A Type 2 Fuzzy C-Regression Method

Asli Celikyilmaz
University of Toronto,
Toronto, Canada
asli.celikyilmaz@utoronto.ca

I. Burhan Turksen
TOBB Economy and Technology University
Ankara, Turkey
bturksen@etu.edu.tr

Abstract

This paper presents a type-2 genetic fuzzy inference system based on fuzzy c-regression method clustering algorithm, to identify uncertainties in hyperplane shaped fuzzy clusters. The uncertainty in learning parameters of the new system is identified by type-2 fuzzy sets. Genetic algorithm is used to optimize the secondary membership grades of the type-2 fuzzy sets. Transductive reasoning, instead of inductive reasoning, is used to develop a local model for every new vector, based on some closest vectors from the given database. This study is novel because it presents a new methodology to identify type-2 fuzzy sets. The results of comparative experiments on financial forecasting problem domain are encouraging.

Keywords: type-2 fuzzy sets, fuzzy c-regression.

1 Genetic Type-2 Fuzzy Regression Method based on Transductive Reasoning

The concept of *type-2 fuzzy set* (T2FS) was introduced by Zadeh [1] as an extension of type-1 fuzzy set (ordinary fuzzy sets) to identify the uncertainties present in fuzzy systems. With fuzzy sets of higher type (e.g. type-2), the fuzziness of the relations is increased to handle inexact information.

A T2FS is identified by a fuzzy membership function (MF) – secondary MF, i.e., membership value. Each data point of this set is a fuzzy set between [0,1] unlike type-1 fuzzy sets, where the membership values are crisp numbers. T2FSs are useful in situations, where it is difficult or uncertain to determine the exact MF of a fuzzy set, primary MFs, viz., they are useful for incorporating uncertainties [2], [3], [4]. Interval T2FS are simplified forms of T2FS, where the secondary MFs are unified, e.g., equal to 1. Interval T2FS identify footprint-of-uncertainty (FOU) as depicted in Figure 1.

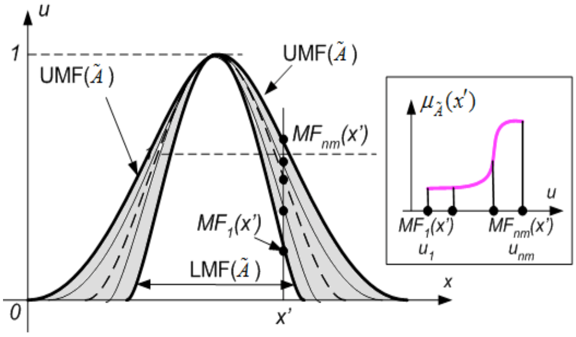

Figure 1: MFs where base-end-points have uncertainty intervals. The insert represents secondary MF of x'.

FOU of a T2FS $\tilde{A}$ is the uncertainty region (2D-region) specified by lower and upper MFs, LMF($\tilde{A}$), UMF($\tilde{A}$). For each data point, x', there can be *nm=2,..,∞* different MFs within this interval. Hence, T2FSs have secondary grades, which sit on top of FOU to form the 3D region.

In different studies, e.g., [6], [7], uncertainties of parameters from imperfect information are investigated using Fuzzy C-means (FCM) clustering algorithm [8]. In particular, the FOU of the interval T2FS are formed based on the level of fuzziness parameter of FCM clustering.

In fuzzy clustering methods, fuzziness is measured by the level of fuzziness parameter, *m*, which determines the degree of overlap between the clusters, viz. structures, granules, etc., identified in the given dataset. In many research, identification of the FOU of MFs of FCM clustering algorithm, e.g., [5], [6], or hybrid clustering algorithms [7] is based on the level of fuzziness parameter. In [6], one fuzziness value is found for each data vector and FOU of the MFs are identified based on resulting fuzziness values of the overall dataset. In [7], FOU is identified dynamically from the dataset based on parameters of a new improved fuzzy clustering method and local fuzzy function structures using genetic algorithms. Interval T2FS are used in these research.

