Applying Interval Type-2 Fuzzy Rule Based Classifiers Through a Cluster-Based Class Representation

J. Navarro, C. Wagner and U. Aickelin
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
Email: psxfjn@nottingham.ac.uk
christian.wagner@nottingham.ac.uk
uwe.aickelin@nottingham.ac.uk

***Abstract*—Fuzzy Rule-Based Classification Systems (FRBCSs) have the potential to provide so-called interpretable classifiers, i.e. classifiers which can be introspective, understood, validated and augmented by human experts by relying on fuzzy-set based rules. This paper builds on prior work for interval type-2 fuzzy set based FRBCs where the fuzzy sets and rules of the classifier are generated using an initial clustering stage. By introducing Subtractive Clustering in order to identify multiple cluster prototypes, the proposed approach has the potential to deliver improved classification performance while maintaining good interpretability, i.e. without resulting in an excessive number of rules. The paper provides a detailed overview of the proposed FRBC framework, followed by a series of exploratory experiments on both linearly and non-linearly separable datasets, comparing results to existing rule-based and SVM approaches. Overall, initial results indicate that the approach enables comparable classification performance to non rule-based classifiers such as SVM, while often achieving this with a very small number of rules.**

I. Introduction

Fuzzy Rule-Based Classification Systems (FRBCSs) have been successfully applied as autonomous classifier and decision support systems in numerous classification problems since they provide a powerful platform to deal with uncertain, imprecise and noisy information while providing good interpretability in the form of IF-THEN rules (e.g.: [1], [2], [3], [4] and [5]).

FRBCs as well as Fuzzy Logic Systems can be classified by the type of Fuzzy Sets used, namely *type*-1 and *type*-2. In this context, performance improvements/advantages of the use of Interval type-2 Fuzzy Sets (IT2 FSs) and their applications over the type-1 counterpart have been found in several applications and fields, such as: IT2 FSs used in fuzzy clustering [6], fuzzy logic control to video streaming across IP networks [7], fuzzy logic modelling and classification of video traffic using compressed data [8] and classification of imaginary hand movements in an EEG-based Brain-Computer Interface [9]). These improvements have been attributed due to the additional degrees of freedom for uncertainty modelling in IT2 FSs and, in the case of classification problems, their capability to define/represent abstract classes.

There are many methods to generate fuzzy rules from known data, including heuristic approaches [10], genetic algorithms [11] [12] [13], neuro-fuzzy techniques [14] [15], data mining techniques [16] [17], and clustering methods [18] [2]. In [2], the use of an algorithm called Subtractive Clustering (based on determining a potential value as cluster center to each point in terms of its distance to the rest of the points in the space) helped to find rules for pattern classification by partitioning the input space into appropriate fuzzy regions for separation of classes.

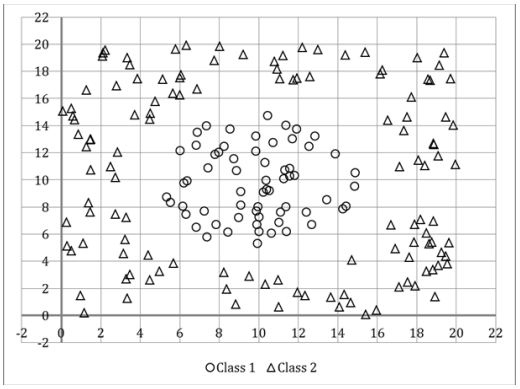

Fig. 1. An example of a non-linear classification problem between 2 classes in which one class is surrounded by a second class of different distribution and density.

The motivation of this paper builds on an a improvement to the IT2 FRBC model proposed in [19] (based on rule generation method [3] derived from Wang and Mendel algorithm [20]) in which, the selection of more clusters for the representation of a single class is suggested but not explored. Building on this, this paper seeks to explore the fuzzy rules extraction method to estimate and select numbers and positions of representative cluster prototypes which together represent classes in linear and non-linear classification problems (such as depicted in the example in Fig. 1).

Nonlinear classification problems have been addressed through different approaches, such as establishing hierarchical

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.253

connections based on the distance between objects (e.g. Decision Trees [21], K-Nearest Neighbour [22]), combining linear classifiers (e.g. Artificial Neural Networks [23]), learning a set of mappings from input-output relationships by changing the feature representation through a kernel (e.g. non-linear SVMs [24]), etc.

The overall aim in this paper is to generate FRBCs which provides good classification performance while maintaining good interpretability and thus, a low number of IF-THEN rules. In order to achieve our goal, we proposed the following developments to the FRBC model introduced in [19].

1) While in [19], the initial centroids representing the classes are calculated as the mean of the training patterns, we proposed the use of SC to identify initial cluster prototypes.
2) The resulting improvement must be able to provide superior performance in particular for non-linear problems, whereas the FRBC model proposed in [19] is suitable for problems with circular distributions of clusters (of different size/density) but, it is not suitable for classification problems involving non-linearly separable classes.

