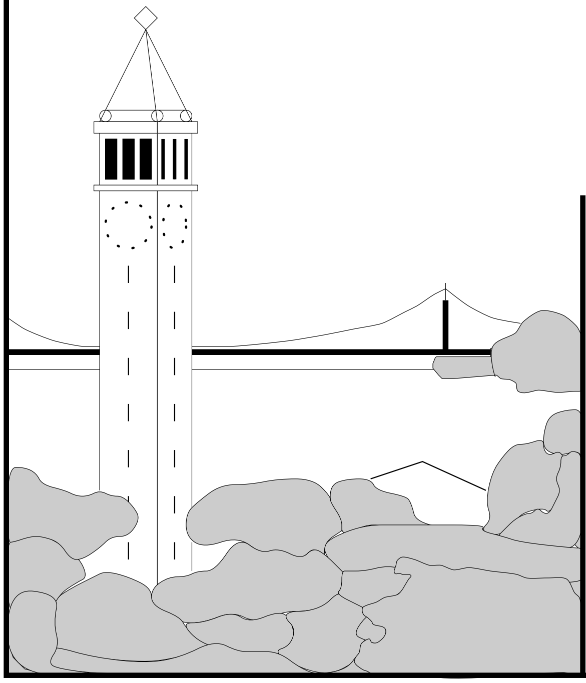
Interpreting Fuzzy Models
The Discriminative Power of Input Features

Rosaria Silipo and Michael R. Berthold



Report No. UCB/CSD-99-1079

November 1999

Computer Science Division (EECS)
University of California
Berkeley, California 94720

1

Interpreting Fuzzy Models
The Discriminative Power of Input Features

Rosaria Silipo and Michael R. Berthold

International Computer Science Institute (ICSI)
1947 Center Street, Suite 600, Berkeley, CA 94704, USA
eMail: rosaria@icsi.berkeley.edu
and
Berkeley Initiative in Soft Computing (BISC)
Dept. of EECS, CS Division, 329 Soda Hall
University of California, Berkeley, CA 94720, USA
eMail: berthold@cs.berkeley.edu

November 1999

Abstract

An important part of the interpretation of a decision process lies on the ascertainment of the influence of the input features, that is, of how much the implemented model relies on a given input feature to perform the desired task. Recently data analysis techniques based on fuzzy logic have gained attention because of their interpretability. Many real-world applications, however, have very high dimensionality and require very complex decision borders. In this case the number of fuzzy rules can proliferate and the easy interpretability of the fuzzy model can progressively disappear.

A method is presented that quantifies the discriminative power of the input features in a fuzzy model. The proposed quantification helps the interpretation of fuzzy models constructed on high dimensional and very fragmented training sets. First, a measure of the information contained in the fuzzy model is defined on the basis of its fuzzy rules. The classification is then performed along one of the input features, that is, the fuzzy rules are split according to that feature's linguistic values. For each linguistic value, a fuzzy sub-model is generated from the original fuzzy model. The average information contained in these fuzzy sub-models is measured and its relative comparison with the information measure of the original fuzzy model quantifies the information gain that derives from the classification performed on the selected input feature. This information gain characterizes the discriminative power of that input feature. Therefore, the proposed information gain can be used to obtain better insights into the selected fuzzy classification strategy, even in very high dimensional cases, and possibly to reduce the input dimension.

Several artificial and real-world data analysis are reported as examples, in order to illustrate the characteristics and potentialities of the proposed algorithm. As real-world examples, the most informative electrocardiographic measures are detected for an arrhythmia classification problem and the role of duration, amplitude and pitch variations of syllabic nuclei in American English spoken sentences is investigated for prosodic stress classification.

*M. Berthold was supported by a grant from the Deutsche Forschungsgemeinschaft, DFG1740/7-1

1 Introduction

In the last years, several algorithms for classification, and pattern recognition in general, have been more or less successfully proposed. Many of these methods were proposed in competition or as an alternative to other data analysis methods on the pure basis of their "better" performances. More recently, this purely numerical performance criterion resulted to be not completely satisfying because of the desire to understand what the decision process is actually performing.

It has also become more and more common to collect and store large amounts of data from different sources [1]. As a consequence, databases with higher dimension and describing more complex problems have been obtained. However, a massive recording of the system's monitoring variables does not grant a better performance of further analysis procedures, if the newly introduced variables do not carry additional information. Moreover, the analysis procedure itself becomes more complicated for very fragmented input spaces and insights about the system's underlying structure become more difficult to achieve.

Because of the more complicated nature of the problems and because of their higher dimension and size, the attention has moved from a pure numerical percentage-based comparison to a knowledge representation problem. For example, if in a given context a data analysis method does not hold as good performance as other methods and offers a more informative representation of the underlying process and a clearer interpretation of the decision process, such technique could represent a better decision support tool for the analyzer than a technique which offers numerically superior performance but is harder to interpret.

The interpretability of the decision process represents a key topic in modern data analysis scenarios and corresponds to the transparency of the model built on the training set to implement a given task. An important part of the interpretation of the decision process lies on the ascertainment of the influence of the input features, that is of how much the implemented model relies on a given input feature to perform the desired task.

A quite common approach for the evaluation of the effectiveness of the input features defines some merit measures, on the basis of a statistical model of the system [2, 1]. Assuming that a large database is available, the probabilities, involved in the definition of these merit measures, are estimated by means of events frequencies. This requires a precise definition of the input parameters and a clear identification of the output classes. In many real world applications, however, estimated frequencies are unavoidably altered by doubtful members of the output classes and by an inaccurate description of the input parameters. In addition, the estimation of a probabilistic model may be computationally expensive for high dimensional input spaces.

One of the most appreciated data analysis techniques for its easy interpretability derives from fuzzy logic. The concept of fuzzy sets was introduced in [3] with the purpose of a more efficient, though less detailed, description of real world events, by allowing an appropriate amount of uncertainty into the model. Fuzzy set theory yields also the advantage of a number of simple and computationally inexpensive available methods, for the modelling of a given training set. If the particular problem has low dimension and not very complex decision borders among the output classes, fuzzy models produce a reasonable amount of fuzzy rules, relatively easy to interpret and with sufficiently reliable performances. Many real-world applications present very high dimensional input spaces and require very complex decision borders. In this case the number of fuzzy rules can proliferate and the easy interpretability of the model can progressively disappear.

In order to quantify the discriminative power of the input features in a fuzzy model, a method is presented that helps the interpretation of fuzzy decision systems constructed on high dimensional and very fragmented training sets. The proposed method consists of an analysis "a posteriori" of the fuzzy system, that was created to model the training set.

Based on fuzzy set theory, some measures of fuzzy entropy have been established [4, 5] as measures of the degree of fuzziness of the model with respect to the training data. All these entropy measures describe the uncertainty of the model in fitting the desired input/output mapping and neglect to describe the decisional structure of the fuzzy model itself. One important descriptive feature of an analysis model consists of the impact of the input features on the final decision process.

The goal of this work is to ascertain how separable the output classes are on a given input dimension of the model and not how well the training data is represented by the model. Thus, an interpretation of the implemented fuzzy model is required with respect to the output classes separability rather than with respect to the training patterns representation. In general, if the training set contains a sufficient number of examples from the different output classes – that is if the fuzzy model is sufficiently general and accurate –, an analysis "a posteriori" of the fuzzy model – determining which input features has the highest influence in separating the output classes – will also reflect information about the training data and the input space. Moreover, the main advantage of analyzing fuzzy rules, instead of fuzzy rules and training data as in [4, 5], consists of the highly reduced computational costs for the same amount of information, provided the fuzzy model faithfully describes the underlying data structure.

At first, a measure of the information available in the fuzzy model is defined on the basis of its fuzzy rules. The classification is then performed along one of the input features x_j and the fuzzy rules are split according to x_j's linguistic values. For each linguistic value, a new fuzzy sub-model is derived from the original fuzzy model. The average information contained in these fuzzy sub-models is measured and the relative comparison with the information measure of the original fuzzy model produces the information gain that derives from a possible classification performed on the considered input feature x_j. This information gain characterizes the discriminative power of input feature x_j. The input dimension with highest information gain defines the most discriminative input feature, according to the analyzed fuzzy model.

Due to the low computational expenses derived from the use of fuzzy models, the proposed information gain generates a simple and efficient algorithm to measure the contribution of each input feature to the discrimination among output classes in the considered fuzzy model. This allows better insights into the adopted fuzzy classification strategy, especially for very high dimensional input spaces, and consequently a possible reduction of the input dimension.

The detection and ranking of the most effective input features for a given task could represent one of the first steps in any data analysis process. In fact, the implementation of a fuzzy model generally requires a short amount of time even for very high dimensional input spaces and so does the corresponding evaluation of the discriminative power of the input features. Whenever a more accurate system's representation is desired, the analysis can continue with the application of more sophisticated and more computationally expensive analysis techniques on the most effective input features, pre-screened on the basis of the proposed fuzzy information gain.

2 Fuzzy Information Measures

2.1 The Average Membership Degree

Fuzzy models represent a particular version of rules set, where a given uncertainty or *fuzziness* is allowed, so that a given input pattern $\vec{x} = (x_1, \cdots, x_j, \cdots, x_n)^T$ belongs somewhat to a certain output class C_i [3]. Thus the set of rules implementing the desired input/output mapping consists of a set of membership functions $\mu_{C_i}(\vec{x}) \in [0,1]$, that associate the input pattern $\vec{x}$ with the output classes C_i $(1 \le i \le m)$.

Given a number m of output classes C_i, and an n-dimensional input space, numerous algorithms exist, which derive a set of N_R fuzzy rules $\{R_k\}$, $k = 1, \ldots, N_R$, mapping the n-dimensional input into the m-dimensional output space. This set of rules models the relationships between the input data and the output classes, so that each input pattern $\vec{x} \in \mathcal{R}^n$ is associated to each output class C_i by means of the membership value $\mu_{C_i}(\vec{x})$. In figure 1 an example is reported with a two-dimensional input space $\{x_1, x_2\}$, two output classes C_1 and C_2 and with trapezoids as membership functions.

The membership function $\mu_{C_i}(\vec{x})$ quantifies the degree of membership of input pattern $\vec{x}$ to output class C_i. The quantity $V(C_i)$ in equation 1 represents the *average degree of membership* of input patterns $\vec{x}$ to output class C_i over the whole domain $D \subset \mathcal{R}^n$.

$$V(C_i) = \frac{\int_{\vec{x}\in D} \mu_{C_i}(\vec{x})\ d\vec{x}}{\int_{\vec{x}\in D}\ d\vec{x}} \tag{1}$$

Considering normalized membership functions, a higher average membership degree to class C_i, $V(C_i)$, indicates a more uniformly distributed class over the input space. An output class represented by a membership function, which takes value $+1$ everywhere on the input space, has average membership degree $+1$. A membership function with average value $V(C_i) = 0$ indicates an output class that is never related with any pattern of the input domain.

The average membership degree $V(C_i)$ over the domain $D \subset R^n$ (eq. 1) represents the first step to quantify the information contained in the set of fuzzy rules $\{R_k\}$. We need now some operator that distinguishes between fuzzy models with only membership functions of one class (no information) from fuzzy models with a very high

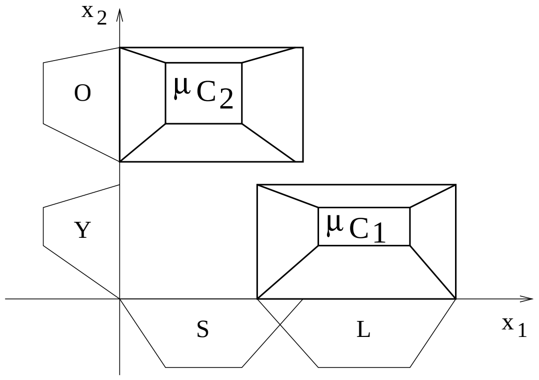


Figure 1: Example of a two-class fuzzy model on a two-dimensional input space.

number of membership functions of different classes (high information). The usual information measures, such as the entropy or the Gini function, from the information theory could be applied to the average membership degree $V(C_i)$ $(1 \leq i \leq m)$. Such information functions, however, require the variable to sum up to 1 across the m output classes C_i, which is not necessarily true due to the non-normalized nature of fuzzy sets.

A solution for this problem was found in [18] by applying the cited information measures to the *relative average membership degree* to output class C_i:

$$v(C_i) = \frac{V(C_i)}{\sum_{j=1}^{m} V(C_j)} \tag{2}$$

The variables $v(C_i)$, with $i = 1, \ldots, m$, now sum up to 1 and the traditional information functions can be applied.

In case the output class C_i is described by Q_i fuzzy rules, C_i^q $(q = 1, \ldots, Q_i)$, the average membership degree of class C_i is given by the average membership degree of the union of these fuzzy subsets of class C_i, each with membership function $\mu_{C_i}^q(\vec{x})$. Thus, to handle more than one fuzzy set per class we need to compute the average membership degree of the union and the intersection of fuzzy sets. This derives straightforward from the usual min/max-definitions of intersection and union of fuzzy sets [6].

