An algorithm for rule selection on fuzzy rule-based system applied to the treatment of diabetics and detection fraud in electronic payment,

Farida Benmakrouha, Christiane Hespel, Edouard Monnier

An algorithm for rule selection on fuzzy rule-based systems applied to the treatment of diabetics and detection of fraud in electronic payment

F. Benmakrouha, C. Hespel, E. Monnier

***Abstract*—Recently, many papers have proposed automatic techniques for extraction of knowledge from numerical data and for minimization of the number of rules. But few works have been developed for design of experiments and datum plane covering. Most of optimization methods make the assumption that datum plane is sufficiently covered. If this assumption no longer holds, we will see that these methods may not work, since it implies that, before optimization, the fuzzy system gives acceptable results. We present in this paper, an algorithm for selection of fuzzy rules based on datum plane covering. We apply this method to two applications :**

- **the problem of treatment of diabetics. Taking the insulin infusion rate as the input and the blood glucose rate as the output, we consider the patient as a black box , whose model has to be obtained from the available measures of inputs–outputs. We dispose of a glycaemia file automatically produced for every person, and an insulin file shared by several persons.**
- **the problem of detecting fraud in electronic payment systems. Along with the development of credit cards, fraudulent activities become a major problem in electronic payment systems. Our model is based on specific and usual customer behavior, and deviation from such patterns is suspect.**

I. INTRODUCTION

Many works have proposed a minimization of the number of rules [2], [3], [4]. Ishibushi in [2] defines candidate rules based on the grade of certainty and applies genetic algorithm for rule selection problem. Alcala and co, in [3], estimate zones with bad rules, with redundant or irrelevant rules and complementary rules, and search a good configuration of rules by removing rules with little importance.

But these methods make the assumption that datum plane is sufficiently covered. In on-line experiments, data don't cover the whole input space and some rules are never learned. This lack of data is pointed out by M Lutaud-Brunet in [8]. Sugeno and Yasukawa underline in [9] how difficult it is to build a fuzzy model when data are scarce and membership functions dont' sweep over all the universe of discourse.

We present, in this paper, a method for selection of fuzzy rules based on on datum plane covering. Our feature selection algorithm takes place in the problematic of trade-off between accuracy and complexity. In [10], Singh and Hirota propose an approximation bound error for Tagaki-Sugeno models. They show particularly that performance is improved with number of membership functions. They also make assumption that consequent values are exactly known. If this assumption no longer holds and there is an insufficient covering of datum plane, training and finer splitting of input space may be inefficient and useless.

- First, we apply our method for modeling the "insulin delivery/glycaemia" behavior of some patient, under insulin infusion, for a given type of insulin, under continuous glucose monitoring. This method belongs to the class of methods that consider the patient as a black box. We construct a collection of linear models that describe the behavior of the glycemia under certain conditions. These conditions can be either defined in advance (for example, during the inter-prandial period, during meals or during physical effort), or determined by a learning process. Each model is thus valid only for a certain period of time. In practice, we can show that our model is valid for at least fifteen minutes. To combine this collection of linear models and represent nonlinear aspects of our problem, we choose a Tagaki-Sugeno(TS) fuzzy model. We validate this model by computing the mean square error, performance index of system, on the totality of measures(1700 points).
  There exist several methods for glycemic identification, based on mathematical, computer science and control theory techniques, open-loop, partially closed-loop and closed-loop techniques. The very first closed-loop regulation method was developed by A. Albisser et al. back in 1974 [11]. Among other methods, we can mention [12], [13], [14], [15], [16]. Unfortunately, in spite of many positive aspects of these methods, none of them was unanimously accepted by the medical community. This is partly due to the lack of the precision of the available data and, especially, to insufficient frequency of glycemic sampling. We note also that most of models proposed are :
  - either linear which do not take into account the nonlinear aspects of almost real-time life problems.
  - or global systems, which is not satisfactory. Since a diabetic does not respond in the same way to equal doses of insulin at different times of the day, it is reasonable to suppose that he is described by different

This work was not supported by any organization

F. Benmakrouha is with the Mathematics and Computer Science, INSA, CS 70839, 35708 RENNES CEDEX 7 FRANCE benma@insa-rennes.fr

C. Hespel is with Mathematics and Computer Science, INSA, CS 70839, 35708 RENNES CEDEX 7 FRANCE hespel@insa-rennes.fr

E. Monnier is with Mathematics and Computer Science, INSA, CS 70839, 35708 RENNES CEDEX 7 FRANCE monnier@insa-rennes.fr

systems at different time instants.

