,” Pattern Recognition Letters, vol. 31, pp. 1437-1444, Aug. 2010.

[2] F. Sebastiani, Text Categorization, A. Zanasi, Ed. Southampton, UK: WIT Press, 2005.

[3] W. Zhang, T. Yoshida, and X. Tang, “Text classification based on multi-word with support vector machine,” Knowledge-Based Systems, vol. 21, pp. 879-886, Dec. 2008.

[4] W. Li, D. Miao, and W. Wang, “Two-level hierarchical combination method for text classification,” Expert Systems with Applications, vol. 38, pp. 2030-2039, Mar. 2011.

[5] A. Sun, E. Lim, and Y. Liu, “On strategies for imbalanced text classification using SVM: A comparative study,” Decision Support Systems, vol. 48, pp. 191-201, Dec. 2009.

[6] D. Miao, Q. Duan, H. Zhang, and N. Jiao, “Rough set based hybrid algorithm for text classification,” Expert Systems with Applications, vol. 36, pp. 9168-9174, July 2009.

[7] L L. Shi, X. Ma, L. Xi, Q. Duan, and J. Zhao, “Rough set and ensemble learning based semi-supervised algorithm for text classification,” Expert Systems with Applications, vol. 38, pp. 6300-6306, May 2011.

[8] V. Mitra, C. Wang, and S. Banerjee, “Text classification: A least square support vector machine approach,” Applied Soft Computing, vol. 7, pp. 908-914, June 2007.

[9] S. Lo, “Web service quality control based on text mining using support vector machine,” Expert Systems with Applications, vol. 34, pp. 603-610, Jan. 2008.

[10] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan, “Automatic classification of Tamil documents using vector space model and artificial neural network,” Expert Systems with Applications, vol. 36, pp. 10914-10918, Oct. 2009.

[11] L. Zhang, Y. Li, C. Sun, and W. Nadee, “Rough Set Based Approach to Text Classification,” in IEEE/WI/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013, p. 245.

[12] J. J. G. Adeva, J. M. P. Atxa, M. U. Carrillo, and E. A. Zengotitabengoa, “Automatic text classification to support systematic reviews in medicine,” Expert Systems with Applications, vol. 41, pp. 1498-1508, Mar. 2014.

[13] L. H. Lee, C. H. Wan, R. Rajkumar, and D. Isa, “An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization,” Applied Intelligence, vol. 37, pp. 80-99, July 2012.

[14] Y. Kılıçaslan, E. S. Güner, and S. Yıldırım, “Learning-based pronoun resolution for Turkish with a comparative evaluation,” Computer Speech and Language, vol. 23, pp. 311-331, July 2009.

[15] A. Çıltık, and T. Güngör, “Time-efficient spam e-mail filtering using n-gram models,” Pattern Recognition Letters, vol. 29, pp. 19-33, Jan. 2008.

[16] Ö. Özyurt, and C. Köse, “Chat mining: Automatically determination of chat conversations,” Expert Systems with Applications, vol. 37, pp. 8705-8710, Dec. 2010.

[17] L. Özgür, T. Güngör, and F. Gürgen, “Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish,” Pattern Recognition Letters, vol. 25, pp. 1819-1831, Dec. 2004.

[18] E. Alparslan, A. Karahoca, and H. Bahşi, “Classification of confidential documents by using adaptive neurofuzzy inference systems,” Procedia Computer Science, vol. 3, pp. 1412-1417, 2011.

[19] A. K. Uysal, and S. Gunal, “The impact of preprocessing on text classification,” Information Processing and Management, vol. 50, pp. 104-112, Jan. 2014.

[20] F. Türkoğlu, B. Diri, and M. F. Amasyalı, Author Attribution of Turkish Texts by Feature Mining, D. –S. Huang, L. Heutte, M. Loog, Ed. Berlin, Germany: Springer-Verlag, 2007.

[21] D. M. Christopher, and S. Hinrich, Foundations of statistical natural language processing, 4th ed., Cambridge, Massachusetts: MIT Press, 2001.

[22] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” Journal of Documentation, vol. 60, pp. 493-502, 2004.

[23] K. S. Jones, “IDF term weighting and IR research lessons,” Journal of Documentation, vol. 60, pp. 521-523, 2004.

[24] J. L. Solka, “Text Data Mining: Theory and Methods,” Statistics Surveys, vol. 2, pp. 94-112, 2008.

[25] J. -S. Xu, and Z. -O. Wang, “Tcblsa: A New Method Of Text Clustering,” in Proc. Second International Conference on Machine Learning and Cybernetics, 2003, p. 63.

[26] W. Zhang, T. Yoshida, and X. Tang, “A comparative study of TF*IDF, LSI and multi-words for text classification,” Expert Systems with Applications, vol. 38, pp. 2758-2565, Mar. 2011.

[27] Y. Yang, and J. O. Pedersen, “Comparative Study on Feature Selection in Text Categorization,” in Proc. ICML-97, 1997, p. 412.

[28] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., M. Jordan, S. L. Lauritzen, J. F. Lawless, V. Nair, Ed. New York, USA: Springer-Verlag, 2000.

[29] T. Joachims, “Text categorization with support vector machines: Learning with many relevant feature,” in Proc. ECML-98, 1998, p. 137.

[30] E. Leopold, and J. Kindermann, “Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?,” Machine Learning, vol. 46, pp. 423-444, 2002.

[31] A. Wang, W. Yuan, J. Liu, Z. Yu, and H. Li, “A novel pattern recognition algorithm: Combining ART network with SVM to reconstruct a multi-class classifier,” Computers & Mathematics with Applications, vol. 57, pp. 1908-1914, June 2009.

[32] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization,” in Proc. CIKM '98, 1998, p. 148.

[33] (2014) Fatih University Computer Engineering Website. [Online]. Available: http://nlp.ceng.fatih.edu.tr/blog/tr/?p=31/

[34] (2014) Zemberek Website. [Online]. Available: https://code.google.com/p/zemberek/

[35] M. Radovanovic, and M. Ivanovic, "Text Mining: Approaches and Applications," Novi Sad J. Math., vol. 38, pp. 227-234, 2008.

[36] (2014) Kemik Website. [Online]. Available: http://www.kemik.yildiz.edu.tr/?id=28/

[37] E. Alpaydın, Introduction to Machine Learning, 2nd ed., London, England: MIT Press, 2010.

International Journal of
Intelligent Systems and
Applications in Engineering

ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

Epileptic State Detection: Pre-ictal, Inter-ictal, Ictal

Apdullah Yayik[1*], Esen Yildirim[2,] Yakup Kutlu[2], Serdar Yildirim[2]

Received 05th September 2014, Accepted 18th December 2014

Abstract: Epileptic seizure detection and prediction from electroencephalography (EEG) is a vital area of research. In this study, Second-Order Difference Plot (SODP) is used to extract features based on consecutive difference of time domain values from three states of EEG (pre-ictal, ictal and inter-ictal), and Multi-Layer Neural Network classifier is used to classify these three classes. The proposed technique is tested on a publicly available EEG database and classified with Naive Bayes and *k*-nearest neighbor classifiers. As a result, it is shown that overall accuracy of 98.70% can be achieved by using the proposed system with Neural Network classifier.

Keywords: *Epileptic State Detection, Second-Order Difference Plot, Neural Network*

1. Introduction

Epilepsy is a chronic disease comprised by repetitive seizures. Approximately 1% of people in the world suffer from epilepsy, and 85% of them live in growing countries [1]. Seizures are resulted from sudden excessive electrical discharge in a group of brain cells. Epilepsy is explained by recurring instant seizures due to the instantaneous development of synchronous firing in the cerebral cortex caused by lasting cerebral abnormality [2]. The electroencephalogram (EEG) signals act vital role in detection of epilepsy and detection and prediction of epileptic seizures [3,4].

Detection of epilepsy is important for diagnosis of epilepsy. Besides, for an epileptic patient, recognizing the period when a seizure is occurring is necessary for the caregiver to prevent serious injuries due to the seizures. Various approaches have been applied in this field in the last decade. Vukkadala and Vijayapriya [1] discussed an automated Neural epilepsy detection system on two classes (awake healthy and pathologic) with features extracted from EEG using Approximate Entropy (ApEn). They have reached 93.3% overall accuracy. In 2011, Shen et al. [5] presented comparison of different kernels (RBF, Linear, Sigmoid and Grid) of SVM classifier on three classes (normal, inter-ictal and ictal) with ApEn features extracted from multichannel EEG signals. Grid SVM kernel resulted in the highest overall classification accuracy of 98.9%. Zainuddin et al. [6] proposed a seizure detection system using statistical features obtained from the discrete wavelet transform and an improved wavelet neural network (WNN). The performance of the classifier is reported as %98.87. Vollala and Gulla [7] presented comparison of Elman and Probabilistic Neural Network classifiers on two classes (epileptic and normal patients) using ApEn features extracted from EEG signals and reached 93.43% overall accuracy with Elman Neural Network classifier. Mercy [8] classified two classes (normal and epileptic) with both SVM and Neural Networks using DWT and Fast Independent Component Analysis and obtained an accuracy of 99.5%. Bayram [9] achieved an 98% overall accuracy in seizure detection with Wavelet Entropy features classified by using SVM.

Epilepsy is a disease which affects the patient only during the seizure and about 70% of the patients can control the seizures with medication. There are numerous studies which show that EEG recordings carry important information prior to the seizure onset [3,4]. A detailed review can be found in [4].

The goal of this study is discrimination of three states of an epileptic patient: pre-ictal, ictal and inter-ictal. For this purpose, EEG database obtained from Children's Hospital Boston (CHB) [10] is used.

Different linear and non-linear classifiers are employed for classification; Naive Bayes, *k*-nearest neighbor and Neural Network. Region parameters of Second-order difference plot are used as features to distribute classes to different data spaces without losing the pattern properties. The results are evaluated with 10-fold cross validation. This paper is organized as follows, in section 2, data used in this study is presented, and then the feature extraction based on second order difference plot is explained. The implementation of Neural Network, Naive Bayes and *k*-nearest neighbor classifiers are next described in section 3. In section 4, effectiveness of the proposed classification and performance analyses are presented, and finally conclusions are discussed in section 5.

2. Proposed Methodology

2.1. Data Description

The EEG Database used in this study is CHB-MIT Scalp EEG Database [10]. The database is collected from 5 males aged between 3 and 22, and 17 females between ages 1.5 and 19.

2.2. Data Processing

Database is divided into three sets: set A, B and C. Each of these sets consists of 20 s long 256 Hz sampled EEG segments from 18 channels. Sets A, B and C are recorded before seizure, during seizure and between seizures respectively (sample recordings are shown in Figure 1).

[1] *Turkish Army Forces, Turkey*
[2] *Department of Computer Engineering Mustafa Kemal University, Turkey*
* *Corresponding Author: Email: ayayik@kkk.tsk.tr; ayayik@mku.edu.tr*

Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

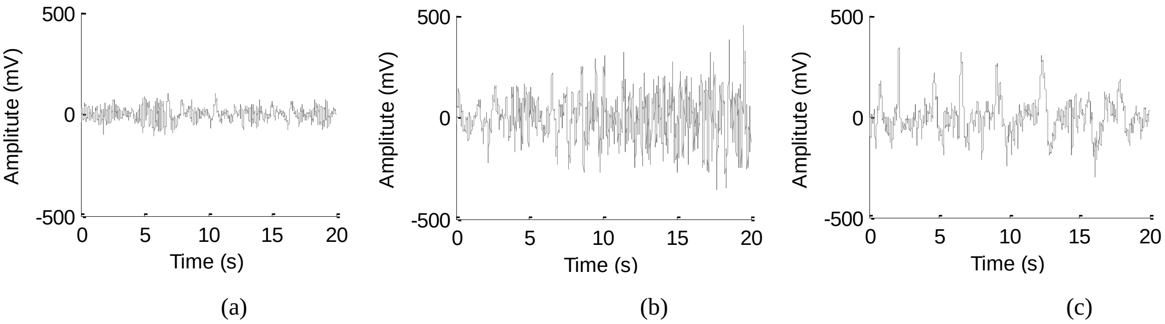


Figure 1. 20 s FP1-F7channel EEG recordings belong to 11 years old female; (a) before seizure (pre-ictal), (b) during (ictal), (c) between seizures (Inter-ictal).

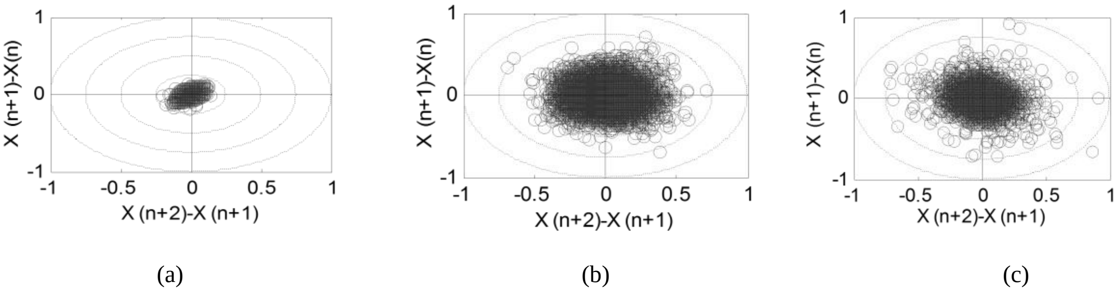


Figure 2. Second Order Different Plot of Multi-Channel EEG (a) before seizure (pre-ictal), (b) during (ictal), (c) between seizures (Inter-ictal).

2.3 Feature Extraction

Second Order Difference Plot (SODP) is a feature extraction method which is formed employing time domain information. The method of SODP can be used as an independent feature extraction tool as well as a supplemental technique to confirm the frequency domain results [11]. If X(t) is the EEG signal, SODP is formed by $X(n+1) - X(n)$ and $X(n+2) - X(n+1)$ points on the plot (Figure 2). In other words, SODP includes scattering of consecutive difference values of points in EEG signal. Thus, the statistical condition of consecutive differences can be observed. Figure 2 shows a sample of the Second-order difference plot of EEG signals before, during and between seizures. The features are extracted from second-order difference plot of the sets A, B and C Region parameters. The SODP is a figure of two-dimensional Cartesian system. The axes of a two-dimensional Cartesian system divide the quadrants, which are four infinite regions numbered from the first to the fourth each bounded by two half-axes. The region numbers and signs of two coordinates are I (+ , +), II (− , +), III (− , −) and IV (+ , −). SODPs are generally divided into different radius of circle regions in order to extract features [12]. Regions of a quadrant in a SODP are shown in Figure 3.

There are four quadrants of a Cartesian coordinate system. Each quadrant has four regions. Therefore, there are sixteen different regions. Each region shows the number of points in SODP. The numbers of points are calculated for each region and used as a feature vector.

3. Classification

3.1 Multi-Layer Neural Network Classifier

ANNs are inspired by biological neural networks. They are generated neuron-like units which are connected together with adjustable weights [13]. Each unit generates an output signal. Among different structures utilized in ANNs, the mostly used one is the multi-layer perceptron (MLP). MLP consists of successive layers each having different number of processing units. The layers are input layer, hidden layer and output layer. The units in each layer are fully connected to units in the next layer. The output of the MLP is the set of units in the output layer. In order to generate a correct output for a given input, the values of weights should be adjusted. The convenient weights are determined under the control of a training algorithm. A variety of training algorithms can be utilized in the network [14]. The main goal of training a network is not to force it to learn the training set perfectly but to generate correct outputs for inputs that are not seen during the training process. In this study; neural network that has 3 layers (input layer 16 that has neurons, one hidden layer that has 9 neurons and output layer that has 3 layers) as shown in Figure 4, sigmoid transfer functions and backpropagation algorithm is used.

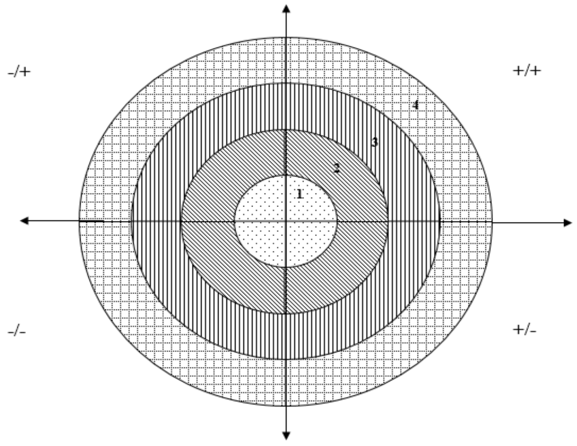


Figure 3. Regions of SODP of EEG

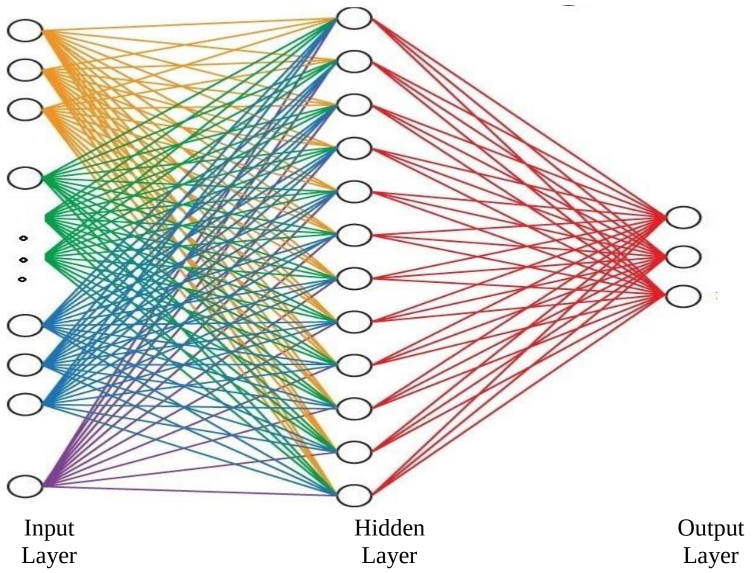


Figure 4. Proposed Multi-Layer NN Classifiers' Topology

3.2 Naive Bayes Classifier

The Naive Bayes classifier is a Bayesian network where the class has no parents and each feature has the class as its only parent. Naive Bayes models have been widely used for clustering and classification in machine learning. It is the simplest form of Bayesian theorem. The Naive Bayes algorithm is based on conditional probabilities. The conditional independence is an assumption in Bayesian theorem [15].

