Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 14, No. 1, March 2014, pp. 1-7
http://dx.doi.org/10.5391/IJFIS.2014.14.1.1

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

On a Novel Way of Processing Data that Uses Fuzzy Sets for Later Use in Rule-Based Regression and Pattern Classification

Jerry M. Mendel
Signal and Image Processing Institute, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

ljfis

Abstract

This paper presents a novel method for simultaneously and automatically choosing the nonlinear structures of regressors or discriminant functions, as well as the number of terms to include in a rule-based regression model or pattern classifier. Variables are first partitioned into subsets each of which has a linguistic term (called a causal condition) associated with it; fuzzy sets are used to model the terms. Candidate interconnections (causal combinations) of either a term or its complement are formed, where the connecting word is AND which is modeled using the minimum operation. The data establishes which of the candidate causal combinations survive. A novel theoretical result leads to an exponential speedup in establishing this.

Keywords: Rule based regression, Pattern recognition, Fuzzy set

1. Introduction

Regression and pattern classification are very widely used in many fields and applications.[1] Both face four major challenges: 1) choosing the variables/features; 2) choosing the nonlinear structure of the regressors/discriminant functions; 3) choosing how many terms to include in the regression model/pattern classifier; and, 4) optimizing the parameters that complete the description of the regression model/pattern classifier.

For Challenge 1, how to choose the variables/features is crucial to the success of any regression model/pattern classifier. In this paper we assume that the user has established the variables that affect the outcome, using methods already available for doing this. For Challenge 4, there are a multitude of methods for optimizing parameters, ranging from classical steepest descent (back-propagation) to a plethora of evolutionary computing methods (e.g., simulated annealing, GA, PS0, QPSO, ant colony, etc. [4]), and we assume that the user has decided on which one of these to use. Our focus in this paper is on Challenges 2 and 3.

For Challenge 2, in real-world applications the nonlinear structures of the regressors/discriminant functions are usually not known ahead of time, and are therefore chosen either[2] as products of the variables (e.g., two at a time, three at a time, etc.), or in other more complicated

Received: Feb. 28, 2014
Revised : Mar. 14, 2014
Accepted: Mar. 24, 2014

Correspondence to: Jerry M. Mendel
(jmmprof@me.com)

[1] A search in Google, on January 22 2014, under regression listed about 20,800,000 results and under pattern classification listed about 19,900,000 results, so, it is beyond the realm of this paper to provide a complete list of articles that have been written about regression and pattern classification. Instead, we refer the readers to, e.g., [1-3].

[2] Linear terms may also be included.

ways (e.g., trigonometric-, exponential-, logarithmic-functions, etc.). Sometimes knowledge about the application provides justifications for the choices made for the nonlinear terms; however, often one does not have such knowledge, and a lot of time is spent, using trial and error, trying to establish the nonlinear dependencies. For Challenge 3, how to determine how many terms to include in the regression model/pattern classifier is also usually done by trial and error, and this can be very tedious to do. In this paper we present a novel method that chooses the nonlinear structure of the regressors/discriminant functions as well as the number of terms to include in the regression model/pattern classifier simultaneously and automatically. This is accomplished using a novel way of pre-processing the given data.

The rest of this paper is organized as follows: Section 2 explains how data can be treated as cases; Section 3 explains how each variable must be granulated; Section 4 describes the Takagi-Sugeno-Kang (TSK) rules that are used for regression/pattern classification; Section 5 presents the main results of this paper, a novel way to simultaneously determine the nonlinear structure of the regressors/discriminant functions and the number of terms to include in the regression model/pattern classifier; Section 6 provides some discussions; and, Section7 draws conclusions and indicates some directions for further research.

2. Data Treated as Cases

A data pair is denoted $(\mathbf{x}(t), y(t))$ where $\mathbf{x}$=$col(x_1, x_2, ..., x_p)$, x_i is the i^{th} variable/feature and $y(t)$ is the output for that $\mathbf{x}(t)$. As is commonly done in the social sciences [5, 6], each data pair is treated as a "case" and index t denotes a data case. Treating data as cases is motivated by a method called fuzzy set qualitative comparative analysis (fsQCA), which was developed by the prominent social scientist Ragin [5, 6], and has been thoroughly quantified by Mendel and Korjani [7, 8].