In particular the average membership degree of the union of two fuzzy rules, C_i^r and C_i^s, can be derived as the sum of the average membership degrees of the two single fuzzy rules, taking into account their intersection only once (eq. 3 and 4).

$$V(C_i^r \cap C_i^s) = \frac{\int_{D \subset \mathcal{R}^n} \mu_{C_i^r \cap C_i^s}(\vec{x})\ d\vec{x}}{\int_{D \subset \mathcal{R}^n}\ d\vec{x}} = \frac{\int_{D \subset \mathcal{R}^n} \min_{r,s} \left\{\mu_{C_i}^r(\vec{x}), \mu_{C_i}^s(\vec{x})\right\}\ d\vec{x}}{\int_{D \subset \mathcal{R}^n}\ d\vec{x}} \tag{3}$$

$$V(C_i^r \cup C_i^s) = \frac{\int_{D \subset \mathcal{R}^n} \mu_{C_i^r \cup C_i^s}(\vec{x})\ d\vec{x}}{\int_{D \subset \mathcal{R}^n}\ d\vec{x}} = \frac{\int_{D \subset \mathcal{R}^n} \max_{r,s} \left\{\mu_{C_i}^r(\vec{x}), \mu_{C_i}^s(\vec{x})\right\}\ d\vec{x}}{\int_{D \subset \mathcal{R}^n}\ d\vec{x}} =$$
$$= V(C_i^r) + V(C_i^s) - V(C_i^r \cap C_i^s) \tag{4}$$

In particular if the two membership functions $\mu_{C_i}^r(\vec{x})$ and $\mu_{C_i}^s(\vec{x})$ do not overlap, which means $\forall \vec{x}$: $\min_{r,s} \left\{\mu_{C_i}^r(\vec{x}), \mu_{C_i}^s(\vec{x})\right\} = 0$, the expressions in eq. 3 and 4 become:

$$V(C_i^r \cap C_i^s) = 0 \tag{5}$$
$$V(C_i^r \cup C_i^s) = V(C_i^r) + V(C_i^s) \tag{6}$$

This result can be extended to a number Q_i of membership functions, by expressing the average membership degree of the union as the sum of their average membership degrees and taking care of including their intersection only once (eq. 7).

$$V(C_i) = V\left(\cup_{q=1}^{Q_i} C_i^q\right) = \frac{\int_{\vec{x} \in D}\ \max_q\ \left\{\mu_{C_i}^q(\vec{x})\right\}\ d\vec{x}}{\int_{\vec{x} \in D}\ d\vec{x}} =$$

$$= \sum_{q=1}^{Q_i} \left[V(C_i^q) - \sum_{h=q+1}^{Q_i} V\left(C_i^q \cap C_i^h\right) \right] \tag{7}$$

If the usual trapezoids are adopted as membership functions, the average membership degree of each fuzzy subset C_i^q becomes particularly simple to calculate [6], as is shown in eq. 8 where h^q is the trapezoid height and $< \vec{a}_i^q, \vec{b}_i^q, \vec{c}_i^q, \vec{d}_i^q >$ are the coordinate vectors of its vertices in the n-dimensional input space.

$$V(C_i^q) = V(< \vec{a}_i^q, \vec{b}_i^q, \vec{c}_i^q, \vec{d}_i^q >) = \frac{\frac{1}{2}\left(\prod_{j=1}^{n}(d_{ji}^q - a_{ji}^q) + \prod_{j=1}^{n}(c_{ji}^q - b_{ji}^q)\right) h^q}{\int_{\vec{x}\in D} d\vec{x}} \tag{8}$$

2.2 Fuzzy Information Measures

There is some a priori information contained in each set of input/output relationships modeling a given system. In statistical theory the basic unit used to characterize the information amount associated with a given output class C_i is its a-priori probability p_i. The definition of probability p_i, however, requires an unambiguous relationship between the input patterns $\vec{x}$ and the output classes C_i, which can not always be guaranteed in real world applications. If multiple classes are allowed for the same input pattern, even with different degrees of membership, the estimation of the probability p_i of class C_i becomes more difficult because of the intrinsic ambiguity of the model. A fuzzy model could be more appropriate for describing such systems. Another advantage of using fuzzy instead of probabilistic models can be their lower computational expenses, because fuzzy models can describe real world problems in a less complex and more concise way.

As we have already described in the previous subsection, in case of fuzzy models the quantity $v(C_i)$ can be taken as the basic unit to quantify the information available in the model. The quantity $v(C_i)$ represents the average membership degree of the input patterns to output class C_i relatively to all the other output classes and is calculated as in eq. 2 according to the fuzzy rules used to model the input-output mapping. With respect to a probabilistic model, the use of the relative average membership degree to class C_i, $v(C_i)$, takes into account the possible occurrence of multiple classes for any input pattern $\vec{x}$ and its calculation is generally easier than the estimation of a probability function.

As in the traditional information theory, the goal is to produce an information measure, that is [1]:

1. at its maximum if all the output classes are equally possible in average on the input space $D \subset \mathcal{R}^n$, i. e. $v(C_i) = \frac{1}{m}$ for $i = 1, \ldots, m$, m being the number of output classes;
2. at its minimum if only one output class C_i exists, i. e. in case $v(C_j) = 0$ for $j \neq i$;
3. a symmetric function of its arguments, because the dominance of one class over the others in terms of relative average membership degree must produce the same amount of information, independently of which the favourite class is.

In order to produce a measure of the global information $I(C)$ of the output space $C = \{C_1, \ldots, C_m\}$ of the fuzzy model, the traditional functions employed in information theory – as the entropy function $I_H(C)$ (eq. 9) and the Gini function $I_G(C)$ (eq. 10) [1, 2] – can be applied to the relative average membership degrees $v(C_i)$ of the output classes.

$$I_H(C) = -\sum_{i=1}^{m} v(C_i)\ \log_2(v(C_i)) \tag{9}$$

$$I_G(C) = 1 - \sum_{i=1}^{m} (v(C_i))^2 \tag{10}$$

and conditions 1, 2, and 3 still hold.

1. If in the considered fuzzy model all output classes have similar relative average degree of membership then the information function is at its maximum.
2. If only one class exists, then the uncertainty is at its minimum and so is the information function.

3. The dominance of one class over the others produces the same amount of information, independently of which is the favourite class. That is, the defined information functions of variable $v(C_i)$ (eq. 9 and 10) are symmetric.

In both cases, entropy and Gini function, $I(C)$ represents the amount of information intrinsically available in the fuzzy model. The classification process aims to extract such information. Not all the input features, however, are effective the same way in extracting and representing this information available in the training set through the fuzzy model. The goal of this paper is to make explicit which dimension of the input space is the most effective in recovering the intrinsic information $I(C)$ contained in the fuzzy model.

3 Fuzzy Feature Merit Measures

A fuzzy merit measure of an input feature x_j should describe the information gain associated with the use of x_j in a given fuzzy analysis. Such information gain could be expressed by means of the information measures described in section 2.2. In particular, it can be defined as the difference between the intrinsic information amount available in the system before – $I(C)$ – and after – $I(C|x_j)$ – *using* that variable x_j for the fuzzy analysis [2]. It remains to define what the use of x_j corresponds to and how to measure the information amount left into the system after input feature x_j has been exploited for the analysis.

3.1 The Key-Points on Input Dimension x_j

Given a fuzzy description of the input space, $\{R_k\}$, the use of input feature x_j for classification purposes corresponds to the definition of an appropriate set of thresholds along x_j that allows the best separation of the input data into the output classes. Intuitively the optimal classification thresholds on a given input dimension j are located at the intersection points of contiguous membership functions of different output classes which is also the optimal decision boundary for the case of equal risks.

Let us restrict our analysis to a one-dimensional problem. In figure 2 an example with two output classes in a one-dimensional space x is reported. Let us choose a discrimination threshold x^* to separate class C_1 and C_2. Every $x < x^*$ is labeled as C_1 and every $x > x^*$ as C_2. Let us call the two labeling regions $\hat{C}_1$ and $\hat{C}_2$. The global degree of falseness (F) of the adopted labeling system is given by the area of $\mu_{C_1}(\vec{x})$ in region $\hat{C}_2$, where a C_2 label is imposed, and by the area of $\mu_{C_2}(\vec{x})$ in region $\hat{C}_1$, where a C_1 label is imposed, as expressed in eq. 11, 12 and 13.

$$F_{C_1} = \int_{C_1 \cap \hat{C}_2} \mu_{C_1}(x)dx \tag{11}$$

$$F_{C_2} = \int_{C_2 \cap \hat{C}_1} \mu_{C_2}(x)dx \tag{12}$$

$$F = F_{C_1} + F_{C_2} \tag{13}$$

The optimal classification threshold x^* refers to the minimum degree of falseness (F) of the global classification process, that is to the minimum intersection areas $C_1 \cap \hat{C}_2$ and $C_2 \cap \hat{C}_1$. After minimizing eq. 13, the optimal threshold x^* is found at the intersection point of the two membership functions, $x^* : \mu_{C_1}(x^*) = \mu_{C_2}(x^*)$.

If trapezoids are adopted as membership functions of the fuzzy model, the optimal thresholds between two contiguous trapezoids of different output classes are going to be located:

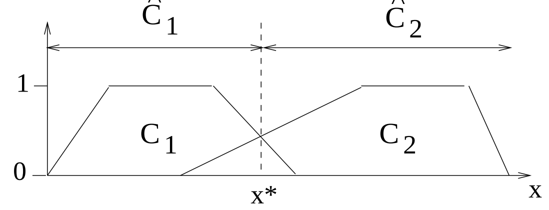


Figure 2: Fuzzy representation of a one-dimensional input space with two output classes.

1. at the intersection of their sides, if the trapezoids overlap only on the sides;
2. in the middle point of the overlapping flat regions ($\mu \equiv 1$, also called *core*), if the trapezoids overlap in the flat regions;
3. in the middle point between the two trapezoids, if they do not overlap anywhere.

3.2 The Information Gain

Let us suppose that input space $D \subset \mathcal{R}^n$ is related to the output classes by means of a number N_R of membership functions $\mu^q_{C_i}(\vec{x})$, with $q = 1, \ldots, Q_i$ membership functions for every output class C_i, for $i = 1, \ldots, m$ output classes, and $N_R = \sum_{i=1}^{m} Q_i$. $I(C)$ be the measure of the information contained in this fuzzy model. The discrimination of the output classes along input feature x_j leads to the definition of a number of optimal thresholds.

A set of cuts is then created on the j-th input dimension, to separate on x_j the $F_j \leq N_R$ contiguous trapezoids related to two different output classes, as discussed in section 3.1. After introducing the upper and lower boundary of x_j range in this set, a number of linguistic value L_k ($k = 1, \ldots, F_j$) can be defined for parameter x_j as the intervals between two consecutive cuts.

To consider $x_j = L_k$ corresponds to isolating one stripe c_k of the input space. In stripe c_k new membership functions $\mu^q(C_i|x_j = L_k)$ to the output classes C_i are derived as the intersections of the original membership functions $\mu^q_{C_i}(\vec{x})$ with the segment $x_j = L_k$. Each stripe c_k is characterized by a local information $I(c_k) = I(C|x_j = L_k)$, which, according to the information functions in eq. 9 and 10 respectively, is expressed as in eq. 14 and 15.

$$I_H(C|x_j = L_k) = - \sum_{i=1}^{m} v(C_i|x_j = L_k) \log_2 (v(C_i|x_j = L_k)) \tag{14}$$

$$I_G(C|x_j = L_k) = 1 - \sum_{i=1}^{m} (v(C_i|x_j = L_k))^2 \tag{15}$$

where:

$$v(C_i|x_j = L_k) = \frac{V(C_i|x_j = L_k)}{\sum_{h=1}^{m} V(C_h|x_j = L_k)} \tag{16}$$

$I(C|x_j = L_k)$ represents the measure of information still available in this part of the model, where x_j falls inside linguistic value L_k. The average of the measures of the information contained in stripes c_k, $I(C|x_j)$, produces a measure of the information still available in average in the fuzzy model after input feature x_j has been exploited for the fuzzy analysis (eq. 17).

$$I(C|x_j) = \frac{1}{F_j} \sum_{k=1}^{F_j} I(C|x_j = L_k) \tag{17}$$

The relative difference between the measure of the information originally available in the fuzzy model, $I(C)$, and the measure of the information still available in the model after the use of input feature x_j, $I(C|x_j)$, as expressed in eq. 17, represents the corresponding information gain (eq. 18).

$$g(C|x_j) = \frac{I(C) - I(C|x_j)}{I(C)} \tag{18}$$

The less effective the input feature x_j is in the original set of fuzzy rules, the closer the remaining information $I(C|x_j)$ is to the original information $I(C)$ of the model, resulting in a lower information gain (eq. 18). The input features producing the highest information gains are the most effective in the adopted model to describe the input space, and therefore the most informative for the proposed fuzzy analysis.

Every input parameter x_j produces an information gain $g(C|x_j)$ expressing its effectiveness in performing the required analysis on the basis of the given fuzzy model. Therefore the proposed information gain can be adopted as a fuzzy feature merit measure.

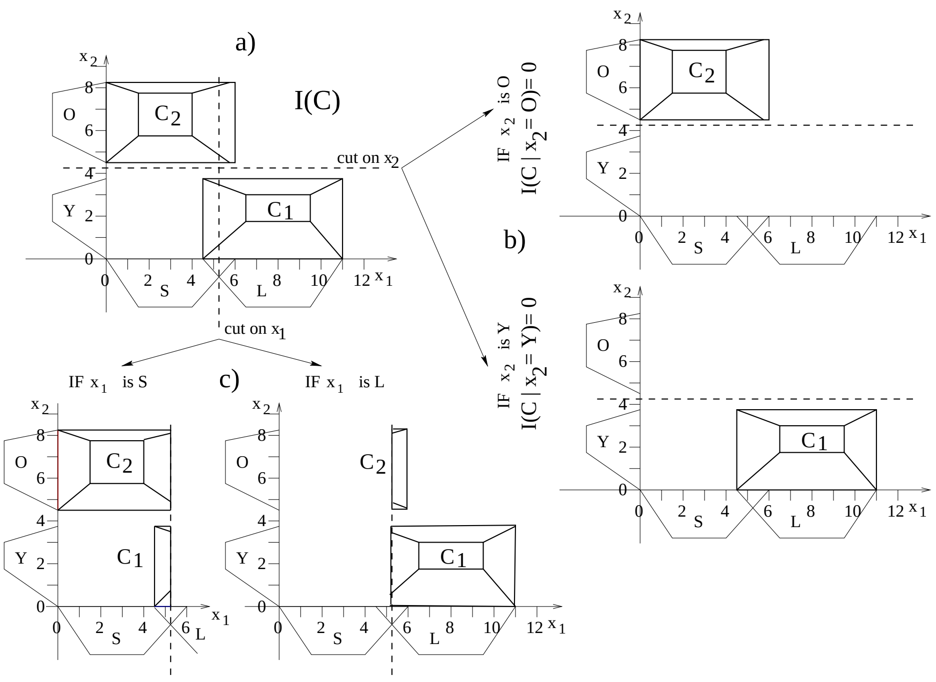


Figure 3: Stripes c_k generated by cutting the original fuzzy model (a) along variable x_2 (b)and x_1 (c).

3.3 An Example

In figure 3 an example is shown with a two-dimensional input space, two output classes, and trapezoids as membership functions. In table 1 the absolute and relative volumes of the two membership functions are reported and based on these values the information intrinsically available in the model is calculated by means of eq. 9 or 10. The discrimination of the two output classes can now be performed along input dimension x_1 or along input dimension x_2.

From the figure, we can easily see that a cut between the two membership functions on dimension x_2 (Fig. 3.b) produces a better separation than a cut on dimension x_1 (Fig. 3.c). That is, the analysis on dimension x_2 should offer a higher gain in information than the analysis on dimension x_1.

To verify this hypothesis, the average information still available in the system, $I(C|x_1)$ and $I(C|x_2)$, is calculated respectively after dimension x_1 and x_2 has been used for the classification. These information measures are reported in table 2 together with the consequent information gains $g(C|x_1)$ and $g(C|x_2)$.

For both choices of $I()$, the entropy or the Gini function, the information gain obtained from cutting along x_1 is smaller than the one obtained by cutting along x_2, that is $g(C|x_1) < g(C|x_2)$ (Tab. 2), as it was to be expected. This indicates that the analysis on variable x_2 extracts more of the information available in the fuzzy model than the analysis carried on variable x_1. The same conclusion could be reached using $I(C|x_1) > I(C|x_2)$, but a measure of merit based on the gain function produces clearer results than the direct use of the information parameter $I(C|x_j)$.

Table 1: The average membership degrees and the information measures for the two-dimensional example in fig. 3.

C_1	C_2	$I_H(C)$	$I_G(C)$
$V(C_1) = 13.0$	$V(C_2) = 12.6$	0.99	0.49
$v(C_1) = 0.51$	$v(C_2) = 0.49$		

Table 2: $I(C|x_j)$ and $g(C|x_j)$ for the example in fig. 3.