3.3 *k*-Nearest Neighbor (*k*-NN) Classifier

In pattern classification, the *k*-nearest neighbor algorithm (*k*-NN) is a non-parametric technique for classifying classes according to nearest training examples in all extracted features. It is a type of sample-based learning. In the machine learning algorithms, the *k*-NN algorithm is the simplest one; an object is classified by a plurality vote of its neighbors, with the features being assigned to the class most common amongst its *k* nearest neighbors [13], [14]. In this study *k* value is preferred as 3.

4. Performance Analysis

4.1 *k*-Fold Cross Validation

Cross-validation is also known as rotation estimation. It is a method to determine how the results of a statistical analysis will generalize to a new data set. In this method, the whole data set is randomly separated into k equal size subsets. One subset is used for testing and all other subsets are used in training. This step is repeated for k times leaving one fold for evaluation each time (Figure 5). This validation method is performed for better approximation error [14].

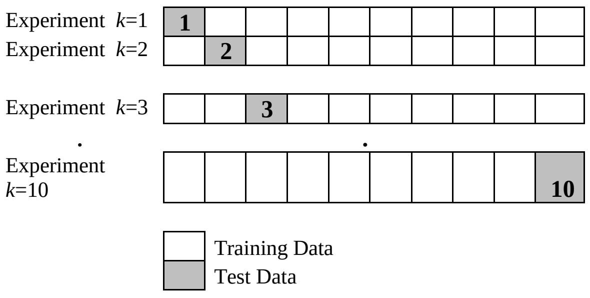


Figure 5. *k*-fold Cross Validation

4.2 Performance Measure

For classification tasks, the terms true positives, true negatives, false positives, and false negatives are used to compare the results of the classifiers. The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to whether that prediction corresponds to the external judgment (sometimes known as the observation). Accuracy is the overall correctness of the model and is calculated as the total number of instances that are correctly classified over the total number of instances. Precision is calculated as the correctly classified positives over total instances predicted as positive. Recall is also known as sensitivity and is calculated as the proportion of true positives (correctly predicted as positive) to the total number of instances belongs to that class. Overall Accuracy, Recall, and Precision are formulated in Table 1.

Table 1. Performance Measurements of Classifiers

$$\textbf{Accuracy} = \frac{TP_b+TP_e+TP_a}{TP_b+TP_e+TP_a+\mathrm{E_{be}}+\mathrm{E_{ba}}+\mathrm{E_{db}}+\mathrm{E_{da}}+\mathrm{E_{ab}}+\mathrm{E_{ae}}}$$

$$\textbf{Recall}\ {}_b= \frac{TP_b}{TP_b+\mathrm{E_{be}}+\mathrm{E_{ba}}} \qquad \textbf{Precision}\ {}_b= \frac{TP_b}{TP_b+E_{db}+E_{ab}}$$

$$\textbf{Recall}\ {}_d= \frac{TP_d}{TP_d+\mathrm{E_{db}}+\mathrm{E_{da}}} \qquad \textbf{Precision}\ {}_d= \frac{TP_d}{TP_d+\mathrm{E_{be}}+\mathrm{E_{ae}}}$$

$$\textbf{Recall}\ {}_a= \frac{TP_a}{TP_a+E_{ab}+E_{ae}} \qquad \textbf{Precision}\ {}_a= \frac{TP_a}{TP_a+E_{ba}+E_{da}}$$

ROC curve is a 2 dimensional graphical plot which demonstrates the performance of a binary classifier by plotting the true positive rate against the false positive rate [15]. Performances of two classifiers are compared by the areas under the ROC curves (AUC). An AUC value of 1 represents a perfect test result where as an AUC value lower than 0.5 is accepted to be worse than random prediction. The Kappa Statistic measures the agreement of prediction with the true class. A value of 1 implies complete agreement [16]. The Mean Absolute Error (MAE) measures the mean magnitude of the errors in a set. The other performance measure is The Root Mean Absolute Error (RMAE) which is a quadratic scoring rule which measures the average magnitude of the error [17].

5. Experimental Results and Conclusions

A method for analyzing multi-channel EEG for detecting pre-ictal, ictal and inter-ictal states using Second Order Difference Plot features are presented here. Recordings that have not common channel information are eliminated. Several classifiers' performances are compared by Precision, Recall, area under the ROC curve, Kappa Statistic, MAE, RMAE and Overall Accuracy metrics. The results are shown in Table 2. One can see that performance parameters of Neural Network classifier are higher than others. The prediction accuracy of EEG signals are 98.70%, 94.71% 95.14% using Neural Network, k-NN and Naive Bayes classifiers, respectively. Table 2 shows details of performance measures of classifiers. In order to validate classifiers performances 10 different test and training datasets, derived by 10-fold cross validation are used and mean values of performance parameters are calculated.

These results are difficult to compare with previous studies. Because previous studies are focused on binary classifications; ictal/normal (non-epileptic) EEG [6–9,19] normal/inter-ictal/ictal [5] or pre-ictal/inter-ictal [20] on different datasets. This study is focused on discriminating three states of epileptic patients; before seizure (pre-ictal state), during seizure (ictal state) or non-seizure. Details of previous studies are described in Table 3. Pachoris and Patidar [21] classified normal and seizure EEG using empirical mode decomposition (EMD) and second-order difference plot (SODP) features with ANN model on Andrzejak Dataset. Their study combines EMD and SODP methods for feature extraction. Our study aims to determine; if a seizure is expected in a near future, if the patient is having a seizure in that moment or the patient is neither in a seizure state nor a seizure is expected soon. For that purpose SODP features of raw signals are used for building a model. This study shows that the state of an epileptic patent can be classified as before, during and between seizures (pre-ictal, ictal and inter-ictal) using SODP features and machine learning algorithms.

The proposed methodology can be very helpful for medical practice. Presented system's accuracy might be improved by applying various feature extraction methods and feature selection algorithms to find the best features that characterize the seizure state.

Acknowledgments

The authors would like to thank Children's Hospital Boston for kindly sharing database freely on internet.

Table 2. Evaluated Performances for All Classifiers

Classifier	*Class*	*Precision*	*Recall*	*ROC Area*	*Kappa Statistic*	*Mean Absolute Error*	*Root Mean Squared Error*	*Overall Accuracy (%)*
Multi-Layer Neural Network	*pre-seizure*	99.00%	99.00%	0.99	0.98	0.01	0.07	**98.70**
	Ictal	97.00%	97.60%	0.99				
	Inter-Ictal	**100.00%**	100.00%	0.99				
***k*-Nearest Neighbor**	*pre-seizure*	94.10%	94.10%	0.95	0.92	0.03	0.18	**94.71**
	Ictal	92.10%	92.10%	0.93				
	Inter-Ictal	98.00%	98.00%	0.99				
Naive Bayes	*pre-seizure*	98.10%	100.00%	0.91	0.95	0.02	0.11	**95.14**
	Ictal	97.30%	94.10%	0.83				
	Inter-Ictal	96.10%	97.00%	0.96				

Table 3. Comparisons with Previous Studies

Study	*Year*	*Database*	*Features*	*Classification Problem (Epileptic States)*	*Classifier*	*Classification Accuracy (%)*
[20]	2009	Freiburg dataset [19]	Wavelet Transform	pre-ictal and inter-ictal	Convolutional NN	71,00
[1]	2009	Individual	Approximate Entropy	normal and ictal	Elman NN	93,33
[5]	2011	National Taiwan University Hospital	Approximate Entropy	normal and inter-ictal and ictal	SVM	98.1
[6]	2012	Andrzejak Dataset [22]	Discrete Wavelet Transform	normal and ictal	ANN	98,87
[7]	2012	Andrzejak Dataset [22]	Approximate Entropy	normal and ictal	ANN	93,43
[8]	2012	Andrzejak Dataset [22]	Fast Independent Component Analysis	normal and ictal	ANN	99,50
[9]	2013	Andrzejak Dataset [22]	Discrete Wavelet Transform	normal and ictal	SVM	98,00
[21]	2014	Andrzejak Dataset [22]	Empirical Mode Decomposition Second-Order Difference	normal and ictal	ANN	95,00
This Study	2014	CHB Dataset [10]	Second Order Difference	pre-ictal, ictal and inter-ictal	ANN	98,70

6. References

[1] V.. Vukkadala, Srinath, Vijayapriya.S (2009). Automated Detection Of Epileptic EEG Using Approximate Entropy In Elman Networks, Int. J. Recent Trends Eng. 1 307–312.

[2] M. Ghanbari, M. Askaripour, N. Behboodiyan (2012). Detection of Epilepsy from EEG Signal during Seizure Using Heuristic Algorithm of Fixed Point Iterations, Res. J. Appl. Sci. Eng. Technol. 4 3584–3587.

[3] F. Mormann, R.G. Andrzejak, C.E. Elger, K. Lehnertz (2007). Seizure prediction: the long and winding road., Brain. 130 314–33. doi:10.1093/brain/awl241.

[4] S. Sanei, J.A. Chambers 2007. EEG Signal Processing, Willey, England.

[5] C.-P. Shen, C.-M. Chan, F.-S. Lin, M.-J. Chiu, J.-W. Lin, J.-H. Kao, et al. (2011) . Epileptic Seizure Detection for Multichannel EEG Signals with Support Vector Machines, 2011 IEEE 11th Int. Conf. Bioinforma. Bioeng. 39–43. doi:10.1109/BIBE.2011.13.

[6] Z. Zainuddin, L.K. Huong, O. Pauline (2012). Reliable Epileptic Seizure Detection Using an Improved Wavelet Neural Network, Australas Med. J. 33–44.

[7] S. Vollala, K. Gulla (2012). Automatic Detection of Epilepsy EEG Using Neural Networks, Int. J. Internet Comput. 506009 68–72.

[8] M.S. Mercy (2012). Performance Analysis of Epileptic Seizure Detection Using DWT & ICA with Neural Networks, Int. J. Comput. Eng. Res. 2 1109–1113.

[9] M. Bayram (2013). EEG sınıflandırma amaçlı bir kompozit sistem, Dicle Univ. J. Eng. Cilt 4, Sayı 1,5-2. 30 5–12.

[10] A.L. Goldberger and coworkers (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, Circ. 101(23)e215-e220. http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[11] D.L.H. and P.C.D. Maurice E.Cohen (1996). Applying Continuous chaotic Modeling to Cardiac Signal Analysis, Eng. Med. Biol. 97–102.

[12] C. Kamath (2012). A new approach to detect congestive heart failure using Teager energy nonlinear scatter plot of R-R interval series., Med. Eng. Phys. 34 841–8. doi:10.1016/j.medengphy.2011.09.026.

[13] C. Bishop (1996)., Neural networks for pattern recognition., 1st ed. NY, USA: Oxford Univ. Press.

[14] S. Haykin (1996). Neural networks: a comprehensive foundation., 2nd ed. New Jersey: Prentice Hall.

[15] S.D. Duda RO, Hart PE(2000). Pattern classification. 2nd ed. Wiley-Interscience.

[16] R. Kumar, A. Indrayan (2011). Receiver operating characteristic (ROC) curve for medical researchers., Indian Pediatr. 48 277–87. http://www.ncbi.nlm.nih.gov/pubmed/21532099.

[17] M. Fauzi, T. Moh, S. Yau, A.B.N. (2007). Classifier, Comparison of Different Classification Techniques Using WEKA for Breast Cancer, IFMBE Proc. Vol. 15. 15 520–523.

[18] G. Ngai, E.C.-H.; Gelenbe, E.; Humber, Inf. ormation-aware traffic reduction for wireless sensor networks, in: Local Comput. Networks, Zurich, n.d. pp. 451 – 458.

[19] Freiburg EEG dataset, (n.d.). https://epilepsy.uni-freiburg.de/freiburg-seizureprediction-project/eeg-database/ (accessed December 12, 2013).

[20] P. Mirowski, D. Madhavan, Y. Lecun, R. Kuzniecky (2009)., Classification of Patterns of EEG Synchronization for Seizure Prediction, Work. ach. Learn. Signal Process.

[21] R.B. Pachori, S. Patidar (2014). Epileptic seizure classification in EEG signals using second-order difference plot of intrinsic mode functions., Comput. Methods Programs Biomed. 113 494–502. doi:10.1016/j.cmpb.2013.11.014.

[22] E.C. Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P (2001). Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E.

International Journal of
Intelligent Systems and
Applications in Engineering

ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

Rainfall estimation based on NAW approach using MSG-SEVIRI images: An application in north Algeria

Fatiha Mokdad*[1], Boualem Haddad[1]

Received 05th September 2014, Accepted 18th December 2014

Abstract: In this work, we will adapt the NAW (Nagri, Adler and Wetzel), precipitation estimation approach to the north Algeria events using the Meteosat Second Generation (MSG) satellite images. The tests are carried out on seven areas of northern Algeria: Sidi Bel Abbes, Oran Port, Algiers Port, Dar El Beida, Bedjaia, Jijel-Achouat and Annaba, in winter 2006. The NAW approach is applied by thresholding to temperature from 253 K. The validation is performed by comparaison the estimated rainfall to in situ measures collected by the National Office of Meteorology in Dar El Beida (Algeria). We use the infrared data (10.8µm channel) of SEVIRI sensor in this study. The results obtained indicate that the NAW approach gives satisfactory results for the rain rates: 4mm/h assigned to the coldest 10%, 2mm/h assigned to the next 40% and 0mm/h given to the remaining 50% of the area defined as cloud. The rain rate 8mm/h assigned to the coldest 10% of the pixels in the cloud applied for the convective clouds observed for tropical regions are not valid for the Algerian climate, especially for the stratiform clouds type.

Keywords: *Precipitation, NAW approach, Stratiform cloud, Convective cloud, Meteosat Second Generation.*

1. Introduction

Precipitation is one of the most relevant meteorological and hydrological quantities. Because of floods caused by torrential rainfall accompanying with extreme weathers that can be the origin of oversize economic and human loss, the quantitative evaluation, forecast and the precision in measurement of precipitation have weighty social and economic key issue in rainfall-rich countries. However, it is not easy to estimate precipitation with height accuracy because its process is not linearly related to the cloud microphysical, thermodynamic, dynamic, and radiative processes [1].

The application of conventional measures of precipitation, which is the use of rain gauges still, is insufficient because the rainfall measurements covering land are very patchy while these over oceans are extremely few. Remote sensing, then, proves the ideal solution for rainfall estimation and monitoring. Radar meteorology has become a discipline in its own right. However, in Algeria, one of the seven ground radar sites is operational. Satellites come to correct these deficits. Thereby, making it possible to monitor the globe with more confidence using various wavelengths measured by various imagers aboard diverse types of satellite such as: GOES, METOP, NOAA, MSG…etc.

Several precipitation estimation techniques based on satellite images are proposed in the literature, where can roughly be split into two main categories: Infrared techniques and microwave techniques.

Microwave techniques sound inside the clouds and rain by using Low Earth Orbiting Satellite images [2]. Different algorithms based on such relationships are proposed [3]-[7]. However, the weak spatial and temporal resolution of the microwave data makes difficult the short-term rainfall estimates.

Contrarily to microwave observations visible and/or infrared data collected from Geosynchronous Earth Orbiting Satellite benefit from the spatio-temporal resolutions and the multi-spectral observations. Infrared techniques are based on the fact that precipitations are likely produced by the convection that is related to cold/bright clouds [2]. These approaches are widely used: [8]-[13]. However, IR techniques measure the cloud-top IR-brightness temperature and do not have a direct physical relationship with precipitations.

To take an advantage from the two approaches, combined methods were developed [14]-[19]. However, these techniques are more complex.

Because, there is no unique model with good performance for all areas in the world [20], also, rainfall is often extremely variable over time and space in the Mediterranean region [21]-[28]. So precipitation estimation is a global challenge of researchers, especially with using IR techniques that are the most adapted to this climate.

In this work, we propose to adapt the NAW (Nagri, Adler and Wetzel) approach [10] to the Algerian climate using Meteosat Second Generation (MSG) images [29]. Thereby is an infrared threshold technique. Recall that, it is originally used for the daily estimation of convective rainfall applied to GOES infrared satellite images and tested in Florida region [10]. However, satellite images delivered from the Spinning Enhanced Visible and Infrared Instrument (SEVIRI) on board MSG allow better characterizing clouds by the means of multi-spectral and multi-temporal images with high spatial resolution [29].

This paper is structured as following: a brief description of the satellites data and the rain gauges sites used in this work are provided in the next section. NAW approach is discussed next. Section IV presents the applications of NAW algorithm to the north Algeria. Section V is dedicated to the analysis and interpretation of NAW technique results involved in this exercise. Conclusions and further works are summarized in the final section.

