# ANFIS -Regression Model for Data Classification

Mrs.S.Dhivya Prabha[1], Dr.M.Deva Priya[2]

[1]M.Phil Research Scholar
PG & Department of Computer Science,
Government Arts College, Coimbatore-18,
Tamil Nadu, India.

[2]Assistant Professor in Computer Science,
Government Arts College, Coimbatore-18,
Tamil Nadu, India.

## Abstract

Data classification is an important area in data mining. Many familiar techniques such as Neural Network, Nearest Neighbors, Decision tree, SVM, etc are available for classification. There are few papers comparing the performance of statistical regression with other models found in the literature. It is learned from the literature that the regression model is well performed like other models. In this research proposed to use Adaptive Neuro Fuzzy Inference System (ANFIS) for data classification. Three real world datasets were collected from UCI machine repository, multiple regressions linear and Hybrid nonlinear classifier models were developed, the accuracy was recorded and the interrelation of attributes with respect to the class memberships were ensured.

The ANFIS Regression model is generated and its results were compared for validating the improved performance and found that the improvement in accuracy of the ANFIS Regression model is significant when compared to traditional regression model. The research illustrates the method of generating linear, hybrid nonlinear regression model and ANFIS Regression model for the given dataset.

## Keywords

ANFIS – Regression Model - ANFIS Applications – Data Classification.

## Introduction

The acronym ANFIS derives its name from adaptive neuro-fuzzy inference system. Using a given input/output data set, the toolbox function anfis constructs a fuzzy inference system (FIS). This adjustment allows the fuzzy systems to learn from data modeling. The architecture and learning procedure underlying ANFIS (adaptive network based fuzzy inference system) is presented, which is a fuzzy inference system implemented in the framework of adaptive networks. By using a hybrid learning procedure, the proposed ANFIS can construct an input-output mapping based on both human knowledge (in the form of fuzzy if-then rules) and stipulated input-output data pairs [19].

System modeling based on conventional mathematical tools (e.g., differential equations) is not well suited for dealing with ill-defined and uncertain systems. By contrast, a fuzzy inference system employing fuzzy if-then rules can model the qualitative aspects of human knowledge and reasoning process without employing precise quantitative analysis.

In a fuzzy inference system, the number of rules is decided by an expert who is familiar with the system to be modeled. In simulation, however, no expert is available and the number of membership functions (MF's) assigned to each input variable is chosen empirically, i.e., by examining the desired input-output data and/or by trial and error. This situation is much the same as that of neural networks; there are no simple ways to determine in advance the minimal number of hidden nodes necessary to achieve a desired performance level.

## Literature Review

A number of surveys, review articles and books especially Suresh Chandra Satapathy, et.al [1], Particle swarm optimized multiple regression linear model for data classification, Incorporating logistic regression to decision-theoretic rough sets for classifications [2], Classification of carotid artery stenosis of patients by neural network and logistic regression [3], Classification and Regression via Integer Optimization [4], Kalman particle swarm optimized polynomials for data classification [5], Semi-Supervised Classification Using Sparse Gaussian Process Regression [6] are excellent sources of literature on the subject.

## ANFIS – An Overview

ANFIS derives its name from adaptive neuro-fuzzy inference system. Using a given input/output data set, the toolbox function anfis constructs a fuzzy inference system (FIS). By using a hybrid learning procedure, the proposed ANFIS can construct an input-output mapping based on both human knowledge (in the form of fuzzy if-then rules) and stipulated input-output data pairs [19].

By contrast, a fuzzy inference system employing fuzzy if-then rules can model the qualitative aspects of human knowledge and reasoning process without employing precise quantitative analysis. However, there are some basic aspects of this approach which are in need of better understanding.

More specifically:

1) No standard methods exist for transforming human knowledge or experience into the rule base and database of a fuzzy inference system.

2) There is a need for effective methods of tuning the membership functions (MF's) so as to minimize the output error measure or maximize performance index.