$x_1 = S$	$x_1 = L$	$x_2 = Y$	$x_2 = O$
$V(C_1\|x_1) = 0.53$	$V(C_1\|x_1) = 13.0$	$V(C_1\|x_2) = 13.0$	$V(C_1\|x_2) = 0.00$
$V(C_2\|x_1) = 12.6$	$V(C_2\|x_1) = 0.53$	$V(C_2\|x_2) = 0.00$	$V(C_2\|x_2) = 12.6$
$v(C_1\|x_1) = 0.04$	$v(C_1\|x_1) = 0.96$	$v(C_1\|x_2) = 1.0$	$v(C_1\|x_2) = 0.00$
$v(C_2\|x_1) = 0.96$	$v(C_2\|x_1) = 0.04$	$v(C_2\|x_2) = 0.00$	$v(C_2\|x_2) = 1.0$
$I_H(C\|x_1) = 0.24$		$I_H(C\|x_2) = 0.00$	
$I_G(C\|x_1) = 0.07$		$I_G(C\|x_2) = 0.00$	
$g_H(C\|x_1) = 0.76$		$g_H(C\|x_2) = 1.0$	
$g_G(C\|x_1) = 0.84$		$g_G(C\|x_2) = 1.0$	

4 Artificial Data Examples

4.1 Fixed Output Classes

In this section some artificial examples are analyzed to show how the information gain defined in section 3.2 characterizes the effectiveness of the input features for the required fuzzy input/output mapping. For all the examples reported in this study, the fuzzy clustering algorithm proposed in [7] has been used to build the set of fuzzy rules approximating the classification task at hand. As usual, trapezoids are used as membership functions.

The first example refers to a three-dimensional two-class problem. The projection of the input space on a two-dimensional plane (figure 4.a) shows a clear discrimination between the two output classes along input feature x_2. Random values for both classes are generated for the third dimension. A set of fuzzy rules is built to discriminate input data belonging to the two different classes. The information of the resulting fuzzy model

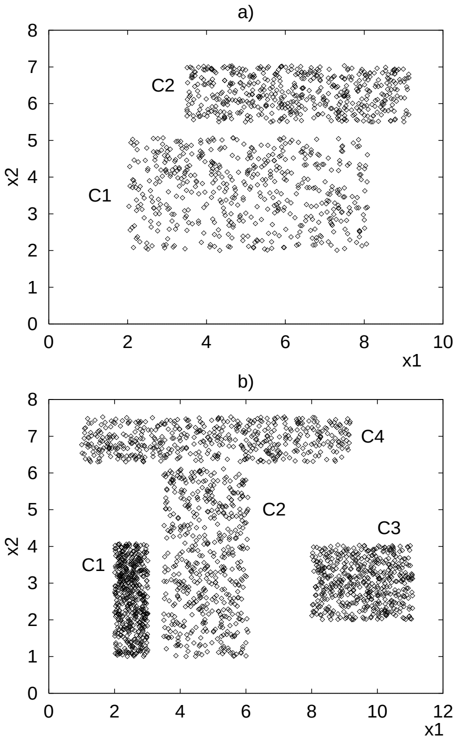


Figure 4: Two-dimensional projections of two three-dimensional input spaces with a) two and b) four output classes. The third dimension consists of random values for all output classes.

Table 3: Information measures – $I(C)$, $G(C)$, $I(C|x_y)$ and $G(C|x_y)$ – and information gains – $g_H(C|x_y)$ and $g_G(C|x_y)$ – from the fuzzy models built on the input spaces described in figures 4.a and 4.b.

	Fig. 4.a			Fig. 4.b		
$H(C)$	0.92			1.87		
$G(C)$	0.45			0.71		
dimension	x_1	x_2	x_3	x_1	x_2	x_3
$H(C\|x_y)$	0.90	0.15	0.92	0.90	1.20	1.87
$G(C\|x_y)$	0.44	0.004	0.45	0.41	0.51	0.71
$g_H(C\|x_y)$	0.02	0.84	0.00	0.52	0.36	0.00
$g_G(C\|x_y)$	0.03	0.91	0.00	0.42	0.28	0.00

is measured according to eq. 9 and 10, as reported in the upper left part of table 3. The average information measures – $H(C|x_y)$ and $G(C|x_y)$ – and the corresponding information gains derived from the use of input dimension x_1, x_2 and x_3 are reported in the first three columns of table 3.

Let us concentrate on the information gain values in table 3. The third dimension (x_3) contributes to the overall classification task with an information gain equal to 0.0, as it was to be expected because of its random values for both output classes. On the opposite, cuts on axis x_2 exploit the highest amount of information from the fuzzy model. An information gain of 1.0 is not reached, because the two output classes are very close and the trapezoids generated by [7] partially overlap. Input feature x_1 has an information gain almost zero, which shows the difficulty in separating the two output classes along input feature x_1 (figure 4.a).

A more complex problem, with a higher number of output classes, is reported in figure 4.b. Also in this case the third input dimension consists of random values for all output classes. Figure 4.b shows that a correct classification of all input data can not be obtained on the basis of only one input feature. Both input features, x_1 and x_2, seem to be necessary for this purpose. The information measures of the corresponding fuzzy model are reported in the upper right part of table 3.

In the three columns on the right in the last two rows of table 3 the information gains associated with the input features are reported. In this case none of the input features has an information gain close to 1.0, which means that a complete separability of the output classes is not achievable on one input dimension alone. Input features x_1 and x_2 present very close values of information gain, showing that they share the responsibility of a correct classification of the input space. Input feature x_2, however, has a lower information gain, due to the fact that only one class can be perfectly separated from the others along x_2, while three output classes can be separated along x_1. Input feature x_3 shows a 0.0 information gain as in the previous experiment (figure 4.a), due to the random nature of its values.

In both examples (Fig. 4.a and 4.b) the input features with highest information gain, both with entropy and Gini function, correspond to those input dimensions potentially producing the most effective cuts among the output classes.

In the described examples in figures 4.a and 4.b, the output classes are convex and therefore relatively easy to model. In order to test the strength of the fuzzy information gain parameter in quantifying the discriminative power of the input features, a pair of more complex examples including concave output classes are generated (fig. 5.a and 5.b).

In figure 5.a the input space described in figure 4.b is extended, so that class C_4 overlaps with class C_2 also on the x_2-axis. Therefore the discriminability of the output classes decreases on both x_1 and x_2 input dimensions. The corresponding information gain values are found in the three columns on the left of table 4. In this case, x_1's information gain decreases only slightly and is still the highest. Indeed x_1 still offers the smallest possibility of confusion among the different output classes in the input space. The decreasing of x_2's information gain is also consistent with the changes to class C_4. x_3 produces a 0.0 information gain, because of its random values like for the previous examples.

In figure 5.b a two-dimensional two-class problem, with one concave shaped data class, is reported. The corresponding information gains, derived from the use of each input feature for the fuzzy classification, are reported in the two columns on the right of table 4. Performing the classification on either x_1 or x_2 does not help in identifying the single output classes. For this reason, x_1 and x_2 give comparable, smaller than 1.0,

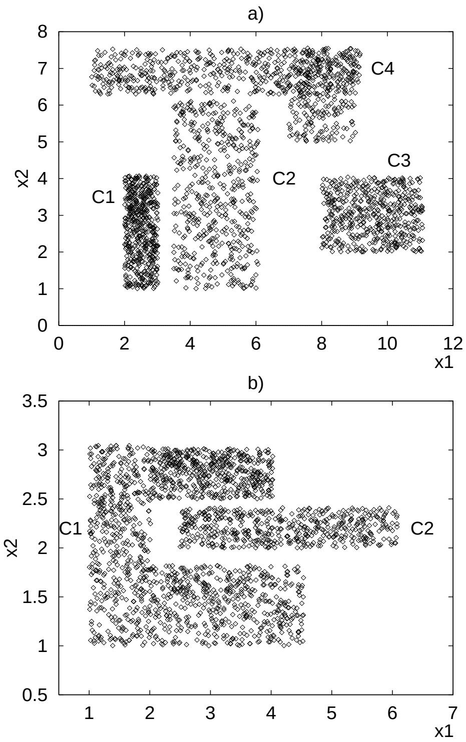


Figure 5: Input spaces with a) four and b) two output classes. The input space in a) is three-dimensional and its third dimension x_3 consists of random values for all output classes. Input space in b) is two-dimensional.

Table 4: Information measures – $I(C)$, $G(C)$, $I(C|x_y)$ and $G(C|x_y)$ – and information gains – $g_H(C|x_y)$ and $g_G(C|x_y)$ – from the fuzzy models built on the input spaces described in figures 5.a and 5.b.

	Fig. 5.a			Fig. 5.b	
$H(C)$	1.85			0.79	
$G(C)$	0.70			0.36	
dimension	x_1	x_2	x_3	x_1	x_2
$H(C\|x_y)$	0.90	1.30	1.85	0.67	0.61
$G(C\|x_y)$	0.42	0.55	0.70	0.29	0.27
$g_H(C\|x_y)$	0.51	0.30	0.00	0.15	0.23
$g_G(C\|x_y)$	0.40	0.22	0.00	0.19	0.26

information gains. The information gain of input feature x_2 is actually slightly higher, because the area of class overlapping on x_2 is smaller than on x_1 and therefore x_2 offers a better discrimination between the two output classes.

4.2 Moving Output Classes

In the previous section, we have shown that the proposed information gain is able to quantify the discriminative power of the input features in a fuzzy model representing artificially produced data. In this section, we want to assess whether changes in the separability of the output classes are reflected into corresponding changes of the information gain. We limit our study to a two-class problem for sake of clarity.

Let us start with a configuration in a two- or three-dimensional input space, where the two output classes are completely separable on at least one input dimension, as in figures 6.a (two-dimensional input space with the two output classes separable along x_1) and 6.b (three-dimensional input space with the two output classes

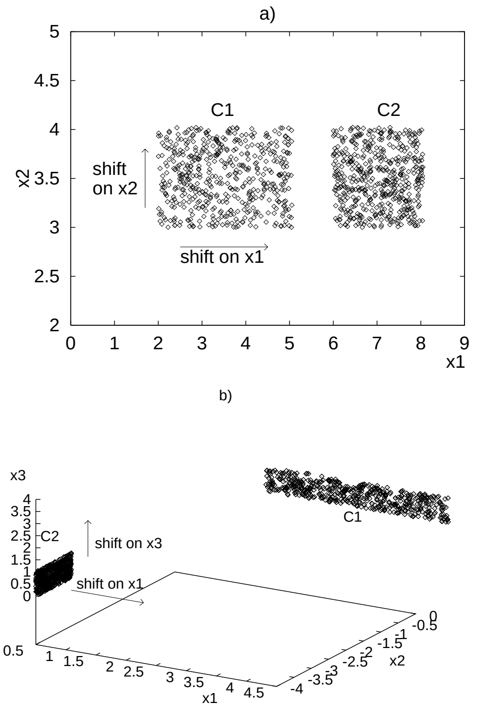


Figure 6: a) Two- and b) three-dimensional input spaces with classes progressively overlapping on one input dimension.

separable along x_1, x_2 and x_3). In the next snapshot, one of the two output classes is progressively shifted towards the other along one of the input dimensions. The information gain is monitored through time, to observe how well the evolution of the input space configuration is described.

The first experiment starts with two output classes on a two-dimensional input space completely overlapping along x_2 and separable along x_1, as described in figure 6.a. The corresponding information gains are reported in the first row of table 5. As it was to be expected, an information gain close to 1.0 describes an almost perfect separability of the two output classes along x_1, while a 0.0 information gain describes the complete overlapping of the two output classes along x_2.

At this point, the patterns belonging to class C_1 are progressively shifted towards class C_2 along the x_1-axis of a step Δx_1, while their x_2 coordinate stays constant. The corresponding information gains are reported in the following rows of the upper part of table 5.

The information gain of input feature x_1 stays very high ($g_H(C|x_1) = 0.85,\ 0.87$ and $g_G(C|x_1) = 0.91,\ 0.93$) as long as the two output classes do not overlap. In fact, the two output classes begin to overlap for $\Delta x_1 = +1.5$ and since then a progressive reduction of x_1's information gain is observed. The minimum value ($g_H(C|x_1) = g_G(C|x_1) = 0.0$) is reached for $\Delta x_1 = +3.5$, where the two output classes overlap completely on x_1. Keeping shifting class C_1 patterns towards bigger values of input dimension x_1, class C_1 begins to part from class C_2. Consequently the separability on x_1 between the two output classes increases and so does the information gain until values close to 1.0 are reestablished, that is, at $\Delta x_1 = +6.0$ when the two output classes do not overlap anymore.

For even bigger shifts Δx_1, the information gain of input feature x_1 is supposed to approach closer and closer the unitary value. However, at $\Delta x_1 = +6.5$ a small decrease in the information gains is observed, even though the two output classes are more separated than for $\Delta x_1 = +6.0$. Indeed the adopted fuzzy learning algorithm builds in this case less steep trapezoids than for closer output classes, because of the non existence of conflict points. When the output classes move farther away, the information gains increase again. Since the distribution of the input patterns along x_2 has not changed, the information gain on x_2 does not change too from the first rows of table 5.

The same experiment is now performed moving class C_1 along input dimension x_2. A progressive delay Δx_2 is applied to the x_2 coordinate of the training patterns belonging to output class C_1, while x_1 is kept

Table 5: Evolution of the information gains based on the entropy, $g_H(C|x_y)$, and on the Gini function, $g_G(C|x_y)$, for both input features, starting with the configuration in figure 6.a. and shifting class C_1 towards class C_2 along x_1 of a step Δx_1.

Δx_1	Δx_2	$g_H(C\|x_y)$		$g_G(C\|x_y)$	
		x_1	x_2	x_1	x_2
-	-	0.93	0.00	0.97	0.00
+0.5	-	0.85	0.00	0.91	0.00
+1.0	-	0.87	0.00	0.93	0.00
+1.5	-	0.44	0.00	0.55	0.00
+2.0	-	0.22	0.00	0.29	0.00
+2.5	-	0.08	0.00	0.10	0.00
+3.0	-	0.02	0.00	0.03	0.00
+3.5	-	0.00	0.00	0.00	0.00
+4.0	-	0.03	0.00	0.04	0.00
+4.5	-	0.08	0.00	0.10	0.00
+5.0	-	0.20	0.00	0.26	0.00
+5.5	-	0.47	0.00	0.58	0.00
+6.0	-	0.91	0.00	0.96	0.00
+6.5	-	0.84	0.00	0.91	0.00
+7.0	-	0.94	0.00	0.97	0.00
-	−1.5	0.93	0.99	0.97	1.00
-	−1.0	0.92	0.42	0.96	0.52
-	−0.5	0.92	0.07	0.96	0.10
-	0.0	0.93	0.00	0.97	0.00
-	+0.5	0.92	0.07	0.96	0.10
-	+1.0	0.93	0.42	0.96	0.52
-	+1.5	0.92	0.99	0.96	1.00
-	+2.0	0.92	1.00	0.96	1.00

constant. The progressive shifting of class C_1 starts this time with the configuration described in figure 6.a and $\Delta x_2 = -1.5$, that is with class C_1 located below class C_2 and perfectly separable from that also on x_2. The corresponding information gains are reported in the first row of the bottom part of table 5. x_1 shows the same information gain as in the initial configuration of the first part of the experiment (Fig. 6.a). x_2 in this case, shows a very high information gain too, due to the complete separability of the two output classes along x_2.

