Data Driven Generation of Fuzzy Systems: An Application to Breast Cancer Detection

Antonio d'Acierno[1,⋆], Giuseppe De Pietro[2], and Massimo Esposito[2]

[1] Institute of Food Sciences - Italian National Research Council
Via Roma 64, Avellino, Italy
dacierno.a@isa.cnr.it

[2] Institute of High Performance Computing and Networking
Italian National Research Council
Via Pietro Castellino 111, Napoli, Italy

Abstract. The detection of diseases often can be formalized as a decision problem that typically has to be solved merging uncertain information; diagnostic tools, intended to aid the physician in interpreting the data, besides attaining the best possible correct classification rate, should furnish some insight into how the problem has been decided. Fuzzy logic is a well known successful attempt to automatize the human capability to reason with imperfect information; fuzzy systems are rule-based so that they can easily provide motivations for their decisions, after having verified some additional conditions.

In this paper we describe a six-steps data driven methodology to automatically build fuzzy systems with a user defined number of rules; almost each step can be approached using several strategies and we thus describe an implementation of the proposed solution. Then, we test our systems on a well known and widely used data set of features of images of breast masses and, having the number of rules varying, we show results both in terms of correct classification rates and in terms of systems' confidence in the obtained decisions. Finally, we select the number of rules that produces the most interpretable and *trustworthy* system; such a system is described in details and tested.

1 Introduction

To increase the change of successful treatments, early detection of almost any disease is a key factor and the *detection* can be often formulated as a binary decision making problem; uncertainty in form of information incompleteness, impreciseness, fragmentariness, not fully reliability, vagueness and contradictoriness often affects these problems [7] so that the ultimate diagnosis can be difficult to obtain even for a medical expert. As a consequence, many computerized diagnostic tools intended to aid the physician in interpreting the data have been developed in the past few decades.

⋆ Corresponding author.

R. Rizzo and P.J.G. Lisboa (Eds.): CIBB 2010, LNBI 6685, pp. 203–214, 2011.

It is widely accepted that a diagnostic tool should possess three characteristics [4]; first, it must attain the best possible performance in terms of correct classification rate (CR) while it would be desirable the system not only provides a diagnosis but also a numerical value (the *confidence* χ) representing the degree to which the system is confident in the solution. It would be also useful if the physician is not faced with a black box that simply outputs answers but the system should provide some insight into how the solution has been derived (*interpretability*). These requirements are often in contrast. Diagnostic tools, however, typically have unequal classification error costs so that straight CR cannot be assumed as a careful measure of the goodness of the classifier; a Receiver Operating Characteristic (ROC) graph [3] has been showed to be a more accurate technique for selecting classifiers based on their performance. We guess that also χ can be used for selecting classifier; in facts, a *good* classifier should be highly confident with correctly classified examples while it should be doubtful with misclassified data points.

Fuzzy logic (a precise logic of imprecision and approximate reasoning [16]) is an attempt at the formalization and mechanization of two remarkable human capabilities: the capability to converse, reason, and make rational decisions in an environment of imperfect information, and the capability to perform a wide variety of physical and mental tasks without any measurement and any computation [16]. Fuzzy logic is a multi-valued logic based on fuzzy set theory [15]; a fuzzy set is a set whose elements have a continuum of grades of membership described using *membership functions.*

A Fuzzy Inference System (FIS) is a system that (tries to) solve a (typically complex and nonlinear) problem by utilizing fuzzy logic methodologies and it is composed (see figure 1) of a fuzzifier (which translates real-valued inputs into fuzzy values), of an inference engine (that applies a fuzzy reasoning mechanism to obtain a fuzzy output), of a defuzzifier (to translate this latter output into a crisp value), and of a knowledge base (containing both rules and membership functions).