While SC provides the potential for such improved performance of the FRBC framework, in particular in non-linear classification problems, SC is highly sensitive to initial parameters which influence the number of clusters generated. As an increasing number of clusters results effectively in an increasing number of rules within the FRBC, we conduct an analysis on a number of publicly available and synthetic datasets, exploring the relationships between good interpretability (i.e. fewer rules) and good classification performance.

In Section II we present an introduction to interval type-2 fuzzy sets, Subtractive Clustering and general structure of Fuzzy Rule-Based Classification Systems. In Section III, we describe the approach used to represent a class through different clusters by taking SC as a method for finding potential representative clusters. In Section IV, the results of different numbers of clusters used to represent the classes in the FRBCs and comparisons against SVMs are being shown by considering 2 classic applications (Iris Plant and Wisconsin datasets) and two synthetic nonlinear classification problems. Conclusions and future work are presented in last section.

II. Background

In order to ease the reader's understanding along the paper, we include a Table of symbols (see Table I) used often in most sections.

TABLE I. Table of symbols

Symbol	Description
$\mathbf{x}$	Pattern
n	number of training patterns
r_a	The radius defining a neighbourhood in Subtractive Clustering
$j = 1, , M$	Number of classes
$k = 1, , c$	Number of rules/clusters
l_j	Number of clusters related to class j
r_j^k	Certainty degree
A	IT2 Fuzzy set
P_i	Potential of x_i pattern
C	Set of M classes
f_Q	Quasiarithmetic mean operator
m_1	Lower fuzzifier
m_2	Upper fuzzifier

A. Subtractive Clustering

Subtractive Clustering is a fast and robust method for estimating the number and location of cluster centers present in a collection of data points [25]. It has been widely used as a method for initial centroid selection in clustering methods such as FCM and Kohonens Self-Organizing Maps [26] and shown to deliver good performance. It was proposed as an extension to the Yager and Filev's mountain method [27] in which each data point is considered as a potential cluster center to be accepted or rejected. In general terms it works as follows: consider a collection of n data points in which the potential as cluster center of the point x_i is defined initially as:

$$P_i = \sum_{j=1}^{n} e^{-\alpha \|x_i - x_j\|^2} \tag{1}$$

where $\alpha = 4/r_a^2$ and r_a is a positive constant. Thus, the measure of the potential for a data point is a function of its distances to all other data points. A data point with many neighboring data points will have a high potential as cluster center. The constant r_a is effectively the radius defining a neighborhood and acts as a threshold to accept or refuse new cluster centers; data points outside this radius have little influence on the potential.

After the potential of every data point has been computed, we select the data point with the highest potential as the first cluster center. Let x_1^* be the location of the first cluster center and P_1^* its potential value. Secondly, the potential of each data point x_i is revised using:

$$P_i = P_i - P_1^* e^{-\beta \|x_i - x_1^*\|^2} \tag{2}$$

where $\beta = 4/r_b^2$ and r_b is a positive constant. Thus, the potential value of each point is reduced by subtracting an amount of potential as a function of its distance from the first cluster center. Also, the data points near the first cluster center will have greatly reduced their potential, and therefore will unlikely be selected as the next cluster center. Commonly, the value of r_b is set to $r_b = 1.25 r_a$.

After the potential of all data points has been revised according to (2) we select the data point with the highest remaining potential as the second cluster center. We then further reduce the potential of each data point according to their distance to the second cluster center. In general, after the *k*th cluster center has been obtained, the potential of each data point is revised similarly and, through an iterative algorithm, subsequent cluster centers are being accepted or rejected until a condition based on the threshold r_a is met. For further details on the algorithm can be found in [25].

B. Interval Type-2 Fuzzy Sets

Fuzzy Set theory was introduced by Zadeh in 1965 [28] and has been successfully applied in numerous fields in which uncertainty is involved. Type-2 Fuzzy Sets were introduced in 1975 [29] and their application has been shown to provide very good results in situations where lots of uncertainties are present [30].

A T2 FS, denoted $\widetilde{A}$, characterized by a *type-2* MF $\mu_{\widetilde{A}}(x,u)$, where $x \in X$ and $u \in J_x \subseteq [0,1]$, i.e.,

$$\widetilde{A} = \left\{ \left((x,u), \mu_{\widetilde{A}}(x,u) \right) | \forall x \in X, \forall u \in J_x \subseteq [0,1] \right\} \quad (3)$$

If all $\mu_{\widetilde{A}}(x,u) = 1$ then $\widetilde{A}$ is an Interval *Type-2* FS (IT2 FS).

Note that in IT2 FSs, since the third dimension (secondary grades) does not convey additional information, the Footprint of Uncertainty [31] is sufficient to completely describe an IT2 FS. This simplification has proven to be useful to avoid the computational complexity of using general type-2 fuzzy sets (T2 FSs) [31] and has been used in several fields, being pattern recognition one of them.