Increasing progressively Δx_2 and moving upwards the C_1 class patterns, the corresponding new information gains are calculated and reported in the following rows in the bottom part of table 5. Even in this case, the progressive overlapping of the two classes along x_2 corresponds to a progressive decreasing of the information gain for input feature x_2, until the two output classes completely overlap ($\Delta x_2 = 0.0$) and the minimum information gain ($g_H(C|x_2) = 0.0$ and $g_G(C|x_2) = 0.0$) is observed. If class C_1 keeps moving upwards, the two output classes begin to separate again and x_2's information gain goes up until a value close to 1.0, when the two output classes do not overlap anymore ($\Delta x_2 = +1.5$).

The described example shows clearly the evolution of the information gain with the progressive overlapping of the two output classes on each input dimension x_1 and x_2.

A similar example, but for a three-dimensional input space, is reported in figure 6.b (initial condition) and table 6. Class C_2 is progressively moved along dimension x_3 (x_1 and x_2 constant) first, and dimension x_1 (x_2 and x_3 constant) then. Even in this case, the progressive overlapping of the two output classes along one of the input dimensions corresponds to a progressive decreasing of the information gain for the same input dimension, while the information gain for the other input features stays constant. On the opposite a progressive separation of the output classes corresponds to an increase of the information gains for the interested input dimension.

The results in tables 6 and 5 show that the proposed fuzzy feature merit measure is able to detect in

Table 6: Evolution of $g_H(C|x_y)$ and $g_G(C|x_y)$, starting with the configuration in figure 6.b and moving class C_2 along x_3 first and x_1 then.

Δx_1	Δx_2	Δx_3	$g_H(C\|x_y)$			$g_G(C\|x_y)$		
			x_1	x_2	x_3	x_1	x_2	x_3
-0.5	-	-	1.00	1.00	0.99	1.00	1.00	0.99
-0.5	-	+0.5	1.00	1.00	0.23	1.00	1.00	0.28
-0.5	-	+1.0	1.00	1.00	0.06	1.00	1.00	0.08
-0.5	-	+1.5	1.00	1.00	0.00	1.00	1.00	0.00
-0.5	-	+2.0	1.00	1.00	0.06	1.00	1.00	0.09
-0.5	-	+2.5	1.00	1.00	0.23	1.00	1.00	0.29
-0.5	-	+3.0	1.00	1.00	0.99	1.00	1.00	1.00
-0.5	-	+3.5	1.00	1.00	1.00	1.00	1.00	1.00
-0.5	-	+4.0	1.00	1.00	1.00	1.00	1.00	1.00
-	-	-	1.00	1.00	1.00	1.00	1.00	1.00
+0.5	-	-	1.00	1.00	1.00	1.00	1.00	1.00
+1.0	-	-	0.99	1.00	1.00	1.00	1.00	1.00
+1.5	-	-	0.74	1.00	1.00	0.81	1.00	1.00
+2.0	-	-	0.08	1.00	1.00	0.11	1.00	1.00
+2.5	-	-	0.03	1.00	1.00	0.04	1.00	1.00
+3.0	-	-	0.02	1.00	1.00	0.02	1.00	1.00
+3.5	-	-	0.03	1.00	1.00	0.04	1.00	1.00
+4.0	-	-	0.06	1.00	1.00	0.08	1.00	1.00
+4.5	-	-	0.40	1.00	1.00	0.48	1.00	1.00
+5.0	-	-	0.99	1.00	1.00	1.00	1.00	1.00
+5.5	-	-	1.00	1.00	1.00	1.00	1.00	1.00
+6.0	-	-	1.00	1.00	1.00	1.00	1.00	1.00

both cases the dimension with maximum information content. An information gain close to 1.0 is shown on those input dimensions where a complete discrimination between the output classes is possible. The more the considered output classes overlap on the given input dimension, the closer to 0.0 the information gain drops to. The fuzziness of the system does not allow an information gain 1.0 when the two output classes are not overlapping but very close to each other. In fact, the representative membership function can extend beyond the physical boundaries of the output classes due to their fuzzy nature. Indeed the membership function slope allows an information gain as 1.0 only when the two output classes are very far from each other. This is due to the inductive bias of the used learning technique [7].

5 Real World Applications

The results in the previous section show the effectiveness of the proposed fuzzy feature merit measure in characterizing the discriminability of the output classes on the different input dimensions for artificially created data. In this section we investigate real world data.

In particular, we concentrate on two experiments. The first is performed on the IRIS database. This database is quite small and the results can not be easily generalized. On the other hand it is a commonly used database, which enables a possible future comparison with other similar techniques. The second validation is based on a database of electrocardiographic signals. Two records are investigated in terms of an arrhythmia classification task. A more general validation including all the records of the database is reported in section 6, where the proposed information gain is compared with a modified version.

5.1 The IRIS Database

The first validation of the proposed fuzzy feature merit measures is performed on the IRIS database. This is a relatively small database, containing data for three classes of iris plants. The first class is supposed to be linearly separable and the last two classes are not linearly separable. The iris plants are characterized in terms of: sepal length (x_1), sepal width (x_2), petal length (x_3) and petal width (x_4).

In [8], where a detailed description of the chosen plants' parameters is produced, the sepal length and sepal width – x_1 and x_2 – are reported to be very similar for all three output classes, i. e. they do not allow a sufficient discrimination of the three iris classes. The first two parameters can thus be considered uninformative. On the opposite, the petal features – x_3 and x_4 – characterize very well the first class of iris (iris setosa) with respect to the other two (iris viriginica and iris versicolour).

Table 7: Information gains $g_H(C|x_y)$ and $g_G(C|x_y)$ of the iris features in the IRIS database.

$I(C)$		x_1	x_2	x_3	x_4
$I_H(C) = 1.44$	$g_H(C\|x_y)$	0.10	0.06	0.82	0.81
$I_G(C) = 0.61$	$g_G(C\|x_y)$	0.10	0.06	0.84	0.79

The fuzzy clustering algorithm [7] is trained by using the whole database as training set. The corresponding information gains for each input features are calculated and reported in table 7. The third and the fourth input parameter exhibit very high information gains, while x_1 and x_2 show almost zero values. These information gain values describe that the resulting set of fuzzy rules concentrates on input features x_3 and x_4 for the discrimination of the three output classes, which corresponds to the characteristics of the database patterns. Hence input parameters x_1 and x_2 could be removed and the fuzzy analysis could be performed solely on the basis of input parameters x_3 and x_4 without a relevant loss of information.

The proposed fuzzy feature merit measures describe the informative character of the input parameters for the considered fuzzy model, which, in this case, agrees with the informative character of the input features for the considered set of data. A sufficient number of examples produces a sufficiently faithful model of the data set and hence a description of the model properties reflects a description of the training set characteristics.

In [9], a statistical correlation measure of the output classes with the input features is reported. Parameters x_3 and x_4 have a very high correlation with the output classes, while x_1 and x_2 are associated with a much lower correlation value. This confirms the hypothesis of a more informative character of the third and fourth input parameter derived from the fuzzy feature merit measures in table 7.

5.2 Arrhythmia Classification

A very suitable area for fuzzy – or in general qualitative – decision systems consists of medical applications. Medical reasoning is quite often a qualitative and approximative process, so that the definition of precise diagnostic classes with crisp membership functions may lead to inappropriate conclusions. One of the most investigated fields in medical reasoning is the automatic analysis of the electrocardiogram (ECG), and inside that the detection of arrhythmic heart beats.

5.2.1 The Physiogical Origins of Arrhythmia

Some cells (the sino-atrial node) in the upper chambers (the atria) of the cardiac muscle (the myocardium) spontaneously and periodically change their electrical polarization, which progressively extends to the whole myocardium. This periodic and progressive electric depolarization of the myocardium is recorded on the human body as small potential differences between two different body locations or with respect to a reference electrode. An almost periodic signal, the electrocardiogram (ECG), that describes the electrical activity of the myocardium in time, is the result. Each time period consists of a basic waveshape, whose waves are marked with the alphabet letters P, Q, R, S, T, and U (Fig. 7). The P wave describes the depolarization process of the two upper myocardium chambers, the atria; the QRS complex all together the depolarization of the two lower myocardium chambers, the ventricula; and the T wave the repolarization process at the end of each cycle. The U wave is often absent from the beat waveshape and its origin is controversial. The heart contraction follows the

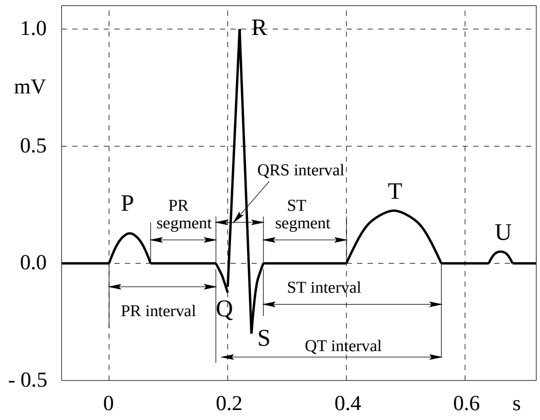


Figure 7: The ECG waveshape.

Table 8: Set of measures characterizing each ECG beat waveshape.

RR	RR interval (ms)
RRa	average of the previous 10 RR intervals
QRSw	QRS width (ms)
VR	Iso-electric level (μV)
pA	Positive amplitude of the QRS (μV)
nA	Negative amplitude of the QRS (μV)
pQRS	Positive area of the QRS (μV * ms)
nQRS	Negative area of the QRS (μV * ms)
pT	positive area of the T wave (μV * ms)
nT	negative area of the T wave (μV * ms)
ST	ST segment level (μV)
STsl	slope of the ST segment (μV/ms)
P	P exist (yes 0.5, no -0.5)
PR	PR interval (ms)

myocardium depolarization phase. Anomalies in the PQRST waveshape are often connected to misconductions of the electrical impulse on the myocardium.

A big family of cardiac electrical misfunctions consists of the arrhythmic heart beats, that derive from an anomalous (ectopic) origin of the depolarization wavefront in the myocardium. If the depolarization does not originate in the sino-atrial node, a different path is followed by the depolarizing wavefront and therefore a different waveshape appears in the ECG signal. Arrhythmia are believed to occur randomly in time and the most common types have an anomalous origin in the atria (SupraVentricular Premature Beats, SVPB) or in the ventricula (Ventricular Premature Beats, VPB).

With the development of automatic systems for the detection of QRS complexes and the extraction of quantitative measurements, large sets of data can be generated from hours of ECG signal. A larger number of measures though does not guarantee a better performance of the classifier, if no significant new information is added. A pre-screening of the most significant measures for the analysis has the double advantage of lowering the input dimension and of improving the classifier's performance by discarding poor quality measures.

The MIT-BIH ECG database [10] represents a standard in the evaluation of methods for the automatic classification of ECG arrhythmic events, because of the wide set of examples provided. The MIT-BIH ECG database consists of forty-eight records, two-channel, 30 minutes long, sampled at 360 samples/s and manually annotated by trained cardiologists.

Two records in particular (200 and 233) are analyzed in this section, because of their high number of arrhythmic beats. QRS complexes are detected and for each beat waveshape a set of 14 measures [11] is extracted by using the first of the two channels in the ECG record (Tab. 8). A two-class problem (normal (N) vs. VPB) is considered for record 200 and a three-class problem (N, VPB, and SVPB) for record 233. Two sets of fuzzy rules are built for the two problems on the two records separately and the discriminative power of their input features is evaluated.

5.2.2 Performance Evaluation After Discarding the Input Features with Lowest Information Gain

At first all fourteen ECG measures from table 8 are used for the classification task. The algorithm described in [7] is used to build a set of fuzzy rules on the first two thirds of each record and the remaining one third is used as test set. The information gains, $g_H(C)$ and $g_G(C)$, of the input features in the resulting fuzzy models are listed in the left columns in tables 9 and 10 for record 200 and 233 respectively. The parameter with highest information gain is marked bold.

For each test pattern the correct answer of the system is defined as the membership degree to the correct output class divided by the sum of all non-zero membership degrees. The percentages of correctly classified test patterns for each output class are defined as the sum of correct answers with respect to the number of test patterns of this output class and are reported on the right columns of tables 9 and 10. Test set beats are labeled as uncertain (unc.) if they are not covered by any rules of the fuzzy model. The percentage of uncertain beats is defined with respect to the total number of beats in the test set.

In the next steps, the ECG measures with smallest information gains are progressively removed from the classification process and the performance of the resulting set of fuzzy rules is evaluated. The right columns of table 9 show the performance changes associated with the gradual removal of the input features with lowest information gain in a two-class classification problem: normal (N) vs. ventricular arrhythmic (VPB) beats. The corresponding information gain values vary throughout the table, because the volumes change when one dimension is removed.

In record 200, ventricular arrhythmia are mainly characterized by morphological alterations in the QRS complex and T wave and by a prematurity degree. There are several different kinds of VPBs, depending on the location of the ectopic focus in the ventricula. This means that VPBs can be grouped in several sub-classes, usually characterized by different morphological alterations, but all occurring prematurely.

The prematurity degree of a beat is usually expressed by a shorter RR interval relatively to the average of the previous RR intervals (RRa). Because of the prematurity of VPBs, the RR interval parameter (RR) exhibits the highest information gain among all the input features, followed by the average RR interval in the previous 10 beats (RRa) (table 9).

Morphologically, VPBs present a larger and higher QRS complex and, up to a lower extent, an altered ST segment. The P wave is usually absent. In the analysis of record 200 (Table 9) only 4-6 ECG morphological features produce a high information gain, that is, are relevant for the classification process. For example, the positive and negative amplitude of the QRS complex (nAmp, pAmp) and the positive area of the T wave (pT) play an important role in the classification procedure. The morphological features, however, are redundant, since the duration and the positive amplitude of the QRS complex are already contained in the positive area. Because of this redundancy, some ECG morphological features, usually considered important by physicians for ventricular arrhythmia diagnosis, show low information gain in table 9.

Some ECG measures produce zero or almost zero information gain from the very beginning, such as the presence of the P wave (P), and the PR interval (PR). This can be due to a real uninformative character of the variable, to an unreliable measurement process or to the redundant information carried by this parameter.

All the estimated discriminative powers in table 9 find positive confirmation in clinical VPB diagnostics. In addition, the redundant or uninformative character of the input features with lowest information gain is proven by the fact that their removal does not affect the system's performance on the test set, as long as at least two of the most significant ECG measures are kept (table 9).

It is interesting to note that the system's performance increases, when the isoelectric level is removed from the input feature set. In fact, the fuzzy rules are constructed on only twenty minutes of ECG signal (two thirds of the record), which can not present a sufficiently high number of VPB examples. Thus, it can happen that some of the ventricular arrhythmic beats are associated to changes in the isoelectric line by the classifier. This is reasonable, because quite often an arrhythmic beat is accompanied by pain, with relative movement of the patient and corresponding change in the isoelectric level of his/her ECG signal. Only a very high number of examples can guarantee a good generalization of the system's rules.

