DEALING THE NONLINEARITY ASSOCIATED WITH THE DATA USING ARTIFICIAL NEURAL NETWORKS

Iuliana IATAN

Technical University of Civil Engineering Bucharest, Romania (iuliafi@yahoo.com)

DOI: 10.19062/1842-9238.2017.15.2.2

Abstract: *The Artificial Neural Networks (ANNs) are well-suited for a very broad class of nonlinear approximations and mappings. The ANN with nonlinear activation functions are more effective than linear regression models in dealing with nonlinear relationships. We are trying to find out how relevant is to use a Fuzzy Neural Network for prediction because it handles well the nonlinearity associated with the data.*

Keywords: *nonlinear, prediction, fuzzy neural network, personality*

1. INTRODUCTION

In this work we propose a specific neural network for predicting personality, by special type of Fuzzy Gaussian Neural Network (FGNN) understanding that it has so special the connections (between the second and third layers) and the operations with the nodes, too.

We shall propose to apply the FGNN for predicting a users' Big Five personality traits (the five factor model of personality) from the public information they share on Facebook. The Big Five traits are characterized by the following: *Neuroticism*, *Extraversion*, *Openness*, *Agreeableness*, and *Conscientiousness*.

To emphasize the performances of our proposed approach for predicting personality we have compared it both with a neural method of regression (like Multilayer Perceptron=MP) and with a non-neural approach Multiple Linear Regression Model (MLRM). The comparison of FGNN and respectively MP versus MLRM marks both the competition nonlinear over linear and of neural over statistical, too.

To test the performance of the neuro- fuzzy prediction achieved based on FGNN we shall use the Normalized Root Mean Square Error (NRMSE). According with the NRMSE criterion, we have achieved that the prediction with FGNN is better than with others two methods both over the training lot and over the test lot, too.

2. RELATED WORK

One distinguish some methods used in prediction with Social Media [12]: Bayes classifier, *K*-nearest neighbor classifier, Artificial Neural Networks, decision trees, model based prediction.

In [3], Golbeck made a Pearson correlation analysis between subjects' personality scores and each of the features obtained from analyzing their tweets and public account data. There are a number of significant correlations here, however none of them are strong enough to directly predict any personality trait. He described later in [4] the results of predicting personality traits through MLRM.

In the case of the MLRM applied in [4] for predicting personality, the optimal parameters were computed using the correlations between each profile feature and personality factor.

More recently, [10] studied the relationship between sociometric popularity (number of Facebook contacts) and personality traits on a far larger number of subjects.

The paper from [9] develops a fuzzy neural network approach to financial engineering; this model was successfully applied to the prediction of daily exchange rates (US Dollar-Romanian Lei). In this work we extend the application domain of fuzzy neural networks, viz. in the field of text mining, to predict personality traits. This FGNN having *M* output neurons is unlike the exchange rate FGNN [9], which uses a single neuron in the last layer to estimate the current exchange rate based on the previous m daily exchange rates.

We are trying to find out how relevant is to use the FGNN for predicting personality because it handles well the nonlinearity associated with the data.

3. BASELINES

The Artificial Neural Networks (ANNs) are well-suited for a very broad class of nonlinear approximations and mappings. The ANN with nonlinear activation functions are more effective than linear regression models in dealing with nonlinear relationships.

A feed-forward neural network is a nonparametric statistical model for extracting nonlinear relations in the data, namely it is a useful statistical tool for nonparametric regression.

A feed-forward neural network with an specific activation function is identical to a linear regression model:

- the input neurons are equivalent to independent variables or regressors;
- the output neuron is the dependent variable;
- the various weights of the network are equivalent to the estimated coefficients of a regression model.

Some advanced neural network techniques are related to more complex statistical methods such as:

1) kernel discriminant analysis
2) *k*-means cluster analysis
3) Principal Component Analysis(PCA).

Some neural networks do not have any close parallel in statistics, such as:

1) Kohonen’s self-organizing maps
2) Fuzzy Gaussian Neural Network.

The regression and correlation are related as the both of them are designed to extract relations between some variables.