Note that there may or may not be a natural ordering of the cases over t. In multi-variable function approximation or pattern classification applications the data have no natural ordering; but in time-series forecasting applications the data cases have a natural temporal ordering. We assume that N data pairs are available, and refer to the collection of these data pairs as S_{Cases}, where $S_{Cases} = \{(\mathbf{x}(t), y(t))\}_{t=1}^{N}$.

3. Preprocessing

To begin, each of the p variables is granulated [9] into a fixed number of terms. Our suggested approach is to begin with only two terms per variable, design the regression model/pattern classifier, and determine if acceptable performance is obtained. If it has not, then increase the number of terms to three (then four, etc.) and repeat this process, stopping when acceptable performance has been obtained. Due to space limitations, we explain our preprocessing only for two terms per variable. Its extensions to more than two terms is straightforward.

For illustrative purposes, we shall call the two terms high (H) and low (L). Each variable x_i $(i = 1, ..., p)$, $x_i \in R^+$ (or $x_i \in R$), is mapped into the membership functions (MFs) of two type-1 fuzzy sets, one each for high and low. There are many different ways to do this, e.g. choose each MF as a prescribed two-parameter sigmoidal or piecewise-linear function.

In order to use the construction that is described in Section 5, it is required that the two MFs must be the complement of one another. This is easily achieved by using fuzzy c-means (FCM) for two clusters [10], (or linguistically modified FCM [LM-FCM] [11]), because it is well known that the MFs for the two FCM clusters are constrained so that one is the complement of the other.

As a result of this preprocessing step, the MFs $\mu_{L_i}(x_i)$ and $\mu_{H_i}(x_i) = 1 - \mu_{L_i}(x_i)$ $(i = 1, ..., p)$ will have been obtained. Note that, if independent MFs are used for L_i and H_i (for which $\mu_{H_i}(x_i) \neq 1 - \mu_{L_i}(x_i)$), then our Section 5 method will use L_i and H_i as well as their complements—four quantities. When, however, $\mu_{H_i}(x_i) = 1 - \mu_{L_i}(x_i)$ the four quantities reduce to two.

4. Rules

Our rules for a rule-based regression model or classifier have the following TSK structure [12]:

$$\begin{aligned} S_v : \text{IF } x_1 \text{ is } A_1^v \text{ ... and } x_p \text{ is } A_p^v, \\ \text{THEN } y_v(\mathbf{x}) = \beta_v, \quad v = 1, ..., R_S \end{aligned} \tag{1}$$

For the rule-based regression model, the β_v play the role of the regression coefficients which are determined by means of optimizing a regression objective function (e.g., minimizing a root mean square error), whereas for a binary pattern classifier the β_v are either +1 or -1, depending upon which class the rule is for. Regardless of whether Eq. (1) is used for regression or pattern classification, observe that the antecedent structure

(x_1 is A_1^v ... and x_p is A_p^v) and the number of rules (R_S) must be specified, after which it is straightforward to convert Eq. (1) into a so-called *fuzzy basis function expansion* [12, 13]. It is the mathematics of this conversion that establishes the nonlinear natures of the regressors/discriminant functions (Challenge 2); but this requires determining R_S and the antecedent structure. We show how to determine these simultaneously and automatically in the next section.

5. Establish Antecedents of Rules and the Number of Rules

The (compound) antecedent of each rule contains one linguistic term or its complement for each of the p variables, and each of these linguistic terms is combined with the others by using the word "and" (e.g., A_1 and A_2 and ... and A_p). We refer to this interconnection as a *causal combination*[3]. Note that in a traditional if-then rule the antecedents only use the terms and not their complements. In our approach (as in fsQCA), protection about being wrong for postulating a term is achieved by considering each term as well as its complement.

