Effectiveness of Artificial Neural Networks in Forecasting Failure Risk for Pre-Medical Students

Jawaher K. Alenezi, Mohammed M. Awny and Maged M. M. Fahmy
Arabian Gulf University
Kingdom of Bahrain

1. Introduction

The registration department in universities usually regulates a set of rules in order to accept new students. These rules are set to select applicants who have the abilities and skills to pursue and succeed in their academic career in a particular field of studies. Acceptance to the College of Medicine and Medical Sciences at the Arabian Gulf University is no exception. The admission body of the college firstly makes sure that qualifications and other particulars of the individual students fulfil the requirements of the college admission according to its rules and conditions. Secondly, they arrange acceptance tests that result in accepting, or rejecting applicants. However, it has been noticed throughout the years that the decisions based solely on the results of these acceptance tests, including an additional personal interview, are not sufficient. A number of students still fail in the first year of education (premedical stage). This creates many problems both to the University and to the students; whether administrative, financial and/or psychological. The Artificial Neural Networks techniques can play a role in this process. However, it should use the same set of entrance rules but with different weightings. This research aims at modelling an effective and intelligent tool that can help the decision makers to have better judgment on possible ability of each individual student to succeed in the premedical stage and hence in his/her subsequent studies in the medical school.

2. Artificial Neural Networks approach

Artificial Neural Networks are sophisticated modeling techniques; capable of modeling extremely complex functions. Cascade Correlation Network (CCN) architecture of Artificial Neural Networks is used to deal with this research problem. It is a supervised learning algorithm developed by Fahlman and Lebiere [1]. It is an attempt to overcome certain drawbacks and limitations of popular Back Propagation Learning Algorithm developed by Rumelhart et al. [2]. It is trained using the Quick Propagation Algorithm (QPA) which is the enhanced version of Back Propagation (BP). It is a convenient approach to solve problems since it is an ontogenic neural network that generates its own topology during training. It overcomes the limitation of BP which is slow at learning from examples. A BP network may require many thousands of *epochs* to learn the desired behaviour from examples. An *epoch* is defined as one pass through the entire set of training examples. Previous studies involve some researches close to this research. Predicting MBA student performance by evaluating the ability of three different models, namely, logistic regression, probability analysis, and neural networks; is reported by Fahlam and Lebiere [3]. The result was that the neural network model performs better than statistical models. Back-Propagation Neural Network was used in selecting surgical residents via the National Residency Matching Program applied to medical students interested in surgery in their fourth year of study. It is used to facilitate the surgical residents' selection process using 36 variables. The study showed that neural networks are capable of producing significantly better results than traditional statistical approaches [4]. Many other studies (Entwistle, 1988; Ardila, 2001; Busato et al., 1999, Furnham et al., 1999) were undertaken in order to try to explain the academic performance or to predict the success or the failure of students; they highlighted a series of explanatory factors associated to the student [5]-[8].

3. Implementation of ANN in forecasting the failure risk of applicants to Medical College

Premedical students' data including their grades at the end of first academic years were collected from both registration department and Medical College Acceptance Committee. Data for the academic years: 2003/2004, 2004/2005, and 2005/2006 were available at the beginning of this research work. Students' academic results of years 2003/2004 and 2004/2005 are used as training/ validation sets. While those of years 2004/2005 and 2005/2006 are used for forecasting (testing).

3.1 Neural inputs and output

The main variables affecting the process of selecting medical students were identified. The main input variables used and involved in the acceptance process are shown in Table 1. The notations column is used for abbreviation.

	Variable	Description	Notation
1	High School Grade	Average grade in high school certificate.	Sec. GPA
2	Science Grade	Average grade of science subjects at high school.	Sec. Sc.
3	English Grade	Student grade at English test carried out at AGU's English department.	ENG
4	MCAT	Medical College Acceptance Test in science provided by AGU medical staff.	MCAT
5	Interview	Measure stability and test the personal quality of applicant	Interview

Table 1. Input variables for ANN forecasting model

The following input variables were added in order to test if they have any significant effect on the forecasted results. The added variables are shown in Table 2, and the Output variable is shown in Table 3.