[1] *Laboratory of Image Processing and Radiation, University of Sciences and Technology Houari Boumediene (U.S.T.H.B.) B.P. 32, El Alia, Bab Ezzouar, Algiers, Algeria.*
** Corresponding Author: Email: f_mokdad@yahoo.fr*

Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

2. Study Area and Dataset

2.1. SEVIRI dataset

We constituted a database consisting of 1920 images collected by the SEVIRI sensor in 10.8µm Infrared channel (C9 channel) during winter 2006. For each 12 spectral channel and every 15 minutes, the High Rate SEVIRI images are acquired, each pixel being encoded on 10 bits. To reduce the computation time and to present better the study area, we take 1100x1100 pixels from the original size image (3712x3712 pixels). These satellite data are coupled to ground measurement covering the national territory.

2.2. Study Area

To validate the NAW approach, we use collected and archived data from the Algerian National Office of Meteorology (Dar El Beida). These data are recorded at ground sites that make daily measurements in 76 operational rain gauges covering the territory. The rain gauge sites chosen in this study are (Figure.1): Sidi Bel Abbes (35 ° 18 N, 2° 61 W), Oran Port (35 ° 7 N, 0° 65 W), Algiers Port (36 ° 76 N, 3° 1 E), Dar El Beida (36 ° 71 N, 3 ° 25 E), Bedjaia (36 ° 71 N, 5 ° 06 E), Jijel-Achouat (36 ° 88 N, 5 ° 81 E), and Annaba (36 ° 83 N, 7 ° 81 E).

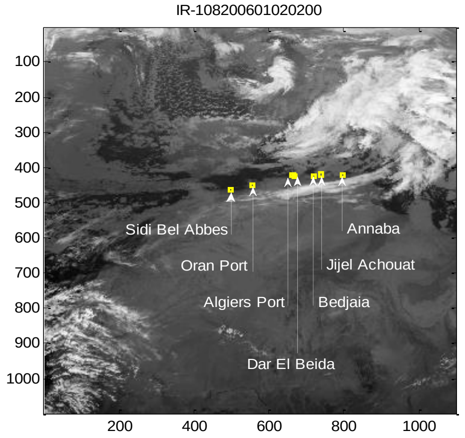


Figure 1. Representation of the seven study areas on the MSG satellite image, IR-10.8 of 02/01/2006 at 0200 h UT

3. NAW Approach

The Nagri, Adler and Wetzel approach (NAW) is an IR precipitation estimation technique based on the threshold temperature. It first delimits clouds of a given image that are colder than 253 K isotherm. Then, for each area which is considered as cloud, it assigns a rain-rate R_1 to the coldest 10%, a lower rain-rate R_2 to the next warmest 40%, and a rain-rate R_3, equal to zero in the first setting of the method, to the remaining 50% of the pixels in the cloud. This means that the threshold attributed for high and low precipitation may not be the same for all types of cloud. The technique was originally calibrated for convective rainfall over Florida, where, the nominal rain-rates assigned are [10]: R_1= 8, R_2= 2 and R_3= 0 mm/h. This distribution of rainfall within a cloud is defined as below:

$$\text{If} \quad T < T_{10}, R_1 = 8mm/h \qquad (1)$$

$$\text{If} \quad T_{10} < T < T_{50}, R_2 = 2mm/h \qquad (2)$$

$$\text{If} \quad T > T_{50}, R_3 = 0mm/h \qquad (3)$$

Where,

R_1, R_2, R_3 are the rain rates assigned to the cloud pixels.

T_{10} and *T_{50}* are the warmest temperatures of the coldest 10% and 50% of the pixels respectively.

Attempts are made to adjust rain rates for mid-latitude [30]-[32]. It is found that the rain rates recalibration allows a better description of non-tropical events. Therefore, according to the authors, NAW approach must, be validated and adapted for the new areas.

4. Results

In this section, we present the intensities of rainfall measured on the ground and estimated by NAW method at seven sites of ONM considered. As an illustration, we are going to limit the study period of 02/01/2006 to 21/01/2006.

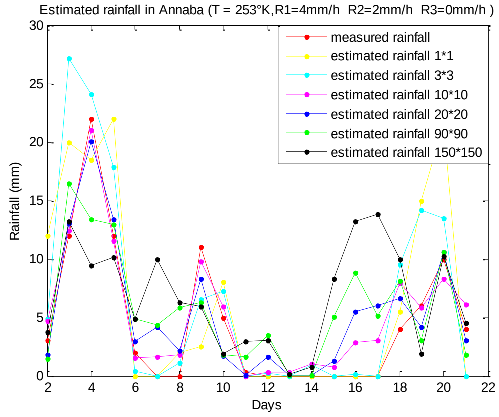


Figure 2. Measured and estimated precipitations accumulation for various windows over Annaba station

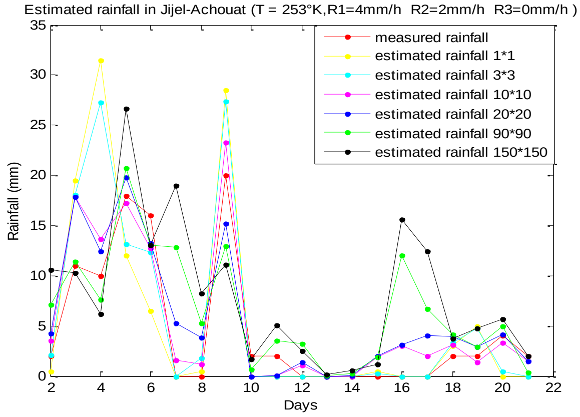


Figure 3. Measured and estimated precipitations accumulation for various windows over Jijel-Achouat station

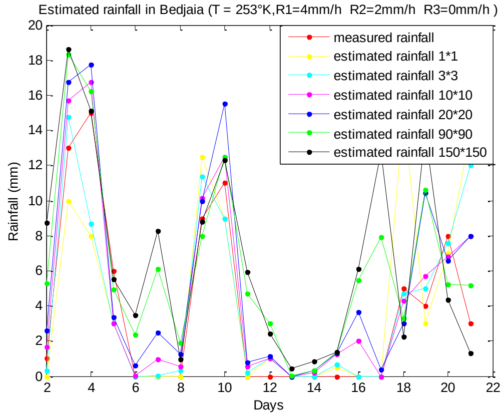


Figure 4. Measured and estimated precipitations accumulation for various windows over Bedjaia station

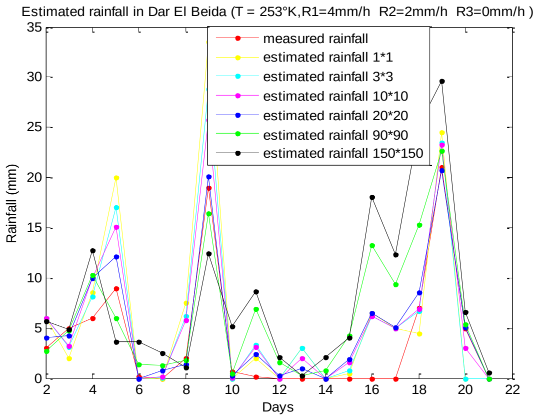


Figure 5. Measured and estimated precipitations accumulation for various windows over Dar El Beida station

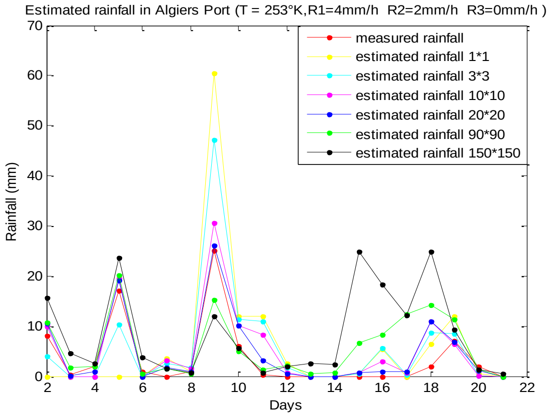


Figure 6. Measured and estimated precipitations accumulation for various windows over Algiers Port station

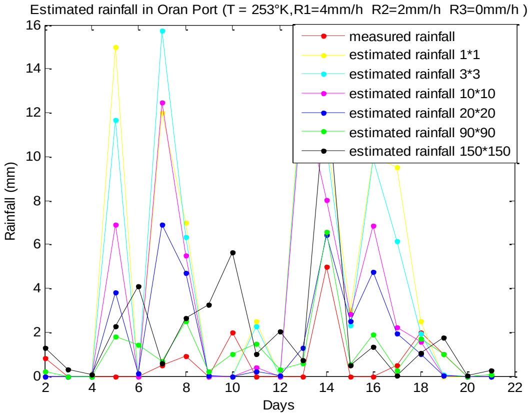


Figure 7. Measured and estimated precipitations accumulation for various windows over Oran Port station

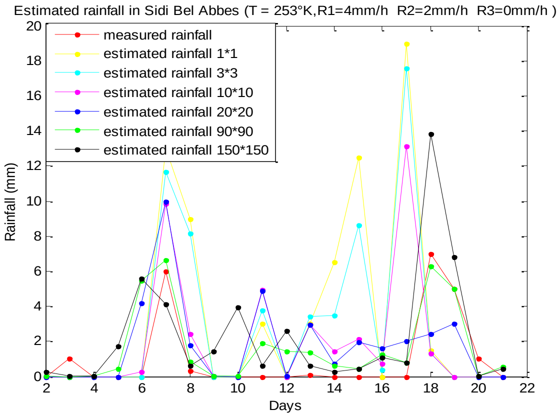


Figure 8. Measured and estimated precipitations accumulation for various windows over Sidi Bel Abbes station

Evaluation of the precipitation estimation approach is carried out by using the Bias (Figure.9), the Root Mean Square Error (RMSE) and the Correlation Coefficient (R) illustrated in (Figure.10) for various windows. The RMSE and the Bias are defined as below:

$$BIAS = \frac{1}{N}\sum_{i=1}^{N}(X'_i - X_i) \qquad (4)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X'_i - X_i)^2} \qquad (5)$$

Where, X_i and X'_i are the ground measured and estimated values for day *i* respectively and *N* is the number of days.

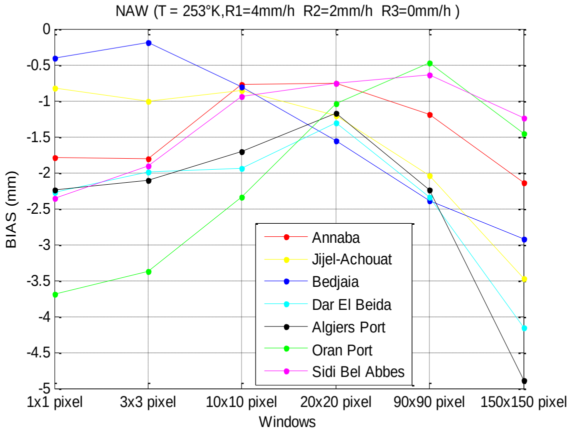


Figure 9. BIAS of different study area, day-by-day comparison, using different windows

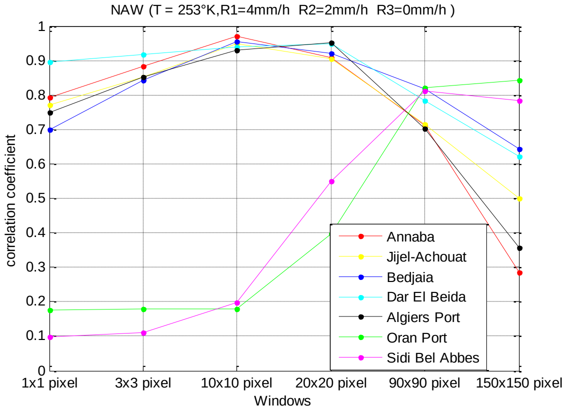


Figure 10. Correlation coefficient of different study area, day-by-day comparison, using different windows

The validation statistics obtained for R, BIAS and RMSE are presented in Table 1. Note that, for each site, only the best window's parameters are presented.

5. Analysis and Interpretation

The Figure.2-Figure.8 present the estimation results and ground measurements for the seven test stations applied to the first two decades of January 2006. They show that the estimate of rain on a daily scale is satisfactory.

The rainfall observed over the north Algeria has some typical distinctive features. Its spatial distribution from west to east has some preferred window's variations of convective rainfall and stratiform rainfall. Standard metrics used for inter comparisons between estimated and ground measurements rainfall: Bias, RMSE and R.

Bias results are generally consistent from -0.4749 through -1.3110. For all sites, it shows negative values. This means that the ground measurements rainfalls are lower than the estimated ones. This is explained by the presence of high clouds (such as the cirrus) that are certainly cold but not precipitants.

Also, it has good statistics results for MRSE and R that are varied from 2.5093 mm to 0.9539 mm and from 0.8106 to 0.971 respectively for the selected windows.

For all tested area, the best estimation is obtained for the rain rates: 4mm/h assigned to the coldest 10%, 2mm/h applied to the next 40% and 0mm/h given to the warmest 50% of the area defined as cloud. However, Nagri, Adler and Wetzel propose 8mm/h, 2mm/h, and 0mm/h assigned to the coldest 10%, the next 40% and the warmest 50% respectively [10], applied to the convective clouds. Therefore, we deduce that the precipitation during this period is also caused by stratiform clouds such as stratocumulus that are relatively warm.

The difference in spatial location between satellite measurements and ground measurements is not negligible. Indeed, the spatial scale of a point of the satellite image at sub-satellite point is 9 km^2 (for all SEVIRI channels except HRV channel), while the reception area of a rain gauge is reduced to just a few square decimetres. This difference can lead to significant errors between the estimated rainfall and in situ measurements.

Similarly, we found that the estimation error can be also related to the analysis window. This error is not due to the method itself, but having assumed that the study area is square, when in reality, it has a definite shape. As the clouds have very high spatial and temporal variability, not taking into account some pixels may lead to a significant error in the estimation of precipitation. To improve the results, we suggest having a mask or a map of the study area to consider only the threshold as a single variable analysis.

We have shown for the NAW approach that the size of analysis windows decrease from the north east to the north west of Algeria .As it was represented in (Figure.2 to Figure.4) corresponding to the east regions (Annaba, Jijel-Achouat and Bedjaia), the more adaptable analysis window is10x10 pixels. Contrarily to the east regions, the most suitable window in west areas: Oran port (Figure.7) and Sidi Bel Abbes (Figure.8) is 90x90 pixels. For the central regions: Dar el Beida (Figure.5) and Algiers port (Figure.6), the low error is measured for the window of 20x20 pixels. These results are logic because the precipitation in Algeria increases from the east to the west and decrease from the north to the south. Also, the convective clouds are more presented in the east contrarily to the stratiform clouds that are more present in the west regions. Similarly, Mesoscale cyclones are more observed in the east areas.

Table 1. The best window's parameters for each station

Regions	RMSE (mm/day)	BIAS (mm/day)	R
Annaba	1.6554	-0.7747	0.9713
Jijel-Achouat	2.4100	-0.8660	0.9472
Bedjaia	1.7535	-0.8107	0.9560
Dar El Beida	2.3473	-1.3110	0.9487
Algiers Port	2.5093	-1.1671	0.9503
Oran Port	0.9539	-0.4749	0.8208
Sidi Bel Abbes	1.4739	-0.6373	0.8106

6. Conclusion

The NAW satellite rainfall estimation method, based on thresholding clouds top temperature is presented and adjusted for the north Algeria. This method is applies to infrared images provided from SEVIRI imager. It has been tested over seven different regions: Three regions in the east of Algeria (Annaba, Jijel-Achouat and Bedjaia), two regions in the west (Sidi Bel Abbes and Oran port) and two regions from the centre (Dar El Beida and Algiers port).

The NAW approach has the merit of being simple and independent of the ground observations. However, its application to the single infrared 10.8µm channel has conceptual limitation. Indeed, the analysis based on single channel of SEVIRI has the inability to exclude the cirrus.

On the other hand, analysis windows show variability from east to west of the northern part of Algeria. The increase in the dataset certainly improves results.

Other applications of the method NAW still need to be conducted to test this method on other climatic regions on south Algeria, where the convective clouds are present in summer.

References

[1] X. Li and S. Gao (2011). Precipitation modeling and quantitative analysis. Springer Dordrecht. Pages. 240.

[2] C. Prigent (2010). Precipitation retrieval from space: An overview. Comptes Rendus Geoscience. Vol. 342. Pages. 380–389.

[3] T.T. Wilheit, A.T.C. Chang, M.S.V. Rao, E.B. Rodgers and J.S. Theon (1977). Satellite technique for quantitatively mapping rainfall rates over oceans. Journal of Applied Meteorology. Vol. 16. Pages. 551–560.

[4] C. Kummerow, Y. Hong, W.S. Olson, S. Yang, R.F. Adler, J. McCollum, R. Ferraro, G. Petty, B.-B. Shin and T.T. Wilheit (2001). The evolution of the Goddard profiling algorithm (GPROF) for rainfall estimation from passive microwave sensors. Journal of Applied Meteorology. Vol. 39. Pages. 1801–1820.

[5] P. Bauer, P. Amayenc, C.D. Kummerow and E.A. Smith (2001). Over-ocean rainfall retrieval from multisensor data of the Tropical Rainfall Measuring Mission. Part I: Design and evaluation of inversion databases. Journal of Atmospheric and Oceanic Technology. Vol. 18. Pages. 1315–1330.

[6] P. Bauer (2001). Over-ocean rainfall retrieval from multisensor data of the Tropical Rainfall Measuring Mission. Part II: Algorithm implementation. Journal of Atmospheric and Oceanic Technology. Vol. 18. Pages. 1838–1855.