C. Fuzzy Rule Based Classification Systems

Pattern classification involves assigning a class C_j from a predefined class set $C = C_1, ..., C_M$ to a pattern $\mathbf{x}$ in a feature space $\mathbf{x} \in \mathbb{R}^N$. The purpose of a Classification System is to find a mapping $D : \mathbb{R}^N \rightarrow C$ optimal in the sense of a criterion that determines the classifier performance.

A FRBC consists of a Rule Base (RB) and a Fuzzy Reasoning Method (FRM) to derive class associations from the information provided by the set of rules fired by a pattern.

1) Fuzzy Rules Structure: Considering a new pattern $x = (x_1, ..., x_N)$ and *M* classes, three basic rule structures are identified in FRBCs according to [3]:

a) *Fuzzy rules with a class in the consequent*: This type of the rule has the following structure:
IF x_1 is A_1^k and ... and x_N is A_N^k **THEN** *Y* is C^k
where $x_1, ..., x_N$ are selected features for the classification problem, $A_1^k, ..., A_N^k$ are linguistic labels modeled by fuzzy sets, and C^k is one of the predefined classes from the set $C = C_1, ..., C_M$
b) *Fuzzy rules with a class and a certainty degree in the consequent*: This type of the rule has the following structure:
IF x_1 is A_1^k and ... and x_N is A_N^k **THEN** *Y* is C^k with r^k
where $r^k \in [0,1]$ is the certainty degree of the *k* rule. The value r^k can also be understood as a rule weight.
c) *Fuzzy rules with certainty degree for all classes in the consequent*: This type of the rule has the following structure:
IF x_1 is A_1^k and ... and x_N is A_N^k **THEN** $(r_1^k, ..., r_M^k)$
where $r_j^k \in [0,1]$ is the certainty degree for rule *k* to predict the class C_j for a pattern belonging to the fuzzy region represented by the antecedent of the rule.

2) Fuzzy Reasoning Method: Given an input pattern $\mathbf{x}$, conclusions can be derived using a FRM based on the RB. The inference process of a FRM for IT2 FSs is summarized below (this is based on the type-1 formulation given in [19]):

1) *Matching degree*. The strength of activation of the antecedent for rule k in the Rule Base (RB) with the pattern $\mathbf{x} = (x_1, ..., x_N)$ is calculated by using a function T which is a *t*-norm (commonly minimum *t*-norm).

$$R^k(\mathbf{x}) = T(\mu_{A_1^k}(x_1), ..., \mu_{A_N^k}(x_N)) \quad (4)$$

2) *Association degree*. The association degree of the pattern $\mathbf{x}$ with the *M* classes is computed in each rule by calculating the product:

$$b_j^k = \left(R^k(\mathbf{x}) \cdot r_j^k \right), \quad (5)$$

where r_j^k stands for the certainty degree provided in rule k for the class j.
3) *Weighting function*. To weight the obtained association degrees through a function g. Commonly, a weighting function boosts big output values and suppresses small output values.
4) *Pattern classification soundness degree for all classes*. An aggregation operator f (for example the Quasiarithmetic mean operator, see below) is used to combine (the positive degrees of association calculated in the previous step) into a single value Y_j, which is the soundness degree of pattern $\mathbf{x}$ with class C_j.

$$Y_j = f\left(b_j^k, k = 1, ..., c\right) \quad (6)$$

The soundness degree Y_j expresses in one value, to what extent the information of the rules have contributed for the classification into a given class j.
5) *Classification*. A decision function h is applied over the soundness degree for all classes. This function determines the class C_j corresponding to the maximum value obtained. This is $C_j = h(Y_1, ..., Y_M)$ such that

$$Y_j = \max_{j=1,...,M} Y_j \quad (7)$$

3) Aggregation Operators: Aggregation Operators are important since their use allows a FRM to consider the information given by all the rules compatible with an input pattern and their selection is dependant of the problem. For these reasons, several Aggregation Operators have been proposed such as the ones described in [3]. Below, we describe the operator used in the original FRM from [19] also used in our experiments.

Let $(a_1, ..., a_s)$ represent the association degrees b_j^k of c rules such that $b_j^k > 0$ and $k = 1, ..., c$ for a given pattern and one class j:

$$f_Q^j(a_1, ..., a_s) = \left[\frac{1}{s} \sum_{l=1}^{s} (a_l)^p \right]^{\frac{1}{p}} \quad (8)$$

where $p \in \mathbb{R}$ and f_Q^j stands for the Quasiarithmetic mean operator applied to the non-zero degrees of association to a given class C_j provided by the weighting function. The behaviour of this operator produces an aggregated value of degree of association between the minimum and the maximum and is determined by the selection of p such that:

- If $p \rightarrow -\infty$, $f_Q \rightarrow$ min,
- If $p \rightarrow +\infty$, $f_Q \rightarrow$ max.

For more detail about its properties see [32].

The implementation of aggregation operators in the FRM provides an inference process capable of using a good combination of the rules information to define class membership. In the next section, a description of our proposed FRBC through a Cluster-Based Representation is presented.