In this paper, we investigate the level of fuzziness, *m,* of particularly fuzzy c-regression model (FCRM) clustering methods [9], instead of conventional clustering algorithms. In building fuzzy inference systems, separate functions are identified for each local input-output relation, which are defined with hyperplanes. Therefore, a better way is to construct hyperplane-shaped clusters.

This paper presents a new type-2 fuzzy inference method, which can identify the optimum secondary MF grades, i.e., weights, of the primary MF grades using genetic algorithms. New data vectors adopt the secondary MF grades obtained from the training samples in their neighborhood. During genetic learning process, each individual in the population encodes these weights for each training vector for each cluster, separately. This is quite cumbersome process when the number of training vectors are large therefore it is simplified in this paper by implementing transductive learning method. Instead of learning the secondary MF grades of the entire training dataset, for each new data point a new set of weights are learnt from fairly less training vectors, which are close to this new vector in distance. Experimental analysis demonstrates the performance of the new approach.

2 Genetic Type-2 Fuzzy Inference (GT2FI) Learning Algorithm

GT2FI, presented here, is a dynamic genetic type-2 fuzzy inference (FI) system akin to Takagi-Sugeno type FI, yet identifies one membership function for the entire antecedent part. We assume that the membership functions of each input variable are not independent, and their interactive affect should be analyzed instead of their individual effect. The first step of GT2FI system is to fuzzy partition the entire dataset into overlapping hyperplanes using FCRM clustering algorithm [9], to be summarized as follows.

Let f_i be a function with *nv* dimensional inputs, $\boldsymbol{x}_k(x_{k,1} \ldots x_{k,nv}) \in X$, $k=1,..,n$ data points and a single output. The representative of *i*th cluster can be expressed as:

$$y_i = f_i(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{a}}_i = \beta_{0,i} + \beta_{1,i} x_1 + \cdots + \beta_{nv,i} x_{nv} \qquad (1)$$

Step 1: Assume c^* hyperplanes as initial cluster representatives, $y_i = \hat{\boldsymbol{a}}_i \boldsymbol{x}$., $i=1,..,c$. For each iteration *t:*

Step 2*:* Calculate the $c \times n$ membership matrix $u_{i,k} \in U^{(t)}$ where $u_{i,k} \in [0,1]$ as follows:

$$E_{i,k} = \left(y_k - \boldsymbol{x}_k{}^T \hat{\boldsymbol{a}}_i^{(t)}\right)^2, \quad 1 \le i \le c$$

$$u_{i,k}{}^{(t)} = \left(\sum_{j=1}^{c} \left(\frac{E_{i,k}}{E_{i,k}} \right)^{1/(m-1)} \right)^{-1} \quad for \quad \sum_{i=1}^{c} u_{i,k}{}^{(t)} = 1 \qquad (2)$$

Step 3*:* If $\|U^{(t)}\text{-}U^{(t-1)}\| \le \varepsilon$, then stop; otherwise go to step 4.

Step 4: Calculate the new cluster representatives at the (*t*+1)th iteration, using *weighted least squares* method as $\beta_i^{(t+1)} = [\boldsymbol{x}^T \boldsymbol{D}_i \boldsymbol{x}]\ \boldsymbol{x}^T \boldsymbol{D}_i y$, and D_i denote the diagonal matrix in $\Re^{n \times n}$ having $u_{k,i}^{(t)} \in U_i^{(t)}$ as its *k*th diagonal element. To estimate one crisp output value, each fuzzy model output values are weighted with MF grades by:

$$\hat{y}_k = \sum_{i=1}^{c} f_i(\boldsymbol{x_k}) u_i(x_k) \Big/ \sum_{i=1}^{c} u_i(x_k) \qquad (3)$$

The MF in (2) depends on the level of fuzziness parameter, $m \in (1,\infty)$, which determines the fuzziness of the resulting clusters. The GT2FS performs the following learning algorithm:

1) ***Execute FCRM Method***. To identify FOU of T2FS, the FCRM is executed for different *levels of fuzziness*, $m^r = \{m^1 \ldots m^r\}$, $r=1 \ldots nm$, given the number of clusters, c^*. For each discrete m^r value FCRM models are represented with local functions $f_i^r(\boldsymbol{x}, \beta_i^r)$ of each cluster, $i=1 \ldots c^*$ and corresponding MFs, $MF_i^r(x) = u_i^r(x)$.