Table 9: Information gains for different ECG beat measures (record 200). The amounts of correctly classified N and VPB and of uncertain beats are expressed in %.

	RR	RRa	QRSw	VR	pAmp	nAmp	pQRS	nQRS	pT	nT	ST	STsl	P	PR	N	VPB	unc.
g_H	**.63**	.33	.12	.06	.27	.39	.13	.18	.28	.03	.32	.11	.00	.03	98.8	93.4	2.7
g_G	**.63**	.35	.13	.07	.28	.42	.13	.19	.29	.04	.32	.12	.00	.03			
g_H	**.69**	.39	.12	.22	.27	.39	.19	.18	.29	.03	.33	.11	-	.03	98.8	93.4	2.7
g_G	**.69**	.42	.13	.24	.28	.42	.20	.19	.30	.03	.33	.12	-	.03			
g_H	**.69**	.40	.12	.23	.27	.38	.20	.17	.29	.03	.33	.11	-	-	98.9	93.7	2.5
g_G	**.69**	.40	.13	.25	.28	.42	.21	.19	.30	.04	.33	.12	-	-			
g_H	**.68**	.67	.09	.29	.29	.29	.24	.05	.35	-	.34	.12	-	-	98.8	95.8	1.8
g_G	**.68**	.70	.10	.32	.30	.31	.25	.06	.36	-	.34	.14	-	-			
g_H	**.68**	.38	.13	.25	.31	.37	.24	-	.28	-	.33	.11	-	-	97.9	98.6	0.7
g_G	**.69**	.42	.13	.27	.32	.40	.25	-	.29	-	.33	.12	-	-			
g_H	**.66**	.36	.11	.25	.31	.35	.25	-	.29	-	.33	-	-	-	98.1	99.0	0.6
g_G	**.67**	.40	.13	.28	.32	.39	.26	-	.30	-	.34	-	-	-			
g_H	**.66**	.30	-	.29	.32	.41	.25	-	.27	-	.35	-	-	-	90.5	99.0	0.6
g_G	**.67**	.35	-	.32	.33	.44	.26	-	.28	-	.36	-	-	-			
g_H	**.61**	.26	-	.21	.33	.40	-	-	.29	-	.36	-	-	-	90.3	99.6	0.3
g_G	**.63**	.30	-	.25	.35	.46	-	-	.31	-	.38	-	-	-			
g_H	**.58**	.26	-	-	.34	.34	-	-	.31	-	.37	-	-	-	95.5	99.6	0.3
g_G	**.61**	.30	-	-	.36	.39	-	-	.33	-	.38	-	-	-			
g_H	**.60**	-	-	-	.30	.31	-	-	.32	-	.39	-	-	-	98.4	99.7	0.2
g_G	**.63**	-	-	-	.32	.35	-	-	.34	-	.40	-	-	-			
g_H	**.55**	-	-	-	-	.31	-	-	.32	-	.38	-	-	-	98.6	99.6	0.2
g_G	**.58**	-	-	-	-	.35	-	-	.34	-	.40	-	-	-			
g_H	**.37**	-	-	-	-	-	-	-	.26	-	.34	-	-	-	98.6	99.6	0.2
g_G	**.42**	-	-	-	-	-	-	-	.28	-	.35	-	-	-			
g_H	.25	-	-	-	-	-	-	-	-	-	**.36**	-	-	-	99.1	96.2	0.1
g_G	.30	-	-	-	-	-	-	-	-	-	**.39**	-	-	-			
g_H	-	-	-	-	-	-	-	-	-	-	.20	-	-	-	92.8	86.8	0.0
g_G	-	-	-	-	-	-	-	-	-	-	.23	-	-	-			
g_H	.29	-	-	-	-	-	-	-	-	-	-	-	-	-	99.7	85.7	0.1
g_G	.35	-	-	-	-	-	-	-	-	-	-	-	-	-			
g_H	-	-	-	-	-	-	-	-	-	-	-	-	.00	.03	91.7	7.1	0.0
g_G	-	-	-	-	-	-	-	-	-	-	-	-	.00	.04			

A big decrease in the system's performance is observed when the duration of the QRS complex (QRSw) is removed. The duration of the QRS complex is considered very informative in clinical practice, as the drop in system performance proves when it is removed from the input vector. In this case, however, the reduced size of the training set transfers erroneously the discriminative power from the QRS width to the VR parameter.

Finally, in the last row in table 9, the performance of the system is reported, when the two input features with lowest information gain (P existence and PR interval) are used. As a result, the percentage of correctly classified VPBs drops to a 7% from over 90%.

The same procedure of progressive removal of the ECG measures with lowest information gain is applied to record 233 for a three-class classification problem: normal (N) vs. ventricular premature (VPB) vs. supraventricular premature beats (SVPB). The corresponding performances are reported in table 10.

With respect to record 200, record 233 presents a new class of premature beats with supraventricular origin

Table 10: Information gains for different ECG beat measures (record 233). The amounts of correctly classified N, VPB and SVPB and of uncertain beats are expressed in %.

	RR	RRa	QRSw	VR	pAmp	nAmp	pQRS	nQRS	pT	nT	ST	STsl	P	PR	N	VPB	SVPB	unc.
g_H	.12	.03	.51	.46	.10	.22	.01	.35	.08	**.63**	.01	.03	.33	.40	99.1	94.0	66.7	1.2
g_G	.14	.04	.53	.49	.11	.24	.02	.37	.10	**.67**	.02	.04	.33	.40				
g_H	.13	.08	.46	.49	.12	.23	-	.35	.07	**.60**	.01	.05	.33	.40	99.1	94.0	66.7	1.2
g_G	.15	.10	.48	.54	.14	.26	-	.38	.09	**.66**	.02	.07	.33	.40				
g_H	.13	.07	.48	.50	.13	.24	-	.35	.07	**.60**	-	.05	.33	.40	99.2	94.0	66.7	1.2
g_G	.15	.09	.50	.54	.15	.27	-	.38	.09	**.65**	-	.07	.33	.40				
g_H	.12	.08	.48	.51	.15	.24	-	.46	.05	**.55**	-	-	.33	.40	99.2	94.7	66.7	1.0
g_G	.14	.11	.50	.55	.17	.28	-	.52	.07	**.62**	-	-	.33	.40				
g_H	.12	.07	.50	.50	.15	.25	-	.50	-	**.52**	-	-	.33	.39	99.2	94.7	66.7	1.0
g_G	.14	.10	.53	.55	.17	.28	-	.56	-	**.60**	-	-	.33	.40				
g_H	.19	-	.40	**.56**	.14	.24	-	.43	-	.47	-	-	.32	.39	99.2	94.7	66.7	1.0
g_G	.24	-	.41	**.62**	.16	.26	-	.49	-	.56	-	-	.33	.39				
g_H	.24	-	.37	**.59**	-	.22	-	.38	-	.39	-	-	.30	.38	99.2	94.7	66.7	1.0
g_G	.29	-	.37	**.64**	-	.24	-	.43	-	.49	-	-	.31	.39				
g_H	.23	-	.40	**.56**	-	-	-	.52	-	.21	-	-	.18	.32	99.2	94.7	66.7	1.0
g_G	.27	-	.42	**.61**	-	-	-	.60	-	.29	-	-	.21	.35				
g_H	-	-	.36	**.62**	-	-	-	.35	-	.26	-	-	-	.38	99.0	95.4	66.7	0.6
g_G	-	-	.37	**.67**	-	-	-	.38	-	.31	-	-	-	.39				
g_H	-	-	.29	**.62**	-	-	-	.25	-	-	-	-	-	.33	99.2	95.8	66.7	0.4
g_G	-	-	.30	**.65**	-	-	-	.26	-	-	-	-	-	.35				
g_H	-	-	.34	**.55**	-	-	-	-	-	-	-	-	-	.20	99.3	95.7	66.7	0.3
g_G	-	-	.34	**.59**	-	-	-	-	-	-	-	-	-	.21				
g_H	-	-	.35	**.52**	-	-	-	-	-	-	-	-	-	-	98.7	95.8	0.0	0.2
g_G	-	-	.35	**.58**	-	-	-	-	-	-	-	-	-	-				
g_H	-	-	-	.36	-	-	-	-	-	-	-	-	-	-	88.9	91.5	0.0	0.2
g_G	-	-	-	.37	-	-	-	-	-	-	-	-	-	-				
g_H	-	-	.33	-	-	-	-	-	-	-	-	-	-	-	61.3	96.1	0.0	0.0
g_G	-	-	.32	-	-	-	-	-	-	-	-	-	-	-				
g_H	-	-	-	-	-	-	-	-	-	-	-	-	-	.40	66.9	22.4	66.7	0.0
g_G	-	-	-	-	-	-	-	-	-	-	-	-	-	.35				
g_H	-	-	-	-	-	-	-	-	-	.10	-	-	-	-	2.3	99.3	0.0	0.0
g_G	-	-	-	-	-	-	-	-	-	.13	-	-	-	-				
g_H	-	-	-	-	-	-	.05	-	-	-	.06	-	-	-	53.3	78.6	0.0	0.0
g_G	-	-	-	-	-	-	.04	-	-	-	.03	-	-	-				

(SVPB) and only a few VPBs. Supraventricular arrhythmia can be differentiated from normal beats mainly by means of the RR interval and from PVBs and normal beats by means of the PR segment, whenever the P wave can be reliably detected. Consequently the analysis of record 233 shows a high information gain also for the PR measure, besides the QRS complex morphological features and the RR interval already used for VPB classification in record 200.

However, none of the ECG measures, if considered individually, produces good performance on the test set for all three output classes (last rows of taqble 10). The SPVB classification relies mainly on the P-R wave distance (PR) and only marginally on the existence of the P wave. On the other hand, the PR parameter results to be quite useless for PVB vs. normal beats classification.

Among the QRS morphological features, the QRS duration (QRSw) exhibits the highest information gain and the system's performance tells us that QRSw is used to characterize VPBs. In addition, the few VPBs of this particular record show a different isoelectric level from normal beats. Because of that, the isoelectric level parameter (VR) shows almost always the highest information gain among the input features. Together with the QRS duration, the isoelectric level can supply the same percentage of VPB vs. normal beats classification as when all the 14 input features are used.

Only the combination of features from the PR segment and from the QRS complex morphology produce a sufficiently reliable classification of all three output classes.

Some ECG measures show a high information gain only if used together with other ECG measures. This is the case for the negative area of the T wave (nT), that characterizes the most evident VPB examples. Its information gain drops drastically when parameters related with the prematurity degree and with the QRS morphology of the beat disappear (Table 10).

5.2.3 Performance Evaluation After Retraining the System Without the Least Informative Input Features

In the previous section, we have shown that the performance of the fuzzy classifier does not change if the input features associated with the lowest information gains are removed from the analysis. It can also be possible that the least informative input features disturb the system's training procedure, leading to a sub-optimal set of fuzzy rules. In order to optimize the system's performance, after the least informative input features are removed, the system is trained again and its performance and information gains are re-evaluated. The corresponding results are shown in table 11 for record 200 and 12 for record 233.

By re-training the system, after the removal of the least informative input features, the system's performances become more stable than in the previous cases. For record 200, for example, the percentages of correctly classified beats drop down only when less than two parameters are considered for the analysis (table 11). In general, the input features with really low information gain are the same as in the previous set of experiments, such as the P wave related measures. On the other side, the most informative ECG measures may vary from one step to the next, but are always related with the beat prematurity and the QRS complex shape. Finally, it is interesting to notice that, even with re-training, the use of the P wave related measures (P and PR) yields very poor performances, close to a random choice (last row of table 11).

Record 233 includes three output classes and a more dishomogeneous class of VPBs. In the previous set of experiments, the iso-electric level was able to locate most of the VPBs, while the PR interval most of the SVPBs. In this case, the re-training guarantees a more general learning process. Here the recognition of SVPBs is still mainly committed to the PR interval and the recognition of the VPBs to the QRS duration, but the RR interval results to be extremely helpful in the recognition of VPBs, as it was to be expected rather than rely on the iso-electric changes of the signal.

The use of the poorest input features for a new training phase slightly improves the system performance, as described by the comparison of the last row of table 12 with the last row of table 10. More in general, from the comparison of tables 9 and 11 and of tables 10 and 12, we can see that re-training optimizes the use of the residual features, leading to better performances with a lower number of input features. It also avoids that particularities of the record, which are no influent if many other input features are present, become too influent once that some of them are removed, as it is the case of the iso-electric changes for VPBs in record 233. In general, however, the input features show similar information gains in the two sets of experiments for both of the evaluated ECG records of the MIT-BIH database.

6 Weighting with a Priori Probability

6.1 Defining the Confidence of Each Membership Function

Depending on the constructive algorithm used to define the set of fuzzy rules on the basis of the training examples, a more or less extensive overlapping of the membership functions is allowed, in order to make the decision process more general. On the other hand such a strategy might lead to conflicts among membership functions for some of the test patterns and as a final consequence affect the system's performance.

Table 11: Information gains (record 200) after re-training the system with the remaining ECG beat measures. The amounts of correctly classified N and VPB and of uncertain beats are expressed in %.