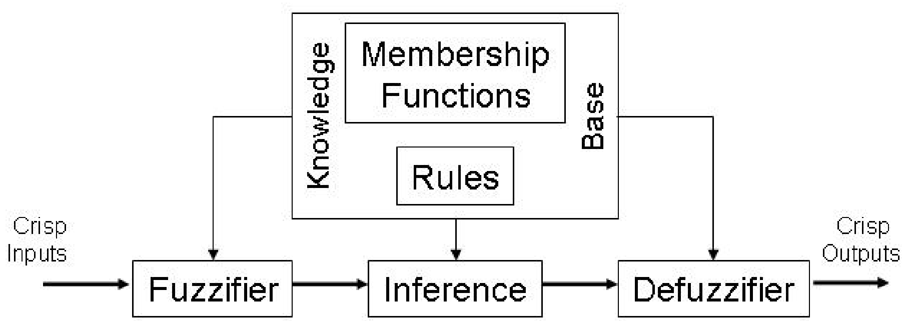

Fig. 1. The basic components of a fuzzy inference system

The inference process is performed by the engine using the rules contained in the rule base, each rule being in the form:

if *antecedent* **then** *consequent*

where the antecedent is a fuzzy-logic expression composed of one or more simple fuzzy expressions connected by fuzzy operators (the fuzzy equivalent of the classical *and*, *or* and *not*), and the consequent is an expression that assigns fuzzy values to the output variables (Mamdani systems [8]), i.e.:

if *service* **is** *good* **then** *tip* **is** *average*

Differently, in Takagi-Sugeno(TS) systems [11], the consequent expresses output variables as a function that maps the input space into the output space, for example:

if *service* **is** *good* **then** $tip = f(service)$

where f (typically) is a first order linear function that becomes a constant in zero-order TS systems.

Fuzzy modeling is the task of identifying the parameters of a FIS so that a desired behavior is attained. When the available knowledge is complete and the problem space is not very large the system can be constructed directly (*knowledge driven approach*) using knowledge elicited from human experts. Alternatively, an emerging solution is represented by *data driven fuzzy modeling*, that is being more and more applied in a wide variety of fields even if the rule base generated automatically from data may not be fully interpretable especially because of redundancy in the rules and in the fuzzy sets. Three conditions can be defined [5] to obtain an interpretable fuzzy model: *(i)* the fuzzy partition must be readable (the fuzzy sets can be interpreted as linguistic labels), *(ii)* the set of rules must be as small as possible, and *(iii)* the if-part of the rules should be derived from a subset of independent variables rather than from the full set.

In this work we describe a methodology to automatically extract the knowledge base and we show some of the results obtained with reference to a well known and widely used data set of features of images of breast masses. In our approach, the number of rules selected is user-defined and the optimal number of rules for the dataset under test has been defined by exploiting not only the correct classification rate but also a confidence based criterion with the final aim of obtaining an highly understandable system with an interesting overall performance.

The paper is organized as follows. Some relevant related approaches available in the literature are reviewed in section 2 while the proposed method is described in section 3. Experimental conditions and results are discussed in section 4, conclusions and open problems being the concerns of section 5.

2 Related Works

The method proposed in [10] to generate TS fuzzy models firstly assumes that fuzzy sets are described by Gaussian membership functions for which centers and widths have to be estimated; the rules are then generated iteratively until a user defined maximum number of rules (R_{max}) is reached or a performance index (usually MSE) is achieved. The fuzzy antecedents of the first rule are evaluated as the mean and the standard deviation of training data while the consequent is evaluated using least squares techniques; rules are iteratively added selecting as center the vector with the worse error (some conditions are introduced to exclude the change of an outlier data point to be considered as a new rule's center). Parameters are also tuned using an hybrid learning algorithm. Even if the presented results show the goodness of the method for problems with a low number of independent variables, the antecedent of each rule is based on the whole set of variables so making the obtained fuzzy model not really interpretable when problems with many independent variables are considered [5].

In [4], 11 feature selection methods and 3 fuzzy modeling methods are combined and tested using two well known medical binary datasets (namely the Wisconsin breast cancer data [14] and the PIMA Indians diabetes data, both available at UCI Repository) and an industrial dataset (welding flaw data). Results for a single run of a stratified fivefold cross validation are presented in terms of average accuracy, also highlighting the top five combinations. Then, only for the best combination, the results are also reported in terms of area under the ROC curve, sensitivity and specificity (among others parameters). With reference to the WBCD, the best average accuracy obtained is 97.17% using three variables (x_{28}, x_{21}, x_{22}); no details are given for the rules in the first rank system while, for the second rank, we have that for each variable (the same as the first rank system) 5 membership functions have been used and 250 rules have been generated so deriving a not fully understandable system.