2) ***Initilize Secondary T2FSs***. Each possible discrete MFs, $\mu_i^r(\tilde{A})$, are randomly assigned initial weights, viz., secondary MF grades $\mu_{i,k}^r(\tilde{A}; x_k) \in [0,1]$,$r=1..nm$, $k=1\ldots n$.(see Figure 2)

These MF grades denote possibilities associated with each m^r at each value of x, $x_k=x'$.

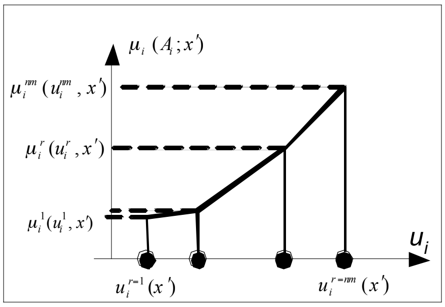

Figure 2: Secondary MF of x' in cluster i formed for each discrete primary MF based on level of fuzziness of FCRM clustering method.

3) ***Genetic Learning Process (GLP)***. Optimum values of the secondary MF grades of T2FSs at each x_k is identified based on genetic learning process. At this point, transductive learning algorithm is implemented to estimate the secondary membership values of data vectors. As a new vector, $\boldsymbol{x'}$ is introduced, a new model is build to estimate its output. The secondary MFs of $\boldsymbol{x'}$ in each cluster is estimated using n_k nearest neighbors from training dataset, which form a sample dataset $x_j \in X_j=\{x_1 \ldots x_{nk}\}$, from the existing dataset X. Each chromosome of $\boldsymbol{x'}$ is encoded using initial weights of each n_k training vector, one for each discrete m^r value for each cluster, as shown in Figure 3.

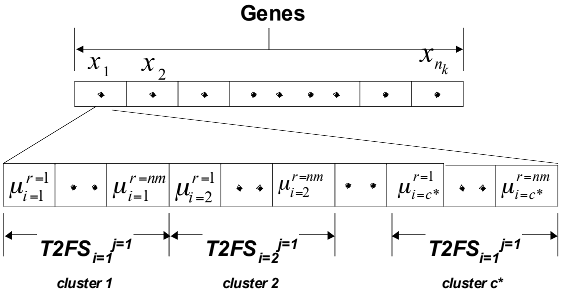

Figure 3: Chromosome structure of x'_k using n_k nearest neighbors.

x_j, $j=1..n_k$, represents each nearest vector to this x' vector and T$2FS_i{}^j$ represents T2FS of jth nearest train-vector in cluster $i=1..c^*$. Each $T2FS_i{}^j$ is identified by set of $u_{i,j}{}^r(x_j)$'s calculated by each m^r. Each chromosome encodes T2FSs of FCRM models for nearest n_k vectors. Herein, genetic algorithms is used to optimize the secondary MF grades of nearest n_k vectors instead of the entire training dataset. A separate genetic learning method is executed for each new x' vector as follows:

Step-1: Initialize each chromosome in the population, chr=1… max#chromosomes.

Step-2: Iterate until max-number of iterations is reached.

2.(i). Update T2FS secondary MF weights in each individual of the population using mutation and crossover operations.

2.(ii) For each chromosome, chr, calculate weighted crisp output values for each nearest train vector, $x_{k,j}$ as follows:

$$\hat{y}_{j,chr} = \sum_{i=1}^{c*} \frac{\sum_{r=1}^{nm} f(x_j;\beta_i^r)u_{i,j}^r(x_j)\mu_{i,j}^r}{\sum_{r=1}^{nm} \mu_{i,j}^r}, \sum_{i=1}^{c*} u_i = 1 \qquad (4)$$

2.(iii). Calculate average performance index of each nearest training vectors, x_j, $j=1..n_k$, of each individual:

$$PI_{chr} = \frac{1}{n_k}\sum_{j=1}^{n_k}\left(y_j - \hat{y}_{j,chr}\right)^2 \qquad (5)$$

2.(iv). Choose surviving individuals based on

$$\underset{chr}{\arg\max}\, PI_{chr} \qquad (6)$$

and go to step 2.(i) if termination condition is not satisfied.