In the case of a linear regression model, of the first order, "the slope of the regression line is the correlation coefficient times the ratio of the standard deviation of *y* to that of *x*."

MLRM is a method used to model multiple linear relationship between a dependent variable and more independent variables.

A major problem with multiple regression consists in the large number of predictors that are available, although only a few of them are actually significant.

The advantage of MLRM is that it can be implemented very easy.

Example 1. We are interested [6] in exploring for a sample of 32 vehicles the relationship between: the number of gears of a vehicle, the overall length (in inches) and its fuel efficiency (measured in miles per gallon).

```
>> x1=[3 3 3 3 3 3 3 3 4 5 3 4 4 3 4 3 3 3 3 3 3 3 3 3 5 4 3 5 3 3 3 3 3];
>> x2=[200.3 199.6 196.7 199.9 194.1 184.5 179.3 179.3 155.7 165.2 195.4 ...
160.6 170.4 171.5 168.8 199.9 224.1 231 196.7 197.6 179.3 214.2 196 165.2 ...
176.4 228 171.5 215.3 215.5 216.1 209.3 185.2];
>> y=[18.9 17 20 18.25 20.07 11.2 22.12 21.47 34.7 30.4 16.5 36.5 21.5 19.7...
20.3 17.8 14.39 14.89 17.8 16.41 23.54 21.47 16.59 31.9 29.4 13.27 23.9 ...
19.73 13.9 13.27 13.77 16.5];
>> X=[ ones(length(y),1) x1' x2'];
>> u=(X'*X)^-1*X'*y'

u =

   36.4857
    3.8272
   -0.1514

>> [xx1,yy1]=meshgrid(1:0.1:7,150:0.5:250);
>> zz=u(1)+u(2)*xx1+u(3)*yy1;
>> plot3(xx1,yy1,zz,x1,x2,y,'om')
```

FIG. 1. Matlab code

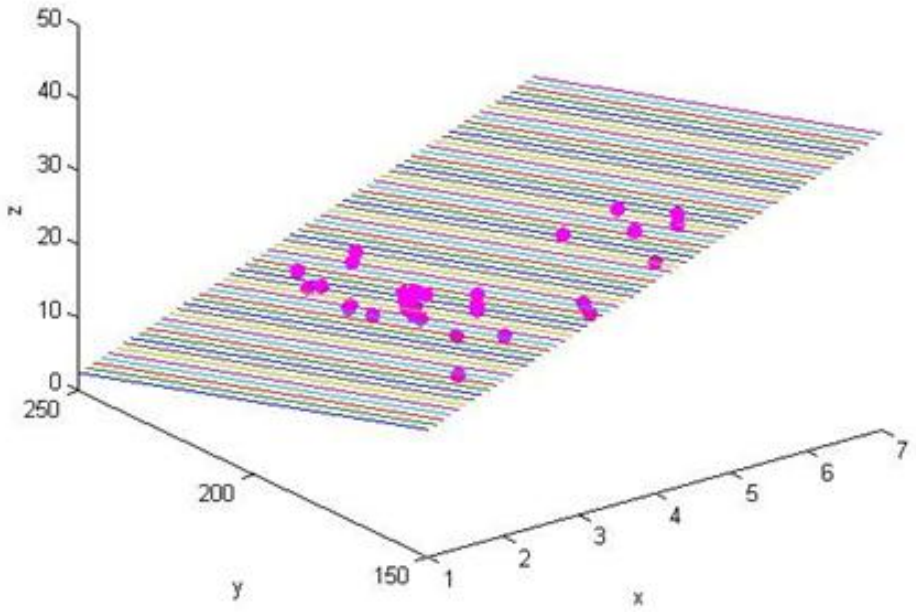

FIG. 2. Multiple linear regression through a scatter plot in space to which a plane of the form z = 36.4857 + 3.8272x + 0.1514y

Our FGNN represents [2] a modified version of Chen and Teng fuzzy neural network, by transforming the function of approximation into a function of classification. The change affects:

- the number of the classes (the number of the neurons belonging to the last layer);
- the equations of the fourth layer, but the structure diagram is similar.

4. FGNN ARCHITECTURE

The four-layer structure of the Fuzzy Gaussian Neural Network (FGNN) is shown in the Fig. 3.