ANFIS can serve as a basis for constructing a set of fuzzy if-then rules with appropriate membership functions to generate the stipulated input-output pairs.

## Data Sets and Experiments

This section presents different results that were obtained during the analysis and testing. In this research mainly used the following software for computational purposes: anfis edit commands and ANFIS GUI in Mat lab prompt.

The various data sets used in the tests and emphasize the behavior of the data classifications. To run the experiments, one has to input necessary parameters such as the desired number of classes and the upper and lower bounds on the datasets and the samples with their features scores.

## Description of Datasets

In this paper three data sets were used for developing the proposed ANFIS Regression model. These data sets named Iris, Balance scale and Tae which cover examples of small, medium and large dimensional data having numerical and categorical attribute values.

Table 1. Description of Datasets

| Dataset | Type of Dataset | Number of Classes | Number of Features | Size of Dataset | Class wise Distribution | | |
|---|---|---|---|---|---|---|---|
| Iris | Numerical | 3 | 4 | 150 | 50 | 50 | 50 |
| Balance Scale | Numerical | 3 | 4 | 625 | 288 | 49 | 288 |
| Tae | Categorical | 3 | 5 | 151 | 49 | 50 | 52 |

Table 1 summarizes these data sets. Experiments have adopted simple integer number encoding scheme for representing the attribute values of few datasets. For others, retained the original representations. Similarly, the class values are also encoded into integer values.

## IRIS dataset

In this research the applied model on well known IRIS data set. This particular data has been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in centimeters on 50 iris specimens from each of three species; Iris Setosa, Iris Versicolor, and Iris Virginica. This data set is one of the best known databases to be found in pattern recognition. In summary, it is characterized by 150 samples and each has four attributes that originally classify them in three classes. As mentioned earlier, in this model, the classes are not predefined; rather a continuous classification attribute is discretized by the ANFIS to form classes. Hence, discard the information related to the three species (classes) and in two instances. To test this method further, three sets of classification attribute values generated.

In the first case will assumed there was no overlapping between the classification attribute scores of samples; for the 3 classes, values were uniformly generated between 1–2, 2–3, and 3–4. In the second case, scores are slightly overlapping (1–2.25, 1.75–3.25, 2.75–4) and finally allowed more overlapping (1–2.5, 1.5–3.5, 2.5–4). Hence, to measure the impact of noisiness in data set on performance of ANFIS regression model.

ANFIS technique is used to develop models for the same data sets. The correct classification percentage of linear, non-linear are compared with ANFIS. In order to generate the training set, validation set, and test sets, each data set is divided proportionately into three parts. Each part has the presence of all classes of the data set. The class wise pattern distributions are done proportionately in each part. The instances in each class in each set are randomly picked up.

### Distribution and Division of datasets

The distribution and division of sets of datasets studied in this work is given below in the Table 2.

Table 2. Distribution and Division of datasets

| Data set | Set 1 | | Set 2 | | Set 3 | |
|---|---|---|---|---|---|---|
| | Class Distribution | Total | Class Distribution | Total | Class Distribution | Total |
| Iris | 16,17,17 | 50 | 16,17,17 | 50 | 16,17,17 | 50 |
| Balance Scale | 96,17,96 | 209 | 96,16,96 | 208 | 96,16,96 | 208 |
| Tae | 17,17,17 | 51 | 16,17,17 | 50 | 16,16,18 | 50 |

Two parts are taken for training and developing the model and the third part is taken for testing. The testing is done in Mat lab prompt. The development of ANFIS model is very straight forward in nature. The training dataset classified by minimizing the mean squared error between estimated value and the desired value. The developments of models are done separately using respective training data sets and the results of correct classifications are reported. For an example when set 1 and set 2 are taken for training, the set 3 is considered for testing. For an illustration for iris dataset the model is developed using set 1 and set 2 and this model is tested using set 3 for computing the correct percentage of classification. This process is repeated in ANFIS regression model through Mat lab prompt. Each simulation is carried for 50 trials.