	RR	RRa	QRSw	VR	pAmp	nAmp	pQRS	nQRS	pT	nT	ST	STsl	P	PR	N	VPB	unc.
g_H	**.63**	.33	.12	.06	.27	.39	.13	.18	.28	.03	.32	.11	.00	.03	98.8	93.4	2.7
g_G	**.63**	.35	.13	.07	.28	.42	.13	.19	.29	.04	.32	.12	.00	.03			
g_H	.21	.18	**.28**	.12	.08	.00	.02	.00	.06	.17	.06	.12	-	.00	99.1	97.6	0.9
g_G	.27	.23	**.32**	.15	.10	.00	.03	.00	.08	.21	.07	.15	-	.00			
g_H	.11	.11	**.17**	.13	.02	.00	.14	.00	.05	.02	.06	.02	-	-	99.1	98.2	0.8
g_G	.13	.15	**.19**	.17	.02	.00	.18	.00	.07	.02	.08	.03	-	-			
g_H	.09	.10	**.43**	.15	.02	.03	.14	-	.05	.05	.07	.04	-	-	99.1	98.3	0.7
g_G	.12	.13	**.47**	.19	.02	.04	.18	-	.07	.07	.09	.05	-	-			
g_H	.05	.08	.14	**.17**	-	.08	.14	-	.04	.05	.03	.11	-	-	99.1	98.3	0.7
g_G	.06	.10	.17	**.22**	-	.10	.18	-	.05	.06	.05	.14	-	-			
g_H	.06	.08	**.14**	.13	-	.08	.13	-	.06	.04	-	.11	-	-	99.0	98.3	0.7
g_G	.08	.11	**.17**	.17	-	.10	.17	-	.07	.06	-	.14	-	-			
g_H	.06	.12	.01	.07	-	.08	.11	-	**.16**	-	-	.06	-	-	99.1	97.6	0.8
g_G	.07	.15	.01	.09	-	.11	.14	-	**.21**	-	-	.09	-	-			
g_H	.08	.08	-	.07	-	.12	.23	-	**.24**	-	-	.04	-	-	98.6	96.5	0.9
g_G	.10	.10	-	.09	-	.16	.29	-	**.28**	-	-	.05	-	-			
g_H	.11	**.37**	-	.18	-	.11	.25	-	.10	-	-	-	-	-	98.1	98.3	0.8
g_G	.13	**.45**	-	.23	-	.14	.31	-	.12	-	-	-	-	-			
g_H	.05	.27	-	.05	-	.16	**.37**	-	-	-	-	-	-	-	98.4	97.9	1.0
g_G	.06	.33	-	.06	-	.21	**.44**	-	-	-	-	-	-	-			
g_H	.25	.38	-	-	-	.11	**.41**	-	-	-	-	-	-	-	98.4	93.7	1.7
g_G	.28	.45	-	-	-	.14	**.49**	-	-	-	-	-	-	-			
g_H	**.25**	.05	-	-	-	-	.15	-	-	-	-	-	-	-	99.0	99.0	0.5
g_G	**.30**	.07	-	-	-	-	.20	-	-	-	-	-	-	-			
g_H	**.31**	-	-	-	-	-	.15	-	-	-	-	-	-	-	99.5	99.0	0.0
g_G	**.36**	-	-	-	-	-	.19	-	-	-	-	-	-	-			
g_H	.30	-	-	-	-	-	-	-	-	-	-	-	-	-	92.8	91.2	0.0
g_G	.33	-	-	-	-	-	-	-	-	-	-	-	-	-			
g_H	-	-	-	-	-	-	.06	-	-	-	-	-	-	-	51.1	73.2	0.0
g_G	-	-	-	-	-	-	.06	-	-	-	-	-	-	-			
g_H	-	-	-	-	-	-	-	-	-	-	-	-	.01	.36	49.7	51.4	0.1
g_G	-	-	-	-	-	-	-	-	-	-	-	-	.01	.43			

Solving this kind of conflicts among membership functions is always based on somewhat arbitrary decisions. One way that can be followed to solve this problem makes use of the number of training patterns covered by each membership function. The idea is that the degree of membership to a given output class is stronger if the corresponding membership function represents many patterns of the training set. The number of patterns of the training set represent the confidence with which a membership function can be trusted. Even more, in case of a conflict among two or more membership functions, the membership degree from the one with higher confidence should be taken into account.

The classification of the test patterns is then performed by means of the membership functions, $\mu^q_{C_i}(\vec{x})$, multiplied by the corresponding number of covered training patterns $N(C^q_i)$, where $\mu^q_{C_i}(\vec{x})$ is the q-th membership function ($q = 1, \ldots, Q_i$) representing output class C_i ($i = 1, \ldots, m$). In this way, trapezoids covering only a few training patterns will have a lower impact on classifying unseen patterns from the test set. In the experiments

Table 12: Information gain for different ECG beat measures (record 233) after re-training. The amounts of correctly classified N, VPB and SVPB and of uncertain beats are expressed in %.

	RR	RRa	QRSw	VR	pAmp	nAmp	pQRS	nQRS	pT	nT	ST	STsl	P	PR	N	VPB	SVPB	unc.
g_H	.12	.03	.51	.46	.10	.22	.01	.35	.08	**.63**	.01	.03	.33	.40	99.1	94.0	66.7	1.2
g_G	.14	.04	.53	.49	.11	.24	.02	.37	.10	**.67**	.02	.04	.33	.40				
g_H	.15	.02	.15	**.15**	.11	.04	-	.11	.14	.02	.11	.06	.00	.16	99.5	94.7	66.7	0.6
g_G	.17	.00	.17	**.18**	.13	.05	-	.13	.17	.01	.14	.05	.00	.09				
g_H	.03	.07	.17	.04	.05	.11	-	.01	.14	.04	.10	.08	-	**.19**	99.5	95.1	66.7	0.5
g_G	.04	.07	.18	.05	.06	.11	-	.02	.17	.04	.12	.09	-	**.16**				
g_H	.08	.07	.13	.14	.07	.07	-	-	.10	.04	.08	.06	-	**.17**	99.5	95.4	66.7	0.4
g_G	.09	.05	.16	.17	.09	.05	-	-	.12	.03	.10	.05	-	**.15**				
g_H	.09	.06	.10	.09	.12	.02	-	-	.06	-	.08	.10	-	**.19**	99.5	94.7	66.7	0.6
g_G	.10	.06	.12	.09	.13	.03	-	-	.07	-	.10	.07	-	**.16**				
g_H	.04	.08	.17	.01	.04	-	-	-	.10	-	.03	.07	-	**.20**	99.3	95.0	66.7	0.5
g_G	.04	.10	.17	.02	.05	-	-	-	.12	-	.04	.07	-	**.24**				
g_H	.13	.12	.13	-	.04	-	-	-	.20	-	.14	.07	-	**.20**	99.3	95.0	66.7	0.5
g_G	.15	.13	.13	-	.05	-	-	-	.24	-	.15	.08	-	**.22**				
g_H	.10	.09	.18	-	-	-	-	-	.14	-	.07	.09	-	**.29**	99.3	94.7	66.7	0.3
g_G	.11	.11	.17	-	-	-	-	-	.16	-	.09	.10	-	**.23**				
g_H	.17	.16	.18	-	-	-	-	-	.11	-	-	.09	-	**.24**	99.6	93.3	66.7	0.4
g_G	.19	.18	.18	-	-	-	-	-	.14	-	-	.10	-	**.21**				
g_H	.12	.02	.15	-	-	-	-	-	.08	-	-	-	-	**.18**	99.6	93.3	66.7	0.4
g_G	.12	.02	.14	-	-	-	-	-	.08	-	-	-	-	**.16**				
g_H	.11	-	.17	-	-	-	-	-	.07	-	-	-	-	**.18**	99.1	94.0	66.7	0.5
g_G	.09	-	.15	-	-	-	-	-	.08	-	-	-	-	**.16**				
g_H	.15	-	**.23**	-	-	-	-	-	-	-	-	-	-	.13	99.3	94.3	66.7	0.4
g_G	.14	-	**.19**	-	-	-	-	-	-	-	-	-	-	.09				
g_H	**.46**	-	.26	-	-	-	-	-	-	-	-	-	-	-	99.3	94.5	66.7	0.1
g_G	**.40**	-	.20	-	-	-	-	-	-	-	-	-	-	-				
g_H	.56	-	-	-	-	-	-	-	-	-	-	-	-	-	98.4	91.2	0.0	0.1
g_G	.59	-	-	-	-	-	-	-	-	-	-	-	-	-				
g_H	-	-	.46	-	-	-	-	-	-	-	-	-	-	-	51.6	85.5	0.0	0.0
g_G	-	-	.40	-	-	-	-	-	-	-	-	-	-	-				
g_H	-	-	-	-	-	-	-	-	-	-	-	-	-	.36	50.3	55.8	33.3	0.0
g_G	-	-	-	-	-	-	-	-	-	-	-	-	-	.26				
g_H	-	-	-	-	-	-	**.49**	-	-	-	-	-	.13	-	50.1	88.3	0.0	0.1
g_G	-	-	-	-	-	-	**.52**	-	-	-	-	-	.18	-				

we have performed, this kind of strategy led to improved performances as will be demonstrated in the next sections.

Adopting this classification strategy, the representational strength of each membership function is involved now in the classification task and must be taken into account for a more faithful measure of the discriminative power of the input features. Consequently, the number of training patterns covered by each membership function, $N(C_i^q)$, is introduced as a weight in the calculation of the corresponding average membership degree, $V(C_i^q)$. The new average membership degree, $\hat{V}(C_i^q)$, of the q-th membership function $\mu_{C_i}^q(\vec{x})$ for output class

Table 13: Set of measures characterizing each ECG beat waveshape.

RR/RRa	prematurity degree
QRSw	QRS width (ms)
pA	Positive amplitude of the QRS (μV)
nA	Negative amplitude of the QRS (μV)
pQRS	Positive area of the QRS (μV * ms)
nQRS	Negative area of the QRS (μV * ms)
Tarea	positive T wave area + negative T wave area (μV * ms)
IVR	Inverted Ventricular Repolarization = (pQRS + nQRS)/ Tarea
ST	ST segment level (μV)
STsl	slope of the ST segment (μV/ms)
P	P exist (yes 0.5, no -0.5)
PR	PR interval (ms)

C_i derives from equation 1 as:

$$\hat{V}(C_i^q) = \frac{\int_{\vec{x}\in D} \mu_{C_i}^q(\vec{x})\ d\vec{x}}{\int_{\vec{x}\in D} d\vec{x}} N(C_i^q) = V(C_i^q)\ N(C_i^q) \tag{19}$$

where $N(C_i^q)$ indicates the number of training patterns belonging to class C_i and covered by membership function $\mu_{C_i}^q(\vec{x})$. The use of $N(C_i^q)$ as a weight reduces the effect of spurious and noisy training patterns both on the system performance and on the information gain measure.

6.2 Arrhythmia Classification on an Extensive ECG Data Set

The ECG arrhythmia classification example, reported in the previous section, is now extended to a more general data set, including a total of thirty-nine records from the MIT-BIH ECG database [10]. A three-class problem, normal vs. VPBs vs. SVPBs, is considered. Twelve more general ECG measures are derived from the original fourteen described in table 8 and are reported in table 13. Two thirds of the selected MIT-BIH records are used as training set and the remaining one third as test set. The output classes in the training set are forced to be equally distributed, by repetition of the examples from the less represented output classes.

At first, a set of fuzzy rules is constructed [7] to discriminate the three output classes by using all twelve ECG measures. The fuzzy information gain for each one of the twelve input dimensions is then evaluated first by using the number of covered training patterns $N(C_i^q)$ as weight, as described in section 6.1, (table 14) and secondly following the original definition of information gain, as described in section 3.2, (table 15). In both tables for each row, that is for each selected set of input features, the highest information gain values are marked bold.

In the last columns of both tables 14 and 15, the percentages of correctly classified beats are reported. The system's performance in the upper part of each row is calculated by weighting each membership function with the corresponding $N(C_i^q)$ and in the traditional way in the lower part of each row.

From the two tables 14 and 15, we can see that introducing $N(C_i^q)$ into the classification procedure improves mainly the percentage of correctly classified supraventricular beats. Indeed there are not many examples of SVPBs in the data set and the resulting membership functions result to have a very small support on the input space, which limits its generalization property. The introduction of $N(C_i^q)$ for classification purposes makes the SVPB membership functions stronger in a possible conflict with larger membership functions from different output classes.

Tables 14 and 15 refer to the same fuzzy system but use two different ways of calculating the information gain associated with the input features. Comparing the first rows from the two tables, the introduction of the weight $N(C_i^q)$ into the information gain calculation changes the distribution of the discriminative power among the input features.

If the weight $N(C_i^q)$ is not taken into account (table 15), the most influent ECG measures seem to be the prematurity degree, RR/RRa, and the PR interval. If the weight $N(C_i^q)$ is considered (table 14), the information

Table 14: Information gains for 12 ECG measures **with** the number of covered training patterns as weight.

RR/RRa	QRSw	pAmp	nAmp	pQRS	nQRS	T	IVR	ST	STsl	P	PR	N	VPB	SVPB	unc.	
.09	.16	**.35**	.02	**.44**	.09	.01	.05	.00	.04	.00	**.36**	.92	.78	.71	.04	weight
												.92	.79	.64	.04	no weight
.06	.14	**.36**	.01	**.45**	.02	.01	.05	-	.03	-	**.36**	.94	.77	.76	.00	weight
												.96	.77	.63	.00	no weight
.04	.16	**.30**	-	**.24**	.04	-	.01	-	.04	-	.15	**.97**	**.74**	**.82**	.00	weight
												.98	.77	.62	.00	no weight
.01	**.13**	**.14**	-	**.14**	-	-	-	-	-	-	**.11**	.93	.81	.69	.00	weight
												.96	.79	.23	.00	no weight
-	.07	.04	-	.04	-	-	-	-	-	-	**.15**	.72	.81	.44	.00	weight
												.97	.74	.00	.00	no weight
-	.05	-	-	.06	-	-	-	-	-	-	**.13**	.68	.80	.64	.00	weight
												.97	.74	.00	.00	no weight
.12	.08	-	-	.18	-	-	-	-	-	-	-	.92	.80	.91	.00	weight
												.96	.79	.51	.00	no weight
-	-	-	-	-	-	-	-	-	-	-	.05	.73	.00	.67	.00	weight
												1.00	.00	.00	.00	no weight
-	-	-	-	.13	-	-	-	-	-	-	-	.71	.47	.19	.00	weight
												.92	.42	.00	.00	no weight
-	.15	-	-	-	-	-	-	-	-	-	-	.78	.75	.04	.00	weight
												.99	.72	.00	.00	no weight
-	-	.10	-	-	-	-	-	-	-	-	-	.80	.28	.00	.00	weight
												.86	.15	.00	.00	no weight
.02	-	-	-	-	-	-	-	-	-	-	-	.93	.10	.71	.00	weight
												.96	.67	.00	.00	no weight
-	-	-	-	-	-	-	-	.16	-	-	-	.35	.29	.64	.00	weight
												1.00	.00	.00	.00	no weight
-	-	-	-	-	-	-	-	-	-	.08	-	.73	.00	.64	.00	weight
												1.00	.00	.00	.00	no weight
-	-	-	-	-	-	-	-	.18	-	.07	-	.28	.00	.87	.00	weight
												.98	.17	.00	.00	no weight

gain of the prematurity degree is much lower, while the information gain of three other parameters related to the QRS complex morphology – its positive amplitude (pAmp), positive area (pQRS) and width (QRSw) – increases dramatically. The PR interval still exhibits a relatively high information gain and the other ECG measures maintain in table 14 a similar value of information gain to the one in table 15.

Let us examine the reasons for such difference. The introduction of the weight $N(C_i^q)$ in the calculation of the information gain takes into account one more property of the distribution of the linguistic values along one input dimension: their relative fragmentation.

If one output class C_k is described on input feature x_j by many more linguistic values than the other output classes C_i, the set of cuts along input dimension x_j will isolate a number of stripes of the input space where $N(C_k^r) << N(C_i^s)$ for $i = 1, \ldots, k-1, k+1, \ldots, m$. The membership function $\mu(C_i^s)$ might even occupy only a marginal region of the stripe, but the weights $N(C_k^r)$ – very small – and $N(C_i^s)$ – very large – mask this situation and produce relative average membership degrees with similar values for the two output classes, C_k

Table 15: Information gains for 12 ECG measures **without** using the number of covered training patterns as weight.