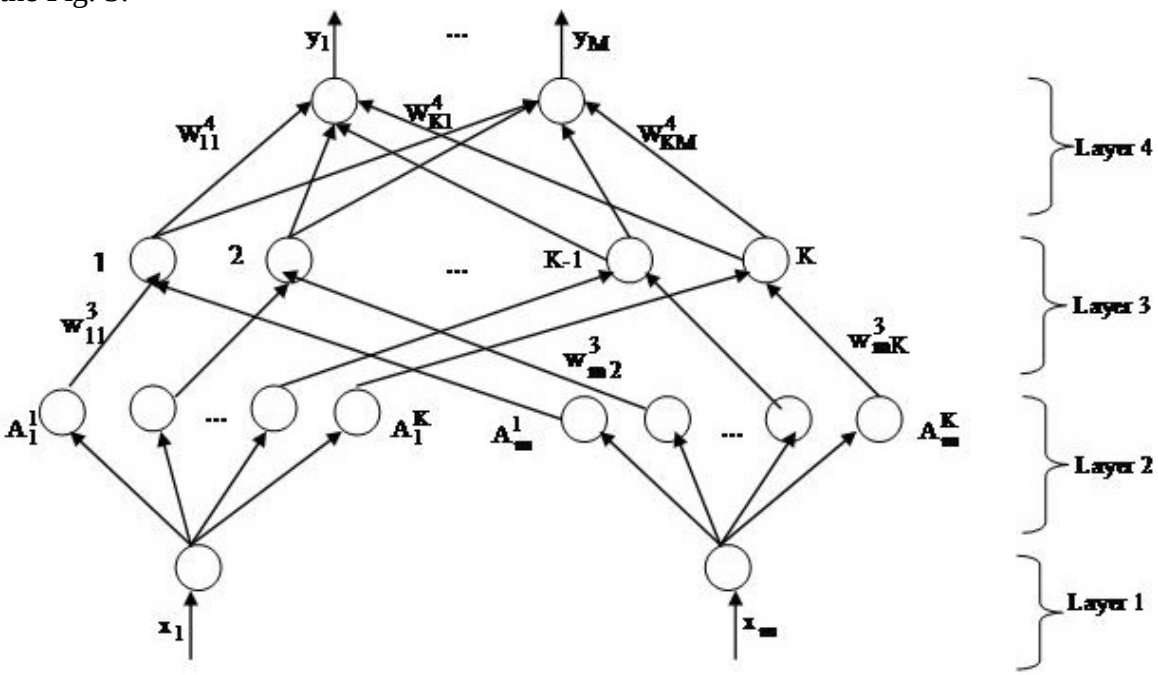

FIG. 3. Structure of FGNN

- *m* means the number of the neurons corresponding to the input layer;
- $X = (x_1, \ldots, x_m)$ represents the vector which one applies to the FGNN input;
- $\{W_{ij}^3\}_{i=\overline{1,m},\ j=\overline{1,K}}$ is the weight between the $(i-1)K + j$ -th neuron of the second layer and the neuron *j* of the third layer, where *K* is the number of the neurons from the third layer;
- $\{W_{ij}^4\}_{i=K,\ j=\overline{1,m}}$ is the connection from the neuron *i* from the third layer and the neuron *j* from the last layer of the FGNN;
- *M* represents the number of the classes;
- $Y = (y_1, \ldots, y_M)$ is the output of the FGNN.

The construction of FGNN is based on fuzzy rules of the form:

$\Re_j$: *If* x_1 *is* A_1^j *and* x_2 *is* A_2^j *... and* x_m *is* A_m^j, *then* y_1 *is* β_1^j, …, y_M *is* β_M^j,

where:

- *m* is the dimension of the input vectors (number of the retained features);
- $j,\ j = \overline{1,K}$ is the rule index;
- *M* is the number of the output neurons (it corresponds to the number of classes);
- $X = (x_1, \ldots, x_m)$ is the input vector, corresponding to the rule $\Re_j$;
- $A_i^j,\ i = \overline{1,m}$ are some fuzzy sets corresponding to the input vector;
- $Y = (y_1, \ldots, y_M)$ is the vector of the real outputs, corresponding to the rule $\Re_j$;
- $\beta_i^j,\ i = \overline{1,M}$ are some fuzzy sets corresponding to the output vector.