[7] R.R. Ferraro, F. Weng, N. Grody, L. Zhao, H. Meng, C. Kongoli, P. Pellegrino, S. Qiu and C. Dean (2005). NOAA operational hydrological products derived from the AMSU. IEEE Transactions on Geoscience and Remote Sensing. Vol. 43. Pages. 1036–1049.

[8] C.G. Griffith, W.L. Woodley, P.G. Grube, D.W. Martin, J. Stout and D.N. Sikdar (1978). Rain estimation from geosynchronous satellite imagery–visible and infrared studies. Monthly Weather Review. Vol. 106. Pages. 1153–1171.

[9] C.G. Griffith (1987). Comparison of gauges and satellite rain estimates for the central United Sates during August 1979. Journal of Geophysical Research. Vol. 92. Pages. 9551–9566.

[10] A.J. Negri, R.F. Adler and P.J. Wetzel (1984). Rain

estimation from satellites: an estimation of the Griffith-Woodley Technique. Journal of Climate and Applied Meteorology. Vol. 23. Pages. 102–116.

[11] R.F. Adler and A.J. Negri (1988). A satellite infrared technique to estimate tropical convective and stratiform rainfall. Journal of Applied Meteorology. Vol. 27. Pages. 30– 51.

[12] I.M. Lensky and D. Rosenfeld (2003). A night-time delineation algorithm for infrared satellite data based on microphysical considerations. Journal of Applied Meteorology. Vol. 42. Pages. 1218–1226.

[13] T. Nauss and A.A. Kokhanovsky (2006). Discriminating raining from nonraining clouds at mid-latitudes using multispectral satellite data. Atmospheric Chemistry and Physics Vol. 6. Pages. 5031–5036.

[14] R. F. Adler, G. J. Huffman and P. R. Keehn (1994). Global Tropical Rain Estimates from Microwave-Adjusted Geosynchronous IR Data. Remote Sensing Reviews. Vol. 11. Pages. 125–152.

[15] C. Kidd, D. R. Kniveton, M. C. Todd and T. J. Bellerby (2003). Satellite Rainfall Estimation Using Combined Passive Microwave and Infrared Algorithms. Journal of Hydrometeorology. Vol. 4. Pages.1088–1104.

[16] F. J. Turk and S. D. Miller (2005). Toward Improved Characterization of Remotely Sensed Precipitation Regimes with MODIS/AMSR-E Blended Data Techniques. IEEE Transactions on Geoscience and Remote Sensing. Vol. 43. Pages. 1059–1069.

[17] G. J. Huffman, R. F. Adler, D. T. Bolvin, G. J. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong, E. F. Stocker and D. B. Wolff (2007). The TRMM Multisatellite Precipitation Analysis (TMPA):Quasi-global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. Journal of Hydrometeorology. Vol. 8. Pages. 38–55.

[18] A. Behrangi, B. Imam, K. L. Hsu, S. Sorooshian, T. J. Bellerby and G. J. Huffman (2010). REFAME: Rain Estimation Using Forward-Adjusted Advection of Microwave Estimates. Journal of Hydrometeorology. Vol. 11. Pages. 1305–1321.

[19] R. J. Joyce and P. Xie (2011). Kalman Filter-Based CMORPH. Journal of Hydrometeorology. Vol. 12. Pages. 1547–1563.

[20] D.A. Vila, L.G.G. de Goncalves, D.L. Toll and J.R. Rozante (2009). Statistical evaluation of combined daily gauge observations and rainfall satellite estimates over continental South America. Journal of Hydrometeorology. Vol. 10. Pages. 533–543.

[21] M.C. Periago, X. Lana, C. Serra and G. Fernandez (1991). Precipitation regionalization: an application using a meteorological network in Catalonia (NE Spain). International Journal of Climatology. Vol. 11. Pages. 529–543.

[22] H. Feidas, G. Kokolatos, A. Negri, M. Manyin, N. Chrysoulakis and Y. Kamarianakis (2009). Validation of an infrared-based satellite algorithm to estimate accumulated rainfall over the Mediterranean basin. Theoritical and Applied Climatology. Vol. 95. Pages. 91-109.

[23] P.T. Nastos (2011). Trends and variability of precipitation within the Mediterranean region, based on Global Precipitation Climatology Project (GPCP) and ground based datasets. Advances in the Research of Aquatic Environment. Vol. 1. Pages. 67–74.

[24] C.M. Philandras, P.T. Nastos, J. Kapsomenakis, K.C. Douvis, G. Tselioudis and C.S. Zerefos (2011). Long term precipitation trends and variability within the Mediterranean region. Natural Hazards and Earth System Sciences. Vol. 11. Pages. 3235–3250.

[25] M. Lazri, S. Ameur, J. M. Brucker, J. Testud, B. Hamadache, S. Hameg, F. Ouallouche and Y. Mohia, (2013). Identification of raining clouds using a method based on optical and microphysical cloud properties from Meteosat second generation daytime and nighttime data. Applied Water Science. Vol.3. Pages.1-11.

[26] P. T. Nastos, J. Kapsomenakis and K. C. Douvis (2013). Analysis of precipitation extremes based on satellite and high-resolution gridded data set over Mediterranean basin. Atmospheric Research. Vol. 131. Pages. 46-59.

[27] F. Lo Conti, K. L. Hsu, L. V. Noto and S. Sorooshian (2014). Evaluation and comparison of satellite precipitation estimates with reference to a local area in the Mediterranean Sea. Atmospheric Research. Vol. 138. Pages. 189-204.

[28] M. Lazri, S. Ameur and Y. Mohia (2014). Instantaneous rainfall estimation using neural network from multispectral observations of SEVIRI radiometer and its application in estimation of daily and monthly rainfall. Advances in Space Research. Vol. 53. Pages. 138-155.

[29] J. Schmetz, P. Pili, S. A. Tjemkes, D. Just, J. Kerkmann, S. Rota and A. Ratier (2002). An introduction to Meteosat Second Generation (MSG). Bulletin of the American. Meteorological Society. in press.

[30] V. Levizzani, F. Porcú and F. Prodi (1990). Operational rain-fall estimation using METEOSAT infrared imagery: an application in Italy’s Arno river basin. Its potential and drawbacks. ESA Journal. Vol. 14. Pages. 313–323.

[31] M. Marroccu, A. Pompei, G. Dalu, G. L. Liberti and A. J. Negri (1993). Precipitation estimation over Sardinia from satellite infrared data. International Journal of Remote Sensing. Vol. 14. Pages. 115–134.

[32] R.Tarruella and J. Jorge (2003). Comparison of three infrared satellite techniques to estimate accumulated rainfall over the Iberian Peninsula. International Journal of Climatology. Vol. 23. Pages. 1757-1769.

International Journal of
Intelligent Systems and
Applications in Engineering

ISSN:2147-6799 www.atscience.org/IJISAE

Advanced Technology and Science

Original Research Paper

Diagnosis of Anemia in Children via Artificial Neural Network

Esra KAYA[1], Mehmet Emin AKTAN*[2], Ahmet Taha KORU[3], Erhan AKDOĞAN[4]

Received 05th September 2014, Accepted 18th December 2014

Abstract: In this paper, a neural network algorithm, which diagnosis of anemia for children under 18 years of age, is presented. The network is trained by using data from hemogram test results from 30 patients and an expert doctor. The network has 5 inputs (HGB, HCT, MCV, MCH, MCHC) and an output. Simulations on 20 different patients show that the artificial neural network detects disease with high accuracy. In this paper, it is shown that anemia diagnosis can be made via neural network methods.

Keywords: *Anemia, Diagnose, Artificial neural network.*

1. Introduction

Anemia is a decrease in number of red blood cells (RBCs) or less than the normal quantity of hemoglobin in the blood. The most common anemia type is iron deficiency anemia among various types [1], [2]. Iron-deficiency affects people at all age but mostly women and children. Iron-deficiency caused by insufficient dietary intake and absorption of iron [1]. Fast growing of children, high iron demand during pregnancy, and menstruation are some physiological reasons that lead to iron-deficiency. Iron-deficiency is a widespread diet problem all over the world. Prevalence of iron-deficiency anemia is higher in low developed and developing countries [2, 3, 4]. Prevalence of anemia is also high in Turkey [2, 4, 5]. As premature and preterm birth increases in recent years, iron-deficiency age decreases [6]. In that manner, hematology includes the study of etiology, diagnosis, treatment, prognosis, and prevention of blood diseases [7].

Anemia diagnosis is being made by a doctor using the data obtained from hemogram blood test results in today's applications. The parameters considered for the detection of disease is listed below.

Table 1. Parameters

HGB: Hemoglobin
HCT: Hematogrit
MCV: Mean Corpuscular Volume
MCH: Mean Corpuscular Hemoglobin
MCHC: Mean Corpuscular Hemoglobin Concentration [8]

There is vast literature about the diagnosis of iron-deficiency anemia. In [9], diagnosis of women's iron- deficiency anemia is investigated by several neural network techniques using Red Blood Cells (RBC), HGB, HCT, MCV, MCH, MCHC datum. These techniques are Feedforward Networks (FFN), Cascade Forward Networks (CFN), Distributed Delay Networks (DDN), Time Delay Networks (TDN), Probabilistic Neural Network (PNN), and Learning Vector Quantization (LVQ). Furthermore, comparative results are presented. In [10], two different neural network model, whose inputs are zinc protoporphyrin (ZPP), Hb, RBC, and MCV, is used to diagnose anemia and the performances of these neural networks are compared.

In this paper, a neural network model, which diagnose anemia for children under 18, is developed. The network architecture is designed to have 3 layers which are input, hidden, and output layers. In this network there 5 inputs and 1 output. Inputs are HGB, HCT, MCV, MCH, MCHC values. In order to train the network real data from 30 different patients are used. Data obtained from another 20 patients are used to verify the trained network. Among total 50 patients, there are 18 female and 32 male. These data is taken from a private hospital. Results are discussed in conclusions section.

2. Materials and Method

2.1. Network Architecture

The network is developed in feedforward multilayer perceptron with 3 layers architecture. Number of neurons in the hidden layer is 100. Tangent sigmoid (TANSIG) is used as activation functions. The results showed that the network is able to diagnose the disease with high accuracy. Network architecture is illustrated in Figure 1.

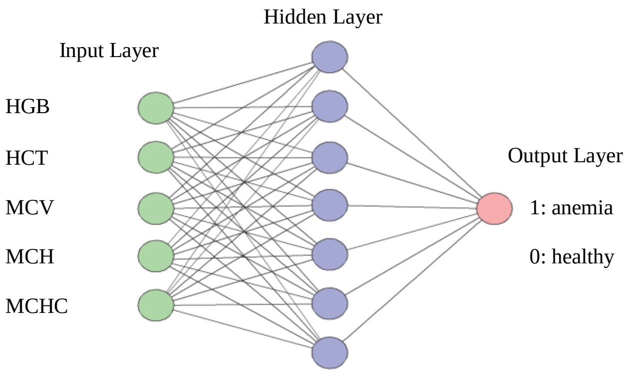


Figure 1. Architecture of the developed artificial neural network

[1] *Yildiz Technical University, Faculty of Control – Automat. Eng. – Turkey*
[2] *Yildiz Technical University, Faculty of Mechanical Eng. – Turkey*
[3] *Yildiz Technical University, Faculty of Mechanical Eng. – Turkey*
[4] *Yildiz Technical University, Faculty of Mechanical Eng. – Turkey*
** Corresponding Author: Email: meaktan@yildiz.edu.tr*

Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

The values to be used as inputs are determined after discussing with experts. According to these discussions 5 inputs are HGB, HCT, MCV, MCH, MCHC. The output is a logical value which is either 0 (healthy) or 1 (anemia). Inputs and outputs of 50 different patients are saved to a computer database. 30 of these 50 data are used to train the artificial neural network. Whereas, the remaining 20 is used to test the network. It is tried to distribute the data such that equal number of healthy and diseased patients are available in each groups. Examples of the training data and test data can be seen in Table 2 and Table 3, respectively.

2.2. Training ANN

The neural network toolbox of Matlab© R2010a environment is used to create, train and test the network. At first, train and test data is normalized between -1 and 1. Corresponding code is,

$$[norm\_train, ps1] = mapminmax(train) \quad (1)$$

$$[norm\_test, ps2] = mapminmax(test) \quad (2)$$

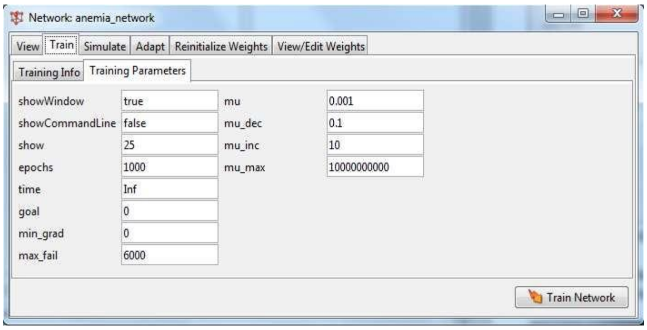
Table 2. Sample train data

HGB *g/dl*	*HCT* *%*	*MCV* *fL*	*MCH* *pg*	*MCHC* *%*	*0: healthy* *1: anemia*
7.27	22.50	84.10	27.10	32.30	1
11.70	35.30	73.10	24.30	33.30	0
7.71	23.40	85.00	28.10	33.00	1
10.90	33.90	76.90	24.80	32.30	0

Table 3. Sample test data

HGB *g/dl*	*HCT* *%*	*MCV* *fL*	*MCH* *pg*	*MCHC* *%*	*0: healthy* *1: anemia*
9.79	31.30	56.70	17.80	31.30	0
12.70	40.30	91.20	28.80	31.50	1
6.47	20.20	68.10	21.80	32.00	1
11.20	33.00	80.00	27.10	33.80	0

Then, training, test, and output data is assigned to variables, so they are ready to be used. Neural Network Toolbox interface can be seen in Figure 2.



Network: anemia_network

showWindow	true	mu	0.001
showCommandLine	false	mu_dec	0.1
show	25	mu_inc	10
epochs	1000	mu_max	10000000000
time	Inf		
goal	0		
min_grad	0		
max_fail	6000		

Figure 2. Neural Network Toolbox interface

Training and test data is transferred as input data, and the output of training data is transferred as target data, and the network is created. By using ”Create Network or Data” interface, network type, training function, number of layers, number of neurons, and transfer functions values are determined. ”Create Network or Data” interface is shown in Figure 3.

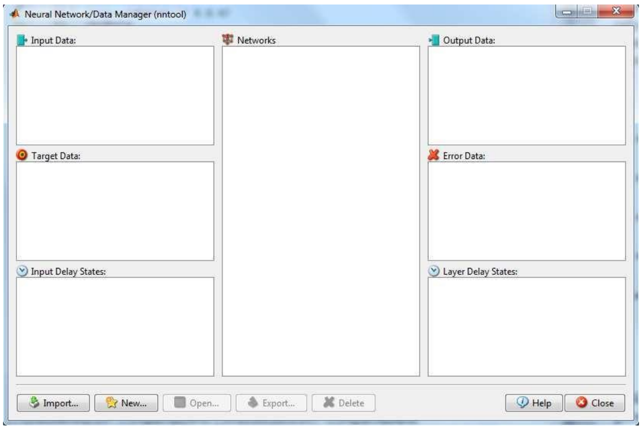
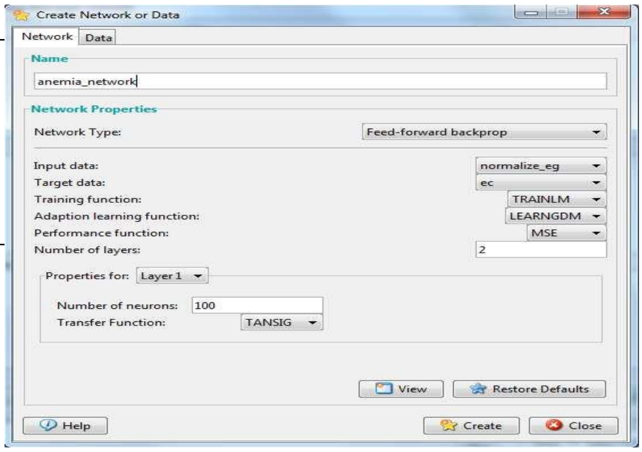


Figure 3. Create Network or Data interface

Once the network is created, the next step is the training. Training parameters are shown in Figure 4.



Create Network or Data

Name: anemia_network

Network Properties

Network Type: Feed-forward backprop

Input data: normalize_eg
Target data: ec
Training function: TRAINLM
Adaption learning function: LEARNGDM
Performance function: MSE
Number of layers: 2

Properties for: Layer 1

Number of neurons: 100
Transfer Function: TANSIG

Figure 4. Training parameters

The training of the network is made by backpropagation method after setting the network properties and the training parameters. Performance of the network, regression, and training state plots are shown in Figure 5, Figure 6, and Figure 7, respectively.

Mean square error is between 10^{-23} and 10^{-26} according to network learning performance plot in Figure 5. We can say that it is a successful training.