RR/RRa	QRSw	pAmp	nAmp	pQRS	nQRS	T	IVR	ST	STsl	P	PR	N	VPB	SVPB	unc.	
.13	.04	.07	.01	.06	.00	.00	.00	.00	.01	.04	**.29**	.92	.78	.71	.04	weight
												.92	.79	.64	.04	no weight
.11	.01	.05	.00	.04	-	-	-	-	.01	.03	.25	**.92**	**.81**	**.74**	.00	weight
												.96	.80	.34	.00	no weight
.03	.01	.07	-	.06	-	-	-	-	-	.30	.11	.93	.80	.72	.00	weight
												.96	.79	.21	.00	no weight
.03	.01	.07	-	.05	-	-	-	-	-	-	.11	.93	.81	.69	.00	weight
												.96	.80	.23	.00	no weight
.01	-	.02	-	.01	-	-	-	-	-	-	.10	.69	.58	.37	.00	weight
												.94	.74	.00	.00	no weight
.03	-	.02	-	-	-	-	-	-	-	-	.10	.71	.49	.34	.00	weight
												.89	.74	.00	.00	no weight
.04	-	-	-	-	-	-	-	-	-	-	.09	.76	.14	.68	.00	weight
												.98	.64	.00	.00	no weight
-	-	-	-	-	-	-	-	-	-	-	.06	.73	.00	.67	.00	weight
												1.00	.00	.00	.00	no weight
.15	-	-	-	-	-	-	-	-	-	-	-	.93	.10	.71	.00	weight
												.96	.67	.00	.00	no weight
-	-	-	-	-	-	.01	.00	-	-	-	-	1.00	.00	.00	.00	weight
												1.00	.00	.00	.00	no weight
-	-	-	-	-	-	.01	-	-	-	-	-	1.00	.00	.00	.00	weight
												1.00	.00	.00	.00	no weight
-	-	-	-	-	-	-	.02	-	-	-	-	1.00	.00	.00	.00	weight
												1.00	.00	.00	.00	no weight

and C_i. The measure of the information contained in the stripe will be very high and, if many of these stripes occur, the information gain for this given input feature will be very low, despite the fact that the different membership functions might be quite well separated.

The introduction of the weights $N(C_i^q)$ penalizes input features with an unbalanced distribution of the output classes, that is with some very fragmented and some very compact linguistic descriptions of the output classes.

Let us suppose that the three output classes are represented by three linguistic values partially overlapping on input dimension x_j, for example $x_j = SVPB$, $x_j = VPB$ and $x_j = N$. Thus, the corresponding average information after using x_j is:

$$I(C|x_j) = \frac{I(C|x_j = SVPB) + I(C|x_j = VPB) + I(C|x_j = N)}{3} \tag{20}$$

where $N(SVPB) = N(VPB) = N(N)$ for an equally distributed training set.

Let us suppose now that more (for example two) non-contiguous linguistic values are necessary to represent one of the three output classes (for example $SVPB^1$ and $SVPB^2$ for $SVPB$) on input dimension x_j. In this case the average information after using x_j, $I(C|x_j)$, will be:

$$I(C|x_j) = \frac{I(C|x_j = SVPB^1) + I(C|x_j = VPB) + I(C|x_j = SVPB^2) + I(C|x_j = N)}{4} \tag{21}$$

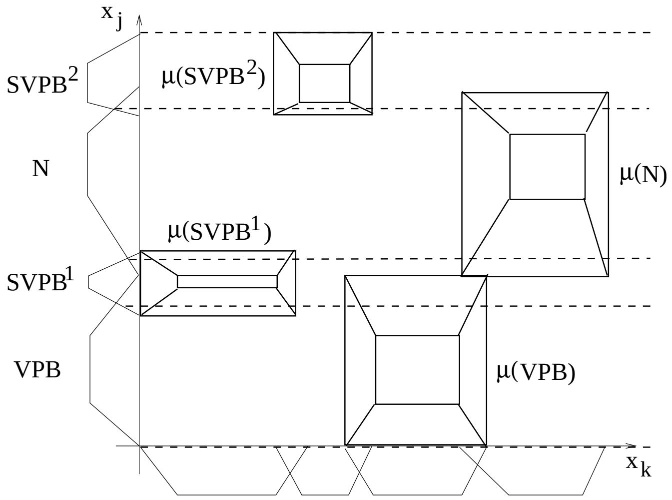


Figure 8: Cutting along input feature x_j with an unbalanced fragmentation of its linguistic values generates stripes with an unbalanced confidence distribution across the membership functions.

where $N(SVPB) = N(SVPB^1) + N(SVPB^2)$. Let us suppose that the cuts between $SVPB^1$ and N and between $SVPB^1$ and VPB originates a stripe containing mainly $\mu(SVPB^1)$ and only small queues from $\mu(N)$ and $\mu(VPB)$ (Fig. 8). The relative average membership degree of $\mu(SVPB^1)$ will be very similar to the ones of $\mu(N)$ and $\mu(VPB)$ if $N(SVPB^1)$ is very small and $N(N)$ and $N(VPB)$ are very high. The corresponding information measure will be very high. If many of these stripes occur, the information gain on this input feature will be very low due to its unbalanced distribution of the $N(C_i^q)$, which is not directly related with the separability of the output classes.

For this reason, the information gain of RR/RRa parameter decreases so much with the introduction of $N(C_i^q)$ in equation 1 and indirectly in equation 2. Parameter RR/RRa, in fact, present a very fragmented linguistic frame for the class of normal beats, referring this class to many different physiological conditions and to many different patients. On the opposite, the linguistic representation of normal beats along pAmp, pQRS and QRSw is more compact, because of its lower variance across conditions.

Such heavy influence of the distribution of the linguistic values on the measure of the discriminative power of input feature x_j may be sometimes desirable, for example when comparing two different fuzzy models with similar performance with the goal of choosing the one with the most balanced set of rules. It is not desirable, however, if only insights about the input features effectiveness are investigated. The goal of the analysis will guide the choice of the information gain to be used. The comparison of the two information gains will give insights about the compactness of the representation of the output classes on each input feature.

By comparing the two tables 14 and 15, it can be noticed that the input features with lowest information gain are more or less the same: the ST-T parameters, the Inverted Ventricular Repolarization index and the P wave existence. Removing them does not change dramatically the system performance on the test set in both tables. The system performance begins to decrease dramatically in table 14 when the prematurity degree (RR/RRa) is discarded and in table 15 when the QRS width is discarded. Both parameters then seem to be critical for the system performance. In table 14 the influence of the prematurity degree on the fuzzy system is partially masked by its very unbalanced description of the output classes, even though it still exhibits a non-negligible information gain. In table 15, the QRS width parameter shows a low but non zero information gain, which is also not negligible if compared with the information gains of many of the other input features.

The second part of the two tables reports the situation – information gain and system performance – of the system when using the input features with highest information gain alone. The prematurity degree, RR/RRa, and the PR interval share the same classification load in table 15, mainly distinguishing between normal and SVPB beats. QRSw, pQRS and up to a lower extent pAmp present a quite high information gain in table 14 because of their balanced representation of the output classes. They mainly share the task of discriminating PVBs from normal beats. Only pAmp supplies a quite low classification performance, when considered alone, despite its high information gain in table 14.

Finally, the bottom part of the tables describes the system performance, when only the input features with lowest information gain form the input vector. The P wave existence supplies interesting percentages of normal vs. SVPBs discrimination, despite the zero information gain in table 14. The ST amplitude has also some discrimination capability, which does not correspond to the zero information gain in table 14. The combination of the two worsens the system performance obtained from the single parameters. In table 15, the T wave area and the IVR parameter are considered as the input features with lowest information gain. These input features either alone or in combination really show their incapability to perform any kind of discrimination among patterns inside the considered fuzzy model.

In conclusion, in this arrhythmia classification problem, the introduction of the number of covered training patterns $N(C_i^q)$ as a weight for the membership functions in the calculation of the information gain might misrepresent the discriminative power of the input features. Input features with the same number of trapezoids for the three arrhythmic classes present higher information gains than input features with a very different number of trapezoids for the three arrhythmic classes. This is not related with the amount of overlapping of the trapezoids and consequently does not allow a faithful description of the discriminative power of the input features. On the other hand, if the information about the kind of representation of the output space on a given input dimension is required, the formulation of the information gain by means of $N(C_i^q)$ could present some advantages. In fact, if the goal is to reject input features with a too fragmented description of some output classes, then such measure could supply interesting hints.

In table 16, the system performance is reported, after removing one of the input features with different information gain in tables 14 and 15 (IVR, P, RR/RRa, QRSw). The system performance is calculated in (b) by multiplying the membership function by the number of covered training patterns and in (a) by using the original membership functions. The results show that the removal of the P wave existence, of the IVR parameter and of the QRS width do not change but actually improve the system generalization property. Only the RR/RRa parameter seem to be necessary to the fuzzy model, because, when it is removed from the input vector and despite the fact that all the other ECG measures are kept, the system performance decreases dramatically. This influence of the prematurity degree parameter can only be deduced from the information gain calculated without $N(C_i^q)$ (table 15).

6.3 Stress Detection in Spoken American English

Prosodic stress is an integral component of spoken language [12], particularly for languages such as English that so heavily depend on this parameter for lexical, syntactic, and semantic disambiguation.

Experimental and descriptive studies [13, 14] indicate that such prosodic information is mainly based on a complex constellation of information pertaining to the duration, amplitude, and fundamental frequency (pitch) associated with syllabic sequences within an utterance. These three parameters assume very different values across the consonant utterances. An investigation of prosodic stress based on the whole syllabic utterance should take into account such differences and provide an adequate normalization to allow meaningful comparisons.

Because large parts of prosodic stress information are carried by the vocalic nucleus [15, 13] and in order to avoid complicated normalization problems, the role of duration, amplitude and fundamental frequency of solely syllabic vocalic nuclei is investigated. Plain unstressed vowels produce comparable measures of amplitude,

Table 16: System performance after removing one of the input features with very low or very high information gain.

	(a) without $N(C_i^q)$ % correct				(b) with $N(C_i^q)$ % correct			
# of input parameters	N	VPB	SVPB	unc.	N	VPB	SVPB	unc.
all 12 input parameters	**92**	**79**	**64**	4	**92**	**78**	**71**	4
without P	92	78	71	2	91	77	78	2
without IVR	93	81	63	2	93	80	77	0
without RR/RRa	92	77	20	0	68	79	46	0
without QRSw	96	79	63	1	95	75	74	1

duration and fundamental frequency. In this case an adequate normalization is required only for diphthongs and lengthened vowels.

Even though it is by now quite generally accepted that prosodic stress depends mainly on the amplitude, duration and pitch of the vocalic nuclei of syllables in spoken English, the role played by each one of these three basic parameters is still controversial. In this section, the concept of fuzzy information gain, described in section 3.2, is applied to the problem of automatic detection of prosodic stress in spoken English, to ascertain the role pertaining to each one of these three basic parameters for reliable stress recognition.

Let us first quantify the three basic parameters, characterizing each vocalic nucleus.

- **Duration.** Inside a speech file, the *duration* of the k-th vocalic nucleus is the number, D_k, of signal samples between its onset and end.
- **Amplitude.** The *amplitude*, A_k, is defined as the Root Mean Square of the D_k signal samples contained in the k-th vocalic nucleus.
- **Pitch.** The *pitch*, P_k, refers to the average value of the fundamental frequency, $f_0(t)$, inside the k-th vocalic nucleus.

Fundamental frequencies $f_0(t)$ are estimated on the basis of the autocorrelation function of quarter of octave spectral channels, calculated on a 25 ms time window centered around time t and overlapping 5 ms with the previous and following time windows [17, 18]. If N_k such fundamental frequencies, corresponding to N_k partially overlapping 25 ms time windows, are detected inside the k-th vocalic nucleus, the corresponding pitch P_k is evaluated as their average value (eq. 22). This technique neutralizes residual outliers reflecting transitions from vowel to consonant and vice versa.

$$P_k = \frac{1}{N_k} \sum_{t=1}^{N_k} f_0(t) \tag{22}$$

The proposed stress characterization focuses on the properties of syllabic vocalic nuclei. Consonants are then discarded before the analysis is performed. Diphthongs, such as "ay", "oy", "er", present a longer duration than plain vowels and, because of that, are divided in three parts. For the same reason, artificially elongated vowels, that are longer than 25 ms or 40 ms, are split into three and five parts respectively. The maximum value of the evidence variable across all the splits is retained for the analysis.

Every speaker appears to use vocalic nuclei with different duration, amplitude and pitch. In order to normalize this variance among speakers, duration, amplitude and pitch are expressed in terms of variance units from the mean value of their probabilistic distributions inside each utterance [18].

6.3.1 Data Description.

To provide a reference platform for the system's performance, the prosodic stress of a portion of the American English component of the OGI Stories Corpus [19] was manually marked by two trained linguists.

The corpus contains 50-60 seconds files of spontaneous speech about any subject. A phonetic transcription of the files is also supplied. Two different subsets of files are extracted from the database and separately annotated in terms of prosodic stress by the two trained linguists. The first subset, annotated by transcriber # 1, includes 83 files, with 49 men and 34 women voices. The second subset, annotated by transcriber # 2, contains 52 files, with 39 men and 13 women voices. 10 files, 5 men and 5 women voices, are common to both subsets. The annotations refer to primary stressed (S+), other minor stressed (S-), and unstressed syllables (N).

The agreement between the two transcribers on the common files of the two OGI data subsets is shown in Table 17 and will be used as baseline for the system's performance.

The first three columns of Table 17 refer to the agreement percentage of transcriber # 1 vs. transcriber # 2, considering only men voice files (M), only women voice files (W) and both together (W+M). The second three columns refer to the same agreement percentage of transcriber # 2 vs. transcriber # 1. Since only a two-level stress automatic classification (stressed vs. unstressed syllables) is implemented, the agreement percentages in table 17 are calculated accordingly. A stressed syllables labeled as S+ (or S-) by one transcriber is considered in agreement if it was labelled as either S+ or S- by the other transcriber. The two transcribers roughly agree in recognizing primary stress (S+: 90-78%) vs. unstressed syllables (N: 84-93%). Much more disagreement exists in recognizing minor stresses (S-: 67-57%).

Table 17: In the first three columns: agreement of transcriber # 1 vs. transcriber # 2. In the last three columns: agreement of transcriber # 2 vs. transcriber # 1. The agreement percentages are calculated on all the common files (W+M), on only the male speakers common files (M) and on only the female speakers common files (W). S+ primary, S- minor stressed, N unstressed vowels.

	Transcr. # 1 vs. # 2			Transcr. # 2 vs. # 1		
	% correct			% correct		
	S+	S-	N	S+	S-	N
W+M	90	67	84	78	57	93
M	93	76	84	81	58	94
W	87	46	85	74	56	92

6.3.2 Fuzzy Classification of Prosodic Stress.

The two different subsets of annotated files from the OGI database are separately used to implement and validate a fuzzy classification system based on [7]. Two different models are obtained on the basis of two different training sets, each one extracted from the data set labeled by one of the two transcribers. The models performances are then evaluated and the two systems analyzed in terms of the discriminative power granted to each input feature.

The information gains calculated with the use of weight $N(C_i^q)$ – the number of training patterns covered by each membership function – are reported in table 20 and 21 for the first and the second transcriber's training set respectively. The same information gains calculated without the use of weight $N(C_i^q)$ are reported in table 18 and 19.

In the first row, the system is trained to distinguish between stressed (S) and unstressed (N) vocalic nuclei on the basis of the corresponding duration, the amplitude, the pitch and their product. The percentages refer to the S+ vocalic nuclei correctly recognized as stressed (under S+), to the S- vocalic nuclei also correctly recognized as stressed (under S-) and to the unstressed vocalic nuclei correctly recognized as unstressed (under N). The same task is performed by the system in the second row of the tables, but without using the product of the three basic parameters.