To begin, 2^p *candidate causal combinations* (the 2 is due to both the term and its complement[4]) are conceptually postulated (we will show below that these causal combinations do not actually have to be enumerated). If, e.g., $p = 6$ there would be 64 candidate causal combinations, or, if $p = 10$, there would be 1,024 candidate causal combinations.

One does not know ahead of time which of the 2^p candidate causal combinations should actually be used as a compound antecedent in a rule. Our approach prunes this large collection by using the MFs that were determined in Section 3, as well as the MF for "A_1 and A_2 and ... and A_p" (obtained using fuzzy set mathematics) and a simple test. The results of doing this are called R_s *surviving causal combinations*.

Let S_F be the collection of the following 2^p candidate causal combinations, F_j ($j = 1, ..., 2^p$ and $i = 1, ..., p$):

$$\begin{cases} S_F = \{F_1, ..., F_{2^p}\} \\ F_j \equiv A_1^j \wedge A_2^j \wedge ... \wedge A_i^j \wedge ... \wedge A_p^j \\ A_i^j = C_i \text{ or } c_i \end{cases} \tag{2}$$

where $\wedge$ denotes conjunction (the "and" operator) and is modeled using minimum and (using Ragin's [5] notation) c_i denotes the complement of C_i. The R_S surviving causal combinations are found from all of the 2^p candidate causal combinations by keeping only those causal combinations whose MF > 0.5 for at least f cases, where f is a threshold that has to be specified ahead of time[5]. A brute force way to do this is to create a table in which there are N rows, one for each case, and 2^p columns, one for each of the causal combinations. The entries into this table are $\mu_{F_j}(t)$ and there will be $N2^p$ such entries. Such a table is called a truth table by Ragin [4, 5]. One then searches through this very large table and keeps only those causal combinations whose MF entries are > 0.5. If $f = 1$ then all such causal combinations, removing duplications, become the set of R_S surviving causal combinations. It is very easy for $N2^p$ to become very large[6] and so this brute force way to carry out this procedure is impractical.

Ragin [5] observed the following in an example with four causal conditions: "...each case can have (at most) only a single membership score greater than 0.5 in the logical possible combinations from a given set of causal conditions (i.e., in the candidate causal combinations)." This somewhat surprising result is true in general and in [8] the following theorem that locates the one causal combination for each case whose MF $>$ 0.5 was presented:

Theorem 5.1 (min-max theorem)**.** [8]: given p causal conditions, C_1, C_2,..., C_p and their respective complements, c_1, c_2, ..., c_p. Consider the 2^p candidate causal combinations ($j = 1, ..., 2^p$) $F_j = A_1^j \wedge A_2^j \wedge ... \wedge A_p^j$ where $A_i^j = C_i$ or c_i and $i = 1, ..., p$.

Proof. Let

$$\mu_{F_j}(t) = \min\{\mu_{A_1^j}(t), \mu_{A_2^j}(t), ..., \mu_{A_p^j}(t)\}, t = 1, 2, ..., N \tag{3}$$

Then for each t (case) there is only one j, $j*(t)$, for which $\mu_{F_{j*(t)}}(t) > 0.5$ and $\mu_{F_{j*(t)}}(t)$ can be computed as:

$$\mu_{F_{j*(t)}}(t) = \min\left\{\max\left(\mu_{C_1}^D(t), \mu_{c_1}^D(t)\right), \ldots, \max\left(\mu_{C_p}^D(t), \mu_{c_p}^D(t)\right)\right\} \tag{4}$$

where $\mu_{e_i}^D(x) = \mu_{c_i}(\xi_i(x))$. $F_{j*(t)}(t)$ is determined from the

[3]The term "causal combination" is borrowed from fsQCA (e.g., [5-7]).

[4]When the two terms are not complements of each other, then there are 2^{2p} candidate causal combinations.

[5]Each of the 2^k candidate causal combinations can be interpreted as a corner in a 2^k-dimensional vector space [3]. Choosing the surviving causal combinations as just explained is interpreted as keeping the adequately represented causal combinations that are closer to corners and not the ones that are farther away from corners.