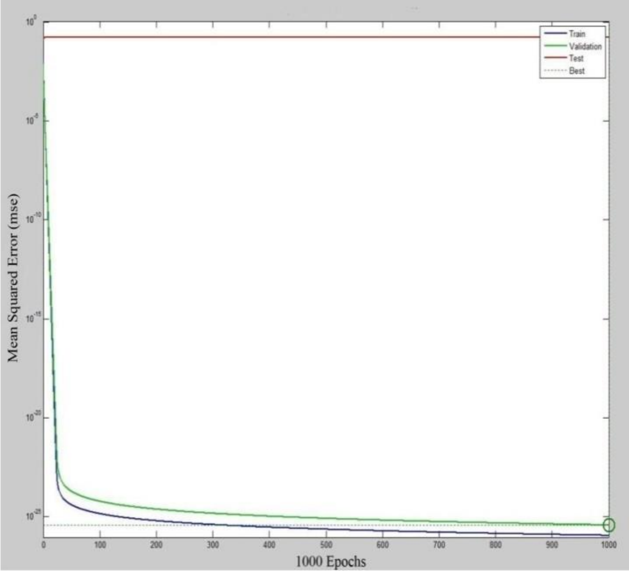


Figure 5. Learning performance of the network

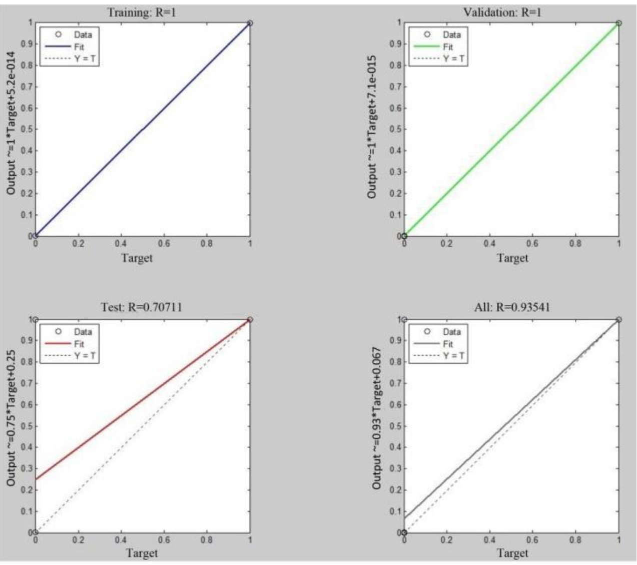


Figure 6. Regressions result

2.3. Testing ANN

The network is tested with 20 data from patients. Once the outputs of the network is compared with hemogram test results, there is only 1 fault detection. Comparison of the network's output with the real diagnosis of patients can be seen in Table 4. The logic 1 means the patient is diseased by anemia and 0 denotes corresponding person is healthy.

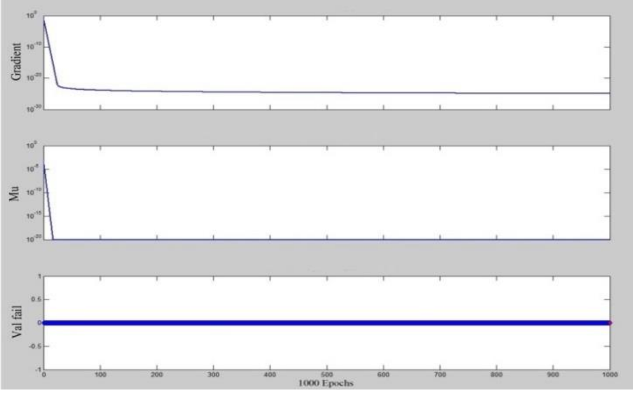


Figure 7. Training state results

Table 4. Test results

Patient	Output	Network Output	Accuracy (%)
P01	1	1	100
P02	0	0	100
P03	0	0	100
P04	0	0	100
P05	1	0	0
P06	0	0	100
P07	1	1	100
P08	1	1	100
P09	1	1	100
P10	0	0	100
P11	1	1	100
P12	1	1	100
P13	1	1	100
P14	1	1	100
P15	0	0	100
P16	0	0	100
P17	0	0	100
P18	1	1	100
P19	0	0	100
P20	0	0	100

3. Conclusions

In present paper, an artificial neural network which diagnose anemia in children under 18 is developed. Some of the data of healthy/diseased decisions made by an expert doctor according to the blood parameters obtained from a hospital are used to train the neural network, while some of them are used to test the performance of the network. In this test process, outputs of the network and the diagnosis of the doctor are compared. For 19 of 20 cases, outputs of the network and the diagnosis of the doctor are matched. Hence, accuracy of the network is %90.909. The developed network architecture is adequate to be used to help doctors in real-life applications.

Acknowledgements

We would like to thank Dr. Fırat Can and Emrah Sarıtaç, M.Sc., for their support and valuable comments

4. References

[1] G.Gedikoğlu and L. Ağaoğlu. Kan hastalıkları, Vol.2. İzmir,Turkey: Nobel Tıp Kitapevleri, 1993.

[2] E.Ç. Eren. Çocuklarda Yaş Gruplarına ve Cinslerine Göre Anemi ve Demir Eksikliği Anemisi Sıklığının İncelenmesi. Aile Hekimliği Uzmanlık Tezi, Turkey 2008.

[3] M. A. Simes and L. Solmonpera. The We Anling Iron for All Or one. Acta Pediatr Scad, 1989, pp. 103-878.

[4] H. Soylu, Ü. Özgen, M. BAbalıoğlu, Ş. Aras and S. Sazak. Iron Deficiency and Iron Deficiency Anemia in Infants and Young Children at Different Socioeconomic Groups in

Istanbul. Turkish Journal of Hematology, 2001, pp. 19-25.

[5] C. Uzel and M.E. Condrad. Absorbtion of Heme Iron. Semin Hematol, 1998,35:27-34.

[6] F. Tiker, A. Tarcan, N. Özbek and B. Gürakan. Çok Düşük Doğum ağırlıklı Bebeklerde Erken Enteral Demir Eksikliği. Çocuk Sağlığı ve Hastalıkları Dergisi, 2005, pp. 14-19.

[7] F. Başçiftçi and H. İncekara. Web Arayüzü ile Hematoloji Laboratuarı Tahlillerinin Değerlendirilmesi için Bulanık Girişli Uzman Sistem Tasarımı. Fen Bilimleri Enstitüsü Dergisi 2011. Pp. 51-55.

[8] Z. Yılmaz and Ş. Ocak . Bulanık Mantık ile Aneminin Belirlenmesi. Çankaya Üniversitesi 1. Mühendislik ve Teknoloji Sempozyumu, Turkey 2008.

[9] Z. Yılmaz and M.R. Bozkurt. Determination of Women Iron Deficiency anemia Using neural Networks. Journal of Medical Systems, 2012, pp. 2941-2945.

[10] J. Ven, M.G. Scholorl and P.C.M. Bartels. Automated result İnterpartion in Anemia Testing Using Artificial Neural Networks. Ned Tijdshr Klin Chem Labgeneesk, 2006, pp. 230-231.

International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Intrusion Detection Forecasting Using Time Series for Improving Cyber Defence

Azween Abdullah *[1], Thulasy Ramiah Pillai [2], Cai Long Zheng [3], Vahideh Abaeian[4]

Received 08h September 2014, Accepted 10th October 2014

Abstract: The strength of time series modeling is generally not used in almost all current intrusion detection and prevention systems. By having time series models, system administrators will be able to better plan resource allocation and system readiness to defend against malicious activities. In this paper, we address the knowledge gap by investigating the possible inclusion of a statistical based time series modeling that can be seamlessly integrated into existing cyber defense system. Cyber-attack processes exhibit long range dependence and in order to investigate such properties a new class of Generalized Autoregressive Moving Average (GARMA) can be used. In this paper, GARMA (1, 1; 1, ±) model is fitted to cyber-attack data sets. Two different estimation methods are used. Point forecasts to predict the attack rate possibly hours ahead of time also has been done and the performance of the models and estimation methods are discussed. The investigation of the case-study will confirm that by exploiting the statistical properties, it is possible to predict cyber-attacks (at least in terms of attack rate) with good accuracy. This kind of forecasting capability would provide sufficient early-warning time for defenders to adjust their defense configurations or resource allocations.

Keywords: *Intrusion forecasting, Predictive modeling, Generalized Autoregressive Moving Average, Long range dependence*.

1. Introduction

Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. A predictive model is made up of a number of predictors, variable factors that are likely to influence or predict future behavior. The end result is both a set of factors that predict, to a relatively high degree, the outcome of an event, as well as what that outcome will be. In marketing, for example, a customer’s gender, age and purchase history might predict the likelihood of a future sale. To create a predictive model, data is collected for the relevant factors, a statistical model is formulated, predictions are made and the model is validated. The model may employ a simple linear equation or can be a complex neural network or genetic algorithm.

There are two main approaches to intrusion detection - traffic and content analysis. Most intrusion detection systems use content analysis. Content analysis looks for signatures within the packet payload and will respond appropriately when a match is found. Through traffic analysis, the interpreter hopes to see patterns in the packet header that may indicate abnormal network behavior. The main advantage of traffic analysis that is possible to get a more accurate interpretation of the data. The disadvantages are that it requires a trained analyst to accurately interpret the data, it is not possible to have close-to-real-time detection, and it requires a large amount of disk space.

Figure 1 depicts the generalized intrusion detection model with forecasting component that we propose in this paper. The model combines the traditional approach to intrusion detection with predictive and self-healing component.

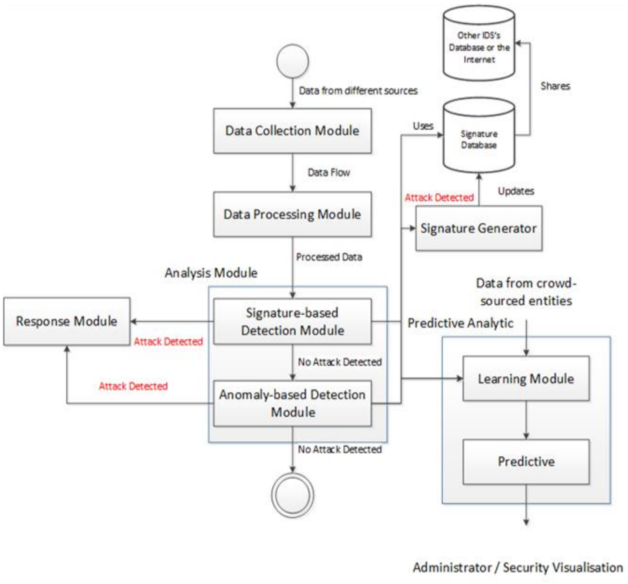


Figure 1. Model deployment and management

The phases of predictive modeling are rather straightforward, and involve activities aimed at ensuring a look into the past through the analysis of various data points will in fact help predict the future. In Section 2, we discussed the various statistical methods for systematically analyzing and exploiting cyber-attack data. In Section 3, we give a brief account of time series models while in Section 4 the estimation of the parameters using Hannan-Rissanen Algorithm and Maximum Likelihood Estimation of the model is discussed. In Section 5 we illustrate the modelling of cyber- attack

[1&2] *School of Computing and IT, Taylors University, Subang Jaya, Selangor, Malaysia*
[3]*Unitar International University, Petaling Jaya, Selangor, Malaysia*
[4] *School of Business, Taylors University, Subang Jaya, Selangor, Malaysia*
* *Corresponding Author: Email: azween.abdullah @taylors.edu.my*
Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

process of a university network using GARMA model. There are 4 types of cyber-attacks, namely DOS, U2R, R2L and PROBE. We have used the cyber-attack which is called PROBE in our discussion. Finally the conclusion is drawn in Section 6.

2. Related Work

Reference [1] found that, for the first time, that Long-Range Dependence (LRD) had exhibited by honeypot-captured cyber-attacks. They have exploited the statistical properties (LRD) to predict cyber-attacks (at least in terms of attack rate) with good accuracy. They proposed the statistical framework for systematically analyzing and exploiting honey-pot captured cyber-attack data. It is called cyber-attack process, which is a new kind of mathematical objects that can model the cyber-attacks. Reference [1] also mentioned that in many cases, attack processes may not be Poisson. They have suggested for characterizing such processes, we need to use advanced statistical methods, such as Markov process, Levy process and the time series methods.

Reference [1] had summarized that 80% network-level attack processes, 70% victim level attack process and 44.5% port level attack processes exhibit LRD. Due to this cyber-attack processes should be modelled using LRD-aware stochastic processes. Reference [1] had added that for attack processes that exhibit LRD, LRD-aware models can predict their attack rates better than LRD-less models do. However, there are LRD processes that can resist the prediction of even LRD-aware models. This hints that new prediction models are needed.

Reference [1] had suggested a future work on the cause of LRD in cyber-attack process. In order to investigate such properties, a new class of Generalized Autoregressive Moving Average which has been introduced by [17] can be utilized. Reference [2] described how far into the future one can predict network traffic by employing Autoregressive Moving Average (ARMA) as a model. Reference [3] had studied anomaly prediction in network traffic using Adaptive Weiner Filtering and ARMA modeling.

Reference [4] suggested 3 categories of econometric models such as time series models exploiting the statistical properties of the data, financial models based on the relationship between spot and future prices and structural models describing how specific economic factors and the behaviour of economic agents affect the future values of oil prices. They have described these three classes of econometric models that have been proposed to forecast oil prices and presented the different and controversial empirical results.

In addition, they also added that it is not possible to identify which class of models outperforms the others in terms of forecasting. Reference [4] had mentioned that there are a number of statistical issues which should be accounted for in the development of an econometric model, namely heteroskedasticity, autocorrelation and non-stationarity. They also added that we have to follow the idea that all relevant information to forecast the oil price is embedded in the price itself. Random walks, Martingale processes and simple autocorrelation models root their justification on this idea.

Reference [5] proposed a proactive security system to forecast Distributed Denial of Service (DDos) attacks. Their study focused on informative forecasts by providing us with identifying the type, time and target of attacks rather than merely forecasting the increase or decrease of attacks. The Honeynets were deployed to collect the raw data necessary to forecast DDos attacks and they analyzed Hflow data gathered from the Honeynets as a first step to estimate intrusion factors. They have chosen regression analysis as the forecasting method. Reference [5] also suggested that several forecasting methods with regression analysis should be considered to improve the accuracy of forecasts in the future.

The previous studies regarding intrusion forecasting lack certain details. Most of the prediction methods merely depend on preceding attack trends [6] - [10]. They do not provide a specific forecasting of the exact type, time and target of the attack. Although these studies are meaningful, we need more specific and concrete forecasts for proactive forecasting systems to be effective. There are also the studies that predict the propagation of attacks and predict the next stage of attacks based on information from network scanning and present attacks [6]-[8]. However, these studies shorten the detection time but they do not guarantee sufficient time to provide a countermeasure to the attack. Reference [6] developed a system called STARMINE. This system visualizes the attack traffic in a 3- dimensional graph using spatial, temporal and logical analysis. This study provided the basis to under-stand the characteristics of attack traffic and to predict intrusion. The forecast depends on the judgment of the individual who interprets the graph. Reference [14] proposed a forecasting method to predict the probability of Internet intrusion using a regression model. The study merely provided a theoretical approach using an econometric model of the intruder and the victim rather than presenting experiments and quantifiable results. However, the study is worth noting because it emphasizes the possibility of specifically forecasting attacks, rather than merely predicting increases or decreases in attack frequency. Reference [11] proposed the Security Operation Center (SOC) framework for the cooperation between ISPs to forecast new attacks. This framework performs statistically automated and manual forecasts using Bayesian network. It quickly detects abnormal events in a high-speed network and selects a target by predicting the type and quantity of the attack. Although the purpose of the prediction is not to prepare for the attack, this study is valuable because it predicts the attack in a spatial rather than a temporal context. Reference [12] used Neurogenetic algorithm to predict attacks within a short time. The purpose of this study is to predict and block attacks just before they occur to improve the effectiveness of IPS. Reference [13] predicted attacks by using graph theory. This study proposed a model that uses system vulnerability to predict the progression of attacks. This study also attempts to shorten the time of intrusion detection. Reference [7] proposed a forecasting method using Bayesian inference, which calculates the increase or decrease of the probability of the next attack based on the number of attacks observed previously. Reference [8] numerically expressed the present security situation, and used time-series analysis to forecast the variation of the security situation due to time. Both works, by forecasting the increase or the decrease of intrusions, serve as a valuable foundation for the field of intrusion forecasting. Reference [9] proposed a method to forecast the increase or decrease of the Bot agents by month that uses Hidden Markov Model (HMM). This study argues HMM is a superior forecasting method for predicting attacks than time-series analysis, because time-series analysis does not precisely represent the hidden characteristics of attacks. Reference [9] proposed the framework of an intrusion forecasting system that is more accurate by using two algorithms, rather than just one algorithm.

This study proved that accuracy is particularly high when they used Markov chain and time-series analysis. These two studies worked to improve the accuracy of forecasting based on the increase or decrease of attacks. Our study is focused on informative forecasts by providing us with identifying the type, time and target of attacks rather than merely forecasting the increase or decrease of attacks. Time Series Modelling

A time series is a set of observations X_t, each one being recorded at a specific time t and denoted by $\{X_t\}$. It can be represented as a realization of the process based on the general model called Classical Decomposition Model, and specified as:

$$X_t = m_t + s_t + Y_t \qquad (1)$$

t = 1; 2…; *n*, where m_t is a trend component, s_t is a seasonal component and Y_t is a random noise component which is stationary [16].

Time series modeling help us to predict data series that are typically not deterministic but contain a random component. The deterministic components, m_t and s_t need to be estimated and eliminated as to make the residue or noise component Y_t to be stationary time series. A non-stationary time series needs to be transformed to a stationary time series, in order to analyze its properties and to use it for prediction purposes [16].