III. FRBC Construction with Cluster-Based Class Representation

In order to explain both the generation of the rules from training data and classification process, we explain all steps as shown in Fig. 2:

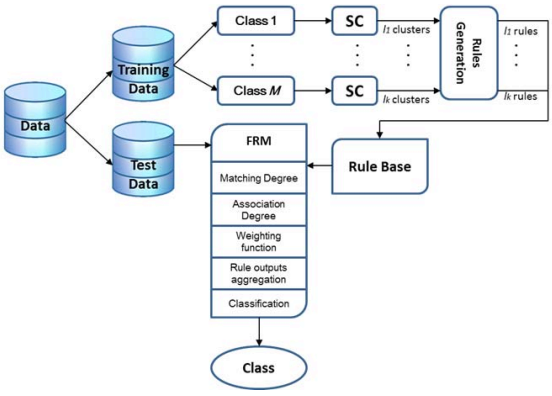

Fig. 2. Rule extraction and classification process

A. Rule Base Construction

Consider a given set of n patterns for training a FRBC with M classes.

1) Separate training dataset per class and perform SC in each training subset.
2) Let the number of clusters found by SC for class C_j be l_j, set the total number of clusters as $c = \sum_{j=1}^{M} l_j$.
3) Calculate the matrix of distances from each i training pattern to each k cluster.
4) By using two fuzzifiers m_1 and m_2, calculate the membership degrees of each training pattern $\mathbf{x}$ to a cluster k such as shown in (9) and (10):

$$\mu_k^{m_1} = \frac{1}{\sum_{q=1}^{c} \left(\frac{d_k}{d_q}\right)^{\frac{2}{(m_1-1)}}} \tag{9}$$

$$\mu_k^{m_2} = \frac{1}{\sum_{q=1}^{c} \left(\frac{d_k}{d_q}\right)^{\frac{2}{(m_2-1)}}}, \tag{10}$$

where d_k is the distance to the kth cluster prototype, and k = 1,..., c. This way, the uncertainty associated with the size and density of the clusters representing a class is being managed through the fuzzifier m, such as shown in (11) for the upper bound of the membership:

$$\overline{\mu}_k = \max\left(\mu_k^{m_1}, \mu_k^{m_2}\right), \tag{11}$$

whereas for the lower bound we consider (12):

$$\underline{\mu}_k = \min\left(\mu_k^{m_1}, \mu_k^{m_2}\right) \tag{12}$$

Thus, a footprint of uncertainty is created by using the highest and lowest primary memberships of a pattern $\mathbf{x}$ to a cluster k.

5) Generate the c IT2 FSs and their respective MFs according to (11) and (12) such that the MF of the kth IT2 FS named $\widetilde{A}_k$ is denoted by

$$\mu_{\widetilde{A}_k} = \left[\underline{\mu}_k(\mathbf{x}), \overline{\mu}_k(\mathbf{x})\right] \tag{13}$$

6) Construct a type-2 fuzzy rule base with c rules, $\widetilde{A}_k$ being the single antecedent in each rule. The certainty degree is defined as

$$r_j^k = \frac{\sum_{c(\mathbf{x}_i)=c_j} U_k(\mathbf{x}_i)}{\sum_{i=1}^{n} U_k(\mathbf{x}_i)} \tag{14}$$

where j = 1,...,M, $c(\mathbf{x}_i)$ denotes the class label of training pattern $\mathbf{x}_i$, n is the number of training patterns and

$$U_k(\mathbf{x}_i) = \frac{\underline{\mu}_k(\mathbf{x}_i) + \overline{\mu}_k(\mathbf{x}_i)}{2} \tag{15}$$

Summarizing the methodology to construct the Rule Base, our proposed construction of the FRBC is performed in one single-pass by splitting the training data per class and applying SC subsequently in order to find representative clusters prototypes of a single class such as illustrated in Fig. 3 where, non-circular dispersions of points are approximated by several circular clusters with uncertain size/density.

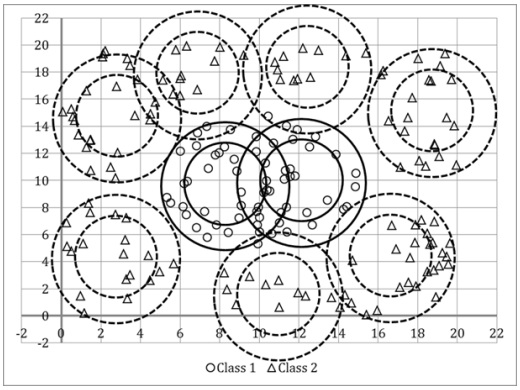

Fig. 3. A possible cluster based representation solution for a non-linear classification problem. The continuum circles stand for the clusters representing the Class 1, whereas the dashed circles stand for the Class 2

Clearly, the resulting clusters do not yet provide a useful answer in terms of classifying the data into the expected number of classes. The final stage of the process is addressed in the following section.