- The second application is the problem of detecting fraud in electronic payment systems.
  Along with the development of credit cards, fraudulent activities become a major problem in electronic payment systems [5], [6], [7]. Our model is based on specific and usual customer behavior, and deviation from such patterns is suspect. Our model is based on specific and usual customer behavior, and deviation from such patterns is suspect. To define a specific profile, we consider transaction amount and shopping time as inputs and suspicion of fraud as output.

II. The algorithm for fuzzy rules selection

A. The Tagaki-Sugeno Model

The model under consideration in this section is a Tagaki-Sugeno model, that consists in a family of linear models mixed together with nonlinear membership functions.

$$y(t) = \sum_{i=1}^{r} \mu_i(z(t))(a_1^i.u_1(t) + a_2^i.u_2(t))$$

r is the number of linear models,
$z(t)$ a vector which depends lineary or not on the state,
$\mu_i(z(t)) >= 0$, $i = 1, \cdots, r$ nonlinear functions verifying the convex sum property.
$u_1(t)$ and $u_1(t)$ are the input. The determination of unknown parameters a_1^i and a_2^i is done by the algorithm of recursive least square.

B. Measure of Datum Plane Covering

We suppose that there exists a learning set $\Omega = \{(\mathbf{x}_j, d_j)\}$, where $\mathbf{x}_j$ is an input vector and d_j, the corresponding output. We also assume that the desired function f is defined in

$$V = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_p, b_p]$$

When designing a fuzzy system, we attribute to each input I r_I modalities (or labels) noted $x_1, \cdots x_l, \cdots x_{r_I}$. We note X^I the variable for the input I of average $\bar{x}^I$ and variance $\sigma^2_{X^I}$. We note Ω^I the corresponding learning set for the input I. Each label x_l of I defines a subset Ω_l^I of Ω^I. We note $n_l^I = card(\Omega_l^I)$ and $n^I = \sum n_l^I$. If we obtain a partition of Ω^I in r_I classes, then $n^I = card(\Omega^I)$.
So, we may define the average (noted $\bar{x}_l^I$) and the variance (noted $\sigma^2_{X_l^I}$) of X^I on the subset Ω_l^I) $(l = 1, \cdots, r_I)$

$$\bar{x}_l^I = \frac{\sum_{\omega \in \Omega_l^I} X(\omega)}{n_l^I}$$

$$\sigma^2_{X_l^I} = \frac{\sum_{\omega \in \Omega_l^I} (X(\omega) - \bar{x}_l^I)^2}{n_l^I}$$

We have an index of connection between the datum plane coverage (for an input I) and the learning set defined by :

$$s^I = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}}$$

where

$$\sigma_Y^2 = \sigma_E^2 + \sigma_R^2$$

and

$$\sigma_E^2 = \frac{\sum_{l=1}^{r_I} n_l^I * (\bar{x}_l^I - \bar{x}^I)^2}{n^I}$$

and

$$\sigma_R^2 = \frac{\sum_{l=1}^{r_I} n_l^I * \sigma^2_{X_l^I}}{n^I}$$

and

$$\bar{x}^I = \frac{\sum_{l=1}^{r_I} (n_l^I * \bar{x}_l^I)}{n^I}$$

σ_E^2 is the variation between the different labels.
σ_R^2 is the variation inner to a label.
This index of connection consists in detecting relationships between the learning set Ω^I and r_I labels. This index is low if the features of these labels are not so different. This gives an information about the repartition of data of learning set Ω^I between membership functions.
We define also a covering level (lcr_i) of a rule labelled i,

$$lcr_i = (\prod_{j=1}^{j=n} (n_{p_j}^j / n^j)$$

where the rule i is :
Rule i : if x_{p_1} is A_{j_1} and x_{p_2} is A_{j_2} and $\cdots$ and x_{p_n} is A_{j_n} then C_i

C. Fuzzy algorithm

Usually, to validate a fuzzy inference system, the mean square error (MSE) is calculated on a test set. If the MSE exceeds a threshold, then training is done, using a gradient method. This consists in modifying C_j at each presentation of examples from the error $(y(\mathbf{x}_j) - d_j)$.
Unfortunately, in case of model invalidation (a MSE too important), we cannot determine never learned rules that cause the gap between the model and the real system. Moreover, if there is an insufficient covering of datum plane, training and finer splitting of input space are inefficient and useless.
With the criterion proposed above, we estimate the datum plane coverage and we are able to isolate inactivated rules. Then, partial remodeling of the fuzzy inference system is possible.