Time series data are usually modelled as Autoregressive Moving Average (ARMA) processes. ARMA processes are widely used in forecasting. The family of standard Autoregressive AR (1) processes generated by,

$$X_t - \alpha X_{t-1} = Z_t \qquad (2)$$

where $|\alpha| < 1$, $\{X_t\}$ is a time series, $\{Z_t\}$ is a sequence of uncorrelated random variables (not necessarily independent) with zero mean and variance σ^2 ,known as white noise and denoted by $WN(0,\sigma^2)$. Using the backshift operator, $B(\text{i.e.} \mathrm{B}^j X_t = X_{t-j}, j \geq 0)$ and the identity operator $I = B^0$, equation (2) can be written as,

$$(I - \alpha B)X_t = Z_t \qquad (3)$$

Reference [17] has introduced Generalised Autoregressive (GAR (1)) model, defined as,

$$(1-\alpha B)^\delta X_t = Z_t \ , \qquad (4)$$

by including an additional parameter $\delta > 0$.

The Moving Average or MA (1) is generated by,

$$X_t = (I - \beta B)Z_t \ , \qquad (5)$$

where $|\beta| < 1$.

The Generalised Moving Average (GMA (1)) model has also been introduced [18]. This model is given as,

$$X_t = (1-\beta B)^\delta Z_t \ . \qquad (6)$$

It has been shown that the additional parameter ± plays an important role in modelling and forecasting [17] & [18].

The standard ARMA (1, 1) can be written as,

$$(I - \alpha B)X_t = (I - \beta B)Z_t \qquad (7)$$

where $|\alpha|, |\beta| < 1$.

Reference [17] also introduced a new, generalised version of (7) with the additional parameters $\delta_1 \geq 0$ and $\delta_2 \geq 0$ satisfying

$$(I - \alpha B)^{\delta_1} X_t = (I - \beta B)^{\delta_2} Z_t \qquad (8)$$

This new class of models known as the Generalised Autoregressive Moving Average (GARMA) Model has been introduced by [17] in order to reveal some hidden features in time series data. These types of models could be used to describe data with different frequency components for suitably chosen indices.

More recently, [19] have considered the GARMA $(1,1;1,\delta)$ model which is defined by,

$$(1-\alpha B)X_t = (1-\beta B)^\delta Z_t \ . \qquad (9)$$

where $-1 < \alpha$, $\beta < 1$ and $\alpha > 0$. In addition, [20] studied the behaviour of the process GARM $(1,1;\delta,1)$. The GARMA $(1,1;\delta,1)$ process is generated by

$$(I - \alpha B)^\delta X_t = (I - \beta B)Z_t \ , \qquad (10)$$

where $-1 < \alpha$, $\beta < 1$ and $\delta > 0$.

The GARMA $(1,1;1,\delta)$ and GARMA $(1,1;\delta,1)$ models can be further generalised as follows.

$$(1-\alpha B)^{\delta_1} X_t = (1-\beta B)^{\delta_2} Z_t , \qquad (11)$$

where $-1 < \alpha$, $\beta < 1$, $\delta_1 > 0$ and $\delta_2 > 0$. This model is denoted by GARMA $(1,1;\delta_1,\delta_2)$ and some properties of this model have been established [20]. All these models have been shown to be useful in modelling time series data.

It is interesting to note that the GARMA model can be further expanded to GARMA $(1,2;\delta,1)$ and it is given as below:

$$(1-\alpha B)^\delta X_t = (1-\beta_1 B - \beta_2 B^2)Z_t \ , \qquad (12)$$

where $-1 < \alpha, \beta_1, \beta_2 < 1$ and $\delta > 0$.

In this paper, we have utilized advanced time series models namely GARMA in improving intrusion forecasting. The objective of this paper is to illustrate the application of GARMA modelling to cyber-attack processes. We illustrate the fitting of GARMA model to the cyber-attack process which has been observed from November 2013 to January 2014 in the university network. The estimation of the model was done using Hannan-Rissanen Algorithm and Maximum Likelihood Estimation.

3. Estimation of Parameters

Estimation of the parameters of the model is the second procedure in time series analysis after model selection in forecasting. The estimation algorithm used in this study requires initial parameter values. A number of preliminary estimation algorithms are available to provide these initial values.

In this section, we discuss the estimation methods that we have employed in this study. Hannan-Rissanen Algorithm is used as the preliminary estimation. In addition, Maximum Likelihood Estimation is also discussed here.

3.1. Hannan-Rissanen Algorithm Estimator

The Hannan-Rissanen Algorithm technique is one of the preliminary techniques used for ARMA (p; q) models where p > 0 and q > 0. ARMA (p; q) is generated by,

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t - \theta_1 Z_{t-1} - \ldots - \theta_q Z_{t-q} \qquad (13)$$

Firstly, a high-order AR (m) model with m > max(p; q) is fitted to the data using the Yule-Walker estimates. If $(\phi_{m1},\ldots,\phi_{mm})$ is the vector of estimated coefficients, then the estimated residuals are computed from the equations

$$Z_t = X_t - \phi_{m1} X_{t-1} - \ldots - \phi_{mm} X_{t-m}, t = m+1,\ldots,n \qquad (14)$$

Secondly, the vector of parameters, $\omega = (\phi_1 ',\ldots,\phi_p ',\theta_1 ',\ldots,\theta_q ')'$ is estimated by minimizing the sum of squares

$$S(\omega) = \sum_{t=m+1+q}^{n} (X_t - \phi_1 X_{t-1}, ..., \phi_p X_{t-p} + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q})^2$$

(15)

with respect to ω. This gives the Hannan-Rissanen Algorithm estimator $\omega = (Z'Z)^{-1} Z' X_n$,

where $X_n = (X_{m+1+q}, ..., X_n)'$ and Z is the $(n-m-q)\times(p+q)$ matrix. ARMA (p,q) model is fitted using the Hannan-Rissanen estimates. See [15] for details. The fitted model is

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t - \theta_1 Z_{t-1} - ... - \theta_q Z_{t-q} \tag{16}$$

Thirdly, ω values can be manipulated to obtain the parameter values for GARMA $(1,1;1,\delta)$. $\omega = (\phi_1, \theta_1, \theta_2, ..., \theta_q)$ is computed using ARMA $(1,q)$ model. The fitted ARMA $(1,q)$ model is

$$X_t - \phi_1 X_{t-1} = Z_t - \theta_1 Z_{t-1} - ... - \theta_q Z_{t-q} \tag{17}$$

The GARMA $(1,1;1,\delta)$ model also can be recorded as below after the expansion of the right side expressions of equation (9),

$$X_t - \alpha X_{t-1} = Z_t - \beta\delta Z_{t-1} ... \tag{18}$$

After comparing the equation (17) and equation (18), we can deduce that the ϕ_1 value is equivalent to α and the θ_1 value is equivalent to $\beta\delta$. The estimation of β and δ are done by assuming that $\beta = \delta$. If the value of $|\beta| > 1$, then we assume $|\beta| = 0.6$ and $\delta = \theta_1 / \beta$.

Hannan-Rissanen Algorithm is used to provide preliminary estimates of the GARMA parameters as such these aforementioned assumptions are made.

The corresponding estimate for σ^2 is given as,

$$\sigma^2 = S(\omega / (n-m-q)) \tag{19}$$

3.2. Maximum Likelihood Estimation

Reference [22] developed the principle of Maximum Likelihood Estimation (MLE). MLE is a popular method of parameter estimation and is an indispensable tool for many statistical modelling techniques.

The maximum likelihood estimates (MLE) for the parameters of the model are obtained by numerically minimizing the function,

$$-2\ln f(x) = T\ln(2\pi) + \ln\left|\sum\ \right| + x'\sum^{-1} x \tag{20}$$

where T is the number of observations, x is the observed vector and ∑ denotes the covariance matrix. The entries of ∑ are given by [19],

$$\gamma_0 = \frac{\sigma^2}{1-\alpha^2}$$

$$\left[\sum_{j=1}^{\infty} \binom{\delta}{j} (-\alpha\beta)^j F(-\delta, j-\delta; j+1; \beta^2) + \right.$$

$$\left. \sum_{j=0}^{\infty} \binom{\delta}{j} (-\alpha\beta)^j F(-\delta, j-\delta; j+1; \beta^2) \right] \tag{21}$$

and

$$\gamma_h = \frac{\sigma^2}{1-\alpha^2}$$

$$\left[\beta^h \sum_{j=1}^{\infty} \binom{\delta}{h+j} (-\alpha\beta)^j \right.$$

$$F(-\delta, h+j-\delta; h+j+1; \beta^2)$$

$$+\alpha^h \sum_{j=0}^{\infty} \binom{\delta}{j} (-\alpha\beta)^j F(-\delta, j-\delta; j+1; \beta^2)$$

$$\left. +\sum_{j=1}^{h} \binom{\delta}{j} \alpha^{h-j} (-\beta)^j F(-\delta, j-\delta; j+1; \beta^2) \right] \tag{22}$$

$h \geq 1$ where

$$F(a,b;c;z) = 1 + \frac{ab}{1!c} z + \frac{a(a+1)b(b+1)}{2!c(c+1)} z^2 + \frac{a(a+1)(a+2)\,\mathrm{b(b+1)(b+2)}}{3!c(c+1)(c+2)} z^3 + ... \tag{23}$$

The initial start-up values for the numerical minimization are the approximate Hannan-Rissanen Algorithm estimates.

4. Experimental Set up and Analysis

To gather raw data for intrusion factors, we have deployed the internal Honeynet and the external Honeynet of our university. The Honeynet provides more detail data such as system logs than statistical data from security organization. Using the Honeynet also has the advantage of collecting valuable raw data. It holds a meaningful correlation, since the location of the installed Honeynet and the attacked network are close by. We use Hflow to integrate and store various types of data such as network flows, IPS logs, and data regarding intrusive activity captured by Sebek. As a first step toward attack forecasting, we analyzed Hflow data gathered from the Honeynets. This analysis used data collected from November 2013 to January 2014.

4.1. Modeling of the Network

In this section, we illustrate the modeling of cyber-attack process of a university using GARMA (1, 1; 1, ±) model. The live data was obtained from the university network through daily observations of the network traffic measured in packets from November 2013 to January 2014. The plot of the time series are shown in Figure 2 and it is clear that it is non-stationary.

In order to achieve stationarity, the data set was twice-differenced at lag 1 and mean corrected using an Interactive Time Series Modelling Package (ITSM) and a plot of this is shown in Figure 3. Plot of the sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) are also shown in Figure 4. From Figures 3 and 4, the time series appears to be stationary.

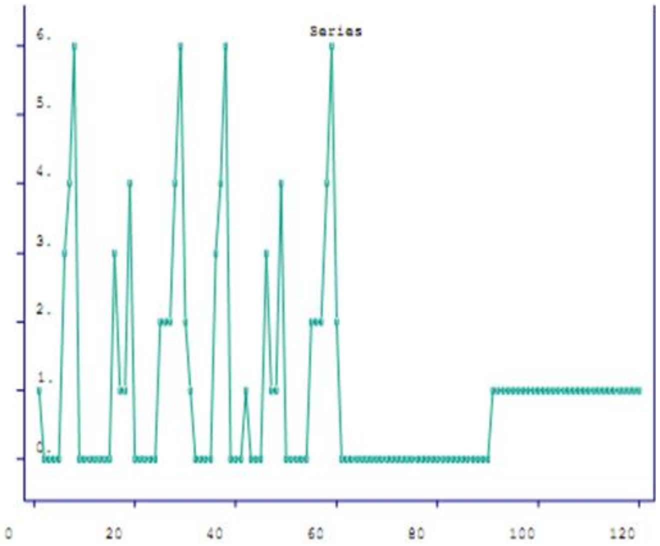


Figure 2. Cyber-attack processes measured in packets from November 2013 to January 2014

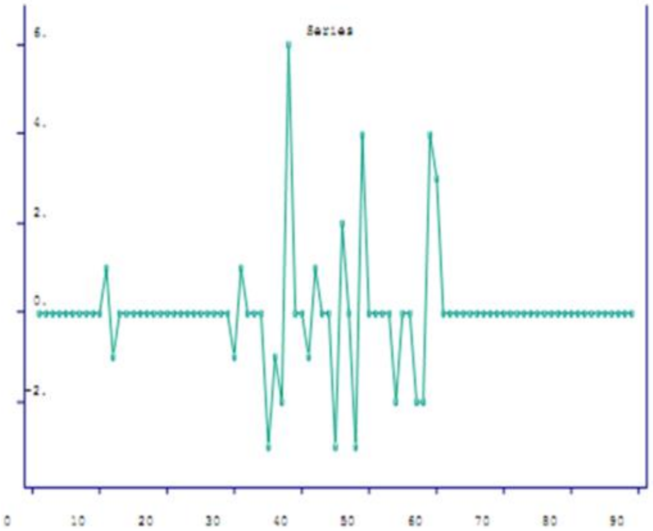


Figure 3. Cyber-attack processes which was twice differenced and mean corrected

Computer programs were written using S-PLUS language to model the stationary network data collected using GARMA (1, 1; 1, ±) model. The estimation of the parameters using Hannan-Rissanen Algorithm and Maximum Likelihood Estimation was done. The results are shown in Table 1 below.

4.2. Experimental Results and Analysis

The Hannan-Rissanen Algorithm estimation is obtained for the GARMA (1; 1; 1; ±) model and the fitted model is

$$(1+0.2425B)Y_t = (1+0.1420B)^{0.1420} Z_t \,, \tag{24}$$

$$Z_t \sim WN(0,1.9221)\,. \tag{25}$$

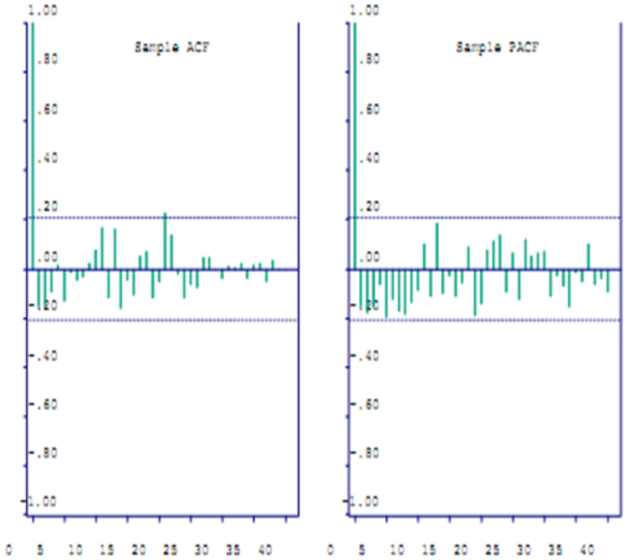


Figure 4. Plot of ACF and PACF of cyber-attack processes

where,

$$Y_t = (1-B^{30})(1-B)(X_t - 0.0112)\ . \tag{26}$$

On the other hand, the GARMA $(1,1;1,\delta)$ fitted models are,

$$(1-0.9999B)Y_t = (1+0.4303)^0 Z_t \,, \tag{27}$$

$$Z_t \sim WN(0, 0.0012) \tag{28}$$

by the maximum likelihood method. Using the above fitted models, point forecasts for the cyber-attack process data set for the next six time periods are shown in Table 1. It can be seen from Table 1 that all the point forecasted values through HRA and MLE estimation give a very close reading to the actual values.

Table 1. Actual and forecast values for cyber-attack process data

Day	*Actual value*	*Forecast value using HRA*	*Forecast value using MLE*
115	1	-0.219332	0.999 999
116	1	-0.219332	0.999 999
117	1	-0.219332	0.999 999
118	1	-0.219332	0.999 999
119	1	-0.219332	0.999 999
120	1	-0.219332	0.999 999

5. Conclusion

The objective of our study in this paper was to illustrate the fitting of GARMA (1, 1; 1, ±) model to cyber-attack process. The estimation of the parameters was done using Hannan-Rissanen Algorithm and Maximum Likelihood Estimation. The point forecast obtained through Maximum Likelihood Estimation very close to the actual value. The performance of the GARMA (1, 1;

1, ±) model in cyber-attack is very good. In future works, more advanced GARMA such as GARMA (1, q; 1, ±) models could be considered to improve the accuracy of forecasts of massive network traffic. In the next paper we would consider the other types of attack namely DOS, U2R and R2L.

Acknowledgements

The project is funded by the Ministry of Education Malaysia under the Fundamental Research Grant Scheme 2013 titled: Predictive Analytic Theory Generation and Foundation for a Novel Bio-inspired Intrusion Prevention and Self-regeneration System for Cyber Defence (FRGS/2/2013/ICT02/TAYLOR /02/1).

References

[1] Z. Zhan, M. Xu and S. Xu, Characterizing Honeypot-captured cyber- attacks: Statistical Framework and Case study, Information Forensics and Security, IEEE Transactions, 8(11): 1775-1789, November 2013.

[2] Sang and S. Li, A predictability analysis of network traffic, Computer Networks, 2012.

[3] M. Celenk, T. Conley, J. Graham and J. Willis, Anomaly Prediction in Network Traffic using Adaptive Wiener Filtering and ARMA Modeling, SMC 2008. IEEE International Conference on Systems, Man and Cybernetics, 3548-3553.

[4] G. Frey, M. Manera, A. Markandya and E. Scarpa, Econometric models for oil price forecasting: A critical survey, CESifo Forum 1/2009.

[5] D. Kwon, J. W. Hong and H. Ju, DDos Attack Forecasting System Architecture using Honeynet, dpnm.postech.ac.kr/papers/.../12/dwkwon/APNOMS2012-

[6] Y. Hideshima and H. Koike , "STARMINE: A visualization system for cyber-attacks," 2006 Asian-Pacific Symposium on Information Visualization, pp. 131-138, February 2006.