B. Inference process

Once the construction of the rules is performed, we have a rule base so further patterns can be used to test it by using the FRM as inference method for classification. Given a pattern $\mathbf{x}$ to classify we follow these steps:

1) Calculate its matching degree in each rule by determining its interval of membership to each cluster k:

$$R^k(\mathbf{x}) = \mu_{\widetilde{A}_k} = \left[\underline{\mu}_{A_k}(\mathbf{x}), \overline{\mu}_{A_k}(\mathbf{x})\right] \tag{16}$$

Note that in these rules there is just one antecedent so, considering (4), the matching degree is the degree of membership in $\widetilde{A}_k$.

2) Considering (5) and (16), calculate its association degree in the interval by using product

$$b_j^k = \left[b_{jl}^k, b_{jr}^k\right] = \left[\left(\underline{\mu}_{A_k}(\mathbf{x}) \cdot r_j^k\right), \left(\overline{\mu}_{A_k}(\mathbf{x}) \cdot r_j^k\right)\right] \quad (17)$$

where b_j^k is the association degree to the j class under the *k*th rule.

3) In this FRM, the weighting function g is $g(x) = x$ such as in [19].

4) Considering (6) and (17), calculate the soundness degree Y_j of pattern $\mathbf{x}$ with class C_j by

$$Y_j = [Y_{jl}, Y_{jr}] = \left[Y_{jl} = f\left(b_{jl}^k\right), Y_{jr} = f\left(b_{jr}^k\right)\right] \quad (18)$$

where $k = 1, ..., C$ and f stands for the Quasiarithmetic mean aggregation operator defined in (8).

5) Finally, the decision function h is applied over the soundness degree interval in all classes. This is performed by considering (7) and (18) so, assign the class C_j to the pattern $\mathbf{x}$ such that the function $h(Y_1, ..., Y_M)$ determines the class corresponding to the maximum value obtained. This is:

$$Y_j = \max_{j=1,...,M}\left(\frac{Y_{jl} + Y_{jr}}{2}\right) \quad (19)$$

Note that, contrary to (7), here we are considering the average of the soundness degree bounds rather than a single value in order to determine the class to be assigned.

IV. Results

In this section, the utility of IT2 FRBCs is demonstrated on four datasets by comparing their results against the ones of a SVM using a radial basis function. We used $m_1 = 1.5$ and $m_2 = 2.5$ as fuzziness parameters for IT2 FSs. These values were chosen based on the Pal and Bezdek [33] study which suggests that the best choice for *m* (based on the performance of some cluster validity indices) is commonly in the interval [1.5, 2.5]. Also in SC applications, normalization to $[0, 1]$ along with a r_a value between [0.4, 0.6] is recommended to get reasonable results but, in this application, different values of r_a are being explored since different numbers of clusters prototypes were found in the data and consequently the number of rules.

A. Non-linear classification problems

1) Circular surrounding: Our first experiment is performed with the synthetic data set shown in Fig. 1, where there are 186 patterns with two features generated randomly in the interval [0, 20]. Then we proceed to label the patterns according to their Euclidean distance to the point (10, 10) so if the distance of a pattern x_i is greater than 7, then the label assigned is *class* 2, otherwise if its distance is lower than 5 it is associated to *class* 1. Thus, we generated a circular distribution of the class 1 with 63 patterns and, on the other hand, we generated 123 patterns for the class 2 surrounding the former class with different sizes and densities. In order to create the FRBC we shuffled and divided the data into 2 sets, 50% to construct the model and the rest for test. This process was repeated over 32 experiments per each r_a value and the SVM as well. In Table II, we show the results of these experiments, where the number of clusters found in the 32 runs is reported as an interval for the minimum and maximum numbers. When SC was not used, we chose the mean of the training patterns in each class as the initial prototype for each cluster (such as made in [19]).

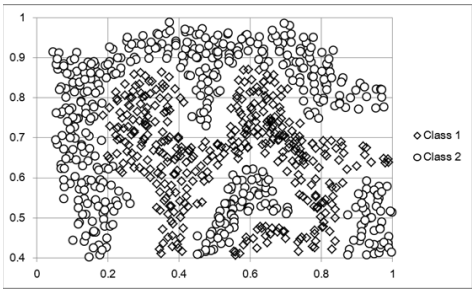

Fig. 4. Two non-linearly separable classes

TABLE II. Cluster-based class representation in a synthetic Non-Linear problem

	r_a	Clusters/rules	Best	Average	Worst	σ
FRBC	none	2	74.19	66.30	58.06	3.52
	0.2	[19,25]	**100**	**99.16**	95.70	1.10
	0.3	[10,14]	**100**	96.94	90.32	2.39
	0.4	[8,10]	**100**	95.23	90.32	2.51
	0.5	[6,8]	98.92	94.72	87.10	2.80
	0.6	[6,7]	98.92	93.75	89.25	2.67
SVM			**100**	97.85	93.55	1.68

As can be seen, the results of using more than one cluster for representing a single class were superior over the use of a single cluster but note that, as we increased the r_a threshold and consequently the number of clusters/rules were reduced (see Section II-A), the average performance of the FRBCs was reduced. Nevertheless, in all cases there was a significant improvement over the original FRBC model (one cluster to represent each class), which in all cases misclassified all patterns of *class* 1 to *class* 2.