The *j*-th fuzzy rule is illustrated in Fig. 4.

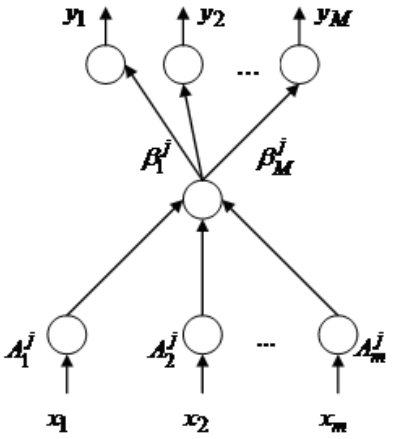

FIG. 4. The *j*-th component of FGNN

As in the case of the other neural networks, the FGNN input layer is a transparent layer, without a role in the data processing; the neurons of the first level (*input level*) do not process the signals; they only transmit the information to the next level.

The neurons of the second layer linguistic term layer (*level 2*) of the FGNN are membership neurons, resulting by the fuzzification of the first layer neurons. Each neuron of this level performs a Gaussian membership function [2], [6], [9]. The FGNN parameters have a physical significance, in the meaning that:

- m_{ij}, $i=\overline{1,m},\ j=\overline{1,K}$ represents the average;
- σ_{ij}, $i=\overline{1,m},\ j=\overline{1,K}$ is the variance

of the membership functions corresponding to some fuzzy sets, *m* being the number of the neurons from the input layer of the FGNN and *K* representing the number of the fuzzy considered rules.

The number of neurons characterizing this level is mK. Each input x_{ki}^2 is transformed by this layer into a fuzzy membership degree.

The third layer of the FGNN is called the *rule layer*. The connections between the membership neurons of the second layer and the rule neurons that characterize the third layer of the FGNN indicate the premise of the fuzzy rules. This layer computes the antecedent matching by the product operation [2], [6], [9].

The last layer of the FGNN is the output layer, which contains the output neurons. The conclusion (the consequence) of the rules is evidenced by the connections between the neurons of the third layer and the neurons of the output layer. This level performs the defuzzification of its inputs, providing *M* non-fuzzy outputs.

The FGNN parameters one initialize according to the *on- line initialization algorithm* [2], [6], [9] and they will be refined during the training algorithm [2], [6- 9].

5. TRAINING ALGORITHM

The training algorithm is of type back- propagation (BP), in order to minimize the error function:

$$E=\frac{1}{K}\cdot\sum_{k=1}^{K}E_k, \tag{1}$$

where:

- ✓ E_k represents the error for the rule *k*:

$$E_k = \frac{1}{2} \cdot \sum_{i=1}^{M} (d_{ki} - y_{ki})^2, \; k = \overline{1, K}, \tag{2}$$

- ✓ $d_k = (d_{k1}, \ldots, d_{kM})$ is the ideal output vector of the FGNN when at its input is applied the vector having the index *k*;
- ✓ $y_k = (y_{k1}, \ldots, y_{kM})$ is the corresponding real output vector of the FGNN ($k = \overline{1, K}$).

The training of this neural network is supervised, namely for of the K vectors from the training lot, we know the set of the ideal outputs. The refining of the FGNN parameters can be divided into two phases, depending on the parameters of premises and respective of conclusions of the rules, as follows:

A) in the part of the premise of the rules, the means and variances of the Gaussian functions one refine.
B) in the conclusions of the rules, the weights relating to the latest layer of FGNN must to be refined, the others being equal to 1.

7. EXPERIMENTAL EVALUATION

We are trying to find out how relevant is to use the Fuzzy Gaussian Neural Network for predicting personality because it handles well the nonlinearity associated with the data. We are also asking if the FGNN proves very good prediction performances over a statistical approach of prediction like MLRM and over a neural network as MP, too.