The proposed procedure is as follows :

1) Step 1 Initialization of the fuzzy system and partition of the space in r labels.
2) Step 2 Tuning of membership functions and elimination of bad rules
3) Step 3 Training of the fuzzy system

In Step 1, we determine the number of r rules, the shape and the number of membership functions. In Step 2, we compute the index of connection and measure the adequation between avalaible data and splitting of input space. We identify also the 'bad rules' i.e the rules insufficiently covered, with a covering level lcr_i below a threshold. In Step 3, training is done.

III. Application to the insulin/glycaemia behavior of diabetics

A. The application

Diabetes is a major chronical disease affecting over 200 million people world wide, and for which no efficient treatment exists so far. Diabetes is responsible for numerous complications which cost up to 60 % of the whole diabetes management budget. Following the invention of implantable insulin pumps and continuous glucose sensors, it became possible to envisage developing an artificial pancreas, a computer program located inside the insulin pump and calculating, from the glycemic values measured by the glucose sensor, the insulin values necessary to maintain the normoglycemia.

B. The available data

The correlated data "'insulin infusion delivery/glycaemia"' has been provided by the team of Pr. Pinget, CHU of Strasbourg. They concern the same person and the same insulin.
The insulin infusion has been done by an intra-peritoneal route and the glycaemia has been checked by a subcutaneous sensor. Measures of glycaemia have been made every five minutes during 7 days, which corresponds to 1700 measures. A bolus is a dose of insulin infused manually, in addition to the basic dose, since postprandial glycemia cannot be regulated satisfactorily. The insulin file contains crude data about basic insulin doses as well as boluses. So, a pretraitement of the insulin file has been necessary to produce a file of insulin delivery for the same person every five minutes.

C. The fuzzy system

$y(t) = \sum_{i=1}^{r} \mu_i(a_1^i.y(t - idecal) + a_2^i.u(t - idecal))$
$r = 12$ linear models, considering that each model is valid about three and half hours.
a_1^i is the glucose rate partitionned into 3 triangular membership functions (weak, middle, high)
a_2^i is the insuline dose partitionned into 1 triangular membership functions (middle)
$\mu_i = \prod_{j=1}^{p}(\mu_i^j)$, $i = 1, \cdots, r$ product of membership degrees
$idecal$ is the time lag between input and its effect, which is specially interesting in our application. We have measures about every five minutes and we admit that the effect of insulin (considered in our application)is fast and noticeable ten (idecal=2) minutes later, up to half an hour(idecal = 6).

D. Experiments and Validation of the model

1) Step 1 Initialization: The learning set is composed of the first measures(280 points) that corresponds to insulin infusion and blood glucose concentration of a patient during a day. We have done, first experiments with 3 triangular membership functions for the first input and 4 for the second input.

2) Step 2 Tuning and elimination of bad rules:

- Tuning using index of connection
  The index of connection gives information about repartition of data between membership functions. This repartition is inadequate if there is a gap between inter and intra variation when available data are evenly distributed. We show variations for the first input in this table. We take 2 membership functions. We see

n_1^l	n_1^2	σ_E^2	σ_R^2	s^I	MSE
45	0	0.0	0.06	0.0	non défini
45	113	0.61	0.04	0.96	non défini

TABLE I
TABLE 1.

  that only 45 (158 in the second case) points, by 280 measures, are covered by membership functions (251 after tuning).
  After tuning, the index of connection s^I for the first input is 0.89 (with σ_E^2 = 0.56 and σ_R^2 = 0.14).
  and for the second input 0.0 (with σ_E^2 = 0.0 and σ_R^2 = 0.66), since there is only one membership function.