Since in rule based systems built starting from numerical data redundancy often exists in form of redundant rules and similar fuzzy sets generating unnecessary structural complexity and decreasing the interpretability of the system, in [1] a simplification method is proposed after rules' extraction (by means of fuzzy clustering associated with a fuzzy partition validity index) and parameters' estimation (by means of a gradient descent algorithm). Results are shown with reference to function approximation, dynamical system identification and mechanical property prediction for hot rolled steels. No result are reported for problems with an high number of variables.

Hierarchical TS fuzzy systems have the advantage that both the number of rules and the number of fuzzy operations involved can be reduced significantly when compared with those requested by single level systems. An automatic way of evolving hierarchical TS fuzzy system using probabilistic incremental program evolution is the concern of [2]; interesting results are shown for some non linear system identification problems (Makey-Glass chaotic time series prediction problem, and the Iris and Wine classification problems).

In [13] it is presented a fuzzy rule based decision support systems for the diagnosis of coronary artery disease (CAD) automatically generated from an initial annotated dataset, by means of a four stage methodology. A set of crisp rules (obtained from a postpruned induction tree based on the well known C4.5 algorithm) are fuzzified using two sigmoidal membership functions (a decreasing one expressing the linguistic term LOW and an increasing one expressing $HIGH$). Rules are then weighted using a likelihood ratio and parameters are optimized using the healed topographical multilevel single linkage algorithm. The reported results clarify that fuzzification and optimization significantly improve the performance of the pruned tree. No information is shown on the obtained rules, so that the overall interpretability of the obtained system cannot be judged.

Genetic algorithms are used in [9] to produce fuzzy systems for an older version of WBCD (444 benign cases and 239 malignant cases, 9 variables) using a fitness function that tries to combine classification performance, the interpretability of the system and a term adding pressure towards systems with low quadratic error. For each variable there are two orthogonal trapezoidal membership functions and the number of rules is assumed to be a user-configurable parameter (limited to be between 1 and 5). After 120 evolutionary runs, the best system has 3 rules for benign cases (one of them is found to be never triggered by any of the input case) and shows a $CR = 98\%$ on the whole data set. Cross validation has been also performed but the choice of learning-set and test-set is performed anew at the outset of every evolutionary run, so deriving not fully generalizable results.

Last, it is worth citing the three stages generic methodology (crisp rules extraction, fuzzification and optimization) proposed in [12] that is able to integrate alternative techniques in each stage. Specific implementations (using decision trees for crisp rules extractions and four optimization strategies) are tested on several well known datasets; on WBCD, the best solution obtains, in a single run of a ten-fold cross validation, a $CR = 95.15$. The number of rules used equals the number of classes, but each rule is not really interpretable since it combines all the crisp rules of a given class.

3 The Proposed Approach

The main components of the implemented system are sketched in figure 2. First, the available data are used to extract crisp rules; since in the current implementation a decision tree is used, each leaf node can be easily translated into a crisp rule parsing the tree from the root to the leaf itself and assuming the tests encountered along the path form the conjunctions of the rule's antecedent, while the class label of the leaf node is clearly assumed to be the rule consequent (zero order TS FIS will be so used).

Here, having in mind to implement a general methodology, we do not adopt any pruning technique, so that we obtain several rules that are likely to over fit the data; for such a reason, we use a well defined stage (the *Selector*) that using some heuristics selects a proper subset of rules for each class (we choose the same number of rules for each class).

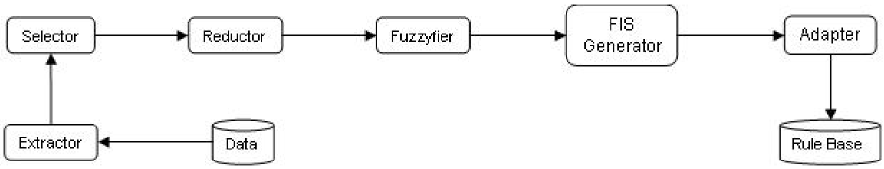

Fig. 2. The proposed approach

The antecedent of the selected i^{th} rule (in the current implementation it clearly corresponds to the path path from the root node to a given leaf) we obtain is in the form of conjunction of conditions:

$$A_i = (x_{1i}\theta_{1i}c_{1i}) \wedge (x_{2i}\theta_{2i}c_{2i}) \wedge \ldots \wedge (x_{ki}\theta_{ki}c_{ki}) \tag{1}$$

where each x_{ji} is the feature used in the node, θ_{ji} is a standard comparison operator $(<, \leq, >, \geq)$ and c_{ji} is a crisp threshold.