Regarding the comparisons with SVM results, the average of the 32 experiments was found comparable to the constructed FRBCs, outperformed only by the FRBC where $r_a = 0.2$.

2) Irregular distribution: Our second experiment is performed with the pattern set shown in Fig. 4 where there are 863 patterns with two features. There are 480 and 383 patterns respectively for class 1 and class 2 over an irregular distribution surrounding the latter class with different sizes and densities.

Similarly as in the previous application, we shuffled the data and divided it into 2 sets in which 50% was used to construct the model and the rest for testing. This process was repeated 32 times. In Table III, we show the results of these experiments, where the number of clusters found for all the classes in the 32 iterations is being reported as an interval for the lower and bigger amounts. Similarly to the previous synthetic experiment, in all cases there was a significant improvement over the original FRBC model which uses a single prototype per class given that, by looking at

TABLE III. CLUSTER-BASED CLASS REPRESENTATION IN A SYNTHETIC NON-LINEAR PROBLEM

	r_a	Clusters/rules	Best	Average	Worst	σ
FRBC	none	2	62.96	56.32	52.08	2.33
	0.2	[23,30]	98.61	95.43	91.20	1.51
	0.4	[10,13]	91.44	85.63	82.41	1.84
	0.6	[5,9]	80.09	71.92	58.80	4.39
SVM			**99.76**	**98.57**	96.98	0.75

the confusion matrices, there were misclassified patterns from *class* 2 to *class* 1. This shows that the results of the use of more than one cluster for the representation of a single class were superior to the use of a single prototype as expected.

Additionally we also found a negative correlation between the r_a value and the number of rules, as well as a positive correlation between the average performance of the FRBCs and the size of the rule base. These correlations, seem to indicate a high importance in the selection of the r_a value in order to control the sensitivity of the FRBC to accept reference points. Regarding the comparisons with SVMs results, the 32 experiments were found superior to the constructed FRBCs, with the FRBC with $r_a = 0.2$ the closest in average.

B. Application to Iris Plant Benchmark

For our next experiment, we used the Iris Plant dataset from the UCI repository of machine learning databases. The Iris dataset is composed of 150 4-dimensional patterns uniformly distributed among three classes. For these experiments we performed the same methodology used in the previous experiments by dividing the dataset into 2 subsets, 50% for training and 50% for testing. Note that for comparison purposes, we followed the approach in [19] as closely as possible. Our FRBCs results are shown in Table IV and also, we include the results reported in the original approach [19].

TABLE IV. CLUSTER-BASED CLASS REPRESENTATION IN IRIS PLANT DATASET

	r_a	Clusters/Rules	Best	Average	Worst	σ
FRBC	none	3	98.67	92.21	85.33	3.06
	0.2	[21,32]	**100**	94.46	89.33	2.38
	0.3	[8,16]	98.67	95.08	90.67	2.07
	0.4	[6,10]	98.67	94.63	85.33	3.01
	0.5	[6,9]	98.67	94.04	86.67	2.85
	0.6	6	**100**	93.54	85.33	3.09
Tang [19]		3	97.33	88.80	72.00	
SVM			**100**	**96.95**	92.00	1.98

As in previous experiments, some improvements were reached by using more than one cluster for class representation although in this case were subtle. By comparing with the results reported in [19], our cluster-based representation seems to improve the classification accuracy. We attributed this improvement to the use of different fuzziness parameters, since in the original paper they were chosen as: $m_1 = 2$ and $m_2 = 5$ and also, they refined the cluster prototypes by using IT2 FCM before constructing the Rule Base.

C. Application to WBCD Benchmark

Finally, we performed experiments in the same fashion with the Wisconsin breast cancer diagnosis (WBCD) database which is the result of the efforts made at the University Of Wisconsin Hospital for accurately diagnose breast masses based solely on an Fine Needle Aspiration (FNA) test. This dataset contains a total of 699 clinical instances, with 458 benign and 241 malignant cases. Each instance has 9 attributes but, 16 of the samples are each missing one of the nine attributes so in common practice, those instances are eliminated as in this experiments as well. In Table V we show these experiments results.

TABLE V. CLUSTER-BASED CLASS REPRESENTATION IN WISCONSIN DATASET

	r_a	Clusters/rules	Best	Average	Worst	σ
FRBC	none	2	98.25	**96.54**	94.44	0.79
	0.2	[112,132]	97.08	95.87	94.15	0.64
	0.4	[107,132]	97.08	91.51	85.38	2.87
	0.6	[93,121]	95.32	92.35	88.30	1.69
	1	[12,35]	97.66	95.92	93.57	0.90
	1.5	[4,5]	**98.54**	96.31	94.44	1.07
SVM			96.48	95.30	93.25	0.85

Contrary to findings in previous experiments, the recommended interval of r_a found several clusters in this dataset which seem to overfit the generated FRBC models, i.e. the use of more rules resulted counterproductive in most of the experiments although, non clear improvements were reached over the results of the FRBCs in which we did not perform SC despite the number of rules constructed. Additionally, by comparing the results of the generated FRBCs against the SVM results we found a slight difference favouring our approach.