The average classification accuracy is reported in the result analysis after 50 such trials for each testing set as shown in Table 3. For example the IRIS average percentage of correct classification in the linear model accuracy is 93.03 (over 50 trials) when set 2 and set 3 are taken to train and develop the model and set 1 is used to test the model. Average percentage of correct classification in the ANFIS model accuracy is 99.33 when set 2 is used for testing while set 1 and set 3 are used for developing the model. In ANFIS Regression model development the classification computed using ANFIS approaches described.

The ANFIS algorithm and procedure is run separately for separate datasets using any two parts of datasets for developing the model and finally the testing is done by the remaining third part. The experiment is carried for 50 simulation runs for each data set and the average results are reported separately for different training sets and tested.

### Overall percentage for correct classification for linear, non linear and ANFIS

Table.3 Overall percentage for correct classification for linear, non linear and ANFIS

| Dataset | Linear model accuracy (%) | Non-linear model accuracy (%) | ANFIS model accuracy (%) |
|---|---|---|---|
| Iris | 93.03 | 94.72 | 99.33 |
| Balance Scale | 69.34 | 71.41 | 78.62 |
| Tae | 38.94 | 42.15 | 46.25 |

The overall average is then computed over for each datasets. For the termination of ANFIS algorithm the maximum iteration is set to 200. These values are arrived at after having several initial simulation runs to get the optimized results. And it is seen that if iteration is set to 200 then optimum results are obtained in each dataset. For each classification problem has considered the class value as fitness value.

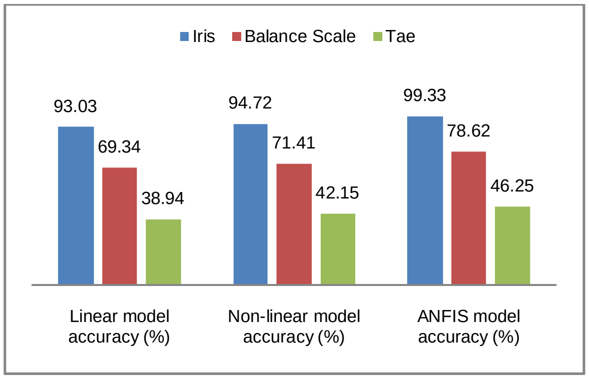

Fig.1 Overall percentage for correct classification for linear, non linear and ANFIS

The ANFIS tries to evolve the values of dataset so that the class value given by the same as the one given in the data set for a given record. The ANFIS Regression is also used for analysis, training and testing. The classification results are shown in Table 3. It can be seen clearly from Table 4 that ANFIS Regression outperforms all datasets as far as percentage of correct classification is concerned. In three datasets Iris, Balance scale and Tae the ANFIS approach performs far superior to traditional approaches.

**Confusion Matrix for three class datasets**

However, in datasets like Tae will do not get substantial improvement over traditional classification. This result shows the effectiveness of ANFIS approach over the classical approach in data classification. To further investigate the performance of ANFIS in classification have calculated the standard deviation of results over 50 trials for each testing dataset. It is clearly seen that ANFIS approach is a good candidate for pattern classification. The standard deviations for each testing set are well within the tolerable limit of computation. To further reinforce claim, the constructed confusion matrices separately for 2-class, 3-class, 4-class, and 7-class datasets for ANFIS Regression results.

The confusion matrices are obtained randomly from one of the 50 trials. The confusion matrix for 3-class problem is given in Table 4. Records correctly classified are shown bold in the confusion matrix.