Depending on the strategy used to extract the rules, each antecedent A_i is generated with some characteristics. For example, when a decision tree is used, as in the current implementation, we have that, in an antecedent, the same feature could appear several times (see figure 3), and, consequently, we have implemented a *Reductor* that translates each antecedent C_i in an antecedent we have named in *standard form* $\tilde{A}_i$: an antecedent is in standard form if each feature appears at most one time. Given the form of the antecedent (equation 1), and $\forall i$, we act as follows. We collect the terms that refer to the same feature and with the same operator; because of the conjunctions among conditions, these sets can be easily simplified as follows (we assume, without loss of generality and for the sake of simplicity, that just two conditions share the same feature):

$$(x_{mi}\theta_{mi}c_{mi}) \wedge (x_{mi}\theta_{ni}c_{ni}) \rightarrow x_{mi} > max(c_{mi}, c_{ni}) \tag{2}$$

if $\theta_{ni} \in \{>, \geq\}$, or

$$(x_{mi}\theta_{mi}c_{mi}) \wedge (x_{mi}\theta_{ni}c_{ni}) \rightarrow x_{mi} < min(c_{mi}, c_{ni}) \tag{3}$$

if $\theta_{ni} \in \{<, \leq\}$.

After this step we have that the i^{th} condition is already in standard form or some features appear in two conditions (greater than a threshold and less than another threshold, i.e. *between* two thresholds); if this is the case we apply:

$$(x_{ji} > c_{mi}) \wedge (x_{ji} < c_{ni}) \rightarrow x_{mi} \in [c_{mi}, c_{ni}] \tag{4}$$

if $c_{mi} < c_{ni}$, otherwise we simply delete the rule.

Once rules are expressed with antecedents in standard form, they can be easily expressed in fuzzy terms; we have chosen to partition the universe of each feature into three intervals: *low*, *medium* and *high*. Then, each condition for each antecedent is obviously translated as follows:

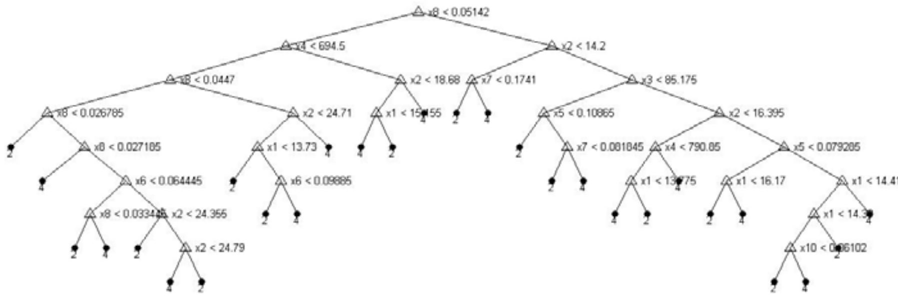

Fig. 3. A decision tree

$$(x_{ji} < c_{mi}) \rightarrow x_{ji} \;\; is \;\; low \tag{5}$$

$$(x_{ji} \in [c_{mi}, c_{ni}]) \rightarrow x_{ji} \;\; is \;\; medium \tag{6}$$

$$(x_{ji} > c_{mi}) \rightarrow x_{ji} \;\; is \;\; high \tag{7}$$

the consequent being the class associated with the leaf node.

In the last step, membership functions have to be adjusted and tuned; a plethora of methods for such a task have been proposed, so here it is just worth to be noted that a back-propagation algorithm is used in the current implementation.

4 Experimental Results

4.1 The Wisconsin Breast Cancer Dataset

Breast cancer is one of the most common cancer among women; the presence of a breast mass is an alert sign, but it does not always indicate a malignant cancer. Fine needle aspiration (FNA) of breast masses is a cost-effective, non-traumatic, and mostly non-invasive diagnostic test that obtains information needed to evaluate malignancy.