- Elimination of bad rules using covering level of rules
  We show, in this second table, covering levels (lcr)for the 12 rules. Among the twelve rules, 9 have a covering

rule	lcr
1	0.15
2	0.0
3	0.0
4	0.001
5	0.22
6	0.0
7	0.0
8	0.002
9	0.5
10	0.0
11	0.0
12	0.005

TABLE II
TABLE 2.

  degree less than 0.01. So, they were labelled "bad" and training was done without these rules. We take 3 triangular membership functions for the first input and 1 for the second input.
  We have decreased the number of rules, while preserving the performance of the model (MSE = 0.06). We show, in this third table, the covering levels (lcr)for the 3 rules.

rule	lcr
1	0.16
2	0.22
3	0.51

TABLE III
TABLE 3.

3) Step 3 Training of the fuzzy system: The mean square error(MSE)is calculated on the totality of the measures(1700 points). We make experiments by changing the parameter $idecal$ of the model, time lag between an input and its effect.

r	idecal	MSE
7	2	0.04
7	3	0.06
7	6	0.16
7	24	1.02

TABLE IV
TABLE 4.

The test of our modeling method shows that we can predict the glycaemia over a long period (7 days), by considering glycaemia and insulin delivery 15-minute (resp 30-minute) before with an error of about 6%(resp 16%), which is a good result compared with current results.
However, we see that results obtained are not so good in the last case, when we consider slow effect insulin (with 2-hours delay). In this case, our model has to be refined, by increasing its order.

IV. Application : Detecting fraudulous behavior of a bank customer

A. The application

Along with the development of credit cards, fraudulent activities become a major problem in electronic payment systems. According to [17], fraud is a million dollar business and it's increasing every year. So, fraud detection is becoming an important issue for research.

B. The fuzzy system

Our model is based on specific and usual customer behavior, and deviation from such patterns is suspect. We suppose that a specific profile was defined by transaction amount and shopping time. We take as inputs deviation from usual amount of transaction and deviation from the usual shopping time. The output is suspicion of fraud .

$$y(t) = \sum_{i=1}^{r} \mu_i (a_1^i . u_1(t) + a_2^i . u_2(t))$$

r is the number of linear models,
$z(t)$ a vector which depends lineary or not on the state,
$\mu_i = \prod_{j=1}^{p} (\mu_i^j)$, $i = 1, \cdots, r$ product of membership degrees
$u_1(t)$ is the deviation from usual transaction amount partitionned in 4 triangular membership functions (small middle big very big)
$u_2(t)$ is the deviation from the usual shopping time partitionned in 4 triangular membership functions (small middle big very big)
$y(t)$ is the suspicion of fraud $\in [0, 1]$
We have simulated a set of data by considering that the suspicion of fraud is more serious if the customer begins suddenly to spend a lot of money more than usual. We have eliminated 7 "bad rules" among 16 without decreasing the performance of the model (MSE= 0.07). We got an index of connection for the two inputs equal to 0.90.

V. Conclusion

We have proposed a measure for datum plane covering of a fuzzy inference system and an algorithm of selection of rules based on this measure. From these experiments, we remark that coverage rate is correct under two conditions :

- (i) datum plane and test interval are not disjoined.
- (ii) the ratio number of rules / card (T) is acceptable.

Our feature selection algorithm takes place in the problematic of trade-off between accuracy and complexity. If there is an insufficient covering of datum plane, we show that training and finer splitting of input space are inefficient and useless. We applied, in this paper, our method to the modeling the "insulin delivery/glycaemia" behavior of some patient. The linear model does not seem to be satisfactory, our idea was to use a combined linear/nonlinear model describing the insulin-infusion-rate/glycemia behavior. Our method gives encouraging results over longer time periods (compared to current results).
However, this modeling has to be compared to other existing modeling methods.
There are promising results for the second application. In fact, there is not a strict boundary between a fraud and a regular behavior. So, suspicion of fraud is intrinsically fuzzy. We have to develop this application and test it with real data.

References

[1] K.Kosaki, H.Ishibuchi, H. Tanaka, *A simple but powerful method for generating fuzzy rules from numerical data*, Fuzzy sets and Systems 86 (1997) 251-270

[2] H. Ishibuchi, T. Murata, *Multi-objective genetic local search for minimizing the number of fuzzy rules for pattern classification problems*, Proceedings of The 1998 IEEE International Conference on Fuzzy Systems, 1100-1105, vol.2

[3] R. Alcala, M.J. Gacto, F. Herrera, *a multi-objective genetic algorithm for tuning and rule selection to obtain accurate and compact linguistic fuzzy rule-based systems*, International journal of uncertainty fuzziness and knowledge-based systems, Vol.15, No.5 (2007) 539-557.