	Variable	Description	Notation
1	Gender	Student's sex	Gender
2	Age	Student's age	Age
3	Marital Status	Student's marital status	Marital
4	Graduate Year/ Batch	Current or previous year graduate (0 or 1)	Batch

Table 2. Added input variables for forecasting model

	Variable	Description	Notation
1	Accumulative GPA	Forecasted student's Grade (Out of 4.00 points)	Ac. GPA

Table 3. Output variable for ANN forecasting model

As there are many variables involved in the acceptance process, Cascade Correlation Network technique was chosen to build the forecasting model. Two types of networks have been built for each data set: One network with only five inputs as in Table 1. The other network with nine inputs is formed by including those in Table 2.

3.2 Forecasting model

Neural network software tool used in this research is "Forecaster XL' (v. 2.3)". It uses the Cascade Correlation supervised *learning algorithm* - trained using Quick-Propagation. It divides the rows of inputs x_i and their related output y_i into training and validation subsets. 83% of rows are used for training, while 17% of rows are used for validating the model. Network training is designed to adjust network weights for maximizing predictive ability and minimizing forecasting error. Validation subset is that part of data used to tune the network topology or network parameters other than weights. For example, it is used to define the number of hidden units to detect the moment when predictive ability of neural network started to deteriorate. Test subset is a part of data set used to test how well the trained neural network forecasts using new data. Test subset is used after the network is trained, and hence ready to forecast. This subset is not used during training and thus consists of a new data to the model. The regular Sigmoidal activation function is used. It is expressed in relation to the input variable "x" as follows: $1 / (1 + \exp(-x))^{-0.5}$, (Where x is the input).

3.3 Testing the performance of the trained model

After developing the model through training and validating, goodness of model fitting is examined statistically. Mean Square Error (MSE) is calculated after each training. The predictive model is then identified as a good one if the *MSE* is sufficiently small i.e. close to 0. The model with minimum *MSE* is identified as the best predictive model. A brief analysis of this study is presented in the following.

4. Results and discussions

The Artificial Neural Networks model is built based on students' historical data. The input variables are identified together with their related output variable (Input / Output variables pair). A typical example of the prepared data Inputs for a particular student are: Gender (F), Age (18), marital (S), Batch (0), Sec. GPA (93.90), Sec. Sc. (91.50), MCAT (60.00), ENG (95.00), and Interview (88.10). A typical related output is Ac. GPA (3.104). The proposed architecture for the Cascade Correlation Network forecasting model is illustrated in Figure 1.

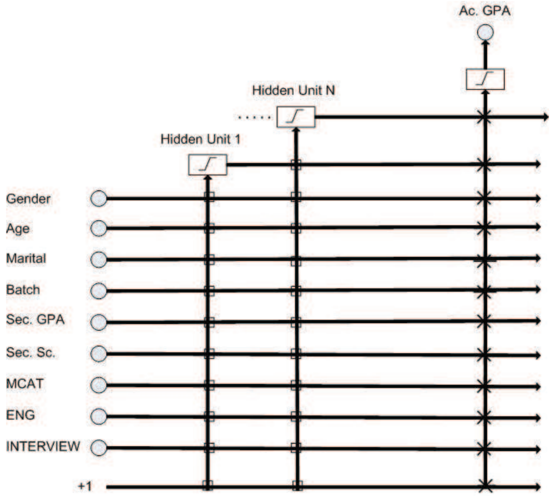


Fig. 1. The architecture of Cascade correlation network to forecast academic performance

Data reprocessing is done by analyzing the input and output variables. Data was cleaned by removing any column of data that was identified as unsuitable for neural network (e.g. that contains text or too many missing values or repeated values). Numeric columns are marked with *numeric* mark. Categorical columns are marked as *categorical*. Table 4 shows input and output variables used in this research work as recognized by the used software.