The Wisconsin breast cancer diagnosis (WBCD) database [14] is the result of the efforts made at the University of Wisconsin Hospital for accurately diagnosing breast masses based solely on an FNA test. Features computed describe characteristics of the cell nuclei present in the image, and ten visually assessed characteristics of an FNA sample considered relevant for diagnosis were identified (see table 1).

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. The diagnostics in the WBCD database were furnished by specialists in the field; the used version of the database consists of 357 benign cases and 212 malignant cases. Having in mind to obtain highly understandable systems, we decided to use just the first 10 variables that can be easily estimated simply having a look at the image.

Table 1. The variables of the Wisconsin dataset

1	radius	mean of distances from center to points on the perimeter
2	texture	standard deviation of gray-scale values
3	perimeter	
4	area	
5	smoothness	local variation in radius lengths
6	compactness	perimeter^2 / (area - 1.0)
7	concavity	severity of concave portions of the contour
8	concave points	number of concave portions of the contour
9	symmetry	
10	fractal dimension	"coastline approximation" - 1

4.2 Experimental Parameters

The proposed approach has been implemented using functions available in the standard version of the R2007a 64-bit version of MATLAB.

In the current implementation, to extract crisp rules we use a full decision tree (without pruning) with a Gini's diversity index as split criterion and a split minimum factor equal to 1. Given a user defined number of rules R (assumed to be even), we select, for each class, the $\frac{R}{2}$ most covering leaf nodes. For each selected node, we derive a standard if-then rule that is then fuzzified.

The outcoming FIS is finally trained using the well know ANFIS [6] algorithm with back-propagation and assuming 500 as the maximum number of epochs and 0.1 as the training error goal. For each FIS we compare the defuzzified output with a threshold τ to classify the sample; the chosen τ is the one that maximizes the CR on the learning set.

4.3 Numerical Results

To test the system we use a ten-fold cross validation that is repeated 100 times and we measure the CR on the learning set (LS), on the test set (TS), and on the full set (FS) for both unadapted FISs (UFIS) and adapted ones (AFIS); averaged correct classification rates ($\overline{CR}$) as the number of rules varies are reported in table 2.

First, it should be noted that also unadapted FISs show an interesting performance ($\overline{CR} > 90\%$ on the TS). An interesting feature is that, with some approximation, $\overline{CR}$ on TS (clearly the most interesting one) increases with the number of rules for unadapted systems while it decreases for adapted systems. The best result we obtained is with 4 rules where we have a $\overline{CR} \approx 93\%$ on the TS. We also show the average number of variables used (V) that increases as the number of rules does.

For each FIS we evaluate the ROC graphs (on the test sets) that are then vertically averaged [3] to compute the area under the curve (AUC, equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance) as the number of rules varies.

Table 2. Averaged correct classification rate as the number of rules varies. Last column shows the average number of variables used.

	UFIS			AFIS			
R	LS	TS	FS	LS	TS	FS	V
2	92.01	90.63	91.32	93.15	92.22	92.69	**5.21**
4	92.14	90.95	91.54	93.48	**92.66**	93.07	6.23
6	91.77	90.45	91.11	93.22	92.34	92.78	7.19
8	92.12	90.81	91.46	92.62	91.68	92.15	7.70
10	92.19	91.00	91.59	92.22	91.13	91.68	8.10
12	92.27	91.13	91.70	92.09	91.02	91.55	8.26

Table 3. The area under the ROC graph as the number of rules varies

R	2	4	6	8	10	12
AUC	0.965	0.969	0.972	0.970	0.969	0.967

Results (table 3) show that the number of rules does not significantly affects the AUC.

At each iteration, and for each sample in TS, we define the confidence χ as:

$$\chi = \frac{|D - \tau|}{\alpha} \tag{8}$$

D being the defuzzified output of the FIS and α a normalizing factor so that $\chi \in [0, 1]$.

Starting from the confusion matrix on the training set of each system, we measure the cases correctly classified (N_{TP} and N_{TN}) with $\chi > 0.7$ and the number of instances incorrectly classified (N_{FP} and N_{FN}) with $\chi < 0.3$. Table 4 shows the results we obtained (in percentage) as the number of rules varies on the test sets. The most trustworthy system (on average, and also in detecting both false positives cases and true negatives ones) is the system with just two rules while, for example, a system with six rules is highly confident in detecting true positives samples.