V. DISCUSSION

In the FRBC model presented, we explored different numbers of rules while seeking for reasonably good performance instead of focusing on optimizing MF parameters such as in other FRBCs applications (e.g. [2]). This exploration was addressed by considering different r_a values for SC given that, as mentioned in Section II-A, the parameter r_a is crucial to determine the sensitivity of the algorithm to accept or refuse cluster centroids. Consequently, we found that r_a acts as a threshold to control the number of rules of the FRBC which can increase the computational cost related to the number of clusters/MFs to analyse and affect the interpretability. Also, we found that the relation between number of rules and performance is not always as expected since we observed that, in some cases, a relatively big number of rules does not necessarily help to improve the FRBC accuracy and affects the interpretability of the FRBCS (according to the discussion about considerations for interpretability in [34]).

Another consideration for interpretability is the *number of antecedents* in the rules. In this approach one rule is created for each cluster prototype found in SC while handling different certainty degrees, as shown in (20) and determining the membership degrees of the training patterns to different clusters. Thus, the created rules keep the structure:

$$R_k: \quad IF \quad x \quad is \quad \widetilde{A}_k, \quad THEN \quad \left(r_1^k, ..., r_M^k\right), \quad (20)$$

Thus, points of reference (class prototypes) are being used to measure the similarity of input patterns to the antecedents representing the points of reference within the feature space. Using type-2 FSs provides the potential to capture more

complex mappings without adding additional rules, a feature which we seek to explore further in the context of FRBCs in the future.

Another important consideration for interpretability is related to the distribution of the membership functions and their *distinguishability*. Here, as consequence of using SC to create the rules, the antecedent's membership functions are distinguishable from the rest in the pattern feature space. Also, the Rule-Base is *consistent* (i.e. non-contradictory rules) since it is generated from the centroids provided by SC which follow certain separability and compactness (dependant on the threshold r_a) so, rule antecedents related to different classes will be reasonably different.

A number of remaining challenges have been identified: the potential for overfitting due to a large number of cluster prototypes - as identified in the case of the Wisconsin dataset. Further, the fundamental relationship of performance and interpretability and how to automatically balance them (which includes measuring the level of interpretability at runtime) remains to be addressed. In other words, as remarked previously, an appropriate selection of the r_a value for SC is required to reach both characteristics of a FRBC: interpretability and good performance.

Still, even in the face of these challenges, the results show that the relative simplicity of the FRBCs through this cluster-based class representation can help FRBCs to become candidates for both, linear and non-linear classification problems in which the advantages of FRBCs (e.g. interpretability of IF-THEN rules) is valuable.

VI. Conclusion

Our main contribution in this work consists in showing the potential of a cluster-based representation in linear and non-linear classification problems while avoiding an space transformation of input data such as required in our reference method (SVM). This was done by starting from an improvement to the IT2 FRBC model described in [19]. This improvement was reached by using distributed clusters of uncertain size due to the use of two fuzziness parameters in order to create a foot of uncertainty represented by an interval. This manner, the initial representation in [19] which (during the experiments) misclassified all patterns of one class to another in the first two datasets of Section IV, was changed for a more appropriate representation.

We implemented Subtractive Clustering along with a FRBC managing IT2 FSs and we found that, the appropriate selection of value r_a is highly dependent of the size/density of the data. We also could determine that, in classification problems, a large number of rules do not necessarily helps to improve the accuracy of the FRBC.

We have presented four experiments while comparing to SVMs and focusing mainly in the potential of cluster-based representation to improve results in different scenarios such as nonlinear classification problems.

As part of future work, we aim to develop a methodology to balance classification performance and interpretability based on application-specific criteria. Further, we will pursue more extensive comparisons with other techniques including other rule based approaches (e.g., ANFIS [35] [36]) and decision tree based classifiers.

References

[1] P. Herman, G. Prasad, and T. M. McGinnity, "Design and on-line evaluation of type-2 fuzzy logic system-based framework for handling uncertainties in bci classification," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 4242–4245.

[2] S. Chiu, "Extracting fuzzy rules from data for function approximation and pattern classification," *Fuzzy Information Engineering: A Guided Tour of Applications.*, 1997.

[3] O. Cordón, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," *International Journal of Approximate Reasoning*, vol. 20, no. 1, pp. 21–45, 1999.

[4] V. López, A. Fernández, M. J. Del Jesus, and F. Herrera, "A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets," *Knowledge-Based Systems*, vol. 38, pp. 85–104, 2013.

[5] H. Ishibuchi and T. Yamamoto, "Rule weight specification in fuzzy rule-based classification systems," *Fuzzy Systems, IEEE Transactions on*, vol. 13, no. 4, pp. 428–435, 2005.