[6]If each variable is described by n_c independent terms, then $N2^p \rightarrow N2^{n_c p}$. In this situation, $N2^{n_c p}$ can easily become enormous, e.g. if $p = 6$ and $n_c = 3$, then $2^{n_c p} = 2^{18}$, or, if $p = 10$ and $n_c = 3$, then $2^{n_c p} = 2^{30}$. Even for $n_c = 3$, the brute force approach way to carry out this procedure is totally impractical or impossible.

right-hand side of Eq. (4), as:

$$\begin{aligned} F_{j*(t)}(t) &= arg\max\left(\mu_{C_1}^D(t), \mu_{c_1}^D(t)\right) \\ &\quad \ldots arg\max\left(\mu_{C_p}^D(t), \mu_{c_p}^D(t)\right) \\ &\triangleq A_1^{j*(t)} \wedge \ldots \wedge A_p^{j*(t)} \end{aligned} \tag{5}$$

In Eq. (5), $arg\max\left(\mu_{C_i}^D(t), \mu_{c_i}^D(t)\right)$ denotes the winner of $\max\left(\mu_{C_i}^D(t), \mu_{c_i}^D(t)\right)$, namely C_i or c_i.

A proof of this theorem is in [8]. When n_c independent terms are used for each variable, replace p by $n_c p$. A numerical example that illustrates the computations can be found in [7].

This min-max Theorem leads to the following procedure for computing the R_S surviving causal combinations[7]:

1. Compute $F_{j*}(t)$ using Eq. (5).
2. Find the J uniquely different $F_{j*}(t)$ and re-label them $F_{j'}(j' = 1, ..., J)$.
3. Compute $t_{F_{j'}}$, where $(t = 1, ..., N)$

$$t_{F_{j'}}(t) = \begin{cases} 1 & \text{if } F_{j'} = F_{j*(t)}(t) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

4. Compute $N_{F_{j'}}$, where

$$N_{F_{j'}} = \sum_{t=1}^{N} t_{F_{j'}}(t) \tag{7}$$

5. Establish the R_S surviving causal combinations F_v^S $(v = 1, ..., R_S)$, as:

$$F_v^S = \begin{cases} F_{j'}(j' \to v) & \text{if } N_{F_{j'}} \geq f \\ 0 & \text{if } N_{F_{j'}} < f \end{cases} \tag{8}$$

where $F_{j'}(j' \to v)$ means $F_{j'}$ is added to the set of surviving causal combinations as F_v^S, and v is the index of the surviving set.

Numerical examples that illustrates this five-step procedure can be found in [8].

In order to implement Eq. (8) threshold $f$has to be chosen. In our works, we often choose $f = 1$. This choice is arbitrary and depends on an application and how many cases are available. Discussions on how to choose f are given in [5-7, 14]. One popular way to choose f is as the integer such that at least 80% of all cases are covered by the set of surviving causal combinations.

From F_j in Eq. (2) and F_v^S in Eq. (8), it follows that $(v = 1, ..., R_S)$:

$$F_v^S(x_1, ..., x_p) = F_v^S(\mathbf{x}) = A_1^v(x_1) \wedge A_2^v(x_2) \wedge ... \wedge A_p^v(x_p) \tag{9}$$

In [8] it is shown that the speedup between our method for determining the surviving causal combinations and the brute-force approach is $\approx O\left(2^{n_c p}\right)$, where n_c is the number of terms used for each variable (assumed to be the same for all variables).

Example: This example illustrates the number of surviving causal combinations for eight readily available data sets: abalone [15], concrete compressive strength [15], concrete slump test [15], wave force [16], chemical process concentration readings [17], chemical process temperature readings [17], gas furnace [17] and Mackey-Glass Chaotic Time Series [18]. Our results are summarized in[8] Table 1. For each problem a two-cluster FCM was applied to all of its cases. The five-step procedure described above was then used to determine R_S.