[7] C. Ishida, Y. Arakawa, I. Sasase, and K. Takemori, "Forecast techniques for predicting increase or decrease of attacks using bayesian inference," 2005 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 450-453, August 2005.

[8] Y. Zhang, X. Tan, and H. Xi, "A novel approach to network security situation awareness based on multi-perspective analysis," 2007 International Conference on Computational Intelligence and Security, pp. 768-772, December 2007.

[9] D.-H. Kim, T. Lee, S.-O.D. Jung, H.-J. Lee, and H.P. In, "Cyber threat trend analysis model using HMM," 2007 International Symposium on Information Assurance and Security, pp. 177-182, August 2007.

[10] S.-H. Kim, S.-J. Shin, H.-W. Kim, K.-H. Kwon, and Y.-G. Han, "Hybrid intrusion forecasting framework for early warning system," IEICE TRANS. INF. and SYST., vol. E91-D, no. 5, pp. 1234-1241, May 2008.

[11] K. Takemori, Y. Miyake, C. Ishida, and I. Sasase, "A SOC framework for ISP federation and attack forecast by learning propagation patterns," 2007 IEEE Intelligence and Security Informatics, pp. 172-179, May 2007.

[12] S.S.S. Sindhu, S. Geetha, S.S. Sivanath, and A. Kannan, "A neurogenetic ensemble short term forecasting framework for anomaly intrusion prediction," 2006 International Conference on Advanced Computing and Communications, pp. 187-190, December 2006.

[13] S. Nanda and N. Deo, "A highly scalable model for network attack identification and path prediction," 2007 IEEE Southeast Conference, pp. 663-668, March 2007.

[14] S.E. Schechter, "Toward econometric models of the security risk from remote attacks," IEEE Security and Privacy, vol. 3, issue 1, pp. 40-44, January-February 2005.

[15] P. J. Brockwell and R. A. Davis, "Time Series: Theory and Methods," New York: Springer-Verlag, 1991.

[16] P. J. Brockwell and R. A. Davis, "Introduction to Time Series and Forecasting," 2nd Edition. New York: Springer, 2002.

[17] M. S. Peiris, "Improving the Quality of Forecasting using Generalized AR Models: An Application to Statistical Quality Control," 2003, Statistical Methods, vol. 5, issue 2, pp. 156-171, 2003.

[18] M. S. Peiris, D. Allen anf A. Thavaneswaran, "An Introduction to Generalized Moving Average Models and Applications," Journal of Applied Statistical Science, vol. 13, issue 3, pp. 251-267, 2004.

[19] T. R. Pillai, M. Shitan and M. S. Peiris, "Time Series Properties of the Class of First Order Autoregressive Processes with Generalized Moving Average Errors,"Journal of Statistics: Advances in Theory and Applications, vol. 2, issue 1, pp. 71-92, 2009.

[20] M. Shitan and M. S. Peiris, "Time series Properties of the class of generalized first-order autoregressive processes with moving average errors," Communication in Statistics-Theory and Methods, vol. 40, pp. 2259-2275, 2011.

[21] T. R. Pillai, M. Shitan and M. S. Peiris, "Some Properties of the Generalized Autoregressive Moving Average (GARMA(1, 1; δ 1, δ 2)) model," Communication and Statistics-Theory and Methods vol. 4, issue 41, pp. 699-716, 2012.

[22] R. A. Fisher, "A Mathematical Examination of the methods determining accuracy of an observation by the mean error and by the mean square error," Monthly Notices of the Royal Astronomical Society 80, vol. 1, pp. 758-770, CP12 in Bennett, 1971.

International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Improving Intrusion Detection using Genetic Linear Discriminant Analysis

Azween Abdullah *[1], Cai Long Zheng [2]

Received 05th September 2014, Accepted 10th October 2014

Abstract: The objective of this research is to propose an efficient soft computing approach with high detection rates and low false alarms while maintaining low cost and shorter detection time for intrusion detection. Our results were promising as they showed the new proposed system, hybrid feature selection approach of Linear Discriminant Analysis and Genetic Algorithm (GA) called Genetic Linear Discriminant Analysis (GLDA) and Support Vector Machines (SVM) Kernels as classifiers with different combinations of NSL-KDD data sets is an improved and effective solution for intrusion detection system (IDS).

Keywords: *IDS, Features selection, Features transformation, NSL-KDD, GLDA, SVM Kernels*

1. Introduction

An intrusion is defined as anything which compromises confidentiality, availability or integrity [14]. User authentications, avoiding programming mistakes, firewalls and data encryptions are first-level defences against cyber-attacks and intrusions. Intrusion prevention is totally dependent on their detection, and detection is a key part of any security tool such as Adaptive Security Alliance, Intrusion Detection System, Intrusion Prevention System, firewalls and checkpoints.

The selection of a suitable data set is the backbone of any efficient intrusion detection approach. The performance of any intrusion detection system (IDS) also depends on the efficiency and accuracy of the data set. If the training data set is precise with optimal contents and rich features, the efficiency of the training as well as test system will be improved. There are many standard pre-built simulated data sets like Darpa's KDD Cup 98, 99, Six UCI db and NSL-KDD etc. KDD-Cup 99 is most widely used as a benchmark data set for training and testing of IDSs. KDD-CUP 99 is built based on the data captured in DARPA'98 which has been criticized by [5], mainly because of the characteristics of the synthetic data. One of the most important deficiencies in the KDD data set is the huge number of redundant records. On analyzing KDD training and test sets, the author found that about 78% and 75% of the records were duplicated in the training and test sets, respectively, which caused the learning algorithms to be biased towards the frequent records, thus preventing them from learning infrequent records which are usually more harmful to networks such as U2R and R2L attacks.

Due to the drawbacks of KDD-Cup 99 which highly affects the performance of evaluated systems and results in a very poor evaluation of intrusion detection approaches, an advanced form of KDD-Cup was proposed, namely NSL-KDD which consists of selected records of the complete KDD data set. Important drawbacks of KDD-Cup are fixed in NSL-KDD data set. Although there are many techniques for intrusion detection such as computational intelligence, soft computing, data mining, this research focuses on using an ensemble of soft computing approaches to improve detection rate and accuracy.

The rest of the paper is organized as follows. In Section 2, related work of IDS is discussed briefly. In Section 3, the proposed model with different phases is discussed and analyzed in detail. Conclusion and future work is mentioned briefly in Section 4.

2. Related Work

Reference [19] adopted the NSL-KDD data set in their research work on feature extraction for intrusion detection using the Linear Discriminant Analysis (LDA) approach. LDA is extensively used as feature dimension reduction approach to find out an optimal transformation that minimizes the within-class scatter and to maximize the between-class distance. Back Propagation Algorithm (BPA) was used to classify attacks into five classes. The Artifical Neural Network (ANN) approach was adopted to compare the performance of the proposed method. In their experiments, 41 features were reduced to only 4 features new space by reducing 97% of the input data and about 94% of the training time as well as same percentage of accuracy in new attack detection [19].

The hybrid approach for feature reduction was adopted by [6] as PCA was not suitable for nonlinear data set as well as for large data set. In their work, the authors preferred Generalized Discriminant Analysis (GDA) over PCA for feature selection. Besides reducing the number of input features, GDA also reduces the training time for classifiers by selecting the most discriminant features. It also increases the accuracy of classification. The anomaly detection approach was used to differentiate between normal data based on normal behavior and attack or intrusive data based on its attack behavior. The Self-Organizing Map (SOM) approach and C4.5 decision tree techniques were applied for classification of reduced feature space. The KDD-Cup 99 data set was applied in this research and 41 features were reduced to 12 features space by GDA. The experimental results showed that GDA outperformed

[1] *School of Computing and IT, Taylors University, Subang Jaya, Selangor, Malaysia*
[2] *Unitar International University, Petaling Jaya, Selangor, Malaysia*
** Corresponding Author: Email: azween.abdullah @taylors.edu.my*
Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

PCA especially for large scale data set by providing a better detection rate as well as reduced training and testing time. Moreover, the C4.5 classifier outperformed SOM for all the attack classes.

An integrated intrusion detection system by [21] was proposed to model and implement an efficient system to reduce false alarms and to increase detection rate. The authors extracted the most important segments from the whole features of data set using Information Gain. To achieve a high detection rate of attacks, the authors introduced a high level of generality while deploying the subset of extracted or selected feature space. Genetic Algorithm (GA) and Radial Basis Functions (RBF) were used to classify known and unknown attacks. GA is based on the principles of genetics and natural selection and has a big potential in the domain of intrusion detection. Each individual in GA is called chromosome. Three basic genetic operations, Selection, Cross over and Mutation are applied sequentially to every individual during each generation. RBF networks are effectively used to prevent from overfitting. The proposed system was deployed using Java and KDD data sets. KDD consists of 41 features out of which 38 were numeric and 3 were symbolic. The performance of proposed system was compared with Hoffman GA rules for intrusion detection. The training time was reduced considerably as only nine features were considered. However due to the random usage of cross over and mutation operations, detection rate was not good for some runs [21].

An efficient intrusion detection system was proposed by [7] using feature subset selection based on MLP. The authors used Principal Component Analysis (PCA) and GA for preprocessing and MLP for feature classification using the KDD-cup data set. LDA outperformed PCA. PCA is not suitable for large data sets [4], hence their work was limited for small-sized data sets and results were not realistic against actual network traffic as there were obvious deficiencies in the KDD-Cup data set.

Reference [8] used PCA for feature reduction and Naive Bayes algorithm for classification to generate a smaller false positive alarms ratio and to increase the detection rate efficiently. The Naive Bayes classifiers used the probabilistic approach to determine attack probability while considering conditional dependency. The 41 features of the KDD 99 data set were reduced to 14 features and 12 major features that had greater Eigen values were identified by PCA. This new feature set contained the explanation for about 80% of the data variability while 98% of the inconsistency can be attributed to 24 features which can be considered as quite a sufficient value [22]. A brief comparison of the different approaches with their results is shown in Table 1.

Table 1. Comparative analysis of existing approaches

Author [Year]	Data set	Architecture	Accuracy
Osareh, and Bita[2008]	KDD	SVM	83%
S. M. Aqil [2010]	KDD	PCA,Naïve Bayes	N/A
Rupali datti [2010]	NSL-KDD	LDA1, ANN, BPA	96.5%
Lakhina et. al. [2010]	KDD-Cup	PCANNA	80.4%
Ahmad et al. [2011]	KDD-Cup	PCA, GA, MLP	99%
Shailendra Singh and Sanjay Silakari [2011]	KDD-Cup	GDA, SOM, C4.5	98%
Rita Ranjani Singh and Neetesh Gupta [2011]	NSL-DD	SOM	95%

3. Proposed Architecture

There were different interdependent phases in the proposed architecture for an efficient IDS. NSL-KDD was selected during the selection of a suitable data set phase. The LDA approach was used for feature transformation and GA for optimum feature subset selection. In the third phase, SVM Kernels was used as the classification approach in this research. After classification, the system was trained and tested according to the standard rules. Figure 1 shows the block diagram for the proposed system.

3.1. Selection of Suitable Data Set

KDD-Cup is the widely used data set for training and testing of IDSs. There are a total of 41 features which are classified into Basic, Content and Traffic features. As a result, some of inherited issues also exist in KDD-Cup like redundancy of similar records and complexity level of data behavior. NSL-KDD is an advanced version of KDD-Cup data set and does not suffer from the shortcomings found in KDD-Cup. The following presents the unique features that helped us pick NSL-KDD over KDD-Cup.

- No redundancy of records
- No duplication records in test data
- Less complexity level
- Affordable records in training and test sets

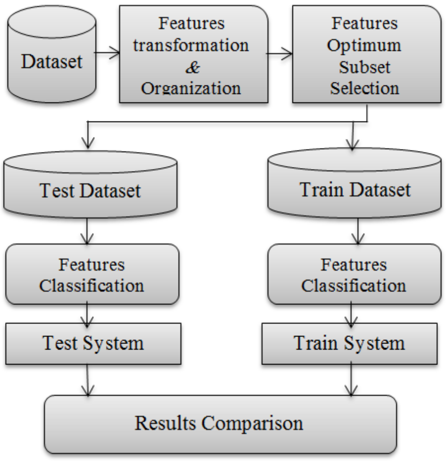


Figure 1. Block diagram of proposed system for IDS

The NSL-KDD features can be classified into the following three groups as shown in Figure 2.

1) Basic features: This category encapsulates all the attributes that can be extracted from a TCP/IP connection. Most of these features lead to an implicit delay in detection.

2) Traffic features: This category includes features that are computed with respect to a window interval and is divided into "Same host" and "Same service" features.

3) Content features: Unlike most of the DoS and probing attacks, the R2L and U2R attacks don't have any intrusion frequent sequential patterns. This is because the DoS and probing attacks involve many connections to some host(s) in a very short period of time; however, the R2L and U2R attacks are embedded in the data portions of the packets, and normally involves only a single connection.

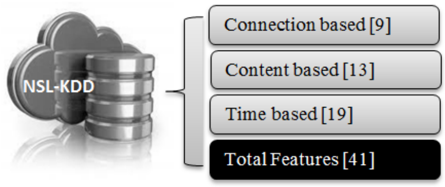


Figure 2. Categories of features in NSL-KDD data set

3.2. Preprocessing of Raw Data set

In most of the existing intrusion detection approaches, raw feature sets are given as input directly to classifiers which causes many problems. In some cases, features are transformed and subset of features is given as input to classifier. In this case, there are also some issues regarding the subset selection scenario. Some major issues in both the above mentioned approaches involve high false alarms, low detection rate and accuracy, loosing important information and many others. A detailed diagram that shows the related issues is shown in Figure 3.

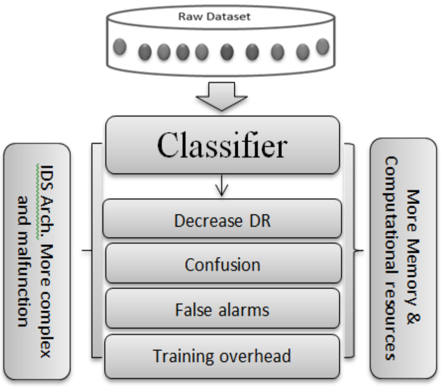

Raw Dataset
Classifier
Decrease DR
Confusion
False alarms
Training overhead
IDS Arch. More complex and malfunction
More Memory & Computational resources

Figure 3. Issues in existing approaches

Instead of directly inserting raw data set to selected classifiers, the raw data set is preprocessed in different ways to overcome different issues like training overhead, classifier confusion, false alarms and detection rate ratios. The preprocessing phase was divided into three sub phases as shown in Figure 4.

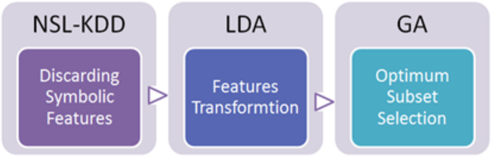


Figure 4. Flow chart of preprocessing steps

3.3. Discarding Symbolic Feature Vectors

There are three kinds of symbolic features (tcp, ftp_data and SF etc.) in the feature space of 41 features. As symbolic values are not of interest to our research, these three feature vectors were discarded to get the following new feature space.

$F(X_m) = X_1, X_2, X_3\ldots\ldots\ldots X_m \qquad$ *where* $m = 38$

3.4. Feature Transformation and Organization

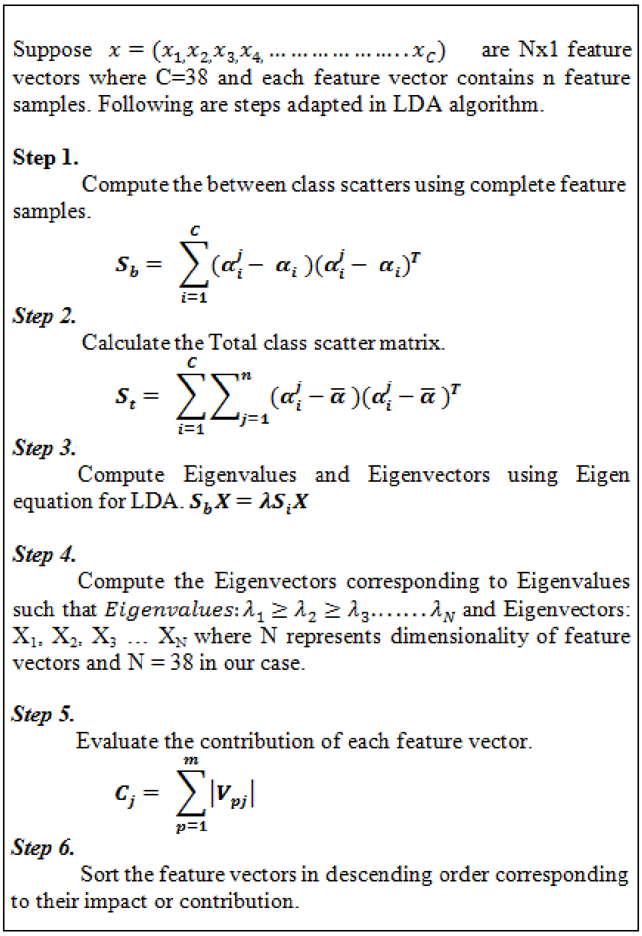
In most of the existing intrusion detection approaches, raw feature sets are given as direct input to classifiers which cause some of the following issues.

- Using raw data set directly for classifiers guzzles more memory space as well as computational resources during the training and testing phases of the system.
- Detection rate decreases in this case.
- The classifier may become confused and generate false alarms.
- Training overhead is increased due to the processing over each input feature even it is not important for the classifier.
- The architecture of IDS becomes more complex.