[6] C. Hwang and F. C.-H. Rhee, "Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means," *Fuzzy Systems, IEEE Transactions on*, vol. 15, no. 1, pp. 107–120, 2007.

[7] E. Jammeh, M. Fleury, C. Wagner, H. Hagras, M. Ghanbari *et al.*, "Interval type-2 fuzzy logic congestion control for video streaming across ip networks," *Fuzzy Systems, IEEE Transactions on*, vol. 17, no. 5, pp. 1123–1142, 2009.

[8] Q. Liang and J. M. Mendel, "Mpeg vbr video traffic modeling and classification using fuzzy technique," *Fuzzy Systems, IEEE Transactions on*, vol. 9, no. 1, pp. 183–193, 2001.

[9] P. Herman, G. Prasad, and T. McGinnity, "Investigation of the type-2 fuzzy logic approach to classification in an eeg-based brain-computer interface," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*. IEEE, 2006, pp. 5354–5357.

[10] H. Ishibuchi, K. Nozaki, and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," *Fuzzy sets and systems*, vol. 52, no. 1, pp. 21–32, 1992.

[11] J. Casillas, O. Cordón, M. J. Del Jesus, and F. Herrera, "Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems," *Information Sciences*, vol. 136, no. 1, pp. 135–157, 2001.

[12] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, "Construction of fuzzy classification systems with rectangular fuzzy rules using genetic algorithms," *Fuzzy sets and systems*, vol. 65, no. 2, pp. 237–253, 1994.

[13] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," *Information Sciences*, vol. 136, no. 1, pp. 109–133, 2001.

[14] S. Mitra and L. I. Kuncheva, "Improving classification performance using fuzzy mlp and two-level selective partitioning of the feature space," *Fuzzy Sets and Systems*, vol. 70, no. 1, pp. 1–13, 1995.

[15] D. Nauck and R. Kruse, "A neuro-fuzzy method to learn fuzzy classification rules from data," *Fuzzy sets and Systems*, vol. 89, no. 3, pp. 277–288, 1997.

[16] M. De Cock, C. Cornelis, and E. E. Kerre, "Elicitation of fuzzy association rules from positive and negative examples," *Fuzzy Sets and Systems*, vol. 149, no. 1, pp. 73–85, 2005.

[17] Y.-C. Hu and G.-H. Tzeng, "Elicitation of classification rules by fuzzy data mining," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 7, pp. 709–716, 2003.

[18] J. A. Roubos, M. Setnes, and J. Abonyi, "Learning fuzzy classification rules from labeled data," *Information Sciences*, vol. 150, no. 1, pp. 77–93, 2003.

[19] M. Tang, X. Chen, W. Hu, and W. Yu, "A fuzzy rule-based classification system using interval type-2 fuzzy sets," in *Integrated Uncertainty in Knowledge Modelling and Decision Making*. Springer, 2011, pp. 72–80.

[20] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 6, pp. 1414–1427, 1992.

[21] A. Ittner and M. Schlosser, "Discovery of relevant new features by generating non-linear decision trees." in *KDD*, 1996, pp. 108–113.

[22] R. Min, D. Stanley, Z. Yuan, A. Bonner, Z. Zhang *et al.*, "A deep non-linear feature mapping for large-margin knn classification," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 357–366.

[23] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[24] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[25] S. L. Chiu, "Fuzzy model identification based on cluster estimation." *Journal of intelligent and Fuzzy systems*, vol. 2, no. 3, pp. 267–278, 1994.

[26] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[27] R. R. Yager and D. P. Filev, "Generation of fuzzy rules by mountain clustering," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 2, no. 3, pp. 209–219, 1994.

[28] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[29] ——, "The concept of a linguistic variable and its application to approximate reasoningi," *Information sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[30] J. M. Mendel, "Type-2 fuzzy sets: some questions and answers," *IEEE Connections, Newsletter of the IEEE Neural Networks Society*, vol. 1, pp. 10–13, 2003.

[31] ——, *Uncertain rule-based fuzzy logic system: introduction and new directions*. Prentice–Hall PTR, 2001.

[32] H. Dyckhoff and W. Pedrycz, "Generalized means as model of compensative connectives," *Fuzzy sets and Systems*, vol. 14, no. 2, pp. 143–154, 1984.

[33] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *Fuzzy Systems, IEEE Transactions on*, vol. 3, no. 3, pp. 370–379, 1995.

[34] Y. Jin, "Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement," *Fuzzy Systems, IEEE Transactions on*, vol. 8, no. 2, pp. 212–221, 2000.

[35] J.-S. R. Jang *et al.*, "Fuzzy modeling using generalized neural networks and kalman filter algorithm." in *AAAI*, vol. 91, 1991, pp. 762–767.

[36] J.-S. R. Jang, "Anfis: adaptive-network-based fuzzy inference system," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 23, no. 3, pp. 665–685, 1993.