Observe that: (1) for three variables (as occurs for wave force, chemical process concentration reading and chemical process temperature readings), the number of surviving causal combinations is either the same number, or close to the same number, as the number of candidate causal combinations, which suggests that one should use more than two terms per variable; and, (2) In all other situations the number of surviving causal combinations is considerably smaller than the number of candidate causal combinations. Although not shown here, this difference increases when more terms per variable are used, e.g., using three terms per variables the candidate causal combinations for the concrete slump test data set is 134,217,728 whereas the number of surviving causal combinations is only 97 [19].

Observe, from the last column in Table 1, that for four of the problems $R_S \geq 25$. We seriously doubt that a human designer could postulate the non-linear structures for so many regressors/discriminant functions. Our method not only shows that so many of them are necessary, it also finds their nonlinear structures.

[7]This procedure is modeled after Step 6NEW in Fast fsQCA, as described in [7].

[8]The entries in this table were obtained by Mr. Mohammad M. Korjani, a Ph. D. student in the Ming Hsieh Department of Electrical Engineering, University of Southern California.

Table 1. Number of surviving causal combinations for eight problems

Problem	Cases	Variables (p)	Two terms per variables[a]	
			Candidate causal combinations (2^p)	Surviving causal combinations (R_s)
Abalone [14]	4,177	7	128	55
Concrete compressive strength [14]	1,030	8	256	73
Concrete slump test [14]	103	9	512	71
Wave force [16]	317	3	8	8
Chemical process concentration reading [17]	194	3	8	8
Chemical process temperature readings [17]	223	3	8	6
Gas furnace [17]	293	6	64	25
Mackey-Glass chaotic time series [18]	1,000	4	16	8

[a] The two terms are low and high, and their fuzzy c-mean membership functions are the complements of one another.

6. Discussion

In Korjani and Mendel [19] have shown how the surviving causal combinations can be used in a new regression model, called variable structure regression (VSR). Using the surviving causal combinations one can simultaneously determine the number of terms in the (nonlinear) regression model as well as the exact mathematical structure for each of the terms (basis functions). VSR has been tested on the eight small to moderate size data sets that are stated in Table 1 (four are for multi-variable function approximation and four are for forecasting), using only two terms per variable whose MFs are the complements of one another, has been compared against five other methods, and has ranked #1 against all of them for all of the eight data sets.

Specific formulas for fuzzy basis function expansions can be found in [12, 13]. Similar formulas for rule-based binary classification can be found in [12].

Surviving causal combinations have also been used to obtain linguistic summarizations using fsQCA [7, 8].

7. Conclusions

This paper presents a novel method for simultaneously and automatically choosing the nonlinear structures of regressors or discriminant functions, as well as the number of terms to include in a rule-based regression model or pattern classifier. Variables are first partitioned into subsets each of which has a linguistic term (called a causal condition) associated with it; fuzzy sets are used to model the terms. Candidate interconnections (causal combinations) of either a term or its complement are formed, where the connecting word is AND which is modeled using the minimum operation. The data establishes which of the candidate causal combinations survive. A novel theoretical result leads to an exponential speedup in establishing this. For specific applications, see [7, 8, 19].

Much work remains to be done in using surviving causal combinations in real-world applications. The extension of the min-max Theorem to interval type-2 fuzzy sets is currently being researched.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] B. Pushpa and R. Vasuki, "A least absolute approach to multiple fuzzy regression using T_w-norm based operations," *International Journal of Fuzzy Logic Systems*, vol. 3, no. 2, pp. 73-84, Apr. 2013. http://dx.doi.org/10.5121/ijfls.2013.3206

[2] C. Ritz and J. C. Streibig, *Nonlinear Regression with R*, New York, NY: Springer, 2008.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., New York, NY: Wiley, 2001.

[4] D. Simon, *Evolutionary Optimization Algorithms*, Hoboken, NJ: John Wiley & Sons, 2013.

[5] C. C. Ragin, *Redesigning Social Inquiry: Fuzzy Sets and Beyond*, Chicago, IL: University of Chicago Press, 2008.