To avoid the above mentioned issues, the LDA approach was adopted to transform original numeric feature spaces into new linear feature spaces. LDA is a high-dimensional data analysis method and suitable for feature transformation to facilitate classification [9]. Its steps are shown in Figure 5. There has been a tendency to use the PCA approach for the feature subset selection or reduction in many different domains like face recognition, image compression as well as intrusion detection [10] but LDA has more benefits and is preferred over PCA due to the following reasons.

- LDA outperforms PCA in case of large data sets [4].
- LDA directly deals with both discrimination within-classes as well as between-classes while PCA does not have any concept of the between-classes structure [1].
- LDA preserves class discriminatory information as much as possible while performing dimensionality reduction [11].

The following are steps involved in feature transformation and organization.



Suppose $x = (x_1 x_2 x_3 x_4 \ldots\ldots\ldots\ldots\ldots\ldots\ldots x_C)$ are Nx1 feature vectors where C=38 and each feature vector contains n feature samples. Following are steps adapted in LDA algorithm.

Step 1.

Compute the between class scatters using complete feature samples.

$$S_b = \sum_{i=1}^{C} (\alpha_i^j - \alpha_i)(\alpha_i^j - \alpha_i)^T$$

Step 2.

Calculate the Total class scatter matrix.

$$S_t = \sum_{i=1}^{C}\sum\nolimits_{j=1}^{n} (\alpha_i^j - \bar{\alpha})(\alpha_i^j - \bar{\alpha})^T$$

Step 3.

Compute Eigenvalues and Eigenvectors using Eigen equation for LDA. $S_b X = \lambda S_t X$

Step 4.

Compute the Eigenvectors corresponding to Eigenvalues such that $Eigenvalues: \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots\ldots \lambda_N$ and Eigenvectors: $X_1, X_2, X_3 \ldots X_N$ where N represents dimensionality of feature vectors and N = 38 in our case.

Step 5.

Evaluate the contribution of each feature vector.

$$C_j = \sum_{p=1}^{m} |V_{pj}|$$

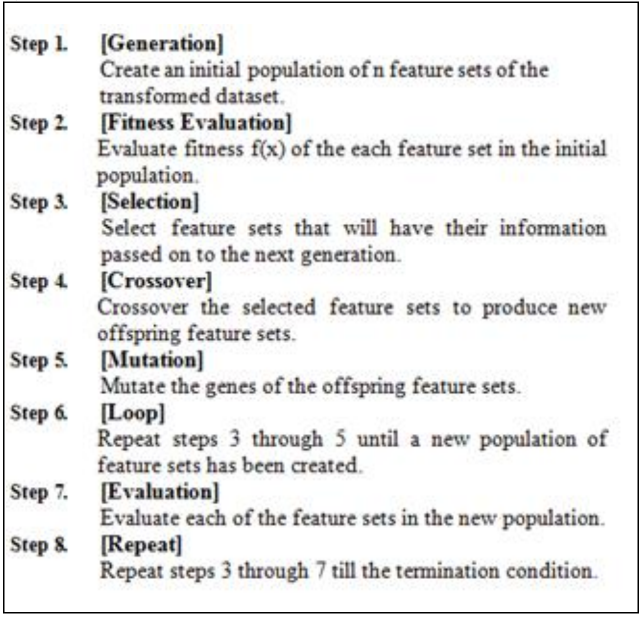
Step 6.

Sort the feature vectors in descending order corresponding to their impact or contribution.

Figure 5. LDA steps for feature transformation

3.5. Optimum Subset Selection

By using LDA for feature transformation, the data set was transformed into a new feature space called linear feature space. This new feature space may also be used as input to the classifier but the classifier becomes biased due to architecture complexity and training and testing efficiency decreases which in turn, increases memory consumption rate and computational cost. GA was applied to select optimal subset of linear features space (Figure 6).



Step 1. **[Generation]**
Create an initial population of n feature sets of the transformed dataset.

Step 2. **[Fitness Evaluation]**
Evaluate fitness f(x) of the each feature set in the initial population.

Step 3. **[Selection]**
Select feature sets that will have their information passed on to the next generation.

Step 4. **[Crossover]**
Crossover the selected feature sets to produce new offspring feature sets.

Step 5. **[Mutation]**
Mutate the genes of the offspring feature sets.

Step 6. **[Loop]**
Repeat steps 3 through 5 until a new population of feature sets has been created.

Step 7. **[Evaluation]**
Evaluate each of the feature sets in the new population.

Step 8. **[Repeat]**
Repeat steps 3 through 7 till the termination condition.

Figure 6. GA specific steps for features subset selection

3.6. Feature Classification

After the selection of the optimum feature subset, the classifier is designed to train and test the features using different Support Vector Machines (SVM) Kernels. The proposed approach was implemented with kernel functions by tuning different parameters including the cost parameter C and other kernel parameters. This was done by selecting parameters using 5x2 cross validation. An overview of the different SVM kernels is shown in Figure 7. The system was trained and tested with the given set of parameters to evaluate the best possible classifier performance on the selected data set.

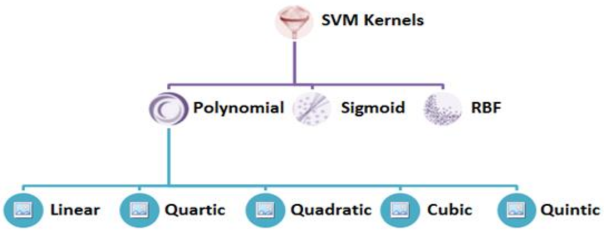
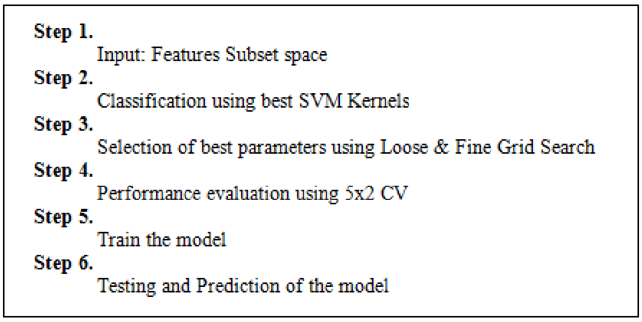


Figure 7. SVM Kernels categories

Figure 8 shows the different steps taken to classify the network traffic into normal or intrusive using SVM kernels.



Step 1.
Input: Features Subset space

Step 2.
Classification using best SVM Kernels

Step 3.
Selection of best parameters using Loose & Fine Grid Search

Step 4.
Performance evaluation using 5x2 CV

Step 5.
Train the model

Step 6.
Testing and Prediction of the model

Figure 8. Classification steps using SVM Kernels

4. Experimental Results

Data sets with 11, 15 and 21 feature vectors were selected for training as well as for testing experiments using the GA approach as the optimum subset from the complete data set of 41 features. Different tools including Net LDA, NeuroSolutions and Matlab were used for this purpose. Table 2 shows the 11 features selected using the GA approach.

Table 2. Optimum features subset from 41 features

No	*Feature Name*	*Type*
1	Duration	Continuous
2	Service	Discrete
3	Count	Continuous
4	dst_bytes	Continuous
5	logged_in	Discrete
6	srv_count	Continuous
7	rv_rerror_rate	Continuous
8	serror_rate	Continuous
9	srv_diff_host_rate	Continuous
10	dst_host_count	Continuous
11	Is_guest_login	Discrete

Network weights were adjusted during the training phase. Confusion matrices were used to verify the training process. The weights of the system were frozen after the training of the system was completed and the system performance was evaluated during the testing phase. The testing phase was divided into verification and generalization steps. The objective of verification was to calculate the learning efficiency of the trained system while the generalization step was used to measure the generalization ability of the trained system using another data set besides the training data set. We selected randomly 10,000 feature samples as the training data set from a total of 125,974 preprocessed linear feature samples while 20% of the training data was used as a cross validation data set. A separate data set of 5,000 was selected randomly from NSL-KDD preprocessed test data set of 22,545 connection records as shown in Figure 9.

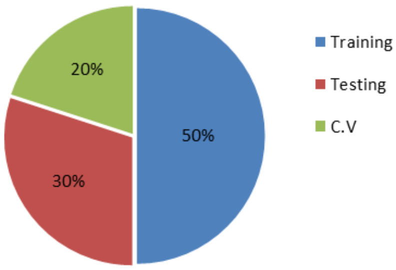


Figure 9. Data set distribution for training, testing & cross validation

We have used several parameters to evaluate the performance of the proposed system which include True Positive, True Negative, False Positive, False Negative, Accuracy rate, Detection rate, Sensitivity and Specificity.

1) Classification Accuracy = 100*(TP+TN) / (TP+FP+FN+TN)

2) Sensitivity: It is the measure of detecting normal patterns accurately.

Sensitivity = (100 * TP / TP + FN)

3) Specificity: It is the measure of detecting intrusive patterns accurately.

Specificity = (100 * TN / TN + FP)

Three different experiments were conducted using different SVM Kernels. Results in Table III reflect that when optimum subset of features is selected, time consumption rate is relatively reduced and accuracy ratio is increased. Since reduced feature space was given as input to the classifier, lesser resources were utilized due to minimum training and learning overheads, hence, computational cost was also minimized. Figure 10 depicts the performance using different subsets.

Table 3. Time & detection rate analysis

No.	Features	Not Selected	Time	Detection Rate
1	11	27	45 h	99.3 %
2	15	23	51 h	99 %
3	21	17	55 h	98.7 %

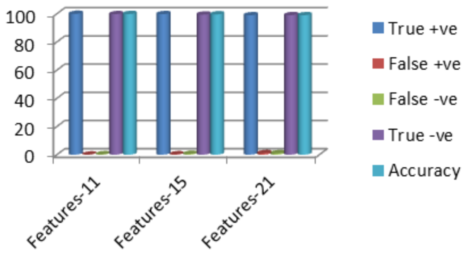


Figure 10. Performance measurements with different features space

The sensitivity and specificity results for different SVM kernels and data feature combinations are shown in Table IV. The graphical analysis for sensitivity and specificity are shown in Figures 11 and 12, respectively. Results in Table 4 clearly show that the RBF kernel performs best for all the recipes of features.

Table 4. Sensitivity & specificity analysis

Cases	*NSL-KDD 11 Features*		*NSL-KDD 15 Features*		*NSL-KDD 21 Features*	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
RBF Kernel	100	99.2	99.1	99.7	98	98.3
Linear Polynomial	99.7	99.3	99	99.3	99.1	98.7
Sigmoid Kernel	98.9	99	99.1	98.6	99.5	98
Quadratic Polynomial	97	97.4	100	95.7	99.1	97.7
Cubic Polynomial	98.4	95.5	90.1	97.1	100	98.1
Quartic Polynomial	99.9	98.7	100	97.5	99.1	98.1
Quintic Polynomial	99	98.5	93.9	98.1	98.3	96.4

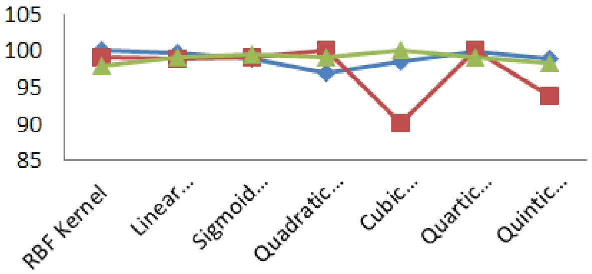


Figure 11. Sensitivity analysis of different feature spaces

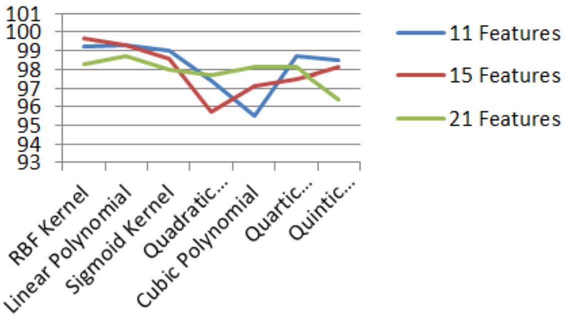


Figure 12. Specificity analysis of different feature spaces

The research results were compared with some existing approaches and are depicted in Figure 13.

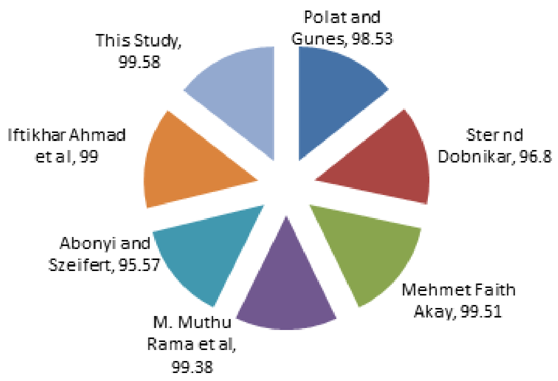


Figure 13. Comparison of new approach with existing approaches

5. Experimental Results

Feature transformation and selection is generally performed using single approach but in our work, the hybrid approach of LDA + GA named as GLDA was adopted to get better results. LDA is preferred over PCA as it outperforms PCA. The advanced form of KDD-Cup called NSL-KDD was used as standard data set. The prominent classification approach SVM with different kernels was used to classify network traffic into normal or intrusive. Our work shows that time consumption rate is relatively reduced whilst accuracy ratio as well as detection rate is increased due to optimum subsets. Since reduced feature space is used as classifier input, minimum resources are utilized and computational cost is minimized due to minimum training and learning overheads.

Our future plan is to design and develop an efficient intrusion detection system for multi class problems by selecting the optimal subset of features.

Acknowledgements

The project is funded by the Ministry of Education Malaysia under the Fundamental Research Grant Scheme 2013 titled: Predictive Analytic Theory Generation and Foundation for a Novel Bio-inspired Intrusion Prevention and Self-regeneration System for Cyber Defence (FRGS/2/2013/ICT02/TAYLOR /02/1).

References

[1] A. Martinez and A. Kak (2001). "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence," vol. 23, no. 2, pp. 228-233,.

[2] China Papers Online (2011). "Study on Application of Hybrid Soft-Computing Technique to Intrusion Detection".

[3] Adel Nadjaran Toosi and Mohsen Kahani (2007) "A new approach to intrusion detection based on an evolutionary soft computing model

using neuro-fuzzy classifiers," Department of Computer, Ferdowsi University of Mashhad, Iran.

[4] Kresimir Delac, Mislav Grgic and Sonja Grgic (2006). "Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set," University of Zagreb, FER, Unska 3/XII, Croatia.

[5] J. McHugh (2000) "Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection," ACM Transactions on Information and System Security.

[6] Shailendra Singh, Sanjay Silakari and Ravindra Patel (2011). An Efficient Feature Reduction Technique for Intrusion Detection System, IPCSIT, Vol. 3.

[7] Ahmad I, Abdullah AB, and Alghamdi (2011). "Intrusion detection using feature subset selction based on MLP," Scientific Research and Essays, Vol 6(34).

[8] S. M. Aqil, M. Sadiq Ali Khan and Jawed Naeem (2010). Efficient Probabilistic Classification Methods for NIDS, IJCSIS, Vol. 8, No. 8, November.

[9] P. Belhumeur, J. Hespanha, and D. Kriegman (1996). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, Proc Fourth Eur Conf Computer Vision, Vol. 1, 1418, pp. 45–58.

[10] M. Turk and A. Pentland (1991). "Eigenfaces for recognition," J Cogn Neurosci 3, 71–86.

[11] K. Baek, B. Draper, J.R. Beveridge and K. She (2002). "PCA vs. ICA: A Comparison on the FERET Data Set," Proc. of the Fourth International Conference on Computer Vision, Pattern Recognition and Image Processing, Durham, NC, USA, 8-14, pp. 824-827.

[12] Chittur A. (2006). "Model Generation for an Intrusion Detection System Using Genetic Algorithms," High school Honors Thesis.

[13] Acohido B. (2009). "Hackers breach heartland payment credit card system", 11 March.

[14] Abraham A. and Jain R. (2008). "Soft computing models for network intrusion detection systems, 15 May.

[15] Sandhya P., Ajith A., Crina G. and Thomas J. (2005). "Modeling intrusion detection system using hybrid intelligent systems. Journal of Network and Computer Applications,".

[16] Ilgun K, Kemmerer R.A. and Porras P.A. (1995). "State transition analysis: a rule-based intrusion detection approach," IEEE Trans Software Eng 21(3):181–199.

[17] Zadeh LA. (1994). "History; bisc during 90's,".

[18] Zadeh L.A. (1998). "Roles of soft computing and fuzzy logic in the conception," design and deployment of information/intelligent systems. In: Kaynak O, Zadeh LA, Turksen B, Rudas IJ (eds) Computational intelligence: soft computing and fuzzy-neuro integration with applications, vol 162. Springer, New York.

[19] Rupali D. (2010). "Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 1072-1078.

[20] Liao Y. and Vemuri V. R. (2002). "Use of k-nearest neighbor classifier for intrusion detection," Computer Security, vol. 21, no. 5, pp. 439-448.

[21] Selvakani Kandeeban S. and Rengan S. R. (2010). "Integrated Intrusion Detection System Using Soft Computing", I. J. Network Security 10(2): 87-92. 2008.

[22] M.Sadiq Ali Khan (2012). "Application of Statistical Process Control Methods for IDS," International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November.

[23] Chittur A. (2006). "Model Generation for an Intrusion Detection System Using Genetic Algorithms," High school Honors Thesis, accessed in.