Genetic learning process identifies the optimum m^r different secondary MF grades, $\mu_{i,j}{}^{(*)r}$, for each discrete primary MF grade, $u_{i,j}{}^r(x_j)$, $r=1...nm$, of each nearest training vector x_j of this x'. In (4) each output-value from each local function f^r is weighted with their secondary MF grades.

3 GT2FI Reasoning

To estimate the output value of a particular vector x' using GT2FI system, we use the weighing formula of equation (4). Firstly, the primary MF grades, $u_i{}^r(x')$ for each m^r value is calculated using equation (2). Since we do not know the actual output value of this new vector, we use actual output values of nearest training vectors, y_j. Secondly, as for its secondary MF grades, the weights of these nearest training vectors obtained from the GLP step are used.

To calculate $u_i{}^r(x')$ grade, the error of x' in each local model using m^r is measured with:

$$\tilde{E}_{i,j}^r(x') = \left\{m = m^r \,\middle|\, \left(y_j - f(x';\beta_i^r)\right)^2\right\} \qquad (6)$$

Next, a separate MF grade, $u_{i,j}{}^r(x')$ using each nearest training vector, x_j, is calculated using equation (2). The secondary MF grades of nearest train vectors obtained from GLP are used to calculate one output value, $\hat{y}'_j$, for this x' using each nearest training vector j, as follows:

$$\hat{y}'_{i,j} = \frac{\sum_{r=1}^{nm} f(x';\beta_i^r) u_{i,j}^r(x') \mu_{i,j}^{(*)r}}{\sum_{r=1}^{nm} \mu_{i,j}^{(*)r}} \tag{7}$$

In (7), the type of the MF is reduced down to type-1 by using model weights captured in GLP step. The type of the fuzzy output, $\hat{y}_{i,j}$, $i=1\ldots c^*$, is further reduced down to type-0 as follows :

$$\hat{y}'_j = \sum_{i=1}^{c} \hat{y}'_{i,j} \tag{8}$$

To calculate a single crisp output value for x', the output values based on nearest j training points, $\hat{y}_j$, from (8) are weighed based on the distance between x' and x_j, $j=1..n_k$, training points, denoted with $d(x'\ ,x_j)$as:

$$\begin{gathered} \eta_j = 1 - (d'_j \Big/ \sum_{s=1}^{n_k} d'_s(x', x_j)) \\ \hat{y}' = \sum_{j=1}^{n_k} \hat{y}'_j * \eta_j \end{gathered} \tag{9}$$

3 Experiments

In this section, we first demonstrate the distribution of secondary MF grades using an artificial dataset. Next, experiments conducted on a financial forecasting problem domain using real datasets is presented using the proposed GT2FI system.

3.1. Distribution of the Secondary MF Grades Using Artificial Dataset

The artificial dataset as shown in Figure 4 contains single input and single output with two local structures; therefore, the number of clusters is set to two. The primary MF grades, $u(x)$ values, are obtained from FCRM model using list of levels of fuzziness parameter m={1.1,1.25,..,2.6} as shown in Figure 4 top-right graph, also the base of the 3D graph , the bottom graph in Figure 4.

The bottom 3-D graph in Figure 4 displays the secondary MF of a single point $x_k=0.5$. The secondary MF values of nearest data points are optimized with genetic algorithms. For the genetic learning process, the initial population size and number of iterations are set to 100 each, and the number of clusters is set to 2. The crossover rate is set at 0.8 and the mutation rate is set at 0.01. Tournament selection with eliticist strategy is employed.