Table 4. Number of cases (in percentage) with an *appropriate* χ

R	N_{TP}	N_{FP}	N_{FN}	N_{TN}	AVG
2	64.16	**44.32**	51.21	**70.92**	**57.65**
4	77.09	34.75	58.54	55.45	56.46
6	**81.98**	32.08	**61.53**	31.82	51.85
8	64.78	34.55	49.68	31.85	45.22
10	54.33	37.29	41.25	44.81	44.42
12	50.62	37.35	40.51	46.15	43.66

4.4 The Selected FIS

The best classification rate is obtained with four rules while, with just two rules, we have the most trustworthy (on average) system; here we decide to favor both the interpretability and the overall confidence of the system and so, as a result, we chose the system with two rules. This system, moreover, shows the best confidence in detecting true negative and false positive cases; figure 4 shows the confidences' distribution obtained on the test sets in the experiments.

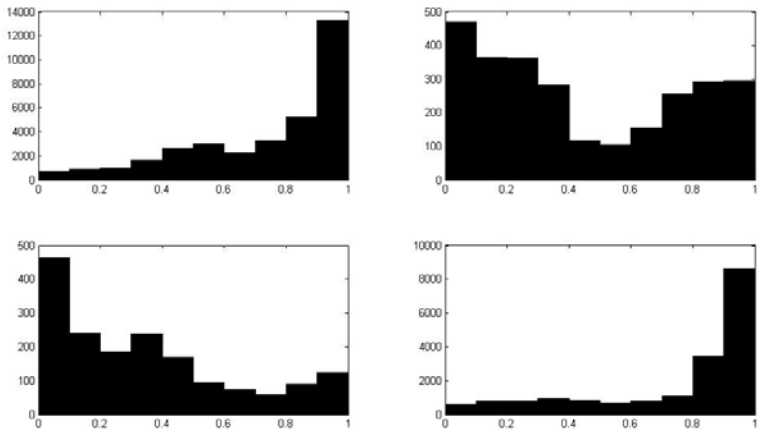

Fig. 4. The distribution of the confidences on the test sets in the experiments by the two rules systems

We then apply our methodology using the full data set and we and obtain the following two rules:

- Rule 1. If (texture is low) and (area is low) and (concave points is low) then (Class is Benign)
- Rule 2. If (texture is high) and (perimeter is high) and (smoothness is high) and (concavity is high) and (concave points is high) then (Class is Malignant)

that use just six variables and at most two linguistic terms (low and high). Membership functions are then adapted; figure 5 (left) shows the root mean square errors obtained, showing that the training process converges quickly and the adapted membership functions are shown in figure 6. The trained system uses a threshold value $\tau = 3.038$ (benign cases are represented with 2 while malignant ones with 4), has the ROC curve shown in figure 5 (right), its area under the curve is 0.972 and the correct classification rate is 93.85 (both on the full data set).

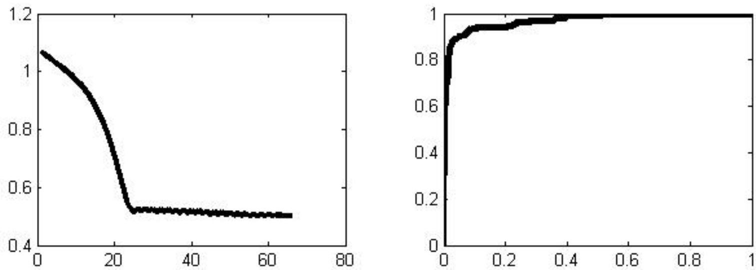

Fig. 5. The training error for the selected system (left) and the ROC curve (right)