INPUT/ OUTPUT Columns	Location	TYPE
Gender	INPUT	Category
Age	INPUT	Number
Martial	INPUT	Category
Batch	INPUT	Number
Sec. GPA	INPUT	Number
Sec. Sc.	INPUT	Number
MCAT	INPUT	Number
ENG	INPUT	Number
Interview	INPUT	Number
Ac. GPA	OUTPUT	Number

Table 4. Types of Input/ Output columns

Table 5 shows an analysis of the data for training process. The data has been classified into training set and validation set. Some columns (input variables) were excluded because it proved that they have no effect on the training process. Specifically, namely; marital and Batch input variables. 'Marital' was excluded since all the students were 'single' and hence it has no contribution to the network output. 'Batch' variable was also excluded since all students enrolled are fresh graduates, and hence the variable is not an informative one.

Analysis parameters	2003/2004 (9) inputs	2003/2004 (5) inputs	2004/2005 (9) inputs	2004/2005 (5) inputs
No. of students Tested	76	76	77	77
Training Set (83%) of total rows	63	63	64	64
Validation Set (17%) of total rows	13	13	13	13
Variables Excluded/ Uninformative or too noisy	Marital Batch	-	Batch	-

Table 5. Analyzing data for training

Different network models have been built using each of the four academic year data as training data. The target is to study the importance of the input to the network being built. The number of training iterations made by each network in order to achieve the minimum possible error of the output result is shown in Table 6. The minimum error reached within few seconds in the software. This is when the percentage of error becomes very low and generalization loss stays within acceptable limit. A performance report then gives the average errors as well as tolerance and number of *Good* and *Bad* forecasts for training and test sets. The quality of the model could be determined by analyzing R-Squared and the Correlation values.

Academic year (no. of variables)	2003/2004 (9)	2003/2004 (5)	2004/2005 (9)	2004/2005 (5)
Number of training iterations	5070	2260	5029	5009
Training stop reason	Generalization loss & Maximum no. of iterations achieved.	Error reduction became too low.	Generalization loss & Maximum no. of iterations achieved.	Generalization loss & Maximum no. of iterations achieved.

Table 6. Neural networks Information

The progressive improvement of network Mean Square Error (MSE) during training is shown in Figures 2 and 3. Figure 2 shows the training MSE using data of the academic year 2003/2004 with 9 input variables, the lowest MSE reached at the early stages of training. Figure 3 shows the training MSE using data of academic year 2003/2004 with 5 input variables. The network restores the structure that resulted in the lowest error (this is at iteration number 1956).

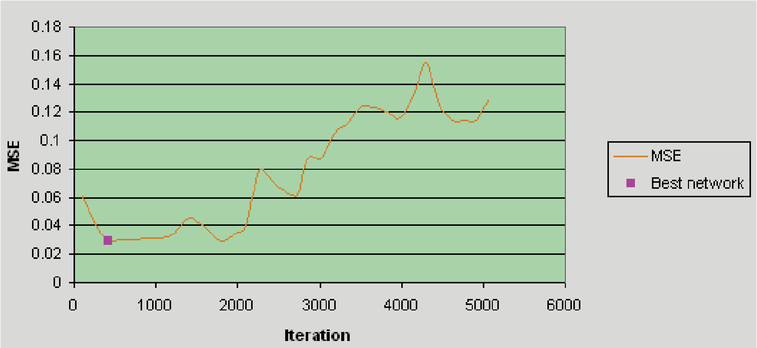


Fig. 2. Training errors (local minima errors) for academic year 2003/2004 with 9 inputs

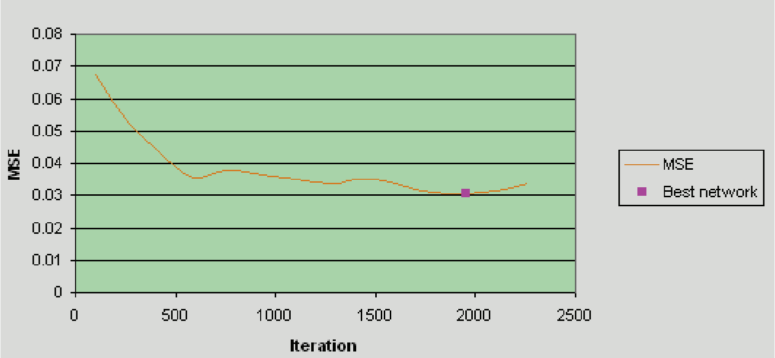


Fig. 3. Training errors (local minima errors) for academic year 2003/2004 with 5 inputs