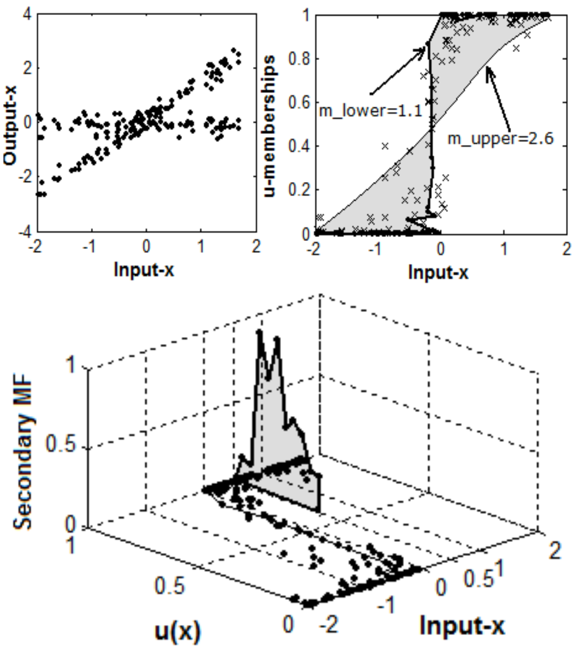

Figure 4: (Top-left) Artificial Dataset, (Top-right) FOU by $m\in$[1.1, 2.6], (Bottom) secondary MF of data point $x'=0.5$.

3.2. Stock Price Estimation Model

Five different Canadian stock prices, e.g., Toronto Dominion (TD), Bank of Montreal (BMO), Enbridge (ENB), Sunlife (SUN), Loblaws (LWS), collected between the years of 2005 and 2007 are used. The datasets are converted into multi-input single-output data mining problem, where the input variables are just the summary values of the stock prices. Among 100 different financial indicators [www.stockcharts.com], we used variables (see Table 1) that model market fluctuations, and focus on when to make buy or sell decisions.

Table 1: Variables of Stock Price Analysis.

Name	Description
EMA	Exponential Moving Average
BB	Bollinger Band
RSI	Relative Strength Index
MACD_t_k	Moving Average Convergence Divergence bwtn *t* and *t+k*

	periods
CMF	Chaikin Money Flow
SMA	Simple Moving Average
PCMA	Present Change of Moving Average - SMA(*t*)-SMA(*t-1*)
SR	Separation Ratio (SMA-Close Value)

There is a continuum between each data vector of the stock prices dataset; therefore, we divided each dataset into two periods. The first period is used for constructing five different training and validation datasets. The last periods are used for testing purposes. A sampling method using an artificial stock price is shown in Figure 5.

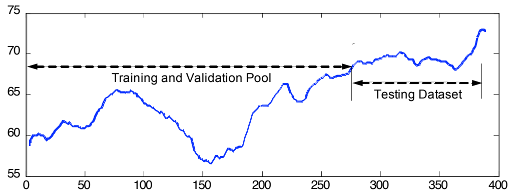

Figure 5: Sampling method used in experiments.

Stock prices collected around 20-22 months are divided into two parts. Approximately data from the first 15-17 months are used to train models and to optimize model parameters. The last 5 months are hold-out for testing model performances. We randomly separated 200 samples for training from the first part, 140 samples for validation of the optimum model parameters again from the first part and 100 samples to test the performance of models from the hold-out part, which has not been used for training or validation purposes. Experiments were repeated with 5 random subsets of above sizes.

The system model performance is measured with Mean Absolute Percentage Error (MAPE),

$$MAPE = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{y_k - \hat{y}_k}{y_k}\right|.100\,,\ \hat{y}_k\text{: predicted-output}$$

MAPE is a commonly used statistical measure of *Goodness of Fit* in quantitative forecasting methods. It produces a measure of *relative overall fit*. To analyze the performance of the new system, the MAPE results of GT2FI models are compared to well-known Adaptive Network Based Fuzzy Inference System (ANFIS) [10], Dynamic Evolving Neuro-Fuzzy Inference System (DENFIS) [11] models and a Type-2 Fuzzy Logic System, which identifies uncertainties based on FCM clustering method *level of fuzziness* parameter [6]. For the benchmark methods, default parameters are used.

For FCRM clustering method, we set the number of clusters to 3 and the boundaries of the level of fuzziness parameter between m_{lower}=1.4 and m_{upper}=2.6. The m interval is discretisized into 10 values. For the genetic learning process, the initial population size and number of iterations are set to 100 each, and the number of clusters is set to 3. The crossover rate is set at 0.8 and the mutation rate is set at 0.01. Tournament selection with eliticist strategy is employed.