[4] D. Chakraborty, N.R. Pal*A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classifier* IEEE Transactions on neural networks, vol.15, January 2004.

[5] T.Guo, G.Y Li *Neural data mining for credit card fraud detection*, 7th International Conference on machine learning and cybernetics, Kunning, 12-15 July 2008.

[6] D. Simic, *Reducing fraud in electronic payment systems*, 7th Conference on operational research, Romania,May 2005.

[7] M.J. Kim, T.S. Kim, *A neural classifier with fraud density map for effective credit card fraud detection*, Proc. International Conference on data engineering and automated learning, LNCS, Springer Verlag, no.2412,pp378-383,2002.

[8] Lutaud M. Identification et Contrle de processus par Rseaux Neuro-Flous. Thse de Doctorat de l'Universit d'Evry Val d'Essonne

[9] Sugeno M.,Yasukawa T. "A Fuzzy-Logic-Based Approch to Qualitative Modeling" , *IEEE Trans. on Fuzzy Systems*, Vol. 1, n^o 1, February 1993.

[10] ,M.G. Singh, X.J. Zeng, *Approximation accuracy analysis of fuzzy systems as function approximators*, IEEE Transactions on fuzzy systems, 4(1):pp44-63, Feb 1996.

[11] A.M. Albisser, B.S. Leibel, T.G. Ewart, et al., *An artificial endocrine pancreas*, Diabetes, 23 (1974), pp. 389-396.

[12] J.L. Selam, P. Micossi, F.L. Dunn et al., *Clinical trial of programmable implantable insulin pump for type I diabetes*, Diabetes Care, 15 (1992) pp. 877-885.

[13] G. Bleckert, U.G. Oppel, and E. Salzsieder, *Mixed graphical models for simultaneous model identification and control applied to glucose-insulin metabolism* ,Comput.Methods Prog.Biomed. 56(1998), no. 2, 141-155.

[14] B. Candas and J. Radziuk, *An adaptive plasma glucose controller based on a nonlinear insulin/glucose model*, IEEE Trans. Biomed. Eng. **41** (1994), no. 2, 116–123.

[15] R.S. Parker, F.J. Doyle III, and N.A. Peppas, *A model-based algorithm for blood glucose control in Type I diabetic patients*, IEEE Trans. Biomed. Eng. **46** (1999), no. 2, 148–157.

[16] Z. Trajanoski and P. Wach, *Fuzzy filter for state estimation of a glucoregulatory system*, Comput. Methods Progr. Biomed. **50** (1996), no. 3, 265–273.

[17] JANS Mieke, LYBAERT Nadine, VANHOOF Koen, *Data Mining for Fraud Detection: Toward an Improvement on Internal Control Systems?*,International Research Symposium on Accounting Information Systems, 7, Milwaukee, 2006.

[18] T. Johansen *Fuzzy model based control: stability,robustness, and performance issues*, IEEE Trans.Fuzzy Syst. 2(3)(1994) 221-234

[19] H. Wang, K. Tanaka, M.Griffin, *An approach to fuzzy control of nonlinear systems: stability and design issues*, IEEE Trans.Fuzzy Syst. 4(1)(1996) 14-23.

[20] S. Cao, N. Rees, G. Feng, *Analysis and design for a class of complex control systems.*, Part I: Fuzzy modelling and identification, Automatica 33(6)(1997) 1017-1028.

[21] J.Gutierrez, F. Fernandez, S. Guadarrama, E. Trillas , "A step towards conceptually improving tagaki-sugeno's approximation",*IPMU'2002*, III:pp 1789-1794, July 2002.

[22] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjberg,and Q. Zhang, *Nonlinear black-box models in system identification: mathematical foundations*, Automatica J. IFAC 31 (1995), no. 12, 1725–1750.

[23] J. Sjberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky, *Nonlinear black-box modeling insystems identification: a unified overview*, Automatica J. IFAC 31 (1995), no. 12, 1691–1724.