We use a data set made available by [3], [11]. The personality test called "The Big Five" (the five factor model of personality) represents the most comprehensive, reliable and useful test of personality concepts. It has emerged as one of the most well-researched and well-regarded measures of personality structure in recent years.

The Big Five traits are characterized [3] by the following:

- ➢ *Openness*: curious, intelligent, imaginative. High scorers tend to be artistic and sophisticated in taste and appreciate diverse views, ideas, and experiences;
- ➢ *Conscientiousness*: responsible, organized, persevering. Conscientious individuals are extremely reliable and tend to be high achievers, hard workers, and planners;
- ➢ *Extroversion*: outgoing, amicable, assertive. Friendly and energetic, extroverts draw inspiration from social situations;
- ➢ *Agreeableness*: cooperative, helpful, nurturing. People who score high in agreeableness are peace-keepers who are generally optimistic and trusting of others.
- ➢ *Neuroticism*: anxious, insecure, sensitive. Neurotics are moody, tense, and easily tipped into experiencing negative emotions.

The data is preprocessed in the following manner: we shall build a data set of 300 vectors, a half of them representing the training lot and the other half being the test lot.

These vectors have 20 components, each of them characterizing a personality trait. Each component means a correlation between the Big Five and individual words. For example [11]:

- *Neuroticism* correlates positively with negative emotion words (e.g. awful (0.26), though (0.24), lazy (0.24), worse (0.21), depressing (0.21), irony (0.21), terrible (0.2), road (-0.2), Southern (-0.2), stressful (0.19), horrible (0.19), sort (0.19), visited (-0.19), annoying (0.19), ashamed (0.19), ground (-0.19), ban (0.18), oldest (-0.18), invited (-0.18), completed (-0.18));
- *Extraversion* correlates positively with words reflecting social settings or experiences (e.g. Bar (0.23), other (-0.22), drinks (0.21), restaurant (0.21), dancing (0.2), restaurants (0.2), cats (-0.2), grandfather (0.2), Miami (0.2), countless (0.2), drinking (0.19), shots (0.19), computer (-0.19), girls (0.19), glorious (0.19), minor (-0.19), pool (0.18), crowd (0.18), sang (0.18), grilled (0.18));
- *Openness* shows strong positive correlations with words associated with intellectual or cultural experience (e.g. folk (0.32), humans (0.31), of (0.29), poet (0.29), art (0.29), by (0.28), universe (0.28), poetry (0.28), narrative (0.28), culture (0.28), giveaway (-0.28), century (0.28), sexual (0.27), films (0.27), novel (0.27), decades (0.27), ink (0.27), passage (0.27), literature (0.27), blues (0.26));
- *Agreeableness* correlates with words like: wonderful (0.28), together (0.26), visiting (0.26), morning (0.26), spring (0.25), porn (-0.25), walked (0.23), beautiful (0.23), staying (0.23), felt (0.23), share (0.23), gray (0.22), joy (0.22), afternoon (0.22), day (0.22), cost (-0.23), moments (0.22), hug (0.22), glad (0.22), fuck (-0.22);
- *Conscientiousness* has strong positive correlations with words like: completed (0.25), adventure (0.22), stupid (-0.22), boring (-0.22), adventures (0.2), desperate (-0.2), enjoying (0.2), saying (-0.2), Hawaii (0.19), utter (-0.19), extreme (-0.19), it's (-0.19), deck (0.18).

We want to predict M components (M being the number of the neurons from the output layer of FGNN) for every vector in order to complete the behavior corresponding to a person.

For evaluation, we use the Normalized Root Mean Square Error (NRMSE) [1].

Following [13], the prediction is considered:

- **excellent** if NRMSE≤0.1;
- **good** if 0.1 < NRMSE≤0.2;
- **fair** if 0.2 < NRMSE≤0.3;
- **poor** if NRMSE > 0.3.

For significance testing we use the three models: Multiple Linear Regression Model (MLRM), Multilayer Perceptron (MP) and Fuzzy Gaussian Neural Network (FGNN).

The three models: MLRM, MP and FGNN have been evaluated using a corresponding test lot, having a number of vectors equal to that of the training lot. The performance of FGNN over MP is based on [6] the fuzzy properties of FGNN, while the MP is a crisp neural network. The comparison [6] of FGNN and respectively MP versus MLRM marks both the competition nonlinear over linear and of neural over statistical, too.