The average MAPE values of the presented system and the benchmark methods and their cross validation standard deviations are displayed in Table 2.

To determine the level of significance of performance of the proposed approach compared to other system modeling tools applied in this paper, a significance test is applied. The results at 95% confidence level indicate that the proposed approach is significantly better that the rest of approaches in 4 out of 5 datasets

Table 2: MAPE results of Stock Price Datasets. Standard deviation of cross validation simples are shown after ± sign.

NAME	**ANFIS**	**DENFIS**	**T2FLS**	**GT2FI**
TD	1.82 ±1.69	1.42 ±0.29	0.45 ±0.26	**0.38** ±0.04
BMO	2.51 ±0.85	0.94 ±0.15	**0.87** ±0.03	0.91 ±0.02
ENB	2.24 ±0.64	1.19 ±0.05	1.21 ±0.08	**0.95** ±0.03
SUN	3.59 ±0.77	0.95 ±0.07	0.86 ±0.06	**0.83** ±0.01
LWS	3.86 ±1.62	1.23 ±0.27	1.09 ±0.23	**0.90** ±0.05

4 Conclusions

In this paper, a new genetic type-2 fuzzy inference system is introduced. Unlike counterparts, hyperplane shaped local structures are identified. The uncertainty interval of primary membership functions (MF) are defined based on upper and lower limits of the level of fuzziness parameter of fuzzy c-regression clustering method. The secondary MF grades are optimized with genetic algorithms. With the implementation of transductive learning method, a new model is constructed with only the training vectors in the vicinity of each new test vector. The algorithm implements a simple type-reduction and does not require defuzzification. The experimental results demonstrate significant performance improvement.

The presented genetic type-2 fuzzy inference identifies secondary MF grades, which are essentially the weights of primary membership grades. Implementing genetic algorithm, the weights are optimized automatically based on the performance improvement strategy. Hence, the optimum model selected by the genetic algorithm, assigns appropriate weights to membership values. During reasoning each local model's affect on the overall outcome is determined by the secondary membership grades.

Acknowledgements

This study is funded in part by Natural Sciences Engineering Research Council of Canada.

References

[1] L.A. Zadeh (1975) The concept of a linguistic variable and its application to approximate reasoning. *Inform. Sci* volume 8, pages 199-249.

[2] I.B. Turksen (1999) Type-1 and Type II Fuzzy System Modeling. *Fuzzy Sets and Systems* volume 106, pages 11-34.

[3] J. Mendel (2001). *Uncertain Rule-Based Fuzzy Logic Systems:Intr. And New Directions.* NJ: Prentice-Hall.

[4] R. John and S Coupland (2007) Geometric type-1 and type-2 fuzzy logic systems. *IEEE Trans. Fuzzy Syst.* volume 15, pages 3-15.

[5] C. Hwang and F.C.-H. Rhee (2007) Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means. *IEEE Trans. Fuzzy Syst.* volume 15, pages 107-120.

[6] O Uncu and I.B. Turksen (2007) Discrete interval type-2 fuzzy system models using uncertainty in learning parameters. *IEEE Trans. Fuzzy Syst.* volume 15, pages 90-106.

[7] A. Celikyilmaz and I.B. Turksen (2007) Improved Interval Valued Fuzzy Reasoning with Evolutionary Computing. In *Proc. JCIS'07 Conf.* 18-24 July, Salt Lake City.

[8] J. Bezdek (1981) *Pattern Recognition with Fuzzy Obj. Func. Algrthms.* NY: Plenum.

[9] R. Hathaway and J. Bezdek (1993) Switching regression model and fuzzy clustering. *IEEE Trans. Fuzzy Syst.* volume 1, pages 195-204.

[10] J-S.R. Jang (1993) ANFIS: Adaptive Network Based Fuzzy Inference System. *IEEE Trans. Syst., Man, Cybrn.* volume 23, pages 665-685.

[11] N.K. Kasabov (2002) DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction. *IEEE Trans. Fuzzy Syst.* volume 10, pages 144-154.