The following rows refer to smaller classification problems, such as S+ vs. N, S- vs. N and S+ vs. S- classification. This should help understand which input feature is the most effective in characterizing each output class. The system performances are obtained by multiplying the membership functions by the corresponding numbers of covered training patterns $N(C_i^q)$. The performances of the fuzzy classifiers without using such strategy are considerably lower and are not reported in tables 20, 21, 18 and 19.

A similar study is reported in [18], where the effectiveness of each basic parameter to a heuristic algorithm is evaluated on the basis of the Reicever Operator Characteristic (ROC) curve.

With respect to the previous arrhythmia classification here only four input parameters are considered rather than twelve. In the previous analysis the high dimension of the problem allowed some redundancy and some of the input features were not adequately exploited, which led to their very low information gains. Here, the lower dimension of the input space requires a more intensive use of the input parameters by the fuzzy clustering algorithm.

In tables 18 and 19 not many information gains have values close to zero. In general, the information gains of amplitude, duration and pitch are comparable, showing that all of them contribute to the definition of the systems' decision process. When a stressed (S+, S- or S+ and S-) vs. unstressed (N) syllables recognition is performed, the product of the three variables present the highest information gain for both transcribers' datasets. This agrees with the results of a similar investigation reported in [18], where the product of the three basic parameters obtains the highest ROC curve on the training set and the best performance on the test set. If such product is not involved in the input vector, the three basic parameters present comparable information gains.

The corresponding performances are slightly lower than the agreement percentages given by the two transcribers, but comparable with the performance of other automatic algorithms [18]. The problem seems to be easier on the first transcriber's dataset, where higher discrimination percentages of stressed (S+, S- or S+ and S-) vs. unstressed (N) syllables are obtained (table 18 compared with table 19).

Table 18: Information gains of the input features characterizing stress in spoken American English, calculated **without** the number of covered training patterns as weight for the first transcriber's training set.

First transcriber's dataset									
classification task		duration	amplitude	pitch	product	S+	S-	N	unc.
S (S+ and S-)	g_H	0.26	0.27	0.21	0.29	0.77	0.60	0.73	0.01
	g_G	0.28	0.29	0.23	0.31				
vs. N	g_H	0.18	0.22	0.16	-	0.80	0.61	0.73	0.00
	g_G	0.20	0.21	0.17	-				
S+ vs.	g_H	0.25	0.34	0.32	0.37	0.76	-	0.71	0.00
	g_G	0.28	0.38	0.34	0.40				
N	g_H	0.08	0.24	0.25	-	0.70	-	0.73	0.00
	g_G	0.10	0.26	0.30	-				
S- vs.	g_H	0.13	0.27	0.25	0.41	-	0.73	0.61	0.01
	g_G	0.15	0.31	0.28	0.46				
N	g_H	0.14	0.18	0.13	-	-	0.65	0.66	0.00
	g_G	0.15	0.22	0.15	-				
S+ vs.	g_H	0.04	0.17	0.01	0.02	0.61	0.59	-	0.04
	g_G	0.05	0.23	0.01	0.02				
S-	g_H	0.19	0.47	0.95	-	0.58	0.59	-	0.02
	g_G	0.20	0.49	0.96	-				

Table 19: Information gains of the input features characterizing stress in spoken American English, calculated **without** the number of covered training patterns as weight for the second transcriber's training set.

Second transcriber's dataset									
classification task		duration	amplitude	pitch	product	S+	S-	N	unc.
S (S+ and S-)	g_H	0.31	0.08	0.59	0.58	0.66	0.54	0.73	0.02
	g_G	0.34	0.08	0.64	0.63				
vs. N	g_H	0.21	0.20	0.22	-	0.82	0.68	0.62	0.00
	g_G	0.22	0.22	0.23	-				
S+ vs.	g_H	0.16	0.24	0.32	0.41	0.62	-	0.78	0.01
	g_G	0.19	0.27	0.36	0.46				
N	g_H	0.32	0.19	0.33	-	0.73	-	0.69	0.00
	g_G	0.35	0.23	0.37	-				
S- vs.	g_H	0.23	0.33	0.25	0.41	-	0.55	0.64	0.02
	g_G	0.25	0.34	0.27	0.44				
N	g_H	0.22	0.34	0.25	-	-	0.56	0.60	0.00
	g_G	0.23	0.37	0.27	-				
S+ vs.	g_H	0.26	0.22	0.20	0.16	0.40	0.54	-	0.02
	g_G	0.31	0.27	0.24	0.19				
S-	g_H	0.25	0.17	0.25	-	0.67	0.50	-	0.01
	g_G	0.28	0.20	0.28	-				

Table 20: Information gains of the input features characterizing stress in spoken American English, calculated **with** the number of covered training patterns as weight for the first transcriber's training set.

First transcriber's dataset										
classification task		duration	amplitude	pitch	product	S+	S-	N	unc.
S (S+ and S-)	g_H	0.37	0.32	0.26	0.44	0.77	0.60	0.73	0.01
	g_G	0.40	0.34	0.28	0.42				
vs. N	g_H	0.23	0.24	0.17	-	0.80	0.61	0.73	0.00
	g_G	0.25	0.23	0.19	-				
S+ vs.	g_H	0.23	0.34	0.32	0.40	0.76	-	0.71	0.00
	g_G	0.26	0.38	0.34	0.43				
N	g_H	0.11	0.25	0.30	-	0.70	-	0.73	0.00
	g_G	0.13	0.28	0.34	-				
S- vs.	g_H	0.16	0.26	0.22	0.40	-	0.73	0.61	0.01
	g_G	0.18	0.30	0.24	0.44				
N	g_H	0.22	0.31	0.25	-	-	0.65	0.66	0.00
	g_G	0.24	0.38	0.29	-				
S+ vs.	g_H	0.21	0.34	0.07	0.09	0.61	0.59	-	0.04
	g_G	0.27	0.43	0.09	0.12				
S-	g_H	0.19	0.44	0.95	-	0.58	0.59	-	0.02
	g_G	0.20	0.44	0.96	-				

Table 21: Information gains of the input features characterizing stress in spoken American English, calculated **with** the number of covered training patterns as weight for the second transcriber's training set.

Second transcriber's dataset									
classification task		duration	amplitude	pitch	product	S+	S-	N	unc.
S (S+ and S-)	g_H	0.20	0.08	0.44	0.56	0.66	0.54	0.73	0.02
	g_G	0.21	0.08	0.48	0.60				
vs. N	g_H	0.30	0.19	0.30	-	0.82	0.68	0.62	0.00
	g_G	0.29	0.21	0.29	-				
S+ vs.	g_H	0.21	0.12	0.17	0.44	0.62	-	0.78	0.01
	g_G	0.23	0.14	0.18	0.47				
N	g_H	0.32	0.19	0.33	-	0.73	-	0.69	0.00
	g_G	0.35	0.23	0.37	-				
S- vs.	g_H	0.23	0.30	0.28	0.51	-	0.55	0.64	0.02
	g_G	0.25	0.32	0.30	0.55				
N	g_H	0.27	0.33	0.29	-	-	0.56	0.60	0.00
	g_G	0.29	0.36	0.31	-				
S+ vs.	g_H	0.32	0.30	0.39	0.26	0.40	0.54	-	0.02
	g_G	0.39	0.37	0.46	0.32				
S-	g_H	0.34	0.20	0.28	-	0.67	0.50	-	0.01
	g_G	0.39	0.24	0.31	-				

The discrimination between the kind of stress (S+ vs. S-) is a much more complicated problem. In general linguists can only reliably distinguish between stressed and unstressed syllables, while the distinction among different levels of stresses can not be reliably performed. The fuzzy system's performances for this task are very low, being close to the random choice. Other automatic algorithms however did not obtain better results [18].

The models constructed for both transcribers (tables 18 and 19) use all the three parameters to separate stressed (S+ and S-) and unstressed (N) vocalic nuclei. In particular, the fuzzy rules rely more on duration and pitch to characterize S+ events and on amplitude to characterize S- events for the second transcriber's dataset (table 19). For the first transcriber's dataset the fuzzy system uses a bit more pitch and amplitude to discriminate S+ from N and only the amplitude for S- vs. N (table 18).

If the number of covered training patterns $N(C_i^q)$ is introduced as a weight in the calculation of the fuzzy information gain, the values in tables 18 and 19 do not change much (tables 20 and 21). This shows that more or less the same fragmentation occurs in representing the output classes on every input feature. Even in this case, the product of the three basic parameters yields the highest information gain as the most discriminative input feature for both transcribers' datasets. When the product is not used, duration, amplitude and pitch present comparable information gains showing their comparable contribution to the final decision process.

7 Related Work

Obviously much work has been done in the area of discovering feature importance, mainly under the umbrella of feature selection. The most closely related methods, however, stem from the area of decision trees, particularly ID3 [20] and it's continuous extension C4.5 [21]. A fuzzy extension to ID3 which requires predefined granulation on all features was proposed in [5]. In contrast to the method discussed in this report, all decision tree algorithms select features which have high discriminant power based on the entire data, which can be quite time consuming for large data sets. The method proposed here investigates only the model which is in general a mere summary of the data. By concentrating on the model's analysis, faster analyses are possible and in addition rule models which were generated from expert rules can be analyzed as well [22]. Another difference to decision tree algorithms is their greedy behaviour. Rather than investigating one cut after each other the proposed method analyses all cuts on one feature in parallel which enables it to find also non-binary splits.

8 Conclusions

Many algorithms already exist to quickly construct fuzzy models from example data. Quite often, however, if the output classes are not easily separable and if the dimension of the input space is too high, the interpretation of the resulting fuzzy model might not be easy to perform. One aspect of such interpretation involves the estimation of the impact that input features have on the final decision process.

In this work, an a-posteriori analysis of the fuzzy model obtained from a given training set is performed, in order to quantify the influence of the input features in discriminating among the output classes, that is their discriminative power.

Using properties of fuzzy logic, it is easy and computationally inexpensive to define a measure of the information contained in the fuzzy model. Such measure is used to quantify the information available in the fuzzy model before and after a given input feature is used for the classification. The relative difference of the two information measures defines the information gain associated with the use of this input feature. The defined information gain provides a quantification of the discriminability among output classes along the analyzed input feature for the considered fuzzy model. This is related with the system's classification performance, only if the fuzzy model is constructed on a sufficiently general set of training examples.

The method's potentiality is illustrated by using artificial and real-world examples. In particular, as real-world examples, the most informative electrocardiographic measures are detected for an arrhythmia classification problem and the role of duration, amplitude and pitch variations of syllabic vocalic nuclei in American English spoken sentences is investigated for prosodic stress detection.

The proposed algorithm represents a computationally inexpensive tool to reduce high-dimensional input spaces as well as to get insights about the implemented decision process. For example, it can be used to determine which input features are exploited by a fuzzy classifier and, in case of a bad performance, the analysis of the most exploited input features can help discover errors in including/normalizing part of the input coordinates. The proposed information gain can also be used to determine the difference, in terms of input features impact, among fuzzy classifiers with different performances.

We believe that especially for large scale data sets in high dimensional feature spaces, such quick approaches to gain first insights into the nature of the data will become increasingly important to successfully find the underlying regularities and to boost the knowledge discovery process for real world problems.

9 Acknowledgements

The authors would like to thank Wei Zong, George Moody, and Prof. R.G. Mark from Harvard-MIT Division of Health Sciences and Technology M.I.T. (Cambridge, USA) for the ECG measures and Steven Greenberg of the International Computer Science Institute (Berkeley, USA) for the measures of the acoustic parameters of vocalic nuclei in spoken American English sentences.

References

[1] M. Berthold, D.J. Hand (Eds), "Intelligent Data Analysis: An Introduction", Springer-Verlag, 1999.

[2] C. Apte, S.J. Hong, J.R.M. Hosking, J. Lepre, E.P.D. Pednault, and B. K. Rosen, "Decomposition of heterogeneous classification problems", *Intelligent Data Analysis*, Vol. 2, n. 2, 1998.

[3] L.A. Zadeh, "A fuzzy-algorithmic approach to the definition of complex or imprecise concepts", *Int. J. Man-Machine Studies*, **8:** 249-291, 1976.

[4] A. De Luca, and S. Termini, "A definition of nonprobabilistic entropy in the setting of fuzzy sets theory", *Information and Control*, **20**(4): 301-312, 1972.

[5] C.Z. Janikow, "Fuzzy Decision Trees: Issues and Methods", *IEEE Trans. Syst. Man and Cyb. PartB: Cybernetics*, **28:** 1-14, 1998.

[6] M. R. Berthold, K.P. Huber, "Comparing Fuzzy Graphs", Proc. of Fuzzy-Neuro Systems, pp. 234-240, 1998.

[7] K.-P. Huber and M.R. Berthold, "Building Precise Classifiers with Automatic Rule Extraction", in Proceedings of the *IEEE International Conference on Neural Networks*, vol. 3, pp. 1263-1268, 1995.

[8] R.A. Fisher, "The use of multiple measurements in taxonomic problems", *Annual Eugenics, II*, John Wiley, NY. **7**:179-188, 1950.

[9] C. Blake, E. Keogh, and C.J. Merz. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998.
(http://www.ics.uci.edu/~mlearn/MLRepository.html)

[10] MIT-BIH database distributor, Beth Israel Hospital, Biomedical Engineering, Division KB-26, 330 Brookline Avenue, Boston, MA 02215, USA.

[11] W. Zong, D. Jiang. "Automated ECG rhythm analysis using fuzzy reasoning", Proc. of Computers in Cardiology, pp. 69-72, 1998.

[12] Lehiste I. 1970. *Suprasegmentals* MIT Press, Cambridge.

[13] Kuijk, D. van and Boves, L. "Acoustic characteristics of lexical stress in continuous telephone speech", *Speech Communication* 27, 95-111, 1999.

[14] Wightman, C.W. and Ostendorf, M. "Automatic labeling of prosodic patterns", *IEEE Transactions on Speech and Audio Processing* 2,469-81, 1994.

[15] Bergem, D. van "Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels", *Speech Communication* 12, 1-23, 1993.

[16] Green, D. M. and Swets, J.A. *Signal detection theory and psychophysics.* New York, Wiley, 1996.

[17] Hess, W. *Pitch determination of speech signals: algorithms and devices.* Berlin, Springer-Verlag, 1983.

[18] R. Silipo and S. Greenberg, "Automatic transcription of prosodic stress for spontaneous english discourse", Proc. of the XIVth International Congress of Phonetic Sciences (ICPhS), 3:2351, 1999

[19] Center for Spoken Language Understanding, Dept. of Computer Science and Engineering, Oregon Graduate Institute. *Stories corpus*, Release 1.0 1995.

[20] J.R. Quinlan, "Induction of Decision Trees", in *Machine Learning*, pp. 81-106, 1986.

[21] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[22] R. Silipo, R. Vergassola, M. R. Berthold, "Expert knowledge and Data Driven Models in Arrhythmia Fuzzy Classification", submitted to *Methods of Information in Medicine* Special Issue on Applications of Intelligent Data Analysis in Medicine.