The resulted network structure is presented in Table 7 for each academic year. Since input data are numerical and categorical; numerical inputs would be recognized as one input in the training set, while categorical inputs would be recognized as two inputs. The table shows the number of hidden units in each network structure. The categorical inputs such as (male/ female) is represented by {-0.4, 0.4}, while the forecasted output (target) is the GPA. Hidden units vary when using a different academic year as training set in order to achieve the lowest error.

(no. of variables)	2003/2004 (9)	2003/2004 (5)	2004/2005 (9)	2004/2005 (5)
Number of inputs	8	5	10	5
Number of hidden units	2	11	5	12
Number of outputs	1	1	1	1

Table 7. Neural networks Information

In academic year 2003/2004 with 9 inputs, the input variables Marital and Batch were excluded as discussed before. Therefore 7 inputs were left. And since Gender is categorical, it had been recognized as 2 inputs and may take 'Male' or 'Female' values. Because of that, the final number of input variables would be 8. In academic year 2004/2005 with 9 inputs, the input variable Batch was excluded, hence, the final number of input variables was 10. In Table 7, the number of hidden units varies among the different networks to achieve the lowest error for each network. Inputs importance (Sensitivity Analysis) is shown in Table 8. Table 9 shows the obtained results in each trained network.

INPUT (no. of variables)	2003/2004 (9)	2003/2004 (5)	2004/2005 (9)	2004/2005 (5)	2005/2006 (9)	2005/2006 (5)
Gender	2.195	Excluded	5.724	Excluded	0.374	Excluded
Age	4.293	Excluded	1.435	Excluded	6.406	Excluded
Martial	0	Excluded	0.175	Excluded	0	Excluded
Batch	0	Excluded	0	Excluded	2.956	Excluded
Sec. GPA	63.070	67.429	8.531	17.121	51.137	42.584
Sec. Sc.	1.207	4.168	52.540	65.124	16.407	10.052
MCAT	16.171	13.492	12.629	3.972	10.658	13.910
ENG	11.913	13.744	13.087	9.859	5.875	21.747
Interview	1.151	1.167	5.879	3.924	6.188	11.706

Table 8. Input Importance percentages for trained networks

The previously trained networks are used to forecast the academic results (Ac. GPA) for a different academic year. The trained network using data of academic year 2003/2004 was used to forecast Ac. GPA for academic year 2004/2005, and then to forecast students' Ac. GPA for academic 2005/2006. The trained network using data of academic year 2004/2005 was then used to forecast the GPA for academic year 2005/2006. A third network was trained using the combined data of both academic years 2003/2004 and 2004/2005, and then used to forecast the Actual GPA for academic year 2005/2006. As seen from Table 8, the Secondary GPA (Sec. GPA) is the most important input in both mentioned years. When comparing the forecasted and actual grades of year 2005/2006, the lowest MSE resulted when trained network using grades of 2003/2004 (as in Table 9 below).

	Forecasted sets (test sets)		
Training sets	2003/2004	2004/2005	2005/2006
2003/2004	-	0.526	0.430
2004/2005	-	-	0.453
2003/2004 and 2004/2005	-	-	0.495

Table 9. Mean Square Error (MSE) for forecasting Ac. GPA for the following academic year