[6] B. Rihoux and C. C. Ragin, Eds, *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*, Thousand Oaks, CA: Sage, 2009.

[7] J. M. Mendel and M. M. Korjani,""Charles Ragin's fuzzy set qualitative comparative analysis (fsQCA) used for linguistic summarizations," *Information Sciences*, vol. 202, pp. 1-23, Oct. 2012. http://dx.doi.org/10.1016/j.ins.2012.02.039

[8] J. M. Mendel and M. M. Korjani, "Theoretical aspects of fuzzy set qualitative comparative analysis (fsQCA)," *Information Sciences*, vol. 237, pp. 137-161, Jul. 2013. http://dx.doi.org/10.1016/j.ins.2013.02.048

[9] A. Bargiela and W. Pedrycz, *Granular Computing: An Introduction*, New York, NY: Springer, 2003.

[10] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Noewll, MA: Kluwer, 1981.

[11] M. M. Korjani and J. M. Mendel, "Challenges to using fuzzy set qualitative comparative analyses (fsQCA) and their solutions: modified-fsQCA," *IEEE Transactions on Fuzzy Systems*, 2013 [submitted].

[12] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*, Upper Saddle River: NJ: Prentice-Hall, 2001.

[13] L. X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least-squares learning," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 807-814, Sep. 1992. http://dx.doi.org/10.1109/72.159070

[14] P. C. Fiss, "Building better causal theories: a fuzzy set approach to typologies in organization research," *Academy of Management Journal*, vol. 54, no. 2, pp. 393-420, Apr. 2011. http://dx.doi.org/10.5465/AMJ.2011.60263120

[15] A. Frank and A. Asuncion, "UCI Machine Learning Repository," Available http://archive.ics.uci.edu/ml/

[16] R. J. Hyndman, "Time Series Data Library," Available http://robjhyndman.com/TSDL/

[17] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed., Englewood Cliffs, NJ: Prentice Hall, 1994.

[18] R. S. Cowder, "Predicting the Mackey-Glass time series with cascade correlation learning," in *Connectionist Models: Proceedings of the 1990 Summer School*, D. S. Touretzky, Ed. San Mateo, CA: M. Kaufmann Publishers, 1990, pp. 117-123.

[19] M. M. Korjani and J. M. Mendel, "Non-linear variable structure regression (VSR) and its application in time-series forecasting," in *Proceedings of FUZZ-IEEE*, 2014, Beijing, China, July 2014.

Jerry M. Mendel received the Ph.D. degree in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, NY. Currently he is Professor of Electrical Engineering and Systems Architecting Engineering at the University of Southern California in Los Angeles, where he has been since 1974. He has published over 550 technical papers and is author and/or editor of ten books, including Uncertain Rule-based Fuzzy Logic Systems: Introduction and New Directions (Prentice-Hall, 2001), Perceptual Computing: Aiding People in Making Subjective Judgments (Wiley & IEEE Press, 2010), and Advances in Type-2 Fuzzy Sets and Systems (Springer 2013). His present research interests include: type-2 fuzzy logic systems and their applications to a wide range of problems, including smart oil field technology, computing with words, and fuzzy set qualitative comparative analysis. He is a Life Fellow of the IEEE, a Distinguished Member of the IEEE Control Systems Society, and a Fellow of the International Fuzzy Systems Association. He was President of the IEEE Control Systems Society in 1986. He was a member of the Administrative Committee of the IEEE Computational Intelligence Society for nine years, and was Chairman of its Fuzzy Systems Technical Committee and the Computing With Words Task Force of that TC. Among his awards are the 1983 Best Transactions Paper Award of the IEEE Geoscience and Remote Sensing Society, the 1992 Signal Processing Society Paper Award, the 2002 and 2014

Transactions on Fuzzy Systems Outstanding Paper Awards, a 1984 IEEE Centennial Medal, an IEEE Third Millenium Medal, and a Fuzzy Systems Pioneer Award (2008) from the IEEE Computational Intelligence Society.