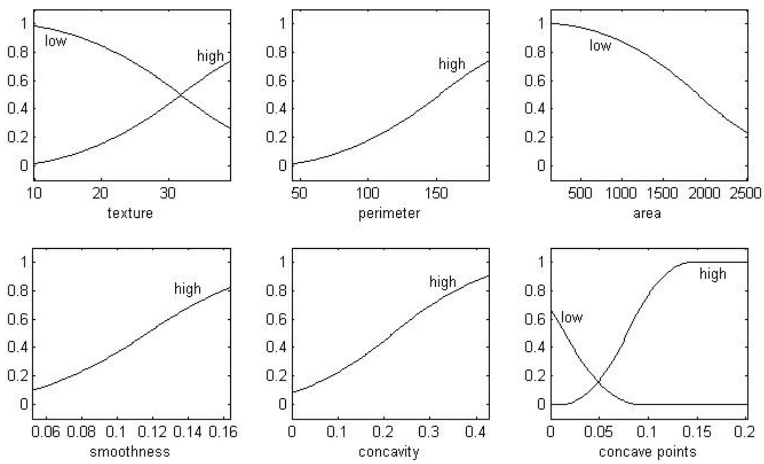

Fig. 6. Themembership functions of the selected system

5 Conclusions and Future Directions

Data driven methodologies in fuzzy modeling are being applied in a wide variety of fields even if attention must be paid when the aim is to obtain interpretable systems. In this paper we describe a six steps general methodology aimed at producing FISs with an user defined number of rules where each step could be carried out in several ways. We test an implementation of the proposed methodology on a well known data set of features of images of breast masses and, having the number of rules varying, we show results both in terms of correct classification rates and in terms of systems confidence in the obtained decisions. Last, using the number of rules able to produce the most interpretable and trustworthy systems, we derive our best system that is described in details.

Concerning our future directions, several questions remain open. First, our methodology needs to be tested with other data sets to be fully considered valuable; we are also planning to test different strategies for extracting rules.

Different techniques to determine the correct thresholds for the FISs (for example choosing the τ that minimizes the mean square error)are also being considered. More promisingly, the optimal threshold could be chosen optimizing an appropriate function that takes into account the unequal classification error costs.

Finally, it is in our opinion worth probing the possibility of using in parallel FISs with different numbers of rules; their predictions could be combined using several strategies based on the confidence showed by each system.

References

1. Chen, M.Y., Linkens, D.A.: Rule-base self-generation and simplification for data-driven fuzzy models. Fuzzy Sets and Systems 142(2), 243–265 (2004)
2. Chen, Y., Yang, B., Abraham, A., Peng, L.: Automatic design of hierarchical takagi-sugeno type fuzzy systems using evolutionary algorithms. IEEE T. Fuzzy Systems 15(3), 385–397 (2007)
3. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006)
4. Ghazavi, S.N., Liao, T.W.: Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine 43(3), 195–206 (2008)
5. Guillaume, S.: Designing fuzzy inference systems from data: An interpretability-oriented review. IEEE Transactions on Fuzzy Systems 9(3), 426–443 (2001)
6. Jang, J.S.R.: Anfis: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man and Cybernetics 23(3), 665–685 (1993), `http://dx.doi.org/10.1109/21.256541`
7. Klir, G., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice-Hall, Englewood Cliffs (1995)
8. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7(1), 1–13 (1975)
9. Pena-Reyes, C.A., Sipper, M.: A fuzzy-genetic approach to breast cancer diagnosis. Artificial Intelligence in Medicine 17(2), 131–155 (1999)
10. Rezaee, B., Zarandi, M.F.: Data-driven fuzzy modeling for takagi-sugeno-kang fuzzy system. Information Sciences 180(2), 241–255 (2010)
11. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics 15(1), 116–132 (1985), `http://www.hi.cs.meiji.ac.jp/~takagi/paper/TS-MODEL.tar.gz`
12. Tsipouras, M.G., Exarchos, T.P., Fotiadis, D.I.: A methodology for automated fuzzy model generation. Fuzzy Sets Syst. 159(23), 3201–3220 (2008)
13. Tsipouras, M.G., Exarchos, T.P., Fotiadis, D.I., Kotsia, A.P., Vakalis, K.V., Naka, K.K., Michalis, L.K.: Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Transactions on Information Technology in Biomedicine 12(4), 447–458 (2008)
14. Wolberg, W.H., Street, N., Mangasarian, O.L.: UCI machine learning repository (1995), `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`
15. Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 3385–353 (1965)
16. Zadeh, L.A.: Is there a need for fuzzy logic? Inf. Sci. 178(13), 2751–2779 (2008)