Table4. Confusion Matrix for three class datasets

| Predicted | Set 1 | | | | Set 2 | | | | Set 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | Total | C1 | C2 | C3 | Total | C1 | C2 | C3 | Total |
| Iris | | | | | | | | | | | | |
| C1 | 0 | 13 | 3 | 16 | 1 | 12 | 3 | 16 | 0 | 15 | 1 | 16 |
| C2 | 0 | 14 | 3 | 17 | 0 | 15 | 2 | 17 | 0 | 14 | 3 | 17 |
| C3 | 0 | 13 | 4 | 17 | 0 | 14 | 3 | 17 | 0 | 14 | 3 | 17 |
| Balance Scale | | | | | | | | | | | | |
| C1 | 60 | 36 | 0 | 96 | 62 | 34 | 0 | 96 | 64 | 32 | 0 | 96 |
| C2 | 0 | 15 | 2 | 17 | 0 | 16 | 0 | 16 | 0 | 15 | 1 | 16 |
| C3 | 0 | 32 | 64 | 96 | 0 | 30 | 66 | 96 | 0 | 33 | 63 | 96 |
| Tae | | | | | | | | | | | | |
| C1 | 0 | 13 | 4 | 17 | 0 | 16 | 0 | 16 | 2 | 14 | 0 | 16 |
| C2 | 0 | 12 | 5 | 17 | 0 | 15 | 2 | 17 | 0 | 16 | 0 | 16 |
| C3 | 0 | 14 | 3 | 17 | 0 | 14 | 3 | 17 | 0 | 15 | 3 | 18 |

The above matrices clearly justify the findings of classification accuracy for ANFIS-based approach. The mathematical models developed using ANFIS Regression is presented below as illustrations for few datasets. These models are randomly picked up from 50 among trials.

The mathematical models shown above clearly indicate the existing interrelationship among attributes of a dataset. They give some rough ideas about the contribution of feature values in predicting the classes. Once the model is derived the class prediction for the unknown data instance can be easily computed from the suggested model in real time manner. The computation cost for testing is also minimal considering only few arithmetic operations like multiplications and additions.

On the data sets like Iris, Balance Scale and Tae, the ANFIS algorithm performed better than the traditional algorithms. This is possibly due to the different set of hyper parameters chosen during training. Note also that these algorithms performed better than the traditional algorithms on majority of the data sets. The ANFIS Regression algorithm performed better than the other algorithms on the Letter data set while the performances of the traditional algorithms were better than the other algorithms on the Iris data set. The high values of standard deviations, observed in some cases, are due to the small sized and poor hyper parameter choices made in one or two realizations. The performances of the proposed ANFIS Regression based algorithms were much better than those of the traditional classifiers.

Test set errors (%) reported are the averages over 10 trials. To study the effect of the size of the data sets on the generalization performance of the classifiers designed using different algorithms. This study conducted many experiments on the UCI machine repository data sets. It is clear from the table that the traditional algorithms performed poorly compared to the ANFIS Regression algorithm. Note also that the performances of the ANFIS Regression and traditional algorithms are comparable.

## CONCLUSION

This work can be extended in several directions. As shown in experimental results, the selection of the shape parameter for the ANFIS function is critical to the performance of the transformed regression algorithms.

This paper proposed an effective ANFIS Regression based multiple regression linear classifier design model for different real data sets in this study. The coefficients of regression model are estimated using ANFIS Regression techniques separately for each dataset. It is found that the regression classifiers using ANFIS Regression performs better compared to traditional approaches. The performance comparisons on many real data sets are presented. The outcome of this approach is a simple linear mathematical model based on ANFIS Regression approach which outperforms standard statistical technique for percentage of correct classification. The resultant mathematical models give a rough insight into the interrelationship among the attributes of dataset.

In this study investigated the ANFIS Regression based classification problem that takes the typical regression data as input to perform classification. This problem formulation could have important applications in real-world business data mining projects, due to the commonality of discrete decisions backed by continuous analytics. This study discussed the drawbacks of reducing such a problem into standard classification or regression analysis.

This paper proposed transformed regression that involves an ANFIS Regression based transform function to unify the classification and regression approaches to this problem. Using simulation data, have shown that the transformed regression generally outperformed both the classification and regression approaches.