5. Conclusion

The target of this research work is to study the effectiveness of Artificial Neural Networks in forecasting failure risk for pre-medical students at the Arabian Gulf University. The model structure for forecasting uses the Cascade Correlation Networks technique, and trained using the Quick Propagation algorithm. Different models with different input variables were built. The data of the academic years 2003/2004, 2004/2005, and 2005/2006 were available. The best accuracy was achieved with the model that had five input variables and trained using the academic data of the year 2003/2004 data to forecast GPA for academic year 2005/2006. MSE is 0.430, and was enhanced to 0.354 when increasing number of training iterations. Statistical analysis is carried out to ensure the validity of the forecasted GPA. Accepted MSE in forecasting the results of the students in the academic year 2005/2006 were reached using trained network with data of academic year 2003/2004. In this case, there is no significant mean differences for those academic years forecasted results with $p<0.000$. ***This proved that ANN is a valid tool for forecasting the results of the students.*** Good result obtained because both data of academic years 2003/2004 and 2005/2006 had the same input importance, therefore there were similarity in the characteristics of their data. It has also been seen that Secondary GPA has the highest influence on the GPA for the premedical students in the academic years 2003/2004 and 2005/2006. A future research work has been suggested to take into consideration added input parameters such as psychological and physiological issues.

6. References

Fahlman, S. E. and Lebiere, C. “The cascade-correlation learning architecture” *Advances in Neural information Processing Systems 2*, D. S. Touretzky, Ed. Morgan Kaufmann Publishers, San Francisco, CA, 524-532, 1990.

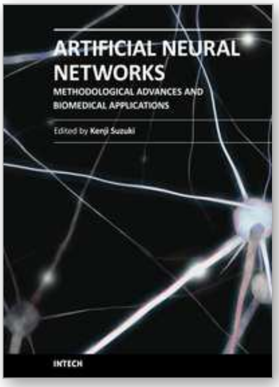
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. “Learning Internal Representations by Error Propagation”, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, 1986.

Naik B. and Ragothaman S., “Using neural network to predict MBA student success”, College Student Journal, 38(1):143-149, 2004.

Aggarwal A. K., Travers S. D., and Scott-Conner C. E. H. “Selection of surgical residents: A neural network approach.” Cybernetics and Systems 31:417-430, 2000.

Entwistle N. , “Motivational factors in students' approaches to learning”, In R.R. Schmeck (Ed.), Learning strategies and learning styles. 21-51. New York: Plenum, 1988.

Ardila A., ‘‘Predictors of university academic performance in Colombia’’, International Journal of Educational Research, vol. 35, pp. 411-417, 2001.


ARTIFICIAL NEURAL NETWORKS
METHODOLOGICAL ADVANCES AND BIOMEDICAL APPLICATIONS
Edited by Kenji Suzuki
INTECH

Artificial Neural Networks - Methodological Advances and Biomedical Applications

Edited by Prof. Kenji Suzuki

ISBN 978-953-307-243-2
Hard cover, 362 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

Artificial neural networks may probably be the single most successful technology in the last two decades which has been widely used in a large variety of applications in various areas. The purpose of this book is to provide recent advances of artificial neural networks in biomedical applications. The book begins with fundamentals of artificial neural networks, which cover an introduction, design, and optimization. Advanced architectures for biomedical applications, which offer improved performance and desirable properties, follow. Parts continue with biological applications such as gene, plant biology, and stem cell, medical applications such as skin diseases, sclerosis, anesthesia, and physiotherapy, and clinical and other applications such as clinical outcome, telecare, and pre-med student failure prediction. Thus, this book will be a fundamental source of recent advances and applications of artificial neural networks in biomedical areas. The target audience includes professors and students in engineering and medical schools, researchers and engineers in biomedical industries, medical doctors, and healthcare professionals.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jawaher K. Alenezi, Mohammed M. Awny and Maged M. M. Fahmy (2011). Effectiveness of Artificial Neural Networks in Forecasting Failure Risk for Pre-Medical Students, Artificial Neural Networks - Methodological Advances and Biomedical Applications, Prof. Kenji Suzuki (Ed.), ISBN: 978-953-307-243-2, InTech, Available from: http://www.intechopen.com/books/artificial-neural-networks-methodological-advances-and-biomedical-applications/effectiveness-of-artificial-neural-networks-in-forecasting-failure-risk-for-pre-medical-students

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821