CONCLUSIONS

The ability to predict personality has implications in many areas:

- like other studies relating to personality and language we adopted the five factor model of personality, which describes the following traits on a continuous scale: *neuroticism*, *extraversion*, *openness*, *agreeableness* and *conscientiousness*;

- ✓ in justice as the **personality rights** are some non- patrimonial civil rights, being regulated in article 58 NCC(New Civil Code); the protection of human personality is regulated by the Constitution of the Romania and the NCC(see the article 252).

To emphasize the performances of our proposed approach for predicting personality we have compared it both with a neural method of regression (like MP) and with a nonneural approach (MLRM), too.

According with the NRMSE criterion, we have achieved that the prediction with FGNN is better than with others two methods both over the training lot and over the test lot, too.

The advantage of the FGNN consists in the fact that, for certain values of the overlapping parameters one achieve very good recognition rates of the test lot.

A major problem with multiple regression consists in the large number of predictors that are available, although only a few of them are actually significant.

REFERENCES

[1] M. Benaddy, M. Wakrim and S. Aljahdali, *Evolutionary prediction for cumulative failure modeling: A comparative study*, in 2011 Eighth International Conference on Information Technology: New Generations, pp. 41-47, 2011;

[2] C. Y. Chen and C.C. Teng, *Fuzzy Neural Network System in Model Reference Control Systems* in Fuzzy Logic and Expert Systems Applications, pp. 285-313, Academic Press, San Diego-Toronto, 1998;

[3] J. Golbeck, C. Robles, M. Edmondson and K. Turner, *Predicting personality from TWITTER*, in IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 149-156, 2011;

[4] J. Golbeck, C. Robles and K. Turner, *Predicting personality with Social Media*, in IEEE International Conference on Privacy, Security, Risk, and Trust, and Proceedings of the ACM Conference on Human Factors in Computing, pp. 253-262, 2012;

[5] I. Iatan, *Neuro- Fuzzy Systems for Pattern Recognition* (in Romanian), Ph.D. Thesis, Faculty of Electronics, Telecommunications and Information Technology- University Politehnica of Bucharest, PhD supervisor: Prof. dr. Victor Neagoe, 2003;

[6] I. Iatan, *Issues in the Use of Neural Networks in Information Retrieval*, Springer-Verlag Berlin Heidelberg, 2016;

[7] V. Neagoe, and I. Iatan, *Face recognition using a fuzzy-gaussian neural network*, in Proceedings of First IEEE International Conference on Cognitive Informatics, ICCI 2002, 19-20 August 2002, Calgary, Alberta, Canada, pp. 361-368, 2002.

[8] V. Neagoe, I. Iatan and S. Grunwald, *A neuro- fuzzy approach to ECG signal classification for ischemic heart disease diagnosis* in the American Medical Informatics Association Symposium (AMIA 2003), Nov. 8- 12 2003, Washington DC, pp. 494-498, 2003.

[9] V. Neagoe, R. Iatan and I. Iatan, *A nonlinear neuro-fuzzy model for prediction of daily exchange rates* in Proceedings of World Automation Congress WAC'04 Seville, Spain, 17, pp. 573-578, 2004.

[10] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski and J. Crowcroft, *The personality of popular FACEBOOK users* in 2012 ACM Conference on Computer Supported Cooperative Work (CSCW 2012), Session: Social Network Analysis, pp. 955-964, 2012.

[11] T. Yarkoni, Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers, *Journal of Research in Personality*, vol. 44, pp. 363-373, 2010;

[12] S. Yu and S. Kak. *A survey of prediction using Social Media*, 2012. Available at http://arxiv.org/ftp/arxiv/papers/1203/1203.1647.pdf

[13] H. Zeng, L. Li, J. Hu, L. Liang, J. Li, B. Li and K. Zhang, Accuracy validation of TRMM multisatellite precipitation analysis daily precipitation products in the lancang river basin of China, *Theoretical and Applied Climatology*, pp. 1-13, 2012.