## REFERENCES

[1] Suresh Chandra Satapathy, Suresh Chittineni, et.al " Particle swarm optimized polynomials for data classification", Applied Soft Computing 9 (2009) 470–476, Vishakapatnam, AP, India, May 2011.

[2] Suresh Chandra Satapathy a, J.V.R. Murthy, "Particle swarm optimized multiple regression linear model for data classification" , Applied Soft Computing (2009) 525-561, JNTU College of Engineering, Kakinada, India, Aug.2011.

[3] Asuncion and D.J. Newman. "UCI machine learning repository", School of Information and Computer Sciences, University of California, Irvine, International Conference on UCI Machine learning, pages 215–235 , June 2007.

[4] Dimitris Bertsimas, Romy Shiod, "Classification and Regression via Integer Optimization", Springer, New York, 2nd edition, April 2006.

[5] Kristin P. Bennett and Ayhan Demiriz. "Semi-supervised support vector machines." In Proceedings of the 12th Annual Conference on Neural Information Processing Systems, pages 368–374, Denver, USA, December 1998. MIT Press.

[6] S. Sundararajan, Shirish Shevade, Amrish Patel, "Semi-Supervised Classification Using Sparse Gaussian Process Regression", saga publications, 1st edition, Feb 2011.

[7] A.C. Tsoi et al, "Comparison of three classification techniques", CART, C4.5 and multilayer perceptrons, Adv. Neural Inform. Process. Syst. 3 (1991) 963–969.

[8] C. Russ, Eberhart, Y. Shi, “Comparing inertia eights and constriction factors in particle swarm optimization” , in: Proceedings of the Congress on Evolutionary Computing, 2000, pp. 84–89.

[9] C.C. Bojarczuk, H.S. Lopes, “ Genetic programming for knowledge discovery in chest pain diagnosis”, IEEE Engineering in Medicine Magazine 19 (4) (2000) 38–44.

[10] Carl Edward Rasmussen and Christopher K. I. Williams. “Gaussian Processes for Machine Learning”. MIT Press, 2006.

[11] D.G. Kleinbaum and M. Klein. “Logistic Regression”. Springer, New York, 2nd edition, 2002.

[12] Dun Liu, Tianrui Li, Decui Liang, “Incorporating logistic regression to decision-theoretic rough sets for classifications”, Xu San publication, 1st edition, November 2010

[13] Estelle R.S. Kone, Mark H. Karwan, “Combining a new data classification technique and regression analysis to predict the Cost-To-Serve new customers”, IEEE Transactions on Systems, April 2011.

[14] Ioan Cristian Trelea, “The particle swarm optimization algorithm: convergence analysis and parameter selection”, Inf. Process. Lett. 85 (6) (2003) 317–325.

[15] J. Adem, W. Gochet, “Mathematical program based heuristics for improving LP generated classifiers for the multi-class supervised classification problem”, European Journal Of Operation Research 168 (1) (2006) 181–199.

[16] J. Adem, W. Gochet, “Mathematical program based heuristics for improving LP-generated classifiers for the multi-class supervised classification problem”, IEEE bio medical service Center, (1994).

[17] J. Fox, “Applied Regression Analysis, Linear Models, and Related Methods”, Sage Publications, Thousand Oaks, CA, USA, 1997.

[18] J. Kennedy, R. Mendes, “Population structure and particle swarm performance, in: International Proceedings of the 2002 Congress on Evolutionary Computation”, IEEE service Center, Piscatawat, NJ, 2002, pp. 1671–1675.

[19] Jang, J.S. R. and C.T. Sun, “Neuro-fuzzy modeling and control”, Proceedings of the IEEE, March 1995.

[20] Jang, J.S. R., “ANFIS: Adaptive-Network-based Fuzzy Inference Systems”, IEEE Transactions on Systems, Man, and Cybernetics, May 1993.